NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF ELECTRICAL
AND COMPUTER ENGINEERING

DIOGO PICANÇO MARTINS
Master in Electrical and Computer Engineering

A NOVEL TOOL FOR SURVIVAL ANALYSIS IN LYMPHOMA PATIENTS

MASTER IN ELECTRICAL AND COMPUTER ENGINEERING

NOVA University Lisbon
September, 2022

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF ELECTRICAL
AND COMPUTER ENGINEERING

A NOVEL TOOL FOR SURVIVAL ANALYSIS IN LYMPHOMA PATIENTS

DIOGO PICANÇO MARTINS

Master in Electrical and Computer Engineering

Adviser: João Paulo Branquinho Pimentão
Assistant Professor, NOVA University Lisbon

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

MASTER IN ELECTRICAL AND COMPUTER ENGINEERING

NOVA University Lisbon
September, 2022

A Novel Tool for Survival Analysis in Lymphoma Patients

Para a minha família.

Acknowledgements

Firstly, I would like to thank my adviser, esteemed professor João Paulo Pimentão and my co-adviser, the esteemed professor Gracinda Diogo Guerreiro that advised and guided me throughout my dissertation. A special thanks to Dr Maria Torrente, who graciously helped with elaborating the dissertation with unmatched specialist knowledge that helped to further the conclusions of the developed work. A special thanks also to my colleagues with whom I developed the project, Mariana Pardal and Bruno Vieira.

I would also like to thank the institution that taught me the competencies that make me the professional I am today, NOVA School of Science and Technology.

Equally important to my formation and education as a person, an enormous thank you to Extrenato Marista de Lisboa and all the staff that made me the man I am today, with solid values that make me a "virtuous citizen".They are always a second home from which I retain unforgettable memories and experiences.

To all my colleagues in HOLOS from whom I have learned extensively, particularly Alexandre Sousa and Francisco António, with whom I worked closely during the progress of this project. Furthermore, to Sofia Rodrigues, who mentored me throughout my stay within HOLOS.

A big thank you to all my old swimming teammates that with me, learned that with perseverance and mental fortitude, everything is achievable when you put your mind and all your effort into it.

To my childhood friends, from "Grupo dos 10" to friends so old that they are part of the family, a big thank you for always being there.

To my university friends and the friends I made in "Missão", thank you for always being there through thick and thin and always coming out better than we got in.

To all my family, my uncle, aunt and cousins, thank you for continuously extending your hand without ever second-guessing the work/difficulty of the request and always making me have a good laugh.

To my grandparents, who constantly have been by my side, helping and advising me throught my life, I am genuinely grateful to have you, especially as a next-door neighbour.

To my mother and father, who always gave me everything and more so I could sucessed not only professionally and personally, a huge thank you. There are not enough words to express my gratitude for all they have done for me, and they were and still are an example of excellence that I aim to be every day. To my sister, who is a large contributor to my "stress" but also an unwavering and loving person, thank you for always being there, even if it is to annoy me as much as I annoy you back.

Lastly, to my best friend and coincidentally my girlfriend Ana Sofia Antunes, for the decade of friendship and six years of relationship, thank you. I would certainly not be where I am today without your support, your perseverance and above all your endurance (towards me).

"It is our choices (...) that show what we truly are, far more than our abilities." (APWBD)

Abstract

Keywords: Lymphoma cancer, Hodgkin Lymphoma, non-Hodgkin Lymphoma, survival analysis, Kaplan-Meier estimator, log-rank test, React, Tool development.

Annually, cancer is responsible for 40% of earlier deaths due to non-communicable diseases, and this number increases at an annual rate of around 1.6%. These alarming values make it essential to study this disease at a global level, to help better the lives of all the affected patients and disseminate prevention when possible.

With the advance in technology and thanks to the influx of patients with digitalised records that suffer from this disease, there is a greater capability to elaborate a study about the possible causes and consequences drawn from the patient's data. Furthermore, the ability to better the patient's quality of life by analysing their data and sensitising them is fundamental in the fight against cancer.

The dissertation focuses on developing a computational tool that enables tha ability to obtain simple statistics, thanks to classical techniques of survival analysis as well as the analysis of lymphoma cancer, both Hodgkin and non-Hodgkin lymphomas that constitute nearly 48% of blood cancers. To determine the factors that influence the study of the received patients' database, a preprocessing is done where the descriptive statistics are obtained using the patients' database information. After that, Kaplan-Meier estimator curves are elaborated to determine the relationship between the studied phenomenon and the different variables present in the database. After taking brief conclusions from the obtained variables and subsequent descriptive analysis, an analysis using the Kaplan-Meier estimator is done. The integration of the achieved results is implemented in a tool that constitutes CLARIFY [1]'s project dashboard.

This dissertation was created in conjunction with the CLARIFY European project, led by the oncology medical team of University Hospital Puerta Hierro de Majadahonda.

[1] https://www.clarify2020.eu/

Resumo

Palavras-chave: Linfoma, Linfoma de Hodgkin, Linfoma não-Hodgkin, Análise de sobrevivência, Estimador de Kaplan-Meier, Teste log-rank, React, Desenvolvimento de ferramentas.

Anualmente, o cancro é responsável por 40% das mortes precoces devido a doenças não transmissíveis, e este valor aumenta anualmente cerca de 1.6%. Estes valores alarmantes fazem o estudo desta doença um foco fundamental a nível global de modo a melhorar a vida de todos os pacientes e disseminar prevenção a quando possibilidade do mesmo.

Com o avançar da tecnologia e graças a um influxo de registos digitalizados sobre pacientes que sofrem este tipo de doença, existe uma maior capacidade de elaborar um estudo sobre as possíveis causas e consequências retiradas a partir dos dados de pacientes que passaram por isso. Para além disso, a capacidade de melhorar a qualidade de vida dos pacientes através da análise dos seus dados e da sensibilização dos mesmos é fundamental para uma constante luta contra o cancro.

Esta dissertação foca-se no desenvolvimento de uma ferramenta computacional que permite aceder de forma simples, a estimativas obtidas a partir de técnicas clássicas de análise de sobrevivência como exemplo de aplicação, foca-se ainda na análise do cancro linfoma tanto Hodgkin como não-Hodgkin, que abrange cerca de 48% dos cancros de sangue. Com o objetivo de averiguar os fatores de risco que influenciam a sobrevivência dos pacientes da base de dados em estudo, é efetuado um pré-processamento dos dados, onde são obtidas estatísticas descritivas da base de dados de pacientes e produzidas estatísticas das curvas de sobrevivência com recurso ao estimador de Kaplan-Meier de modo a determinar a relevância das variáveis analisadas da base de dados em relação ao acontecimento analisado. A integração dos resultados obtidos através do estimador Kaplan-Meier será integrada numa ferramenta que por sua vez fará parte do *Dashboard* do projeto do CLARIFY [2].

Esta dissertação foi criada em conjunto com o projeto europeu, Clarify, liderado pela equipa médica de oncologia do Hospital Universitário Puerta Hierro de Majadahonda

[2] https://www.clarify2020.eu/

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence 2, 19, 24, 25
AIDS	Acquired immunodeficiency syndrome 9
cHL	Classic Hodgkin lymphoma 5, 6, 10
CRF	Cancer-related fatigue 1
DBMS	Database Management System 22, 42
EBV	Epstein-Barr 8
ECOG	Eastern Cooperative oncology group xiii, xvi, 11, 16, 17, 59, 60
EHR	Electronic Health Register 18, 19
ELT	Extract-Load-Transform 21
ETL	Extract-Transform-Load 21, 22, 23
FLIPI	Follicular Lymphoma International Prognostic Index 12, 18
GA4GH	Global Alliance for Genomics and Health 19
GCP	Google Cloud Platform 27
HCV	Hepatitis C virus 9
HIV	Human Immunodeficiency Virus 5, 6, 8, 9
HL	Hodgkin lymphoma 1, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 16, 43, 55, 60, 61, 63, 68, 71, 75, 77, 79, 81, 83, 85, 86, 94, 95, 98, 100, 106, 107, 108, 111, 112, 114
HSCT	Hematopoietic stem cell transplant 13
HUPHM	Hospital Universitario Puerta de Hierro Majadahonda 2, 123
IPI	International Prognostic Index xvi, 11, 12, 17, 64

KDD	Knowledge Discovery in Databases 22, 23, 24
LdcHL	Lymphocyte-depleted classic Hodgkin lymphoma 6, 8
LDH	Lactate Dehydrogenase xv, 17, 74, 75, 112, 113
LrcHL	Lymphocyte-rich classic Hodgkin lymphoma 6
MALT	Mucosa-associated lymphoid tissue 9
MALT-IPI	Mantle Cell Lymphoma International Prognostic Index 18
MCCHL	Mixed cellularity classic Hodgkin lymphoma 5, 6, 8
MIPI	Mantle Cell Lymphoma International Prognostic Index 12, 18
ML	Machine Learning 2, 19, 25, 26
NCD	Non-communicable disease 1
NHL	Non-Hodgkin lymphoma 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 43, 55, 60, 61, 63, 71, 75, 79, 81, 83, 85, 95, 96, 98, 101, 103, 107, 109, 111, 113
NLPHL	Nodular lymphocyte-predominant Hodgkin lymphoma 5, 6, 10
NPL	Natural Language Processing 25
NSCHL	Nodular Sclerosis classic Hodgkin lymphoma 5
PHR	Personal Health Register 18
RDBMS	Relational Database Management System 21, 22, 42
SCID	Severe combined immunodeficiency disease 9
SLCG	Spanish Lung Cancer Group 123
SRA	Single Page Application 27, 28
UI	User Interface xv, 116, 118, 121, 125
WHO	World Health Organization 6

1

Introduction

The first chapter presents the motivation and background behind the dissertation and the larger project it is inserted into, CLARIFY. It also denotes the relevance of the theme of the thesis at hand.

1.1 Background and Motivation

Every day humanity is faced with large amounts of diseases and health problems that can affect how we interact and live as humans. Despite the ever-pending discovery of new conditions, there is a constant problem for global health whose number of diagnostics has increased over the last decades, cancer.

Cancer is considered a Non-communicable disease (NCD); NCD's are more commonly known as chronic diseases, which constitute 71% of yearly deaths [1]. Cancer is also responsible for four in every ten premature deaths caused by Non-communicable disease [2].

Even though the number of diagnosed cases increases yearly, the death rate has been gradually decreasing [3] due to better treatment and earlier diagnoses due to preventive initiatives worldwide. In the United States of America, the overall death ratio has fallen 1.8% in men, 1.4% in women annually since the early 2000's up to 2017, and 1.4% every year in children from 2013 to 2017[4].

These statistics mean that a more significant number of people are undergoing treatment or deemed as cured every year, making essential the task of assuring their future well-being and the ability to diagnose and prevent the relapse of the patients based on their habits as of the treatment that they are subjected.

Even though the treatments such as chemotherapy, radiotherapy, immunotherapy and even surgery are meant to help the patients, this does not mean that there are no side effects from them, some are temporary, such as nausea as a side effect of chemotherapy [5] and others are lasting such as chronic Cancer-related fatigue (CRF) that manifested in 16% of males who survived testicular cancer and 24% for men who survived Hodgkin lymphoma (HL) compared to 10% of the general non-cancerous population [6] as well as

33% of the studied population after being considered cured express some cancer-related pain [7]. The need to find new ways to help cancer survivors combined with the need to prevent and diagnose a relapse makes the study of historical data of the utmost importance.

Alongside medical and data analysis methods to help people who are suffering or recovering from cancer is it also of particular importance that they have a social support group made up of family and friends there to help [8] as well as professional help as depressive symptomology is a common factor when comparing healthy people to people who have had cancer [9]. It is also pivotal that all oncological patients are accompanied throughout their social lives since a majority of patients that underwent a study about "Social problems in oncology" [10] express having some social repercussion as an effect of cancer, such as unemployment, financial issues and even in their home lives, the effects are shown are even more significant in people under 40 that tend to have less structured and consistent lifestyles.

This dissertation was developed in association with Holos SA for the European Union's Horizon 2020 approved and financed project CLARIFY. Holos is a part of the twelve entity consortium coordinated by the Medical Oncology Department in the *Hospital Universitario Puerta de Hierro Majadahonda*. The dissertation will study patients diagnosed with lymphoma cancer and their survivability.

1.2 Problem and Proposed solution

As stated in the previous section, cancer is a complex multidimensional problem. This way is fundamental that every aspect of the process, from diagnosis to treatment and post-treatment, is developed as much as possible to help every patient that is undergoing or has undergone any scenario caused by cancer.

In addition to economic, social and personal problems caused by daily coexistence with cancer, cancer patients and survivors tend to be concerned about their follow-up care once they are deemed survivors and no longer constitute part of the active cancer care [11]. The uncertainty of recurrence and the general ambivalence towards the future make the patients and their next of kin vulnerable [12]; thus, it is of the utmost importance that individualised plans, and support structures are implemented as a way to improve the overall quality of life for the patients [13].

The study of the data will help better comprehend the disease and possible patterns that emerge when dealing with survival. This theory is supported by the use of quantitative results and statistics. As medicine aims to evolve and not only coexist but also include and reap benefits from technology, the possibilities for crossovers seem endless. Artificial Intelligence (AI) and Machine Learning (ML) algorithms is an ever-growing reality.

The dissertation aims to perform the survival analysis with the help of the Kaplan-Meier estimator to provide every patient with a customised itinerary for follow-up cancer survivors. It also focuses on creating a tool that could easily and quickly provide the results of said analysis for any user of the project at hand.

The goal is, together with the survival curves, that help not only the patient but the doctor to visualise the overall survival data, along with awareness-raising campaigns to help further the life of cancer survivors and help them improve their general quality of life.

1.3 Contributions

The dataset was graciously provided by the Medical Oncology Department at Puerta de Hierro University Hospital in Madrid, Spain. A total of 994 patients diagnosed with HL and Non-Hodgkin lymphoma (NHL) were included. An anatomopathological confirmation by a hematopathologist was considered mandatory to include patients in the study. The following variables were collected from the clinical records according to the standard hospital protocol: demographic data from each patient, clinical and pathological features, prognostic factors and treatments received, type of response and survival, as well as the detection of specific causes of death. All patients had to undergo an adequate extension study prior to initiation of treatment and during follow-up. In all cases, the stage was determined according to the Ann Arbor classification, later modified in 1988 at the Cotswolds meeting. Information on the progression of disease or relapse, re-treatment, and death, as well as its causes, was verified through the analysis of medical health records, death certificates, or National Institute of Statistics (INE) records. The cause of death associated with HL included progression of the disease independently of other causes of death. Non-HL-related deaths were included in those causes independent of lymphoma, secondary to the toxicities of chemotherapy treatments and the development of secondary malignancies. The Ethics and Clinical Research Committee of Puerta de Hierro University Hospital reviewed and approved the study.

2 State of the art

It is fundamental to understand the basic concepts of the industry at study to be able to add value to the accrescent medicine field that is oncology.

This section focuses on the preceding insight of lymphoma cancer as well as the different types of lymphomas. Furthermore, the concepts of Healthcare Data and security involving it coupled with defining Machine Learning will also be analysed in this section.

2.1 Lymphoma Cancer

Lymphoma is a type of haematological cancer more commonly known as blood cancer. This group of cancer occurs when lymphocytes develop unorderly at a rapid rate. Lymphocytes are a type of white blood cell that constitute the immune system's immunological memory [14], which aids the recognition of previous occurrences, such as infections so that the immune system can respond adequately.

The lymphocytes are constituted by the B cells, T cells and NK cells. B cells are responsible for the making of antibodies. In contrast, T cells can be either a helper T cell that is responsible for informing the need of antibodies by the host or cytotoxic T cells whose function is to kill germs and organism cells that are deemed not normal, for instance, cancerous cells [15]. NK cells or natural killer cells are in charge of killing cells that are infected by viruses and detecting early cancer signs [16].

Lymphoma occurs when the lymphocytes do not die when they are supposed to or when they split in an anomalous way; they more commonly occur in lymph nodes in the neck or the pubic area, but it could appear virtually in any part of the human body. The symptoms, the treatment and the behaviour of the lymphoma itself are dependent on the type of lymphoma it is. As there are a wide variety of lymphomas, they have been grouped into two main categories Hodgkin lymphoma and Non-Hodgkin lymphoma [17], as can be observed by the simplistic schematic presented below in the image 2.1. According to the Leukaemia and Lymphoma Society, using data from the American Cancer Society, Lymphoma is statistically the most prominent blood cancer in 2021, with 48% of new cases [18]

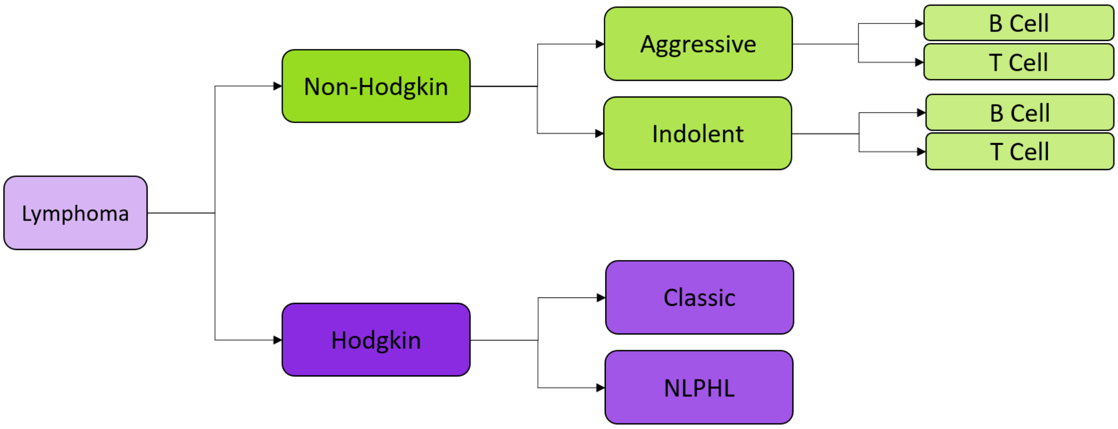


Figure 2.1: General classification of Lymphoma cancer based on Lymphoma action
Based on [17]

2.1.1 Hodgkin Lymphoma

Hodgkin lymphoma is a uncommon monoclonal lymphoid neoplasm with high cure ratios. Both clinical and biological studies have separated this kind of lymphoma into two categories: Nodular lymphocyte-predominant Hodgkin lymphoma (NLPHL) and Classic Hodgkin lymphoma (cHL), as figure 2.2 demonstrates.

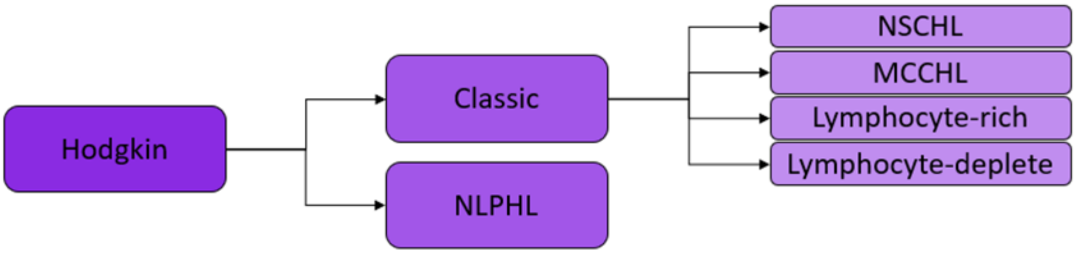


Figure 2.2: Hodgkin lymphoma subtypes
Based on [17]

According to the American Cancer Society, Classic Hodgkin lymphoma is responsible for 9 in 10 cases of HL [19]. The cHL's cancer cells are also known as RED-Sternberg cells, a mutation of B lymphocytes (B cells) [20][21][22]. There are four subtypes of Classic Hodgkin lymphoma [19]:

- Nodular Sclerosis classic Hodgkin lymphoma (NSCHL) - represents approximately 70% of cHL and is more incident in younger generations. It usually affects the neck and chestal region ;

- Mixed cellularity classic Hodgkin lymphoma (MCCHL) - most common after NSCHL, frequent in patients who are Human Immunodeficiency Virus (HIV) positive, children and the elderly, tends to appear in the upper body

- Lymphocyte-rich classic Hodgkin lymphoma (LrcHL) - a very rare subtype, usually occurs like MCCHL in the upper body and affects few lymph nodes
- Lymphocyte-depleted classic Hodgkin lymphoma (LdcHL) - normally affects older people who are HIV positive. It is considered the fiercest of the HL, and it can be found in the liver, bone marrow or stomach.

Nodular lymphocyte-predominant Hodgkin lymphoma is a highly uncommon type of cancer impacting the lives of 0.1 to 0.2 people per hundred thousand people in the USA, about 5% of all Hodgkin lymphoma cases [23].

The NLPHL is a type of HL that takes effect on the B cells and is characterised for its different pathology compared to cHL; the cells in the lymph nodes that suffer from this condition are usually called "popcorn cells" due to their format.

This kind of cancer is more frequent in males than females; this fact can be corroborated by the German Hodgkin Study Group's study on which, from 471 patients with NLPHL 75.8% were men, with only 24.2% of the patients being women. In terms of age demographics, there is no one age that NLPHL could appear, but it is more likely to appear between the age group of 30 to 50 years old, as confirmed by the same study where the median of ages of the patients involved was 39 years old [24].

As Classic Hodgkin lymphoma is responsible for nearly 95% of all HL, the dissertation will focus exclusively on Classic Hodgkin lymphoma when it comes to analysing HL[25][26].

2.1.2 Non-Hodgkin Lymphoma

NHL is a group of malignant neoplasms originating from the lymphoid tissues, mainly the lymph nodes. These tumours may result from chromosomal translocation, toxins, infections, and chronic inflammation.

This group of malignant neoplasms encompasses various subtypes, each with different epidemiologies, etiologies, immunophenotypic, genetic, clinical features, and therapy responses, making the overall classification complex. To help with classification NHL can be split into two groups, 'aggressive' and 'indolent'. This classification is built on the prognosis of the diseases. The most recent World Health Organization meets the previously stated classification of lymphoid neoplasm [27]. It is important to refer that the terminology High-Grade and Low-Grade can be interchangeably used with previously defined classification of aggressive and indolent, respectively. As a result of this classification, we can define the diagram in the figure 2.3 presented below.

Furthermore, the World Health Organization aggregates NHL based on five different characteristics [28][29]:

- Grade of the NHL – aggressive or indolent;
- The type of white blood cell infected – B cell or T cell;

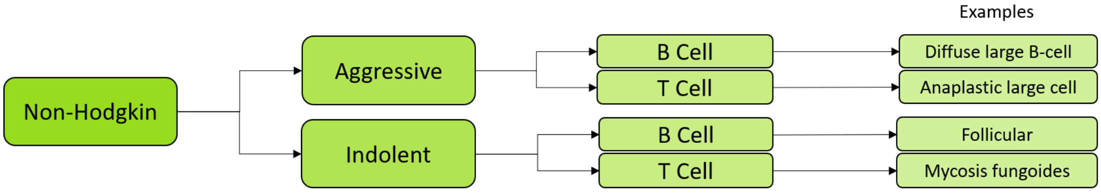


Figure 2.3: Classification of Non-Hodgkin Lymphomas
Based on [17]

- The appearance of the lymphoma under microscopic observance;
- Type of protein markers constitute the surface of the lymphoma – analysed through an immunohistochemistry test;
- If there is a genetic change in the cells – studied by cytogenetics.

The most frequent mature B cell neoplasms are Follicular lymphoma, Burkitt lymphoma, diffuse large B cell lymphoma, Mantle cell lymphoma, marginal zone lymphoma and primary CNS lymphoma. On the other hand, the most recurring mature T cell lymphomas are adult T cell lymphoma along with Mycosis fungoides. Treatment regarding NHL varies greatly, depending on tumour stage, grade, typology of lymphoma and several different patient factors such as symptoms, age and performance status[30].

2.1.3 Grade

As stated before, Non-Hodgkin lymphoma can be graded as either "aggressive" or "indolent".The natural history of these tumours demonstrates significant differences between both.

Aggressive lymphomas, also known as high-grade lymphomas, tend to develop rapidly and are flagged for treatment as soon as detected [29]. They also have specific B symptoms such as night sweats, fever and weight loss; it can also end up being the source of death of the patient within a few weeks if it stays untreated. The most common type of high-grade lymphoma is diffuse large B cell lymphoma which affects the B lymphocytes [31]. Burkitt lymphoma, precursor B and T cell lymphoblastic leukaemia/lymphoma, adult T cell leukaemia/lymphoma, and other peripheral T cell lymphomas are other examples of aggressive lymphomas[32].

Indolent lymphomas, also known as low-grade lymphomas, on the other hand, present with waxing and waning lymphadenopathy for many years and tend to develop at a prolonged rate, and their treatment might be deemed as not urgent or even not necessary [29]. Despite that, it should be accompanied by a health professional, so its growth is controlled. The most common type of indolent lymphoma NHL is follicular lymphoma [33]. Other types of indolent lymphomas are chronic lymphocytic leukaemia/small lymphocytic lymphoma, and splenic marginal zone lymphoma are considered indolent lymphomas

It is important to note that unrelated to the lymphoma's grade; it could spread to other areas of the human body different from the lymph nodes if not treated or accompanied correctly.

2.1.4 Stage

Both Hodgkin lymphoma and Non-Hodgkin lymphoma use the same system for staging, called Lugano classification [34].

Lugano classification was established in 2011 to set a standard for staging HL and NHL [35]; an abridged version of this standard is found in 2.1, presented below.

Stage	Description
I	There is either one lymph node or one extranodal focus involved
II	Two or more nodes involved (lymph or extranodal) exclusive to one side of the diaphragm
III	Two or more nodes involved (lymph or extranodal) on both sides of the diaphragm
IV	Spread into organs independent from the lymphatic system, no longer related to lymph nodes

Table 2.1: Lugano classification, based on [35]

In addition to the primary stage classifier, in Stage II, the Arabic numerals refer to the number of affected regions, II-2 for instances. On the other hand, in stage III, Arabic numerals are used to divide the stage itself into two. Stage III-1 is associated with the spleen, splenic hilar, celiac, or portal nodes. In contrast, stage III-2 is related to patients whose lymphoma's are located in the iliac, mesenteric nodes, inguinal or paraaortic [35].

The additional modifiers can indicate the symptoms, such as "A" for asymptomatic and "B" for symptomatic; this modifier only applies to HL [34]. The type of first involved where "E" stands for extranodal and "N" for nodal. If there is a tumour with a size bigger than 10 centimetres, it is considered bulky and, as such, classified with an "X" modifier [36].

2.1.5 Aetiology

The precise aetiology of Hodgkin lymphoma is undetermined. Nevertheless, Hodgkin lymphoma is increased in Epstein-Barr (EBV) infection, autoimmune diseases, and immunosuppression. In addition, there is also an indication of familial predisposition in Hodgkin lymphoma. EBV is more frequent in the MCCHL and LdcHL subtypes of Non-Hodgkin lymphoma. Notably, the loss of immune surveillance proposed a possible disease aetiology in EBV-positive diseases.

No other virus has been found to contribute to disease pathogenesis significantly.

Conditions such as hematopoietic cell transplantation, immunosuppression secondary to a solid organ, treatment with immunosuppressive drugs and HIV infection have a greater risk of developing Hodgkin lymphoma. Patients with HIV are more commonly associated with a more advanced stage, rare lymph node sites, leading to worts prognoses. Studies found that same-gender siblings of patients with HL are ten times more likely to develop HL. These conclusions suggest that there is a gene-environment interaction part in the predisposition of Hodgkin lymphoma [37][38][39].

Non-Hodgkin lymphoma are associated with numerous factors, such as environmental factors, states of chronic inflammation, infections, and immunodeficiency. Several viruses have also been attributed to different types of NHL [40]. Some examples are presented below

- Human T-cell leukaemia virus type 1 induces chronic antigenic stimulation and cytokine dysregulation, resulting in unrestrained T-cell or B-cell spur and propagation;
- Helicobacter pylori infection is related to an enhanced risk of gastric Mucosa-associated lymphoid tissue (MALT) lymphomas, a primary type of gastrointestinal lymphoma;
- Epstein-Barr virus, a DNA virus, is the cause of certain types of NHL, such as an endemic variant of Burkitt lymphoma;
- Hepatitis C virus (HCV) leads to clonal B-cell expansions. Diffuse large B cell lymphoma and splenic marginal zone lymphoma are several examples of subtypes of NHL owing to the Hepatitis C virus.

Some drugs are associated with NHL, such as phenytoin, digoxin and TNF antagonist. In addition, hair dye, pesticides, dust, organic chemicals, wood preservatives, phenoxy-herbicides, solvents, radiation exposure and chemotherapy are also linked with the NHL[32][40].

The congenital immunodeficiency states such as Severe combined immunodeficiency disease (SCID), Wiskott-Aldrich syndrome and induced immunodeficiency states are linked with an increased risk of NHL Furthermore, patients with Acquired immunodeficiency syndrome (AIDS), can also have primary CNS lymphoma. Autoimmune disorders similar to Sjögren syndrome, Hashimoto thyroiditis, rheumatoid arthritis, Hashimoto's thyroiditis and Celiac Diseases are similarly connected with a heightened risk of NHL is associated [41].

2.1.6 Epidemiology

Worldwide, 65950 new cases of HL are diagnosed each year, accounting for 0.5% of all cancer diagnoses [42]. Hodgkin lymphoma is an uncommon disease with an estimated frequency ratio of 2.6 cases per 100,000 people in the United States of America. The disease embodies 11% of all lymphomas in the United States of America. In Spain, this disease has an incidence rate higher than that of the world population (2.1 and 2.5 cases per 100,000 person-year in women and men, respectively, in Spain, compared to 0.7 and 1.1 cases worldwide)[43]. Besides, the current Spanish trend is that new diagnoses are increasing (12.5-14.7% more), according to the Spanish Network of Cancer Registries (REDECAN)[44].

It has a bimodal distribution where most affected patients are between ages 20 to 40, and there is another peak from age 55. This type of disease affects more males than females, particularly in paediatrics, where 85% of events occur in young males.

Nearly 70% of incidence in Classic Hodgkin lymphoma subtypes is due to nodular sclerosis cHL, while 25% is down to mixed cellularity cHL. Nodular lymphocyte-predominant Hodgkin lymphoma is responsible for nearly 5% of Hodgkin lymphoma, and less than 1% of overall Classic Hodgkin lymphoma is due to lymphocyte-depleted cHL.

Despite improvements in survival in recent decades, surviving patients with Hodgkin's lymphoma have higher mortality than the average population, even 20 years after diagnosis. In recent decades, the survival of these patients has significantly increased due to the improvement in diagnostic techniques, advances in molecular biology and the development of therapeutic options [45].

This fact leads to a need to monitor the occurrence of late complications, such as the appearance of second tumours or cardiovascular events, with a consequent impact on the morbidity and mortality of these patients5. However, there is little information regarding the impact of new treatments regarding late toxicities, adapted to the patient's risk, with minor extensions in the radiotherapy field and shorter chemotherapy schemes.

Non-Hodgkin lymphoma is common in ages 65 and 74, with the median age being 67 years. Geographically there are variations in the occurrence of individual subtypes. For instance, follicular lymphoma is more frequent in Western countries, whilst T cell lymphoma is more common in Asia. Overall, NHL is the fifth most common diagnosis of paediatric cancer in children below 15 years old. It is responsible for roughly 7% of childhood cancers in the developed world [41].

2.1.7 Symptoms, Diagnosis and Complications

As previously stated, both HL and NHL have similar symptoms such as fever, regular night sweats, sudden loss of weight and an energetic deficiency, alongside the swelling of lymph nodes. A patient with one or more pending symptoms should do preemptive diagnosis examinations to determine if the symptoms indeed lead to lymphoma [46][47].

Hodgkin lymphoma diagnosis is based on routine checkups and medical history of the patient. In case of abnormalities, a lymph node biopsy is made to confirm the existence of cancer. The node selection for the biopsy depends on whether it is nodal or extranodal [48]. It is vital to do a histopathological study to diagnose the lymphoma to the full extent [49].

In contrast, in Non-Hodgkin lymphomas, the diagnosis is made differently. When a patient is suspected of having an NHL, it is crucial to analyse the subtype of the NHL. The subtype of the tumour is determined based on a biopsy of the lymph node, the tissue or both. Nextly the stage of the lymphoma is determined based on results of previous or additional exams, for example, other biopsies. Similarly to HL diagnoses, it is pivotal that histopathological study is made to classify the tumour correctly [50].

Life-threatening emergent complications of NHL ought to be considered during evaluation and the initial workup. Initial detection and prompt therapy are critical for these situations, which may interfere with and delay treatment of the underlying NHL. These can consist of nausea, febrile neutropenia, fatigue, vomiting, hyperuricemia, decreased urination, numbness, and tingling of legs and joint pain, hepatic dysfunction and venous thromboembolic disease[30][32].

2.1.8 Prognosis

Hodgkin lymphoma prognosis depends on prognosis factors. Only the disease stage is considered relevant in risk assessment and stratification. As such, the five-year overall survival for stage 1 or 2a is about 90%, conversely to stage 4 that five-year overall survival of roughly 60%[25].

Non-Hodgkin lymphoma differs primarily depending on histopathology, the magnitude of involvement, and the patient's circumstances. The primary prognostic tool used to determine after a standard treatment the overall survival is called IPI. There are also different variants of this tool that could aid the prognosis phase. IPI can be evaluated by taking into consideration the following factors

- The age of the patient is over 60;
- Values for the serum LDH are more significant than usual;
- The clinical stage is considered stage III or stage IV;
- ECOG's performance status is equal to or greater than two;
- More than one extranodal involvement.

Each of the aforementioned factors is given one point, so the total score reflects the sum of each score on a scale from zero to five. Depending on the IPI score, the NHL in question is considered

- Low risk – adverse factor is between zero and one;
- Intermediate risk – adverse factor is two;
- Poor risk –adverse factor greater than 3.

There is an increased risk for patients with congenital or acquired immunodeficiency states, leading to a poorer response to therapy[32].

There are altered iterations of IPI for most of the NHL, so the prognosis assessment is better. Some examples of this are the Mantle Cell Lymphoma International Prognostic Index (MIPI) for mantle cell lymphoma and Follicular Lymphoma International Prognostic Index (FLIPI) for follicular lymphoma.

A worse prognosis is usually attributed to patients with aggressive T or NK cell lymphomas, whilst patients with low-grade lymphomas have improved survival, which usually can range between 6 to 10 years. Despite that, they can progress into high-grade lymphomas.

2.1.9 Treatment / Management

Treatment of Hodgkin lymphoma depends on the stage of the disease, the presence or absence of prognostic factors and the histologic characteristics [51]. The objective of treatment for patients with Hodgkin lymphoma is to cure the disease by controlling short and long-term problems. The International Prognostic Factors Project for Advanced Hodgkin's lymphoma recognises seven variables for patients with advanced disease:

- Age older than 45 years
- Stage-IV disease
- Male gender
- WBC greater than 15.000/mL
- Lymphocytes less than 600/mL
- Albumin less than 4.0 g/dL
- Haemoglobin less than 10.5 g/dL

Risk stratification categorises patients as low risk or high risk for recurrence. The response to therapy is determined by a PET scan and is used to optimise therapy. Hodgkin lymphoma's initial treatment depends on subgroup treatment.

There are three treatment subgroups: patients with the early-stage disease with favourable prognostic factors, patients with the limited-stage disease who have those with advanced-stage disease and unfavourable prognostic factors. Patients who are within early-stage (stage I to IIA) [52] that have favourable prognostic are treated with a short duration of chemotherapy, typically two cycles of ABVD (doxorubicin, bleomycin, vinblastine, and dacarbazine) followed by restricted involved-field radiation therapy (IFRT) [25].

Patients with the limited-stage disease but with unfavourable features such as bulky mediastinal disease, elevated ESR and extranodal extension are treated with a lengthier course between 4 to 6 cycles of chemotherapy followed by a higher dosage of IFRT.

Advanced-stage (stage IIB to IV) patients are risk-stratified by a distinct scoring method, the International Prognostic Score (IPS). Differing on IPS, several chemotherapy regimens, for instance, escalated BEACOPP and Stanford V, can be used, although the standard of care is ABVD for most patients.

In general, radiation is not favourable for patients. In spite of the elevated cure rate of initial therapy, nearly 10% of patients with Hodgkin lymphoma are refractory to initial treatment. Almost 30% of patients will relapse after an initial complete remission, and relapsed patients usually do Hematopoietic stem cell transplantation.

Non-Hodgkin lymphoma treatment is based on the stage, histopathological features, type, and histopathological features. The most conventional treatment includes chemotherapy, stem cell transplant, radiotherapy, immunotherapy and, in extraordinary cases, surgery. Chemoimmunotherapy, like rituximab combined with chemotherapy, is the most common treatment method.

The primary treatment is radiation for the initial stages, stages I and II. On the other hand, lymphomas with stage II and bulky disease, stage III or even stage IV, are treated with chemotherapy, immunotherapy, targeted therapy, and in some cases, radiation therapy [40].

In general, radiation therapy is usually recommended for:

- Early stages (stages I and II): Radiation therapy is given isolated or accompanied by chemotherapy [53].

- Advanced and aggressive lymphomas: Chemotherapy is the primary treatment. Nevertheless, radiation can be used for palliation, like lymphadenopathy causing urinary/gastrointestinal tract obstruction and pain.

- Radiation therapies have several weeks and are primarily administered five days a week if used as treatment, while palliative radiation is generally shorter [52].

- Surgical intervention may be needed for some NHL patients. In those cases, Hematopoietic stem cell transplant (HSCT) or even a splenectomy [54].

2.1.10 Toxicity and Treatment Side Effect Management

At present, one of the main objectives in managing these patients is to reduce the side effects of the treatment while maintaining its effectiveness. Long-term survivors risk developing certain complications that appear years after treatment, such as cardiac diseases, secondary tumours, hypothyroidism, or sexual dysfunction [55]. The extended follow-up shows that with a mean follow-up of 21 years after treatment, 94% of the patients had at least some toxicity related to the treatment.

- Second neoplasms: Secondary cancers are a normal cause of morbidity and mortality. The most frequent secondary malignancy following treatment of patients with Hodgkin lymphoma is lung cancer. Patients with Hodgkin lymphoma treated with chemotherapy and radiotherapy have a higher risk of developing solid and haematological malignancies than the average population, and these have a worse prognosis. After the relapse of Hodgkin lymphoma, secondary neoplasms are these patients' leading cause of death. The risk of developing a second neoplasm will depend on the dose and field of radiotherapy received, the type of chemotherapy and dose, the age of the patient, the history of smoking and recently, it has also been associated with their genetic predisposition. Radiotherapy is the main factor associated with secondary neoplasms, especially when administered in childhood. However, the risk remains even when low doses of radiotherapy have been used in paediatric patients. Some specific types of secondary neoplasia should be kept in mind: lung cancer presents frequent comorbidity in Hodgkin lymphoma patients, along with breast cancer in women. In addition, other tumours have been described, such as soft tissue sarcomas, bone tumours, papillary thyroid cancer, melanoma and non-melanoma skin cancer, gastrointestinal tumours, and mesothelioma.

- Cardiovascular toxicity: Cardiovascular morbidity and mortality are significantly more critical among survivors of Hodgkin lymphoma than in the general population. Treatment with thoracic radiotherapy is associated with a higher incidence of arrhythmias, myocardial infarction and coronary artery disease, pericarditis, myocarditis, pericardial effusion and tamponade, and sudden death. Anthracyclines are the chemotherapeutic drugs that present the most significant toxicity, manifesting as electrocardiographic changes, arrhythmias or cardiomyopathy leading to congestive heart failure.

- Endocrine toxicity: Thyroid changes affect 50% of patients treated for Hodgkin lymphoma. Hypothyroidism is a frequent complication and accounts for 90% of the thyroid pathology that patients present. The most significant risk of hypothyroidism occurs during the first years after treatment. Risk factors include dose and radiotherapy fields, female gender, older age, combined chemotherapy treatment, and radiotherapy time duration. More than 50% of these patients present subclinical hypothyroidism, detected by elevation of TSH with normal thyroid function. In addition, Gonadal dysfunction may occur in these patients. Patients with Hodgkin lymphoma are usually of childbearing age and will receive treatments that may affect fertility, usually temporarily. Most male patients with advanced stage Hodgkin lymphoma have inadequate semen quality before treatment. Chemotherapeutic agents can also cause female gonadal involvement, which is why cryopreservation is recommended before initiating treatment.

- Infections: Patients with Hodgkin lymphoma are at greater risk of infections due to the deficient immune status they present, which is due to the lymphoma itself and diagnostic and therapeutic processes, such as radiotherapy, chemotherapy, laparotomy and splenectomy staging. This immunodeficiency state occurs even before treatment, and although lymphopenia is present in all stages, it is more frequent in more advanced stages.

- Psychological complications: Anxiety and depression are significant problems in cancer survivors and wives compared to the control population. The prevalence of depression is estimated at 11.6%, and 17.9% of anxiety affects patients and close relatives equally.

For NHL, subsequent complications after treatment vary depending on whether surgery or radiation was used as an additional procedure as well as the chemotherapy used. Some frequent undesirable events caused by the chemotherapy treatment are neutropenic fever, myelosuppression and immunosuppression. Neutropenia can lead to a more considerable risk of infections from fungi, viruses or bacteria. Managing the side effects depends on the degree of neutropenia and if it is febrile.

Several chemotherapeutic agents may provoke vomiting and nausea. Anti-emetic serotonin receptor antagonists and/or benzodiazepines, along with other agents, are ordinarily utilised for treatment and prophylaxis [32][56].

Anthracycline can lead to cardiotoxicity, particularly doxorubicin. In opposition, Dexrazoxane has shown considerable benefits in anthracycline-induced cardiotoxicity. Radiotherapy can also lead to heart failure, but the system differs from chemotherapy since the left ventricular ejection fraction is usually preserved. Long-term fatigue is a recurrent symptom in nearly 66% of survivors of NHL. Fatigue usually improves the year after treatment completion, although many patients continue to experience fatigue for considerable periods after treatment (months/years).

The overall risk of developing second malignancy rises in NHL long-term survivors. The possibility of developing a second malignancy varies depending on the subtype of NHL as well as the treatment received by the patient in question. The threat of developing the myelodysplastic syndrome and acute myeloid leukaemia is elevated. The risk of acquiring lung cancer and cutaneous melanoma was boosted among follicular lymphoma survivors. Patients can manifest squamous cell carcinoma of the head, neck, and breast cancer depending on the area of radiation. hypothyroidism can be caused by the application of radiation to the neck and mediastinum.

NHL survivors risk developing endocrine abnormalities, for instance, gonadal dysfunction and hypothyroidism. Radiation therapy and cytotoxic agents can cause gonadal dysfunction in male and female patients. Consequently, fertility preservation is fundamental and can come as an option for freezing (cryopreservation) of either embryos, oocytes, or spermatozoa.

2.1.11 Relapse

Firstly, it is essential to distinguish refractory from relapsing. Refractory refers to a patient that has undergone through treatment, but the treatment is not considered successful and, as such, is stable but still at large. On the other hand, relapse means that the treatment was considered successful, but six or more months later, the treatment stopped responding to the patient, and the lymphoma reappeared.

Even though relapse in HL is not very common, there is still ten to twenty per cent of the cases that relapse or refractory. High dose chemotherapy is applied following surgical intervention to transplanted autologous stem cells for those cases. If there are further relapses, palliative chemotherapy is made to reduce the size of the HL [57].

Since there are many different types of NHL's, relapse treatments vary. If a relapse occurs, two-thirds of the patients who relapsed are cured by second-line chemotherapy chased by a high dose consolidation [58]. The international prognostic index for B-cell lymphoma helps predict survival based on age, Aan arbour disease stage, number of nodes affected, serum lactate dehydrogenase and haemoglobin level [59], as shown in 2.2.

Feature	**Ranges**
Age	≤ 60 or more
Performance Status	0-1 and 2-4
Lactate dehydrogenase level	Normal and Elevated
Extranodal sites	1 and ≥ 2 extranodal sites
Stage (Aan Arbor System)	I-II and II-IV

Table 2.2: Parameters that help predict relapse, based on [60]

This prediction is validated by the study led by the "German Low-Grade Lymphoma Group" and presented for the comparison value between high-risk group (67%), intermediate-risk(92%) and low-risk(90), a p-value of 0,001 [61].

2.1.12 ECOG Performance Status

The ECOG performance status scale was created with the intent to create a standard that could be used to assess the progression of the disease along with its direct impact on the patients[62][63]. This scale not only describes the patient's physical ability but also the ability to perform self-care and the overall functioning capability of each patient.

This scale to define the performance status of each patient is frequently used in the elaboration of clinical studies and studies of new treatments.

The following table 2.3 presents the scale and classification for the ECOG performance status.

Grade	ECOG Performance Status
0	Completely active, able to continue all pre-disease performance with no restriction
1	Physically restricted in strenuous movement but ambulatory and able to carry out work of a light or sedentary nature
2	Ambulant and with the capability of all self-care but unable to indulge in any work activities; less or equal to nearly 50% of waking hours
3	Limited capability of self-care; confined to bed or chair in excess of 50% of awake hours
4	Completely disabled; cannot continue on any self-care; Bed or chair bound
5	Deceased

Table 2.3: ECOG performance status scale based on [62]

2.1.13 IPI

The IPI or International Prognostic Index is an index designed to estimate a patient's risk depending on the IPI group attributed to them. It was developed to analyse patients with large B-cell lymphomas, a heterogeneous group of the most common type of B-cell lymphoma [64][65].

For each element of the list presented below, the patient manifests the correspondent IPI score of the giving patient is incremented by one.

The final scores are accounted for the classification of the patient according to the IPI.

- Age greater than 60 years old;
- ECOG performance status is greater between 2 and 4 ;
- Cancer is in stage III or IV ;
- LDH is high;
- There is more than one nodal site

Considering the sum of the presented values in the list of parameters above, the patient can be classified into one of four different IPI groups [66]. These IPI groups are divided based on the risk of the patients. Table 2.4 presents the distinct grouping for theInternational Prognostic Index [67].

Number of Risk Factors	IPI Group
0-1	Low
2	Low intermidiate
3	High intermidiate
4-5	High

Table 2.4: IPI grouping according to number of risk factors [68]

MALT-IPI, MIPI and FLIPI are identical indexes focusing on marginal zone lymphoma, mantel cell lymphoma and follicular lymphoma, respectively [69][65].

2.2 Healthcare Data

Before any data analysis to predict the probability of survival, it is vital to have data about the disease itself. Consequently, this section will elaborate on the standards for healthcare data management systems, the privacy and security measures necessary to deal with sensitive data, and the ethics behind them.

2.2.1 Healthcare Data Standards

Data in healthcare have different shapes and sizes, as a medical record could be anything between a paper registry and an electronic report. As a consequence of globalisation and data sharing, standards for health information must be created to help preserve and analyse the data itself. To solve this issue, the Electronic Health Register (EHR), was created in 1960 to help format the healthcare data [70].

There is also an alternative method for having patients have their registry of health-related elements, the Personal Health Register (PHR).

Regardless of the benefits that could accrue from using standards such as EHR, there is still pushback throughout the slow process of converting traditional records to electronic platforms. Some of these barriers might be the non-availability of compatible systems with storing/managing this kind of data or the clinicians' lack of training towards this new technology [71].

2.2.2 Healthcare Data Privacy and Security

Generally, there is a need for carefulness when dealing with personal information regarding health-related information. In that case, the need for security and privacy is of the utmost importance.

This vital ability to have control of the disclosure of the patients' personal medical information is perceived as at risk by people when talking about EHR, making standardising health records even more complicated [72].

To assure the security standard [73], guidelines to access EHR and PHR are strict and should be followed carefully. The patient or tutor determines the accessibility of the personal information if the patient is a minor. Secondly, it is determined to be qualified to examine the data, which could be the clinical staff that accompanies the process or clinicians from the laboratory or the healthcare provider.

Lastly, the information can be accessed for research purposes, anonymously, and there is where this dissertation obtains the data to study and better help the overall general health [74].

2.2.3 Healthcare Data Ethics

As observed in the previous subsections, data management and usage in healthcare can be a delicate problem. Besides the technological preventions presented previously, it is fundamental that there is accountability between the patient and the health institution and between the institution and its partners [75].

The Global Alliance for Genomics and Health (GA4GH), developed the "Accountability policy", which aims to provide a framework to control motoring and any data misuse and endorse a transparent and prompt process to help the utilisation of the data safely and respectfully [76].

2.3 Data treatments in healthcare

Derived from the amount of data influx that occurs from the EHR's there is a need to manage and process the data. Saving and managing data is usually handled by data warehousing, enabling data integration. The data integration, alongside the amount of data and better hardware and networking, makes this environment a great application of Big Data, Artificial Intelligence and Machine Learning to help process and research the objective at hand.

2.3.1 Big Data

Big data is the study of amounts of data so large that the general approach to storage and development is not applicable. To define Big Data as only data size-dependent is not entirely correct since despite being a critical element of Big data is not the only one that defines it [77].

Laney defined three dimensions for big data, more commonly known as the "3 V's of Big Data" [78]. These V's stand for

- Volume – the amount of data dealt with, terabytes to exabytes of ever-increasing information ;
- Variety – the heterogeneous nature of the data, since data can be structured, semi-structured or unstructured, and there is a need to analyse and correlate it;
- Velocity – the necessity to optimise the time of analyses so the data does not lose its value.

As mentioned above, the big data dimensions were not sufficient to describe the scope of big data, so in 2013, Dr Mark van Rijmenam proposed the fourth and fifth dimensions [78] to complement the previously defined V's. These extra dimensions were

- Veracity – that deals with the dependability of the data analysed;

- Value – deals with the extraction of vital and relevant information from the data.

This last dimension is seen as one of the most critical dimensions of big data since if there is no value in the data studied, there is no real purpose in doing so. The new approach of the five V's is represented in the diagram 2.4 below.

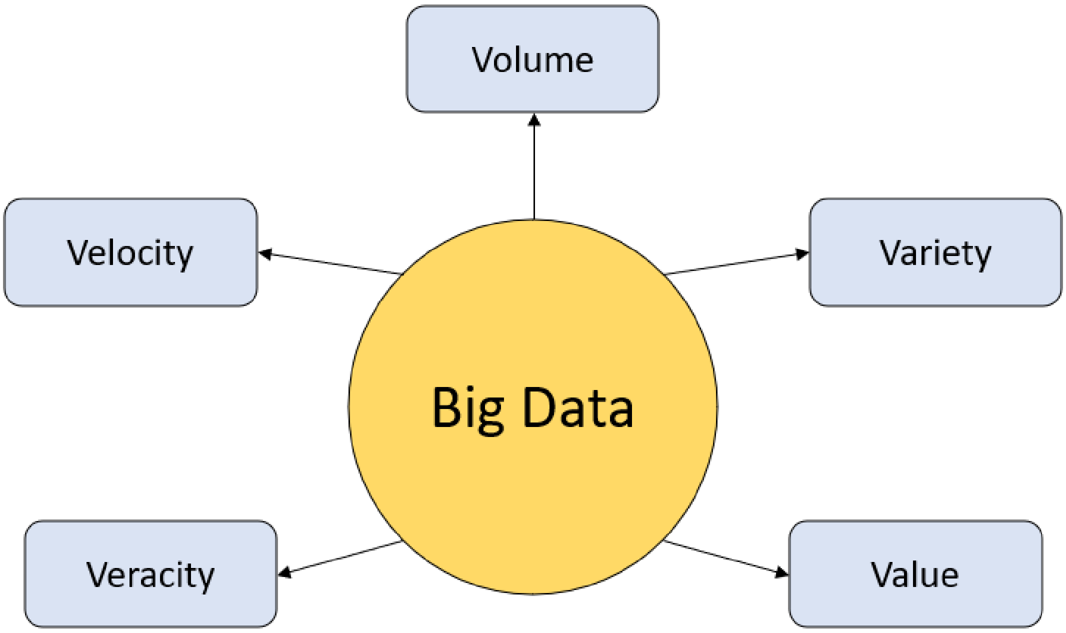


Figure 2.4: The five V's of big data adapted from Dr Mark van Rijmenam's views of Big Data, based on [79]

Despite the five fundamental dimensions presented above, with the evolution of big data, more and more dimensions were added, and some authors defend the 10 V's of Big Data [77]. This constant change of definitions is a testament to the not static world of Big data and its ever-evolving perspective.

These characteristics make big data a great asset in the medical industry since it can develop medicines and help predict a medical condition based on patient history and historical data [80], like the survival probability researched in this dissertation.

2.3.2 Extract-Transform-Load, ETL

Data analyses are usually done on data warehouses, but the data needs to be retrieved from the sources before any analysis. As stated before, the data does not necessarily follow a structured and organised pattern and, as such, need to be treated before entering the data warehouse. This process is called loading, and it consists of consolidating data from different sources into one unique location, be it physical or in the cloud. There are different methods of loading data, of which there will be presented two and one further explained.

Extract-Load-Transform (ELT) is the process where the data is is extracted from the sources and loaded directly to the data warehouse without any previous analysis [81]. This method is considered flexible since all the data extracted from the sources are available in the data warehouse.

In contrast, Extract-Transform-Load (ETL) is when the data is retrieved from the sources and suffers a transformation before being loaded into the data warehouse. The transformation process can also be interpreted as the first approach to clean the data to have better quality data in the warehouse. The transformation step assures the filtering of the data as well as data cleansing and validation. It is also responsible for the correct input for the data warehouse tables [82]. A simplistic view of ETL functions is presented below on the figure 2.5.

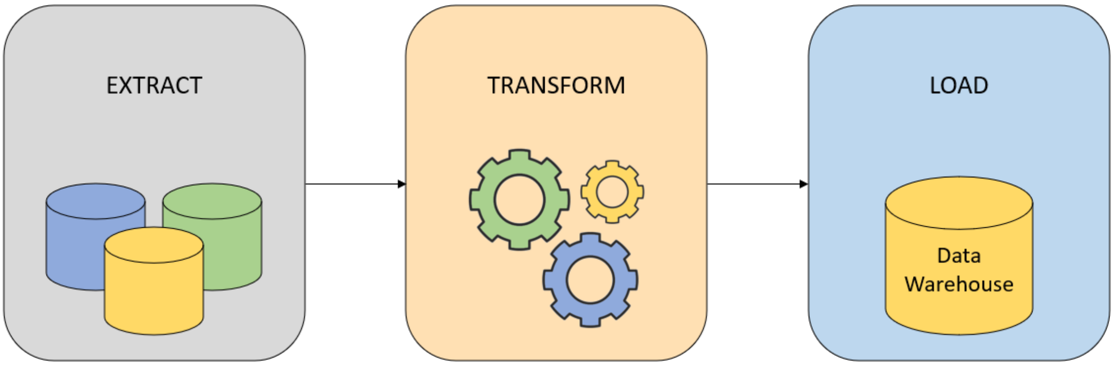


Figure 2.5: ETL method
Based on www.astera.com/wp-content/uploads/2019/07/ETL-e1563879776366.jpg

2.3.3 Relational Database Management System

The received data will be stored in an Relational Database Management System (RDBMS), a relational database management system. This system aims to enhance the method of storing data. Instead of the traditional database management system, the RDBMS aims to give the user an easier-to-understand and more intuitive platform to store the data[83].

The RDBMS stores the given data in interconnected tables to maintain its accuracy and integrity, consistency, and general security.

Some examples of the benefits of using of Relational Database Management System system instead of Database Management System are presented below [84].

- Distributed Databases – allows the implementation of distributed databases which is an impossibility in regular Database Management System (DBMS) ;
- User Quantity – permits more than one simultaneous user ;
- Data Storing Capabilities – RDBMS, unlike DBMS, can handle large amounts of data ;
- Relational model – that permits establishing a relation between tables.

In addition to the examples presented above, RDBMS are significantly more secure since it allows security measures, unlike the regular DBMS [64].

2.3.4 Knowledge Discovery in Databases and Data Mining

Knowledge Discovery in Databases (KDD) is a user interactive and iterative technique that aims to understand possibly valuable patterns in the data at hand. This process is either trivial nor does it guarantee beneficial results, and it may require multiple iterations and different approaches to recognise any pattern in the data used [85]. Despite being a common term, data mining only constitutes part of the more extensive process.

The KDD uses the data in the data warehouse that previously went through an ETL process, which facilitates the KDD's selection, preprocessing, and transformation phase.

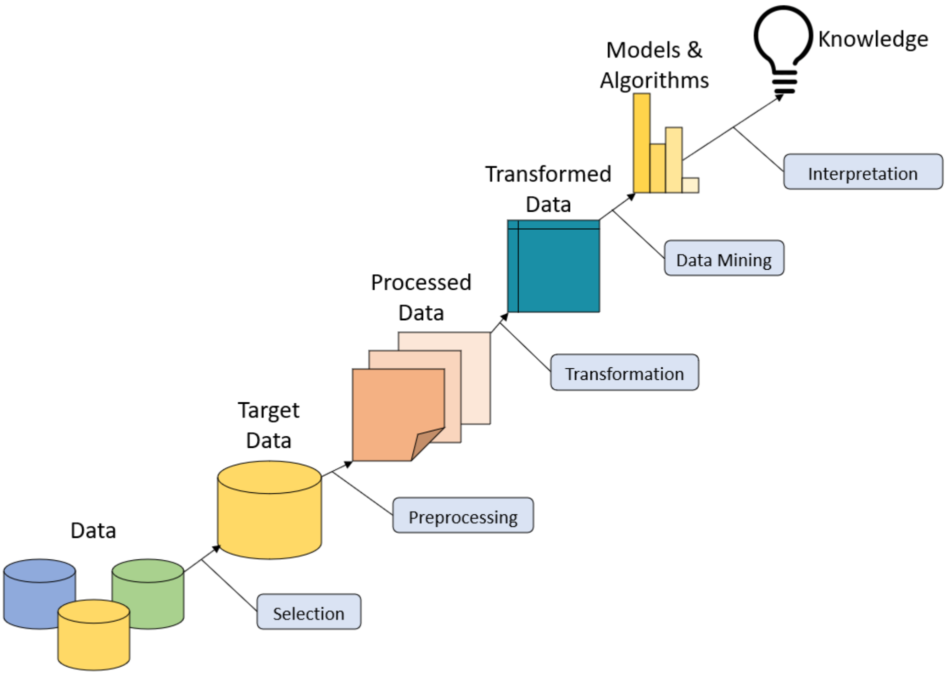


Figure 2.6: Diagram of the KDD method
Based on [85]

The KDD method is divided into five steps, as shown by figure 2.6 above. These steps, as shown above, are not necessarily linear since if there is a need to come back to any of the previous steps, there is that option since any combination of these steps is possible. The five steps that constitute the KDD method are [85]:

1. Selection – choosing the data that will be used to find patterns, can be one data sets multiples or even just sections (in the case of this study, this step is initially done in ETL);

2. Preprocessing – This step is also known as the cleaning of the data. It clears any non-existent values or fills the empty spaces according to the missing value's approach by the user and corrects or eliminates any incongruences of the dataset (in the case of this study, this step is initially done in ETL);

3. Transformation – used as a reducer. Refines the data, so there are only relevant information to avoid problems in the data-mining phase (in the case of this study, this step is initially done in ETL);

4. Data Mining – One of the most prevalent steps for data analysis. This step where the algorithms and study models are applied to the data to see if there are any emerging patterns;

5. Interpretation – Can or cannot be the final step. This step constitutes an analysis of the previously obtained results are prevalent or not to the study at issue.

2.3.5 Types of Streams of Analysis

Before deciding the model to use in a KDD method analysis, it is essential to decide which stream of analysis adequates more to the objective of determining the probability of survival in lymphoma cancer patients. Streams of analyses express the type of the result of the study in question. The four streams of analysis are defined as [77]:

- Descriptive analysis – describes an event and its occurrence in the data studied. This analysis is fundamental for any Big data study to give more knowledge and acuity about the subject. If there is no context to the object studied, there is no way of knowing the study's objective. Uses techniques of data mining or data agglomeration;
- Diagnostic analysis - describes the reason why a determined event occurred. Uses techniques of data mining, decision trees, data discovery and correlations ;
- Predictive analysis – helps to predict an event and the prospect of it occurring. It uses data to fill the knowledge gap with an estimation. Uses regression analysis, multivariable analysis, pattern matching and data-mining;
- Prescriptive analysis – tries to predict the optimal outcome, which means taking into account all the available data to advise and develop a decision of what to do. It also considers the effect of future decisions that could impact the outcome. Uses artificial intelligence, machine learning and data-mining.

Both the predictive and prescriptive analyses require model training and learning algorithms, which will be approached in the following section [77].

As stated in the definitions of the different streams, it is of the utmost importance to start studying the probability of survival in lymphoma cancer patients with a descriptive analysis followed by predictive analysis.

2.3.6 Artificial Intelligence

Artificial Intelligence (AI) is ever more present in everyday life. Due to the complexity of the data in healthcare, it is only natural that AI is starting to be used more commonly to help better solve problems in health.

AI differs from standard programming techniques since, in regular programming, it elaborates a function so the input parameter can output what it wants. On the other hand, AI is given an input and output and, in a "black box" manner, creates a function that could interpolate the input and output. This different perspective is enticing for problems with no certainty for any patterns emerging or effective way to determine them [61].

The diagram presented in figure 2.7 summarises the classification in AI from computer science as a whole.

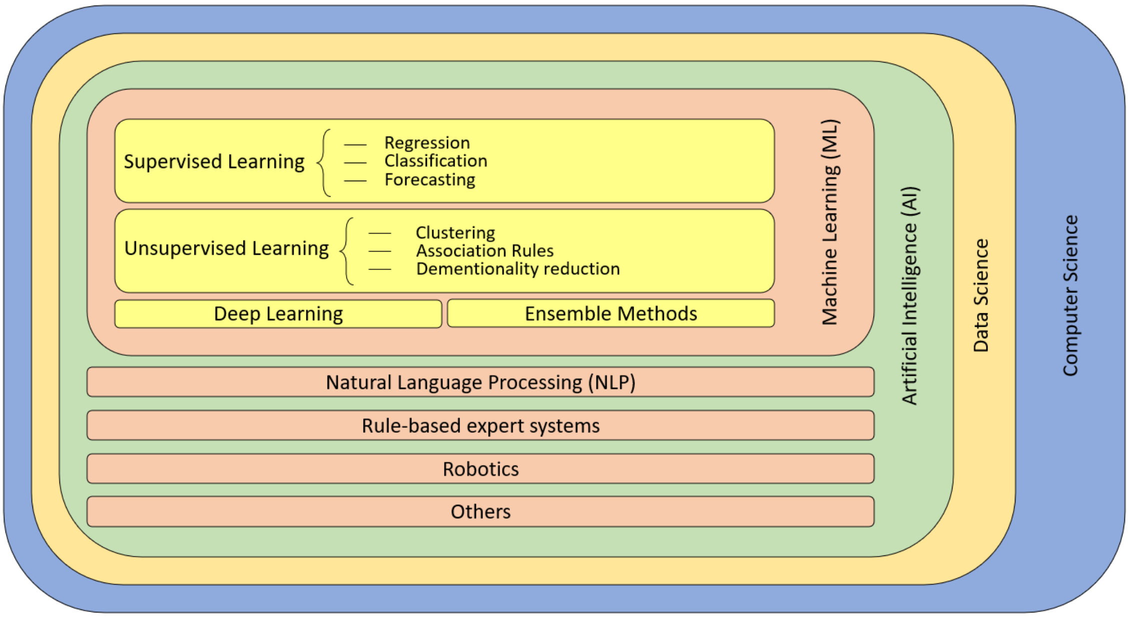


Figure 2.7: Classification of Artificial intelligence
Based on [77]

Artificial Intelligence is considered part of data science, and it englobes a diversity of other subjects, some of which are used in healthcare, such as Natural Language Processing (NPL), which will be explained in the following subsection; NPL that focus on not only understanding the language using text analysis but also translation and speech recognition; Rule-based expert systems that were used as a decision making helping assistance through conditional statements, amongst other areas of AI [86].

2.3.7 Machine Learning

Machine Learning (ML) is considered a subdivision of AI [62]. ML algorithms aim to mimic human decisions with machines; in other words, ML ventures to adapt without being directly programmed to do so. There cannot be any tampering with the ML algorithm to be correctly done since if there is any alteration, the model can be exploited. Machine Learning (ML) is considered a subdivision of AI whose purpose is to study and analyse statistical models and algorithms to imitate human decisions using machines. ML is used as a way to program machines to do specific tasks without explicitly programming them to do so [77][87]. ML algorithms and statistical models are divided into five main categories[77][88] :

- Supervised learning;
- Unsupervised Learning;
- Semi-supervised Learning;
- Reinforcement Learning;

- Ensemble Methods.

This vast field of data science has many different purposes. However, in this dissertation, the use of ML is oriented towards predictive analysis with supervised learning regressions.

2.3.7.1 Supervised Learning

Supervised learning is similar to ordinary human knowledge, where experience gives a better judgment call. Since experience is not a computer ability, this is done with labelled data [89]. The model is trained with this labelled data that is amended if the model's prediction is incorrect; this process occurs until the coveted accuracy is reached and checked posteriorly in the testing phase [77]. Regressions, classifications and forecasting are some examples of supervised learning.

2.3.7.2 Unsupervised Learning

When the data is unlabelled and with large amounts of missing values, the solution uses unsupervised learning. In unsupervised learning, the model interprets the data and finds any correlation or pattern to deduce the results. To help detect patterns, methods like reducing data redundancy, organising the data, and extracting rules can significantly help improve the model [77]. Some examples of unsupervised learning are clustering association rules between others [90].

2.3.7.3 Semi-supervised Learning

Semi-supervised learning combines labelled and unlabeled data that can have a possible desired output. The main objective of semi-supervised learning is to try to understand and correlate labelled and unlabeled data to reach a result [77][91]; this approach can be very time and cost adequate since labelled and sorted data is uncommon and expensive. It is a hybrid between supervised and unsupervised learning.

2.3.7.4 Reinforcement Learning

Unlike the other types of learning, reinforcement learning uses an agent to achieve the desired goals. This dynamic environment uses an algorithm that punishes and rewards based on whether the agent's actions are towards the goal; this process occurs throughout numerous iterations and finishes once the optimal achievable solution is found [77]. It is predominantly used in the robotics field.

2.3.7.5 Ensemble Methods

Unlike the other types of ML, ensemble methods do not focus only on one singular approach to the problem at hand. Ensemble methods retrieve a multitude of models and weight their results to obtain a more robust model [92]. The predominant types of the ensemble are decision trees and random forest, which is closely related to decision trees.

2.3.8 Cloud Computing

For the computing of received data and Big data processing as a whole, there is an eminent necessity for high-power computing machinery that can work dispersedly and quickly, and the solution is Cloud Computing [93].

This solution's extensive resources enable the maintenance and efficiency of storing and elaborating the needed computations to the data in parallel. The execution of these processes in parallel reduces the overall cost of time needed to achieve the wanted results. Cloud computing platforms can be used for different means, from data reception to data management and process execution, giving them the dispersion and versatility that permits a wide range of uses for a large number of different developers [94].

Depending on the service used, the type of integrated cloud can be different and may fit better some user prerequisites than other competing platforms. The service can be constituted by one of three types of cloud, the private cloud, which authorised personnel can only access due to the fact of not being open to the public; the public cloud, which can be accessed by the general population and is managed by the responsible enterprise; and lastly hybrid cloud that is a combination of both of the previous reality by offering some services to the general public and other maintaining private [95].

The platform used in the project is the google based Google Cloud Platform (GCP)[1], which is a public cloud service that offers data management, storage, and application development along with Artificial intelligence and Cloud functions. The variety, in conjunction with the pricing, are the driving reasons to make the platform of choice for medium and minor projects.

2.3.9 React Library

React[2] is a Meta developed open-source JavaScript library that focuses on creating front-end user interfaces on webpages. React is a small library that is simple to use and that, in a not overwhelming way, permits the user to write code analogous to HTML in a JavaScript environment[96].

React is a declarative library which helps in the writing of simpler code as well as more straightforward interpretation by the project developers. It is also a component-based library that encompasses every used element independently, enabling the construction of complex but modular user interfaces [97]. The React library is also an Single Page Application (SRA) which enables the developed applications to be quicker in loading time since only the elements change instead of the entirety of the webpage. Since the SRAs only load the page for a singular time, the overall bandwidth consumption is lower

[1] https://cloud.google.com

[2] https://reactjs.org/

than a conventional website. SRAs also entail better user interference since the overall ambience of the application is more unilateral, creating a better user experience [98].

The React library, as stated, is a small library that does not engulf every single tool usually present in JavaScript frameworks, so the use of complementing tools to achieve the proposed objective is the developer's responsibility.

Regardless of the straining of the tool, this feature is a feature since new tools can be developed at any time, expanding the utility, creativity, and visual elements at a steady rate [99].

All these features work towards a better user and developer environment, making the library in tone and very appropriate for the project at hand.

2.4 Survival analysis methods

This section will present the models used to analyse the survival in lymphoma cancer.

2.4.1 Kaplan-Meier estimator

The survival function empirical estimator calculates the ratio of patients that have survived more than an expected period of time for cases where there is no censorship [100]. The expression of this estimator is present in equation (2.1) below[101].

$$\hat{S}(t) = \frac{Number\ of\ Observations\ >\ t}{n} \qquad t \geq 0 \tag{2.1}$$

On the other hand, in 1958, Kaplan and Meier introduced the Kaplan-Meier estimator, a non-parametric survival function, for censored observations [102]. This estimator is responsible for the probability of survival for a particular time interval.

The Kaplan-Meier estimator is a univariate analysis, only analysing the effect of one variable in survival time at any given time.

The survival function proposed by Kaplan and Meier is presented in the following equation (2.2).

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{2.2}$$

For the aforementioned equation 2.2, *n* is the number of patients, *di* represents the number of dead individuals within the analysed population, and *ni* symbolises the number of patients that are at risk of death. Furthermore, *t(1) to t(i)* signifies the individual death tolls inward of the analysed population [101].

2.4.2 Log-rank Test

As stated in the previous section, the Kaplan-Meier survival function is used to plot the lifetime distribution of multiple groups of patients. This distribution allows the visualization and understanding of the effects of every possible variable on the survival time of the patients. Depending on the results for different values of each variable, the variable itself may be considered significant or not. Determining whether the variable in question is, in fact, of statistical relevance, statistical tests can be done to confirm its usefulness.

The log-rank test is the most common test to determine whether or not different curves have statistical relevance. Instead of proofing the difference between the Kaplan-Meier curves, the Log-rank Test analyses the similarity between each curve [103]. This method helps to determine whether the variable is significant to the survival event.

The Log-rank Test is a test that compares the Kaplan-Meier curve as a whole. This test consists on a chi-square large-sample test that analyses the observed number of patients versus the expected number of patients. As the log-rank test tests the null hypothesis of similarity for over two groups of patients, $g \geqslant 2$, the log-rank statistics is given by the following equation (2.3) [104].

$$\sum_{i}^{g} \frac{(O_i - E_i)^2}{E_i} \overset{a}{\sim} x_{g-1}^2 \tag{2.3}$$

Where O_i represents the number of the observed events, and E_i represents the number of expected events for each group of the overall sum.

The log-rank test is powerful when analysing non-proportional curves as long as the curves do not cross one another. If the curves cross, the log-rank test cannot detect significant differences between curves. Consequently, it is of utmost significance to do the log-rank test to check the statistical independence of the Kaplan-Meier curves to determine whether or not they could be used to attain information relevant to the survival of the patients[101].

2.5 Related Work

Before developing work on any given area, it is fundamental to study the forefront technology and achievements of that area in question. Consequently, this subsection is responsible for presenting the avant-garde of the related projects with a similar base or objective to the dissertation.

The work elaborated is inserted in the CLARIFY project as previously stated and follows the previously elaborated work on long-term survival in the early stages of non-small-cell lung cancer patients [105]. Whilst the cancer type and overall conclusion differ from the precedent work. It is essential to note the study area and methods utilised to optimise and grasp the complete project in the best way possible.

More specifically, in the area of the study of survival prediction in lymphoma cancer, there are numerous studies from a medical point of view regarding the survival in lymphoma cancers, but as to be expected, not on the probability of long-term survivability.

Regarding predicting the survival in cancerous patients, there is a panoply of studies that can be correlated to the objective of the dissertation at hand. In 1998 Abrey, DeAngelis, and Yahalom performed a long-term survival analysis for patients of central nervous system lymphoma as well as treatment-related toxicities for the same specific lymphoma [106]. In conjunction, in 1981, Fisher et al. studied the prediction of long-time survivability for diffuse mixed lymphomas using Cox regression [107]. More recently, in 2008, Schulz et al. presented a study on the treatment of Hodgkin lymphoma patients with Rituximab using the Kaplan-Meier estimator to study the survival probability of the analysed patients [108].

This precedent created in the field of study by the aforementioned studies and the lack of identical studies that engulf the analysis of both Hodgkin and non-Hodgkin lymphomas in the area creates the perfect scenario for the writing of this dissertation.

3

Data Managing and Initial Analysis

The third chapter portrays the format of the received data from the project partner as well as the tools used to analyse and retrieve knowledge from the data itself. In addition, this chapter will also present how the conclusions postulated in future chapters will be presented in the final form to the CLARIFY project itself, along with detailed reports done along the process.

3.1 Data Reception and Description

This subsection aims to represent the overall data received from the project partners and an overview of the content.

3.1.1 Data Reception

Firstly, there is a need to have the data to analyse. As such, the format of the received data and the method of storing and managing it is important to establish, so the process of analysing, filtering, and cleaning the data is streamlined.

These preparation and precautions are critical when dealing with medical records. Such is the heterogeneity and sensitivity of the data that it is essential to analyse and interpret the format and condition of the data relayed for analysis.

The data analysed throughout the dissertation belongs to two different datasets, one for each type of lymphoma, Hodgkin and Non-Hodgkin, composed of patients treated at Puerta de Hierro Majadahonda University Hospital.

The data for both lymphomas typology arrived in one of two formats, comma-separated values, CSV, a commonly used standard for exporting reading data, or in an Excel open XML spreadsheet, XLSX. The data was accompanied by a dictionary that defined the mapping and meaning of each variable. In addition to mapping, there is a need to analyse essential fields such as date of birth or death and other vital fields to ensure the overall dataset's congruence. This process is prolonged and will keep the data accurate across the board.

An identification number not related to personal information was used to refer patients in a phase of analysis of the related content.

3.1.2 General Data Description

As stated in the previous subsection, the data was received as two different datasets, one for Hodgkin lymphoma and the other for Non-Hodgkin lymphoma. Despite their differences, there is a large majority of similarly provided variables to analyse, and there are also some differences and exclusivities regarding particular descriptive variables and conditions.

This subsection will not only present and explain the different types of variables grouped up by subject matter but also explain the needed mapping and consequent categorisation of themselves.

The division of variables by their content is:

- Demographics – general information of the patient's overall description;
- Antecedents – previous diseases and personal history;
- Diagnosis – the subtype of cancer of the patient and respective tests and blood test results;
- Treatments – kind of treatment aplied to the patients;
- Relapse – Patients' relapse if occurred;
- Late Toxicities and pathologies – toxicities due to treatment and pathologies of the diseases;
- Second Tumours – typology of the second tumour;
- Molecular Biology – results of the studies performed;
- Transformation – Non-Hodgkin lymphoma exclusive; details of the possible transformation of the diagnosed NHL lymphoma.

The Hodgkin lymphoma dataset comprises 382 patients, of which 155 are women and 227 are men, whilst the Non-Hodgkin lymphoma dataset comprises 562 patients, of whom 244 are women and 318 are men.

3.2 Data Preparation and Data Handling Used Frameworks

3.2.1 Data Preparation

Both the Hodgkin lymphoma and the Non-Hodgkin lymphoma datasets are vast and complex and, as such, are both in dire need of data treatment. The HL dataset had 1314 variables to analyse, while the NHL had 1408 variables. This large amount of variables is due to the possible fifteen lines of treatment that would be possible to analyse. A large amount of variables makes not only the data preparation as well as the data usage even more complex and intricate.

Consequently, the medical team that accompanied the project and, under their expertise, decided on a subset of variables considering the study's objective. The choices were based on medical knowledge by professionals in the area studied. Therefore with the intention of extracting knowledge from the data received, the dissertation will only present the treatment done to the subset of variables deemed as influential and crucial by the medical specialists.

Despite that, all the different variables, even those not included in the dissertation, were correctly treated and stored in the databases with the objective of if, in the future, there is a need to analyse any other variable other than the subset analysed during the elaboration of the dissertation the bulk of the work is already completed. This complete analysis also applies to the following subsection despite only the chosen variables being portraited.

The data mapping is grouped by the different areas of study as stated in the previous section, and each section contains the used variables in the study.

Since preparing the data is supposed to simplify the received data and make it more workable, removing patients is majorly unnecessary. Despite that, and as a consequence of incongruencies in some patients, two Non-Hodgkin lymphoma patients were removed due to the lack of data.

3.2.1.1 Demographics

All the demographic variables are the same for both types of lymphomas and are presented in table 3.1.

	Variable Values
Gender	0, Male 1, Female
Smoker	0, Non smoker - Less or equal to 100 cigarettes in the patients lifetime 1, Previous smoker - Stopped for more than 1 year 2, Current smoker 3, Unknown
Living Situation	1, Alive without disease 2, Alive with disease 3, Dead 4, Lost follow-up
Cause of death	1, First tumor 2, Second tumor 3, Other
Date of Birth	Datetime
Age Survival	Interval of time between Date of Birth and Date of Death
Date of Death	Datetime
Survival Days	Interval of time survived in Days
Survival Months	Interval of time survived in Months
Survival Years	Interval of time survived in Years

Table 3.1: Demographic variables and correspondent mapping

3.2.1.2 Antecedents

Table 3.2 beneath expresses the variables that represent the antecedent of each patient. The antecedent is defined as any previous history of the patient.

	Variable Values
History Rheumatic Diseases	1, Yes 0, No -1, Unknown
Rheumatic Disease Type	Sjoegren syndrome, Rheumatoid arthritis, Others, Lupus, Unknown
Personal History Type	No personal history, Hyperthension, Others, Dyslipidemia, Tuberculosis, Diabetes, Hepatitis, Nephropathy, Peripheral arterial disease, Deppresive syndrome/anxiety, Former alcoholism, Asthma, Obesity, Epoc, Heart disease
Studies Performed Type	Hiv, Evb lgm, Hbv hbsag, Ebv lgg, Hcv, Cmv lgg, Hbv hbsac, Hbv hbctotal, Cmv igm, Others, Unknown, No studies performed

Table 3.2: Antecedents variables and correspondent mapping

3.2.1.3 Diagnosis

Table 3.3 below states the variables that describe the diagnosis made to the patient when the lymphoma was detected. It also presents some results to test that helped the diagnosis.

It is crucial to note that some of the variables presented are exclusive to one of the types of lymphoma. This characteristic is shown by the labels adjacent to the variables.

			Variable Values
Date of Diagnosis			Datetime
Age Diagnosis			Interval of time between Date of Birth and Date of Diagnosis
History of Rheumatic Diseases			1, Yes 0, No -1, Unknown
Simple Initial Stage			I, Stage I II, Stage II III, Stage III IV, Stage IV Other Unknown
Histology	NHL	B Cell Low Grade	1, Lymphoplasmacytic lymphoma 7, Splenic marginal zone B-cell lymphoma 8, Follicular lymphoma 11, Small cell linfocitic lymphoma 14, Monocytoid B-cell lymhoma 15, Extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue (MALT) 21, Nodal Marginal Zone Lymphoma 0, Other -1, N/D
		B Cell High Grade	2, T-cell/histiocyte-rich large B-cell lymphoma 3, Burkitt's lymphoma 4, Primary effusion lymphoma 6, Diffuse large B-cell lymphoma 10, Lymphoblastic B-cell lymphoma 12, Mantle cell lymphoma 13, Primary mediastinal large B-cell lymphoma 18, High Grade B Lymphoma: NOS 19, Plasmablastic Lymphoma 20, Primary large B-cell lymphoma of the CNS 0, Other -1, N/D
		T Cell	1, Chronic lymphocytic leukemia of T-cells 2, Adult T-Cell Leukemia/Lymphoma 3, Angiocentric lymphoma 4, Anaplastic T-cell lymphoma 5, Linfoblastic T-cell lymphoma 6, Subcutaneous panniculitic T-cell lymphoma 7, Angioimmunoblastic T-cell lymphoma 8, Gamma / Delta T-cell lymphoma 9, Monomorphic epitheliotropic intestinal T-cell lymphoma 10, Peripheral T-cell lymphoma, NOS 11, T-cell large granular lymphocytic leukemia 12, Aggressive NK-cell leukemia 13, Primary cutaneous T-cell lymphoma 14, Enteropathy-associated T-cell lymphoma 15, Hepatosplenic T-cell lymphoma 16, Mycosis Fungoides 17, Sézary syndrome -1, N/D 0, Other
	HL		1, Mixed Celularity 2, Lymphocyte Depletion 3, Nodular Scerosis 4, Lymphocyte Predominance 5, Rich in Lymphocytes 0, Not specified
Grade Lymphoma	NHL		1,Low 2,High 3,Very High -1,Unknown
Primary Organ Type			Bone, Lung,Liver, Central nervous system, Gastric, Intestinal, Head and neck, Spleen, Breast, Bone marrow, Testicular, Other, Unknown

Table 3.3: Part 1, Diagnosis variables and correspondent mapping

	Variable Values
Extranodal Involvement	1, Yes 0, No -1, Unknown
Extranodal Involvement Type	Bone, Bone marrow, Breast, Lung, Liver, Intestinal, Spleen, Gastric, Testicular, Head and neck, Central nervous system, Unknown, Other
Biopsy	1, Yes 0, No -1, Unknown
Biopsy Result	1, Positive 0, Negative -1, Unknown
IPI	0 1 2 3 4 5 6,Other -1,Unknown
Beta 2	0, Normal 1, High -1, Unknown
Performance Status	0 1 2 3 4 -1,Unknown
Altered Enzymes Type	GPT GOT GGT Alkaline phosphatase Bilirrubin Unknown
Bulky	1, Yes 0, No -1,Unknown
White Blood Cell Count	0, Normal 1, Low 2, High -1, Unknown
Lymphocyte Count	0, Normal 1, Low 2, High -1, Unknown
LDH	0, Normal 1, High -1, Unknown

Table 3.4: Part 2, Diagnosis variables and correspondent mapping

3.2.1.4 Treatments

Table 3.5 underneath indicates the variables that express the treatment that each patient could receive. As mentioned in the previous section, the dataset contains 15 lines of treatment. As such, table 3.5 below only represents the first one since the comprehensive mapping of each line of treatment is identical.

	Variable Values
Treatment Received	1, Yes 0, No -1,Unknown
Number of treatments	1, 2, 3
Type of Treatment	Monoclonal Antibodies, Surgery, Clinical Trial, Immunotherapy, Chemotherapy, Radiotherapy, Targeted Therapy, Watch and Wait, Others
Profilactic treatment during chemotherapy/tergeted therapy?	1, Yes 0, No -1,Unknown
Radiotherapy Location	Infradiaphragmatic, Supradiaphragmatic, Others, Unknown
Type of Radiotherapy	2, Local 3, Total 4, Subtotal -1, Unknown
Response to first treatment	1, Complete 2, Stable 3, Partial 4, Progression 0, Unknown
Type of subsequent gastrointestinal tumor	1, Colorrectal 2, Gastric 3, Esophagus 4, Pancreatic 5, Rectal 6, Others
Developed toxicities caused by treatment?	1, Yes 0, No -1, Unknown
Tranfusions Required	1, Yes 0,No -1,Unknown
Type of tranfusion	Blood, Plateletes

Table 3.5: Treatment variables and correspondent mapping for performed studies

3.2.1.5 Relapse

The variables that constituted this category express the location of the lymphoma relapse and the confirmation that the relapse occurred in the exact location as the original tumour.

	Variable Values
Number Relapse	0 to 9 Relapses
First Relapse Location	Disseminated, Infradiaphragmatic, Supradiaphragmatic, Other, Unknown
Previous Disesase in First Relapse Location	1, Yes 0, No -1, Unknown

Table 3.6: Relapse variables and correspondent mapping

3.2.1.6 Late Toxicities and pathologies

Due to the treatment performed on patients, some secondary effects such as late toxicities and pathologies can appear in patients deemed as treated. This subsection presents the possible toxicities and pathologies that a given patient may manifest.

	Variable Values
Treatment Associated Late Toxicity	1, Yes 0, No -1, Unknown
Treatment Caused Toxicity Type	Unknown, Chemotherapy, Radiotherapy, Monoclonal antibodies
Treatment Associated Late Disease	1, Yes 0, No -1, Unknown
Treatment Caused Disease Type	Renal failure, Neurological , Cardiovascular, Endocrine, Digestive, Unknown, Others
Treatment Caused Cardiovascular Disease Type	Congestive heart failure, Coronary syndrome/heart attack, Valvular disease, Unknown, Others
Treatment Caused Pulmonary Disease Type	Pneumonia, Pneumonitis, EPOC, Others
Treatment Caused Digestive Disease Type	Intestinal Ischemia, Malabsorption, Colitis, Gastrointestinal Bleeding, Intestinal Obstruction, Other
Treatment Caused Renal Disease Type	Renal failure , Cystitis, Other
Treatment Caused Infectious Disease Type	CMV, Fungi , Bacterial , Varicela Zoster Virus, Tuberculosis, Febrile Neutropenia, Other

Table 3.7: Late Toxicities and pathologies variables and correspondent mapping

3.2.1.7 Second Tumours

This subsection sheds light on the type of second tumours that could affect both HL and NHL patients.

	Variable Values
Tumour previous to lymphoma	1, Yes 0, No -1,Unknown
Type of previous tumor	Breast Cancer, Lung Cancer, Gastrointestinal Tumor, Head and Neck Tumor, Thyroid Tumor, Genitourinay Tumor, Second Lymphoma, Sarcoma, Carcinoma of Unknown Origin, Cutaneous,Hematological, Other
Type of previous gastrointestinal tumor	1, Colorrectal 2, Gastric 3, Esophagus 4, Pancreatic 5, Rectal 6, Others
Type of previous head and neck tumor	1, Cavidad oral 2, Central nervous system 4, Oropharynx 5, Hypopharynx 6, supraglottis 7, Glottis 8, Subglottis 3, Others
Type of previous genitourinary tumor	1, Cervical 2, Endometrial 3, Ovarian 4, Prostate 5, Bladder 6, Renal 7, Others
Second Tumour After Diagnosed Lymphoma	1, Yes 0, No -1,Unknown
Type of subsequent tumor	Breast Cancer, Lung Cancer, Gastrointestinal Tumor, Head and Neck Tumor, Thyroid Tumor, Genitourinay Tumor, Second Lymphoma, Sarcoma, Carcinoma of Unknown Origin, Cutaneous,Hematological, Other
Type of subsequent gastrointestinal tumor	1, Colorrectal 2, Gastric 3, Esophagus 4, Pancreatic 5, Rectal 6, Others
Type of subsequent head an neck tumor	1,Cavidad oral 2, Central nervous system 4, Oropharynx 5, Hypopharynx 6, supraglottis 7, Glottis 8, Subglottis 3, Others
Type of subsequent genitourinary tumor	1, Cervical 2, Endometrial 3, Ovarian 4, Prostate 5, Bladder 6, Renal 7, Others

Table 3.8: Second tumours variables and correspondent mapping for performed studies

3.2.1.8 Molecular Biology

This section expresses the tests possibly partaken by each patient and the corresponding test results.

	Variable Values
Studies performed: BCL_2, BCL_6, CD3, CD4, CD5, CD7, CD8, CD10, CD20, CD30, CD56, Ebers, Ki67, MUM_1	1, Positive 0, Negative -1,Unknown

Table 3.9: Molecular Biology variables and correspondent mapping for performed studies

	Variable Values
Studies results: BCL_2, BCL_6, CD3, CD4, CD5, CD7, CD8, CD10, CD20, CD30, CD56, Ebers, Ki67, MUM_1	1, Positive 0, Negative -1,Unknown

Table 3.10: Molecular Biology variables and correspondent mapping for performed studies

3.2.1.9 Transformation

The transformation category is exclusive to Non-Hodgkin lymphoma, expressing the possible transformation the NHL could suffer.

	Variable Values
Transformation	1, Yes 0, No -1, Unknown
Debut Tranformation	1, Normal medical examination 2, New symptoms

Table 3.11: Transformation variables and correspondent mapping

3.2.2 Frameworks for Data Handling

This subsection presents and explains the frameworks used to handle the data from both datasets and future use for any of the given frameworks.

The process of not only storing, importing data and manipulating is a complex system. The database was created in the MySQL framework whilst loading the data using Pentaho. The data manipulation and mapping, as well as the calculation for the initial statistical analysis and other analyses, used Python as a base and various Python libraries.

3.2.2.1 Python

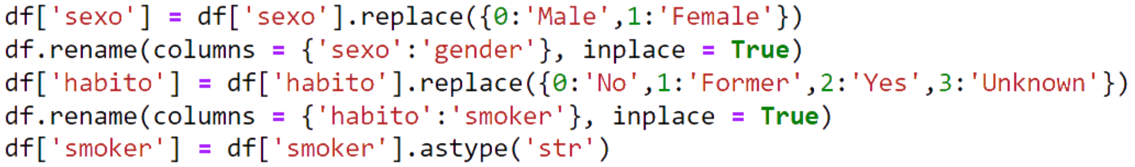
A large portion of the developed code is based on the Python language. This subsection will focus on the treatment work performed to both datasets before exporting them to the respective databases and the libraries used in the subsequent Python applications.

The following were the two main Python libraries used for the initial mapping and filtering to insert the data into the respective databases.

- Python Pandas [1] – an open source library whose objective is to provide data structures and analyses to facilitate the work when dealing with large amounts of data;

- Numpy [2] – also an open source library which is a package that helps computer science realisation on Python can be used in combination with Python pandas for instances but can also be used individually.

With the help of the aforementioned libraries and the provided dictionaries, each dataset variable was correctly mapped. The figure below represents an excert of the procedure necessary to completely match the data with the dictionary values.



```
df['sexo'] = df['sexo'].replace({0:'Male',1:'Female'})
df.rename(columns = {'sexo':'gender'}, inplace = True)
df['habito'] = df['habito'].replace({0:'No',1:'Former',2:'Yes',3:'Unknown'})
df.rename(columns = {'habito':'smoker'}, inplace = True)
df['smoker'] = df['smoker'].astype('str')
```

Figure 3.1: Example of the variables' mapping

All this populating of the dictionary's values helps interpret the forward calculated results, presenting a coherent and easier-to-interpret set of results and variable values.

Addedly, the division of the original data into three smaller datasets is performed using the pandas' library. The original datasets are divided not only due to a restriction on the database engine used ahead but also to facilitate the manoeuvring and manipulation of the general data.

On top of this, Python is also used to create the graphics in the following subsection and to develop the Kaplan-Meier curves and cloud functions presented in the following chapters.

The used libraries for further development explained in the upcoming chapters are hereinafter.

- MatplotLib [3]/Seaborn [4] – open source libraries that focus on constructing graphics and their appearance;

[1] https://pandas.pydata.org/docs/
[2] https://numpy.org/
[3] https://matplotlib.org/
[4] https://seaborn.pydata.org/

- Scikit Learn [5] – machine learning open source library contains the creation, fitting and validation of machine learning models ;
- Lifelines [6] – survival analysis library that enables the creation of statistical survival analysis;
- Docx [7] – library to create Microsoft Word documents.
- XlsxWriter [8] – library to create Microsoft Excel tables and documents.

In addition, complementary libraries to the previously presented were used to achieve the capabilities of the libraries mentioned above.

The wide availability of resources and libraries makes Python one of the most supported coding languages when dealing with statistical analyses, data manipulation, and widespread general use. This environment helps in the creation of models making them more accessible and with further documentation on the different subjects.

3.2.2.2 MySQL

After the initial analysis and mapping using Python, as present in the previous subsection, the treated data needed a database in which to be stored.

Rather than saving the dataset in a similar format as it was provided, the data was stored in a MySQL database. MySQL database is an RDBMS and, as such, is a more advanced and refined management system than the DBMS.

The MySQL RDBMS is tabular, allows multiple accesses simultaneously and can establish relationships between the stored data. RDBMS also provide multiple data security levels, which is fundamental to health data since the information stored is very sensitive.

In addition to the previously given advantages of RDBMS, the higher versatility of a database along with the search to maintain data integrity and the possibility of simultaneous cooperative work to be developed in the database in question, making the relational database an obvious choice to store the data. One final reason to transfer the data into a database is to have the ability to execute queries on the saved data, which simplifies the search and filtering of pretended data.

Due to the limitation of the engine used for the MySQL database, InnoDB and easier manipulation of the data, each original dataset was divided into three separate tables interconnected by their primary key.

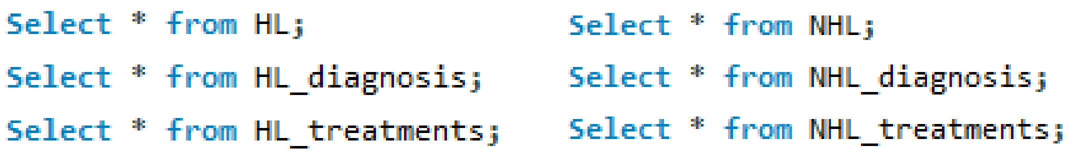
The data division separates it into either treatment table, diagnosis table and a table that engulfs the remaining categories. The figure 3.2 below represents examples of possible queries for all three tables for each dataset.

[5] https://scikit-learn.org/stable/

[6] https://lifelines.readthedocs.io/en/latest/index.html

[7] https://python-docx.readthedocs.io/en/latest/

[8] https://xlsxwriter.readthedocs.io/



```
Select * from HL;
Select * from HL_diagnosis;
Select * from HL_treatments;
```

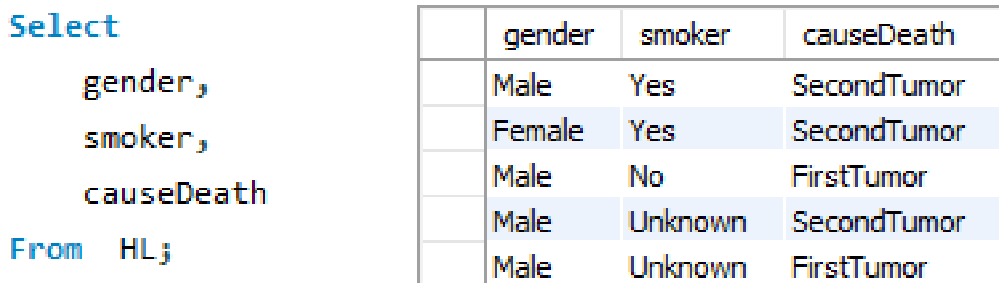
(a) Hodgkin Lymphoma

```
Select * from NHL;
Select * from NHL_diagnosis;
Select * from NHL_treatments;
```

(b) Non-Hodgkin Lymphoma

Figure 3.2: Examples of queries to the MySQL database

Another advantage of the ability to query the data is the size wanted result. Since not every variable is constantly represented, the result is quicker to be obtained, easier to interpret and lighter to send as JSON for instances.



```
Select
    gender,
    smoker,
    causeDeath
From  HL;
```

(a) Executed query

	gender	smoker	causeDeath
	Male	Yes	SecondTumor
	Female	Yes	SecondTumor
	Male	No	FirstTumor
	Male	Unknown	SecondTumor
	Male	Unknown	FirstTumor

(b) Result of the query

Figure 3.3: Example of a restricting query and its result

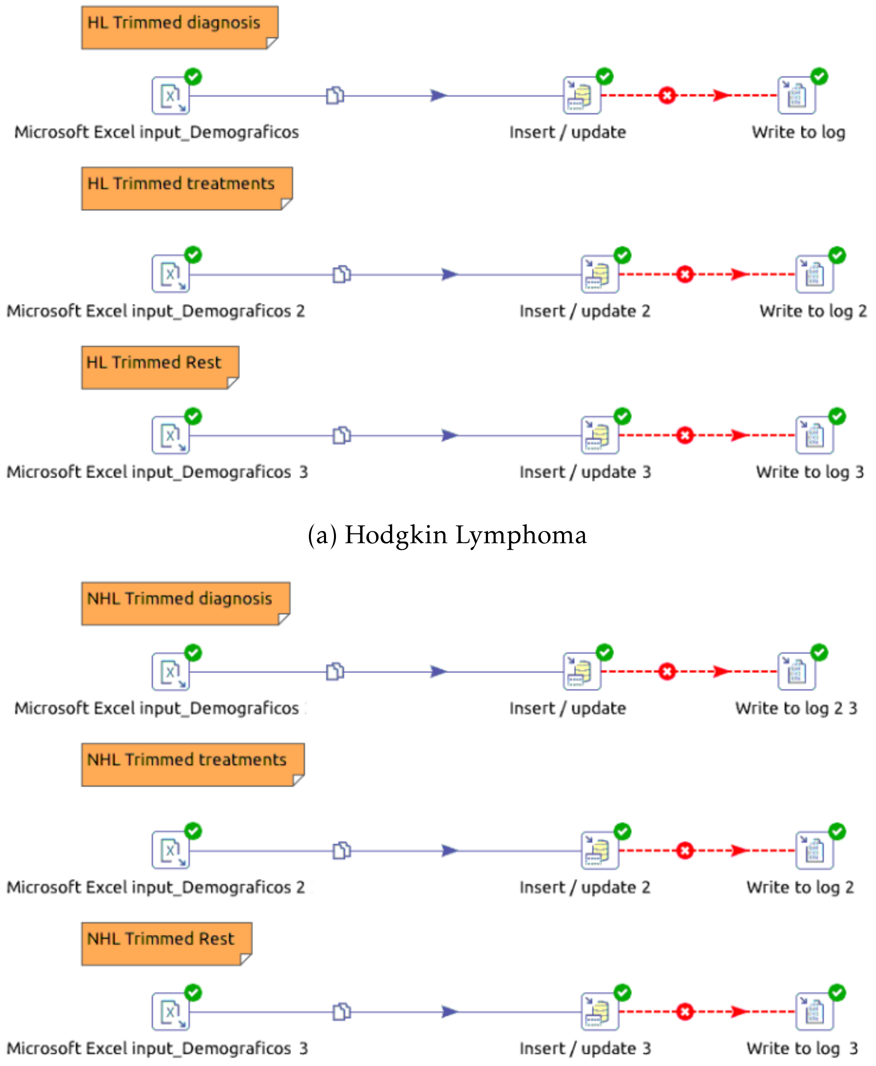
The above figure 3.3 represents the possibilities of restricting the obtained data as well as the obtained result.

3.2.2.3 Pentaho Data Integration

The Pentaho data integration framework was used to load the previously created relational databases with the processed data. The framework was chosen based on the low integration time and the vast versatility concerning database access.

The choice of using Pentaho is also supported by its parallel execution capability that furthers the process's optimisation.

As the Load in the ETL process, two different transformations were created to achieve the data loading, one for each type of lymphoma. The transformations are presented below in image 3.4a in the case of HL and image 3.4b in the case of NHL.



(a) Hodgkin Lymphoma

(b) Non-Hodgkin Lymphoma

Figure 3.4: Age since diagnosis bar plot

As previously stated, the files used as input were altered versions of the original dataset that were divided because of the large number of variables each dataset contained.

The use of Pentaho, a parallel loading to multiple tables of the database, optimises the time spent in the process. This optimisation also means that if there is a need to update or reload any specific tables, that could be achieved without overwriting the entire database.

3.3 Descriptive Statistical Analysis

This section illustrates the initial descriptive data analysis coupled with the tools created to help elaborate and validate this analysis. The initial analysis variables were selected per the medical expert's decision based on medical information along with perceived statistical relevance.

This list of variables is a shortened version of the presented variables in the previous section.

The descriptive analysis' elaboration aims to present an overall look at the studied data along with the properties of each studied variable.

In addition to the univariate initial descriptive analysis, this section also contains an overall age demographic analysis with barplots of the time since diagnosis and survival years, along with scatter plots that describe the data dispersion of the analysed phenomenon.

It is also essential to note that all the graphics presented in the current section are colour matched to each cancer's ribbon colour, which is lime green for Hodgkin Lymphoma and violet for Non-Hodgkin Lymphoma. Furthermore, it is of the utmost importance that the different types of lymphoma are not compared since, as stated previously, the cancer subtypes are not comparable and are considered independent.

3.3.1 Age dispersion using barplots

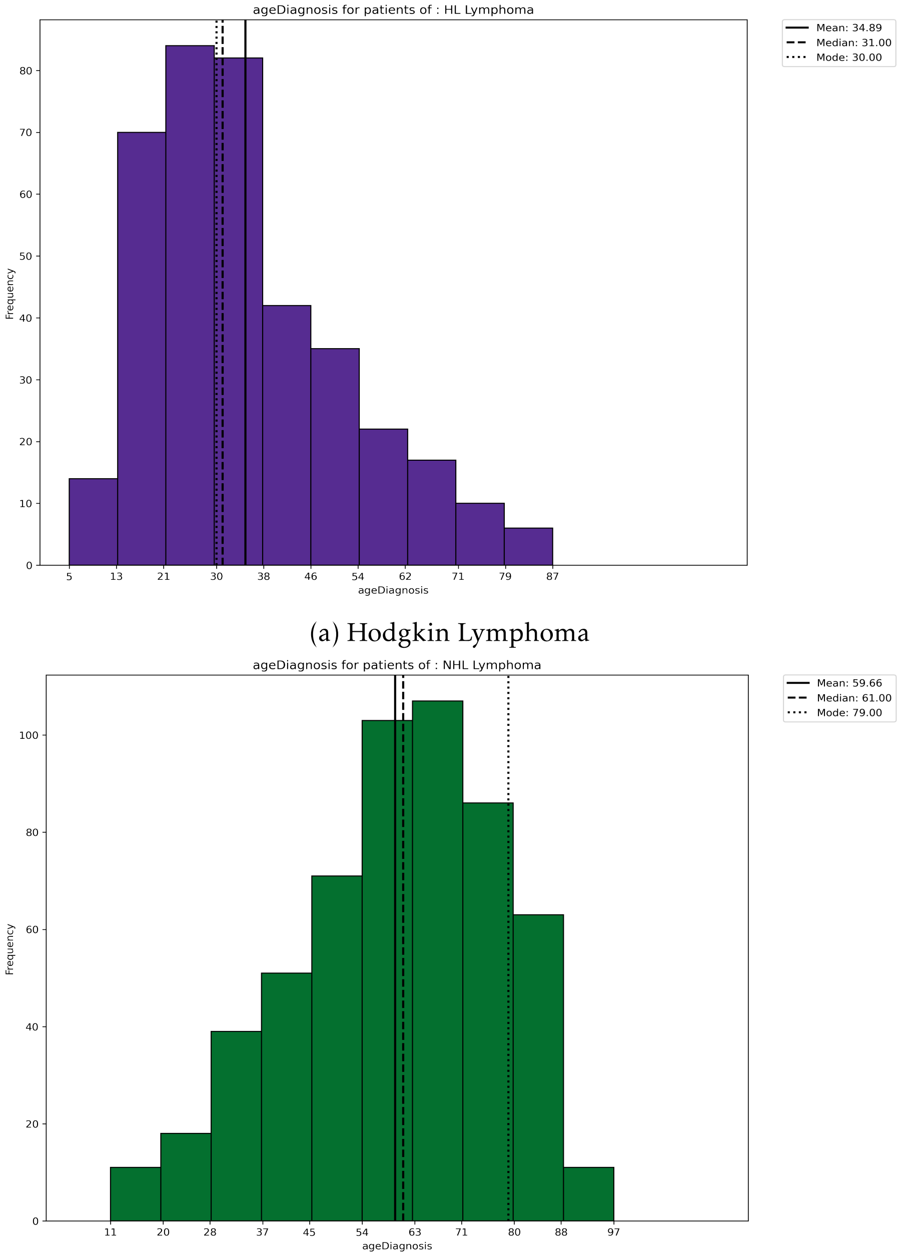
To analyse the age dispersion in both datasets, the use of barplots was the most appropriate graphical representation to incorporate the time since diagnosis, described by the variable "ageDiagnosis" and the variable that describes the time of survival in years "survivalYears" are numeric variables as such.

As both describe a period between dates, the possible values for either variable is a constant range starting from zero.

A histogram is one of the most significant ways to represent a numeric continuous variable distribution and demonstrate significant amounts of data, as is the case.

3.3.1.1 Age since diagnosis – "ageDiagnosis"

The analysis of this variable is important to view how long the patients in the study have been diagnosed.



(a) Hodgkin Lymphoma

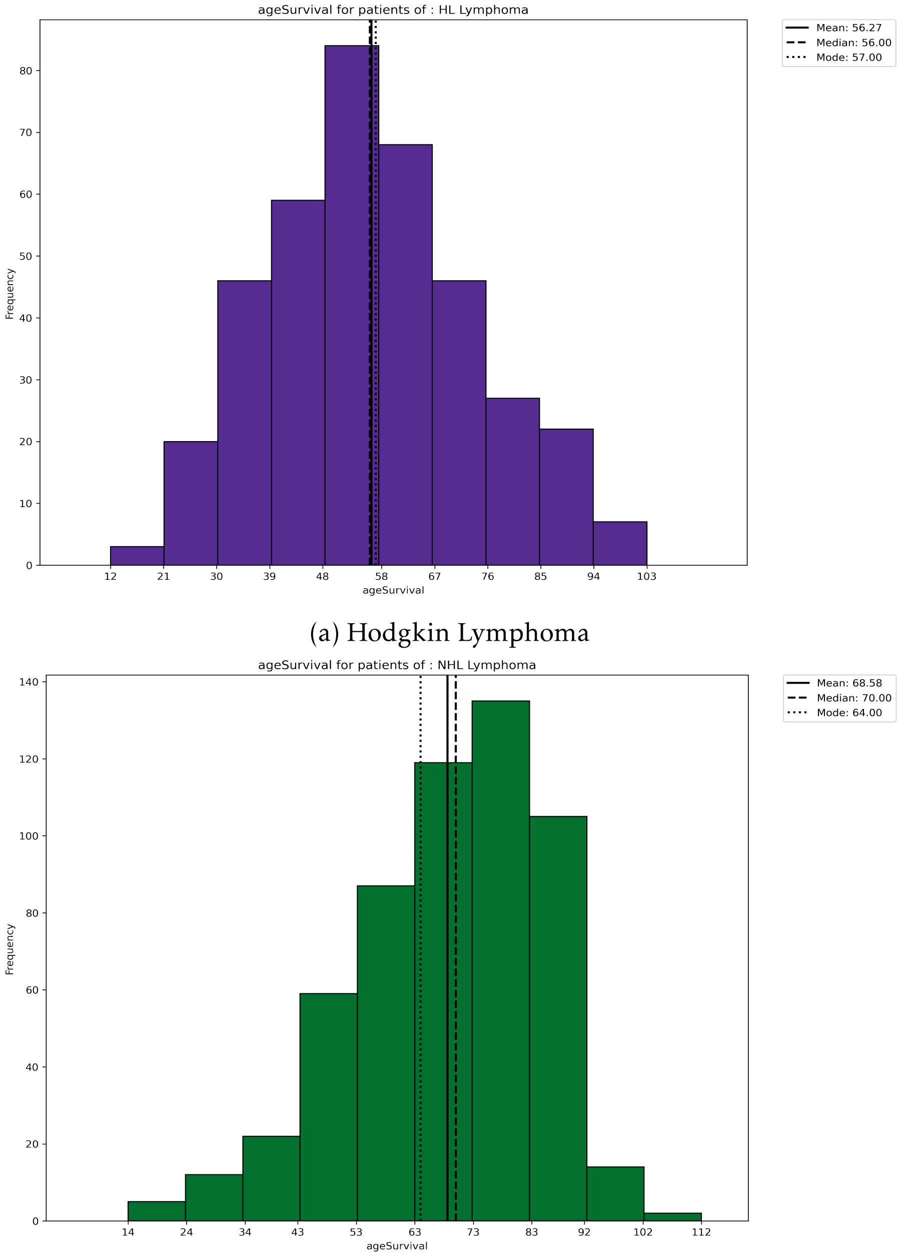
(b) Non-Hodgkin Lymphoma

Figure 3.5: Age since diagnosis bar plot

The results presented above in figure 3.5 are the dispersion of both NHL and HL. As seen in figure 3.5a, the majority of patients have been diagnosed between 21 and 30 years prior, whilst in non-Hodgkin 3.5b, the majority of patients have been diagnosed between 63 to 71 years before.

3.3.1.2 Age – "ageSurvival"

The barplots in the image below 3.6 represent the dispersion of patients' ages in both types of lymphomas.



(a) Hodgkin Lymphoma

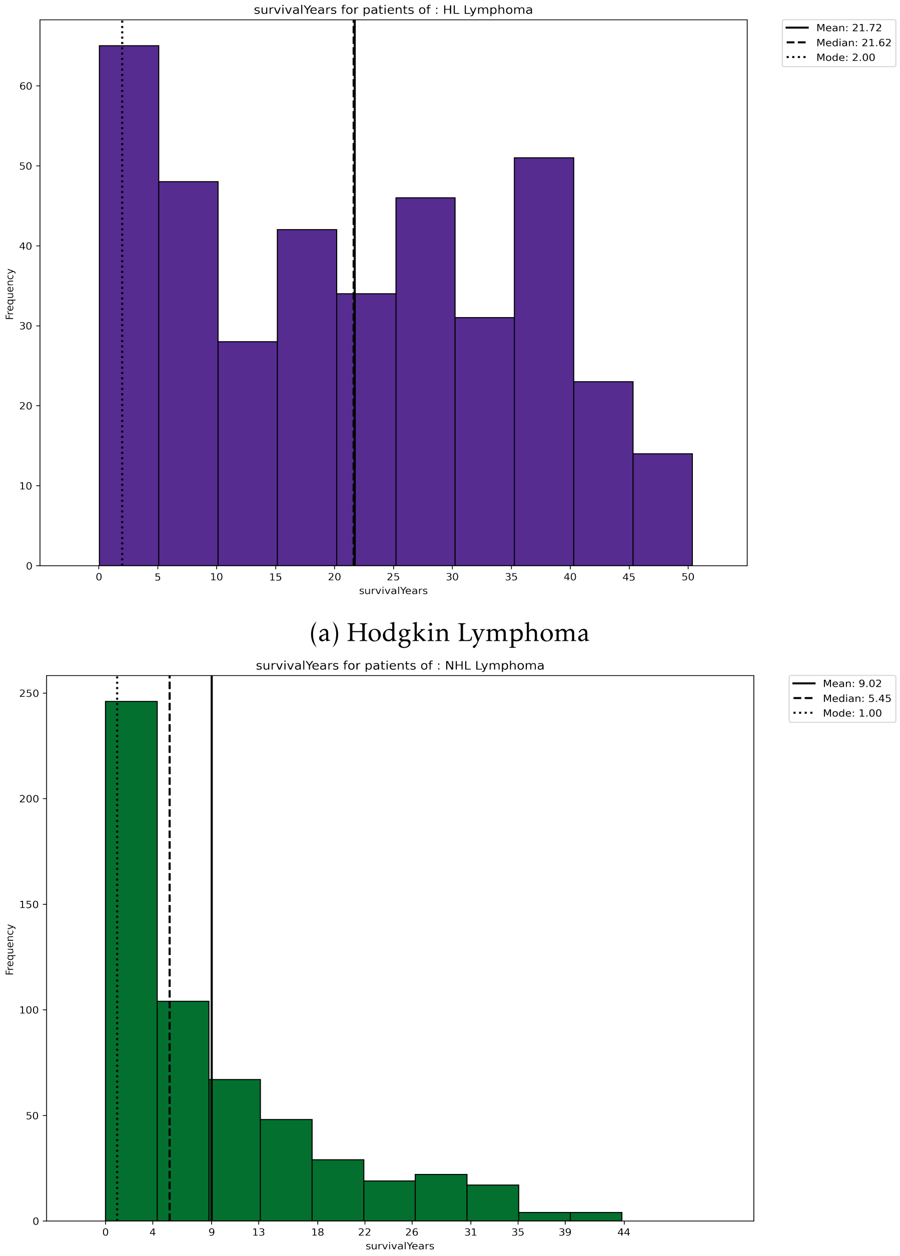
(b) Non-Hodgkin Lymphoma

Figure 3.6: Histogram of the age of the patients

The majority of Hodgkin lymphoma patients in the dataset are 48-58 years of age, whilst in non-Hodgkin lymphoma, the age of the patients is significantly higher as the majority of occurrences in this type of lymphoma is between 73-83 years. For the non-Hodgkin lymphoma patient with a survival age of 102, the choice to maintain it despite being considered an outlier or possibly to correspond to false information was made in accordance with the medical team that accompanied the project.

3.3.1.3 Survival years – "survivalYears"

Figure 3.7 below represents the distribution of the variable survivalYears.



(a) Hodgkin Lymphoma

(b) Non-Hodgkin Lymphoma

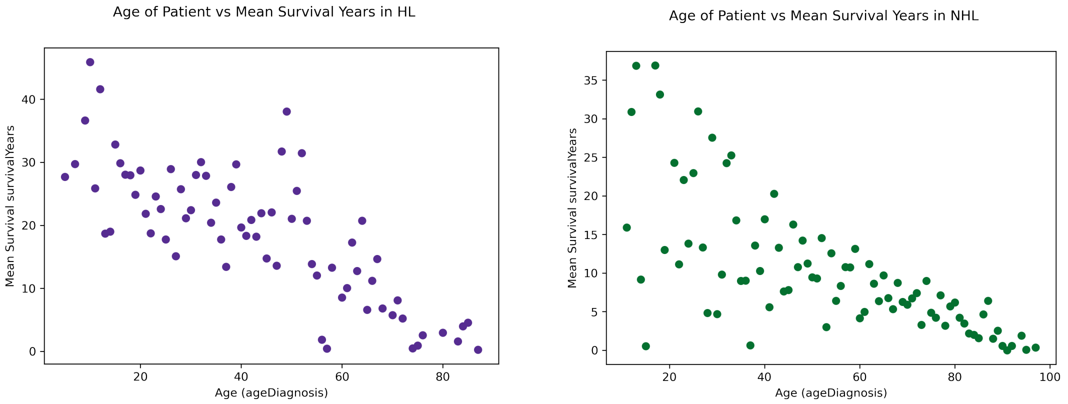
Figure 3.7: Survival years histogram

In the histogram 3.7a of Hodgkin lymphoma, there is no clear pattern as to which survival age is more predominant, as there are many spikes in no particular recursive interval. On the other hand, the non-Hodgkin lymphoma histogram in figure 3.7b has a majority of survival intervals occurring within the first 4 years. This distribution indicates that based on this overall view of dispersion of both survival years analysis, the non-Hodgkin has a majority of deaths within the first decade.

3.3.2 Scatter Plots

The scatter plots were used to represent the mean survival age of both the time since diagnosis, using the variable ageDiagnosis, and the age of the patients using the variable ageSurvival.

3.3.2.1 Age since diagnosis – "ageDiagnosis"



(a) Hodgkin Lymphoma

(b) Non-Hodgkin Lymphoma

Figure 3.8: Scatter plots of the age of the patients dependent on the mean survival age in years

Figure 3.8 above shows that both lymphomas establish a diminishing number of alive patients as the years pass. The relation is more relevant in the Hodgkin lymphoma scatter plot than in the non-Hodgkin lymphoma scatter plot.

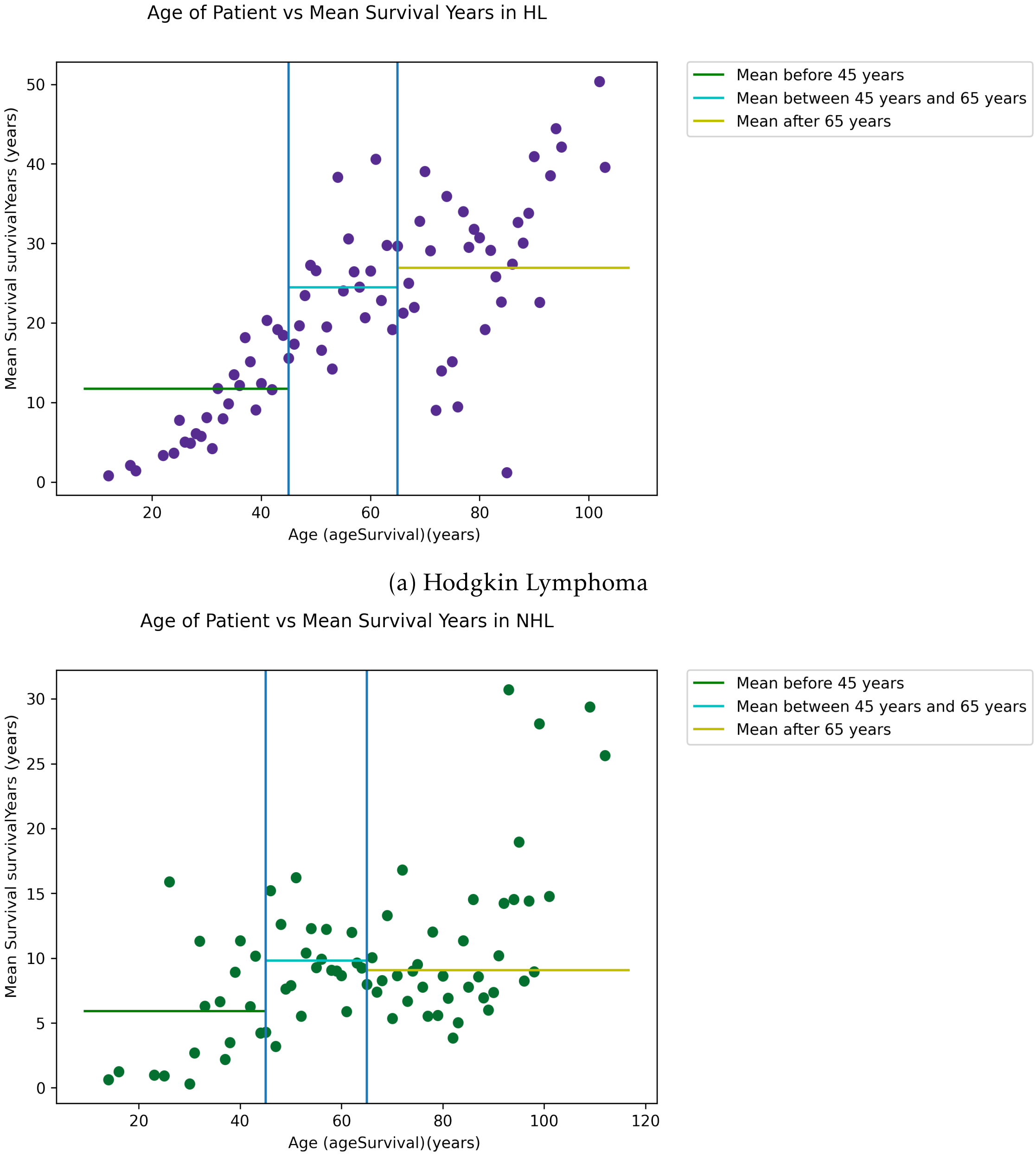
In the non-Hodgkin lymphoma scatter plot in figure 3.8b, it is clear that between the age of zero and the age of 40 years old, there is no relation between the time of survival and the time since diagnosis since the variation between patients is wide. Between the ages 70 and 80, there is a clearer image of the previously stated relation between the number of patients that survived longer.

The Hodgkin lymphoma scatter plot in figure 3.8a presents an overall higher variation of the entire age range but is more homogeneous within the type of lymphoma.

3.3.2.2 Age – "ageSurvival"

In contrast to the scatter plots present previously, the scatter plots representing the patient's age (ageSurvival) have an additional study layer: the stratification of a continuous variable.

This stratification consists of three intervals, the first being up to 45 years of age, the second being between 45 and 65, and the last including more than 65 years old. To help the visualisation of this stratified variable, vertical lines were added to the plot along with the mean of each age group.



(a) Hodgkin Lymphoma

(b) Non-Hodgkin Lymphoma

Figure 3.9: Scatter plot between time since diagnosis of the patient and mean survival age

In the plots in figure 3.9 presented above, the Hodgkin lymphoma scatter plot 3.9a shows a proportional relation between the survival age and the patient's age in the first sector of ages until 45. The other two sectors, ages between 45 and 65 and ages above 65, show the dispersion of the data. All three age groups have a similar mean in plot 3.9b that represents non-Hodgkin lymphoma and are also very dispersed, without any noticeable correlation.

3.3.3 Individual Variable Analysis

This section focus on presenting the results of the univariate initial descriptive analysis. As mentioned before, these variables were chosen following a specialist medical team consultation.

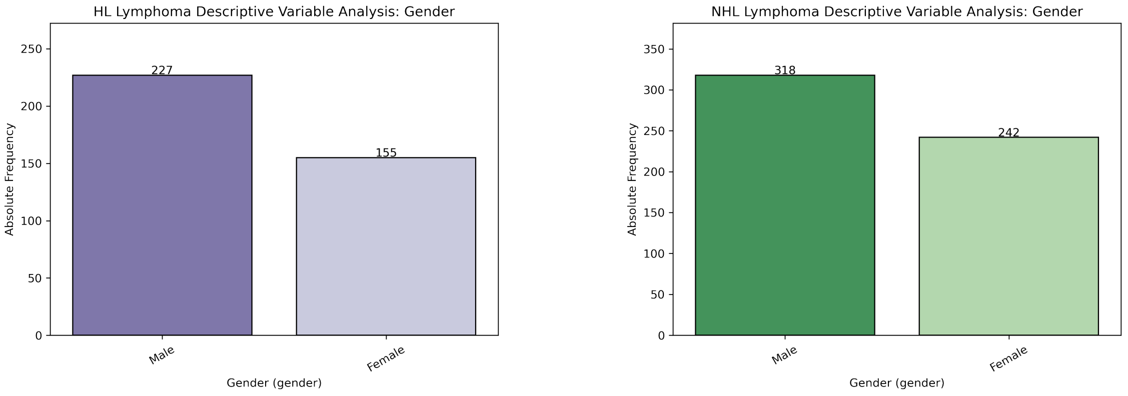
It is essential to reinforce that these variables are not the only ones that constitute the supplied dataset.

For each variable, there will be presented three different plots: the first one will exhibit an overall count based os the distribution of the variables' values; the second one will also be an overall count that takes into consideration the state of living of each patient (if the patient is dead, alive or the follow up was lost); the last plot will be a boxplot that expresses in a standardised way the data distribution, showing the minimum the maximum the median and the first and third quartiles.

Following the plots to complement the already given information extracted from the data, each variable will have a table that expresses the overall dispersion, the mode median and mean for survival time of the subgroup as patients and the age since diagnostic. The choice to once again exhibit the absolute and relative count is to help interpret the other presented results without the need to search for the rest of the information, consolidating it in either plot or table.

3.3.3.1 Gender

The gender of a patient can be either male or female. Presented below in figure 3.10 is the distribution of male and female patients for both Hodgkin and non-Hodgkin lymphoma.



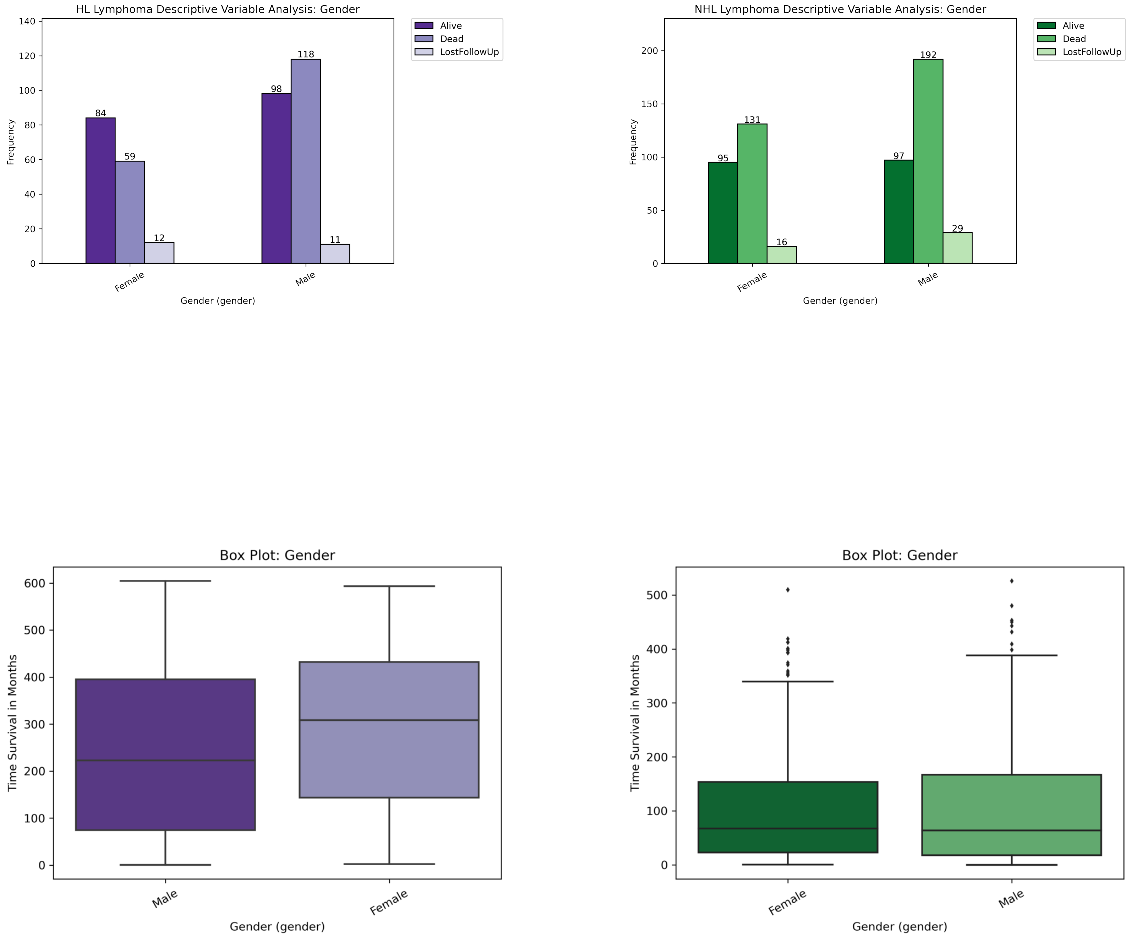


Figure 3.10: Gender descriptive graphs for both types of lymphoma

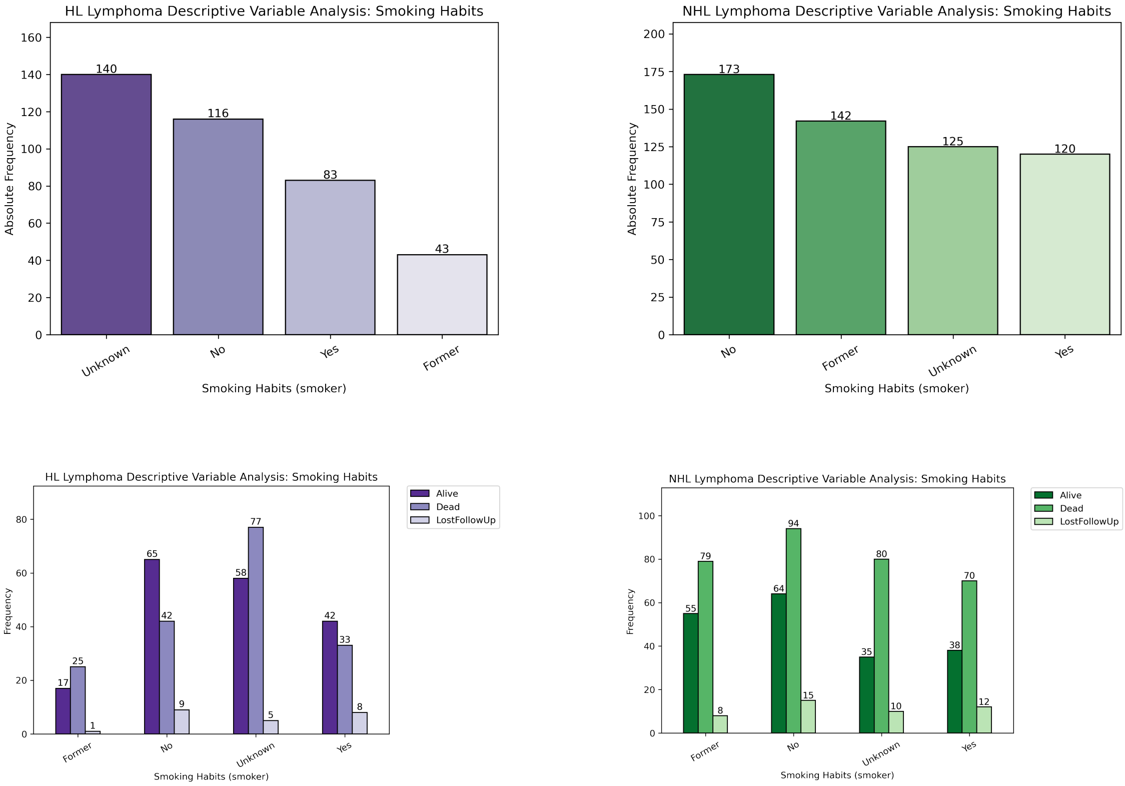
HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Gender	Male	227	59,42%	18,57	19,75	33	37	36,17
	Female	155	40,58%	25,72	24,60	30	21	33,01
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Gender	Male	318	56,79%	5,32	9,11	59	58	57,49
	Female	242	43,21%	5,61	8,91	63	63	62,50

Table 3.12: Gender distribution for Hodgkin lymphoma and non-Hodgkin lymphoma

We can consider that there is a good diversity within the gender since, in table 3.12 above, Hodgkin lymphoma is nearly 59% male and 41% female, while non-Hodgkin lymphoma patients are approximately 57% male and 43% female. Even though the overall data is not overfit either gender, from the data received, men are more prominent in either type of lymphoma, at least within the dataset.

3.3.3.2 Smoking Habits

A patient can be in one of four possible situations regarding the smoking habits variable. A patient's smoking habits can be either a smoker, a non-smoker (which entails people who have smoked up to 100 cigarettes in their lifetime), a former smoker (a patient who has stopped smoking for at least more than one year) or an unknown status. Considering this classification, plots in figure 3.11 below show the data dispersion within their smoking habits.



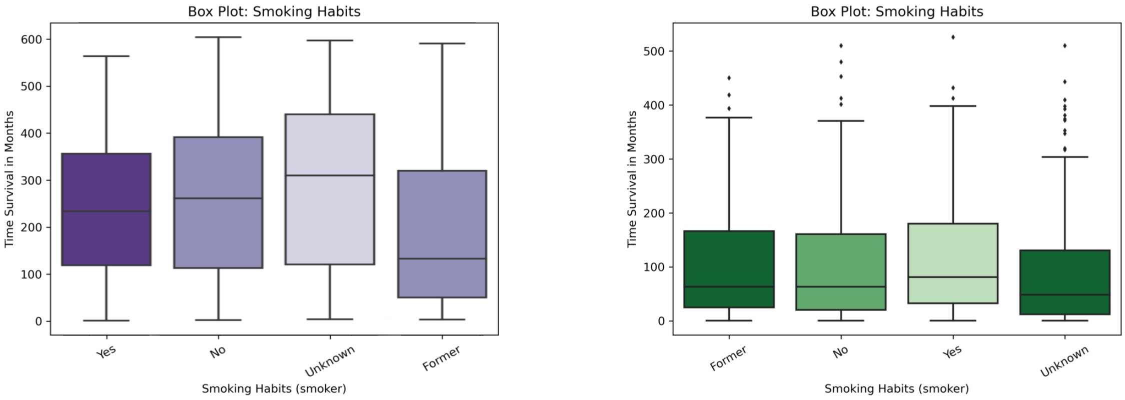


Figure 3.11: Smoking habits descriptive graphs for both types of lymphoma

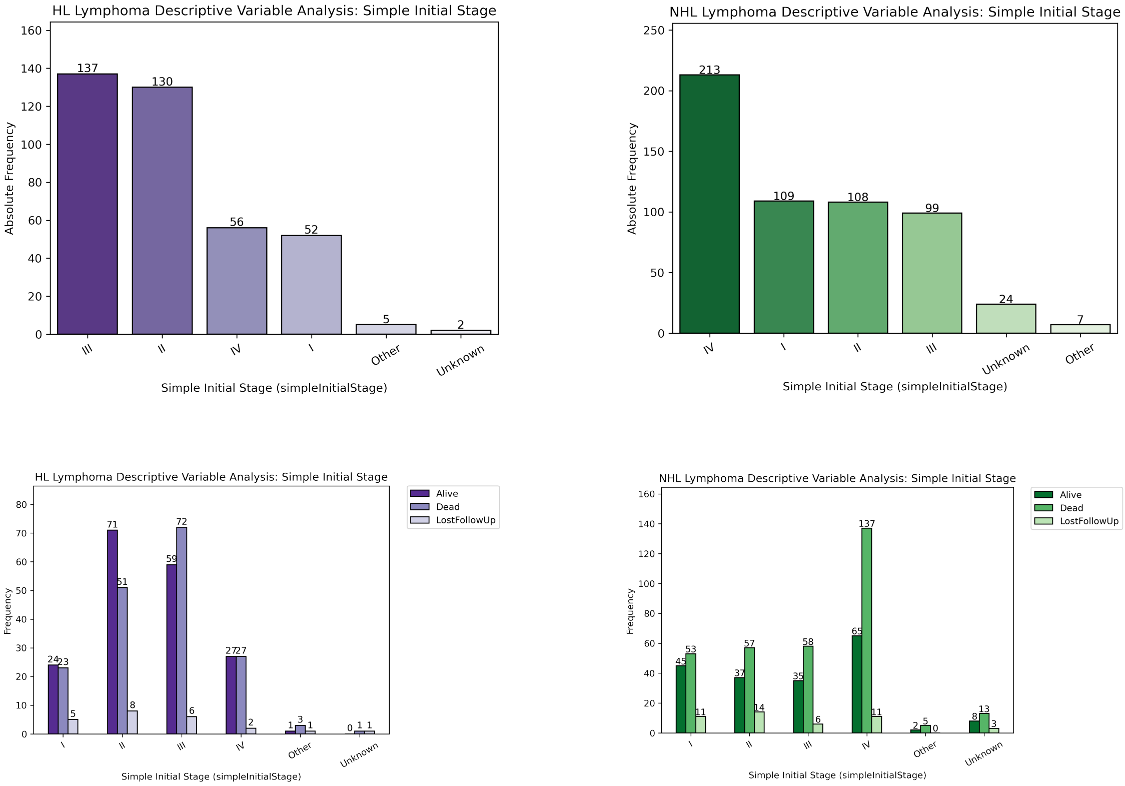
HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Smoker	Yes	83	21,73%	19,61	20,56	31	29	34,49
	No	116	30,37%	21,93	21,23	28	21	31,96
	Unknown	140	36,65%	25,81	24,21	33	30	35,18
	Former	43	11,26%	11,07	17,16	40	48	42,60
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Smoker	No	173	30,89%	5,28	8,96	62	72	60,73
	Yes	120	21,43%	6,75	10,47	50,5	52	48,93
	Unknown	125	22,32%	4,04	8,11	66	81	63,59
	Former	142	25,36%	5,25	8,67	66	71	63,96

Table 3.13: Smoker distribution for Hodgkin lymphoma and non-Hodgkin lymphoma

Observing either plots in figure 3.11 or table 3.13, there is a large subgroup of the patients that are not classified regarding their smoking habits, 37% in HL and 22% in NHL. Despite that, the rest of the population that constitutes the dataset represents a large enough subgroup that the variable could be of possible importance when determining the probability of survival of lymphoma patients, as is analysed in the following chapter.

3.3.3.3 Initial Stage

The staging in both lymphomas obeys, as stated in the previous subsection 2.1.4 to a predetermined classification. As a help to the analyses, the staging was grouped by the primary stage classifier, and as a consequence of that, the patient can have stage I, II, III, or IV lymphoma, as seen in the following plots 3.12.



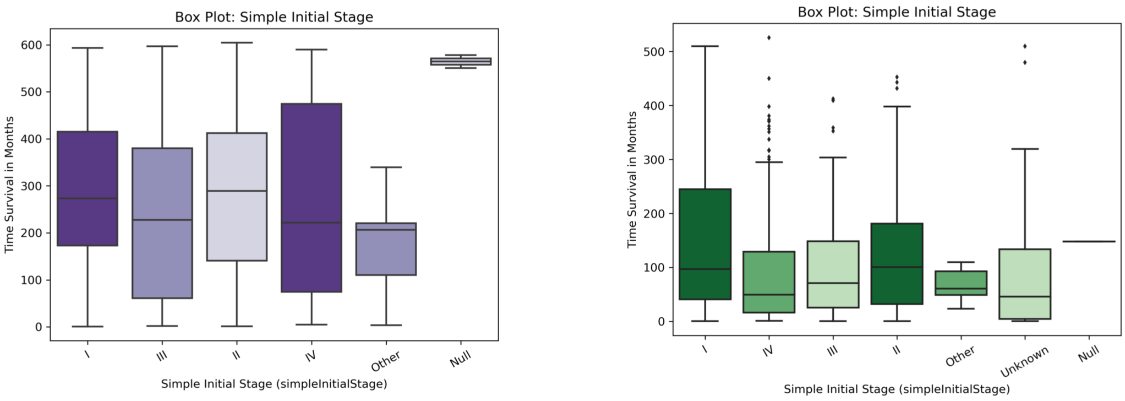


Figure 3.12: Initial Stage descriptive graphs for both types of lymphoma

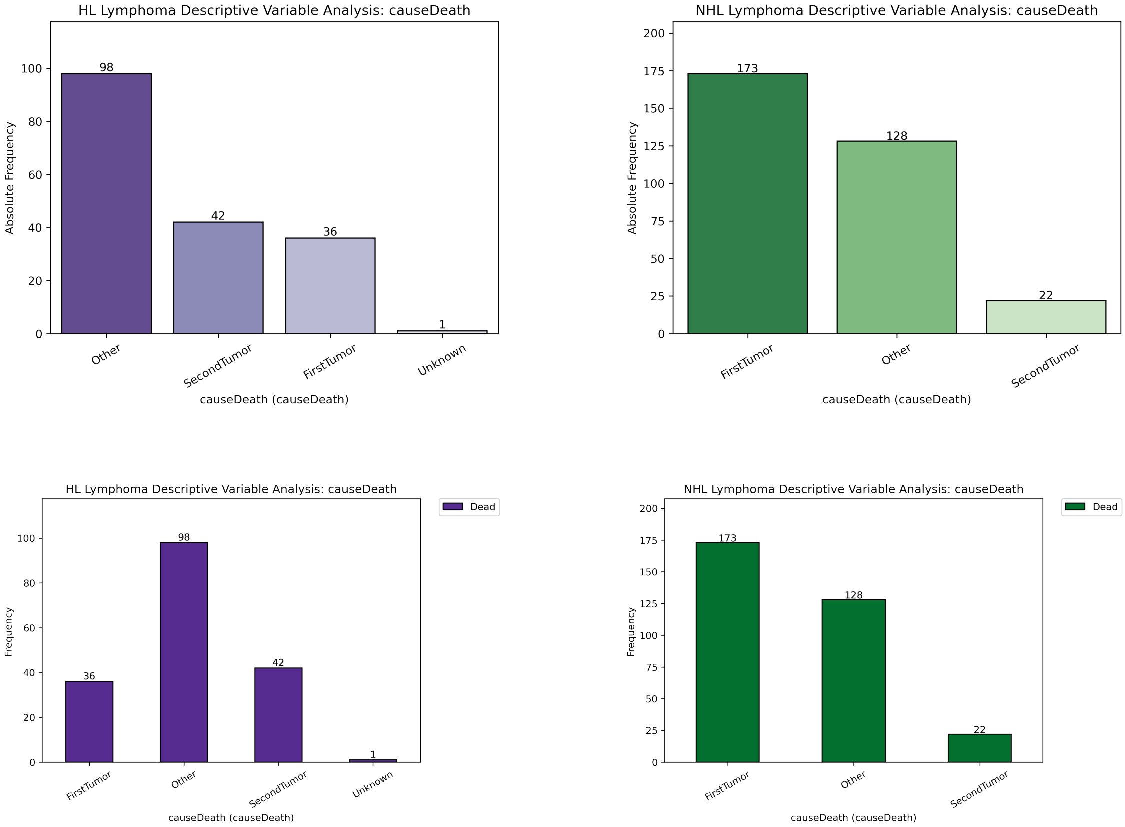
HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Initial Stage	I	52	13,61%	22,79	23,20	30	30	34,88
	II	130	34,03%	24,08	23,17	30	23	32,88
	III	137	35,86%	18,98	19,52	31	21	35,22
	IV	56	14,66%	18,48	22,10	36	19	39,02
	Other	5	1,31%	17,22	14,68	22	21	35,2
	Unknown	2	0,52%	47,03	47,03	26	10	26
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Initial Stage	I	109	19,46%	8,09	11,71	61	63	59,58
	II	108	19,29%	8,35	10,36	58	49	58,98
	III	99	17,68%	5,91	8,35	60	75	59,81
	IV	213	38,04%	4,12	7,39	62	71	59,53
	Other	7	1,25%	5,05	5,67	62	43	61,57
	Unknown	24	4,29%	4,06	8,99	60,5	82	63,04

Table 3.14: Initial stage distribution for Hodgkin lymphoma and non-Hodgkin lymphoma

When observing table 3.14, it is apparent that in Hodgkin lymphoma, there is a predominance of initial stages II and III with an approximate combined percentage of 70%, whereas, in non-Hodgkin lymphoma, the leading stage is stage IV with 38% of patients, while the other stages have a seemingly even distribution between 17% to 20%.

3.3.3.4 Cause of Death

The variable cause of death deals, as its name states, with the cause of decease of the patients. The patient's cause of death can be classified as the first tumour if the death is related to the first tumour, second tumour or other if the cause of death is not related to either the first or the second tumour. The results of the descriptive statistic regarding this variable can be seen in the following plots in figure 3.13.



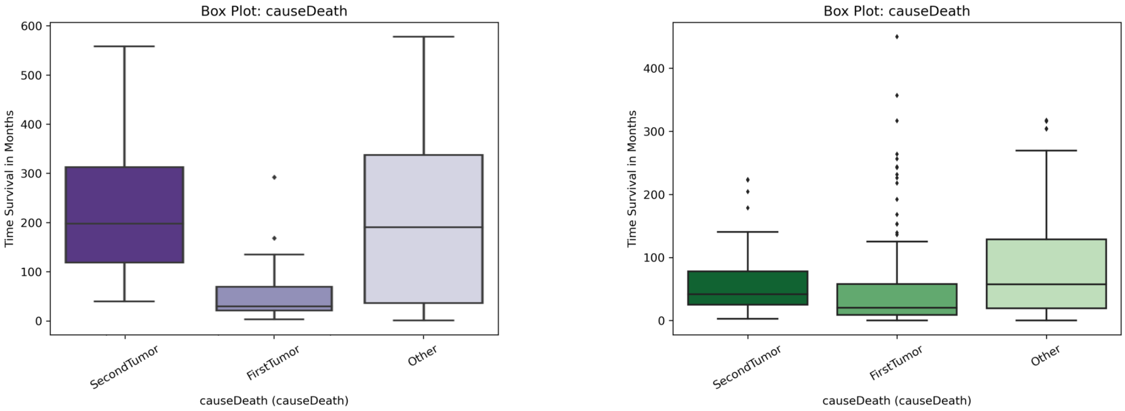


Figure 3.13: Initial Stage descriptive graphs for both types of lymphoma

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Cause of Death	FirstTumor	36	9,42%	2,45	4,56	37	19	41,56
	SecondTumor	42	10,99%	16,47	19,17	34,5	37	34,48
	Other	98	25,65%	15,85	16,25	34,5	21	40,66
	Unknown	206	53,93%	28,45	27,84	30	23	31,06
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Cause of Death	FirstTumor	173	30,89%	1,68	4,06	64	63	63,11
	SecondTumor	22	3,93%	3,48	5,99	65,5	42	61,82
	Other	128	22,86%	4,80	7,01	70	79	67,54
	Unknown	237	42,32%	11,55	14,00	54	54	52,68

Table 3.15: Cause of Death distribution for Hodgkin lymphoma and non-Hodgkin lymphoma

Table 3.15 contains a more significant number of unknown classifications in both lymphomas since, in the tables, the patients that are alive or have their follow-up lost are included, and since they are alive, the variable cause of death is logically unpopulated and therefore classified as unknown

3.3.3.5 ECOG Performance Status

The ECOG performance status can be classified according to the previously present scale 2.1.12. The following plots in figure 3.14 represent the division of the dataset for both lymphomas.

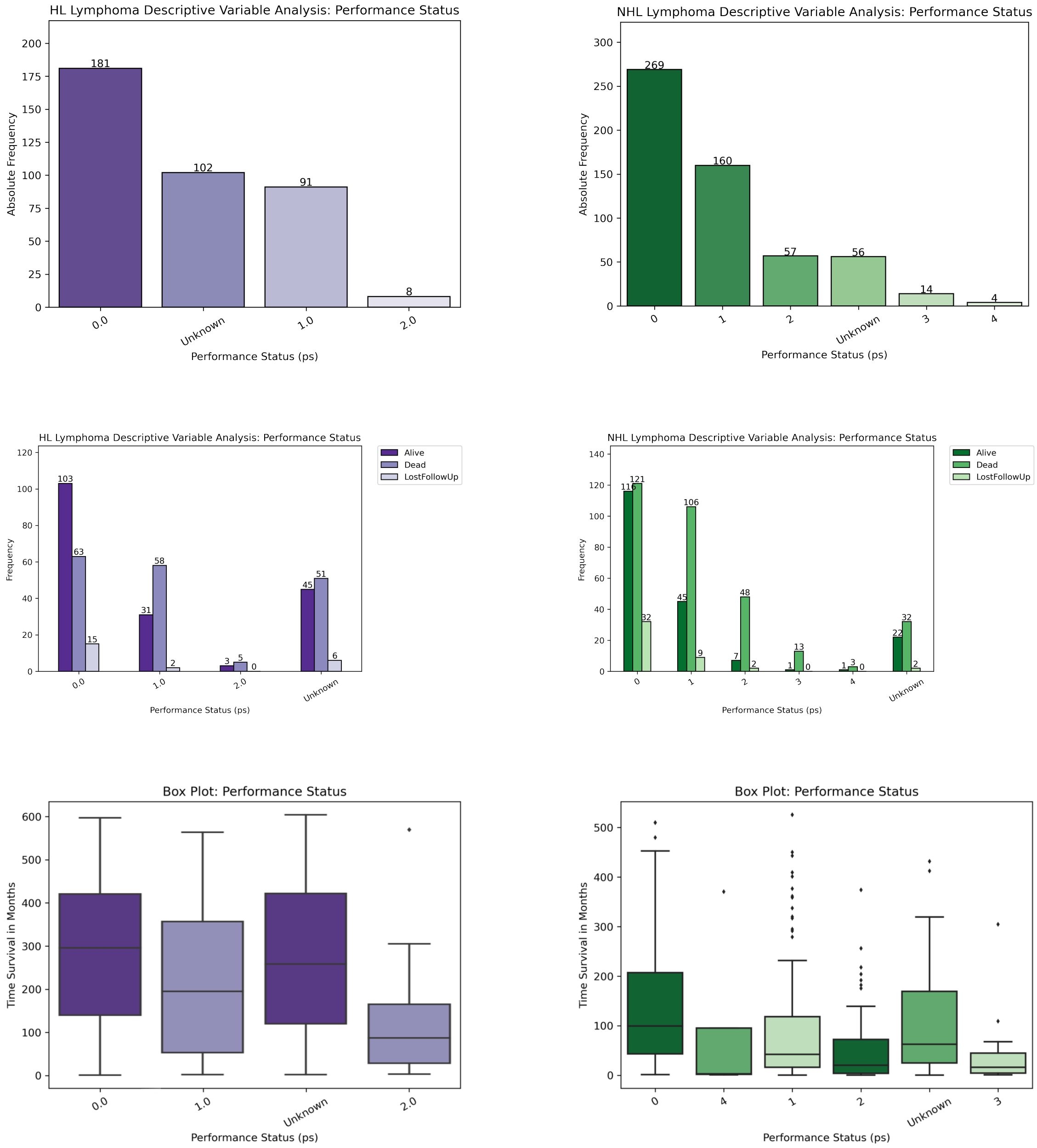


Figure 3.14: ECOG Performance Status descriptive graphs for both types of lymphoma

HL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
Performance Status	0.0	181	47,38%	24,46	23,45	29	29	31,48
	1.0	91	23,82%	16,25	18,27	37	21	38,89
	2.0	8	2,09%	7,25	12,78	66	70	61,38
	Unknown	102	26,70%	21,67	22,44	32,5	18	35,28
NHL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
Performance Status	0.0	269	48,04%	8,26	11,47	57	58	55,59
	1.0	160	28,57%	3,51	7,08	66	71	62,66
	2.0	57	10,18%	1,67	4,82	69	64	66,72
	3.0	14	2,50%	1,33	3,78	71,5	21	69,07
	4.0	4	0,71%	0,25	7,86	72	12	61,25
	Unknown	56	10,00%	5,21	8,49	66,5	68	61

Table 3.16: ECOG Performance Status distribution for Hodgkin lymphoma and non-Hodgkin lymphoma

In both lymphomas, according to table 3.16, the most prevalent performance status is 0 with around 47% and 48% in HL and NHL of all cases, while the second most common occurrence is performance status 1 with roundly 24% and 29% of HL and NHL respectively.

3.3.3.6 Histology

As stated before in 2.1.6, the classification for histology depends on the type of lymphoma. Hodgkin lymphoma will be classified within the "cellularity". Conversely, non-Hodgkin lymphoma will be classified as either T-cell histology or B-cell histology, the last one being divided into low and high grades. The decision to agglomerate the NHL histology into only three classifiers was made to convey the pretended classification established by the medical team. This group by histology also helps the future models since there is more data to extrapolate possible results. The results of the histology plots in figure 3.15 are presented ahead.

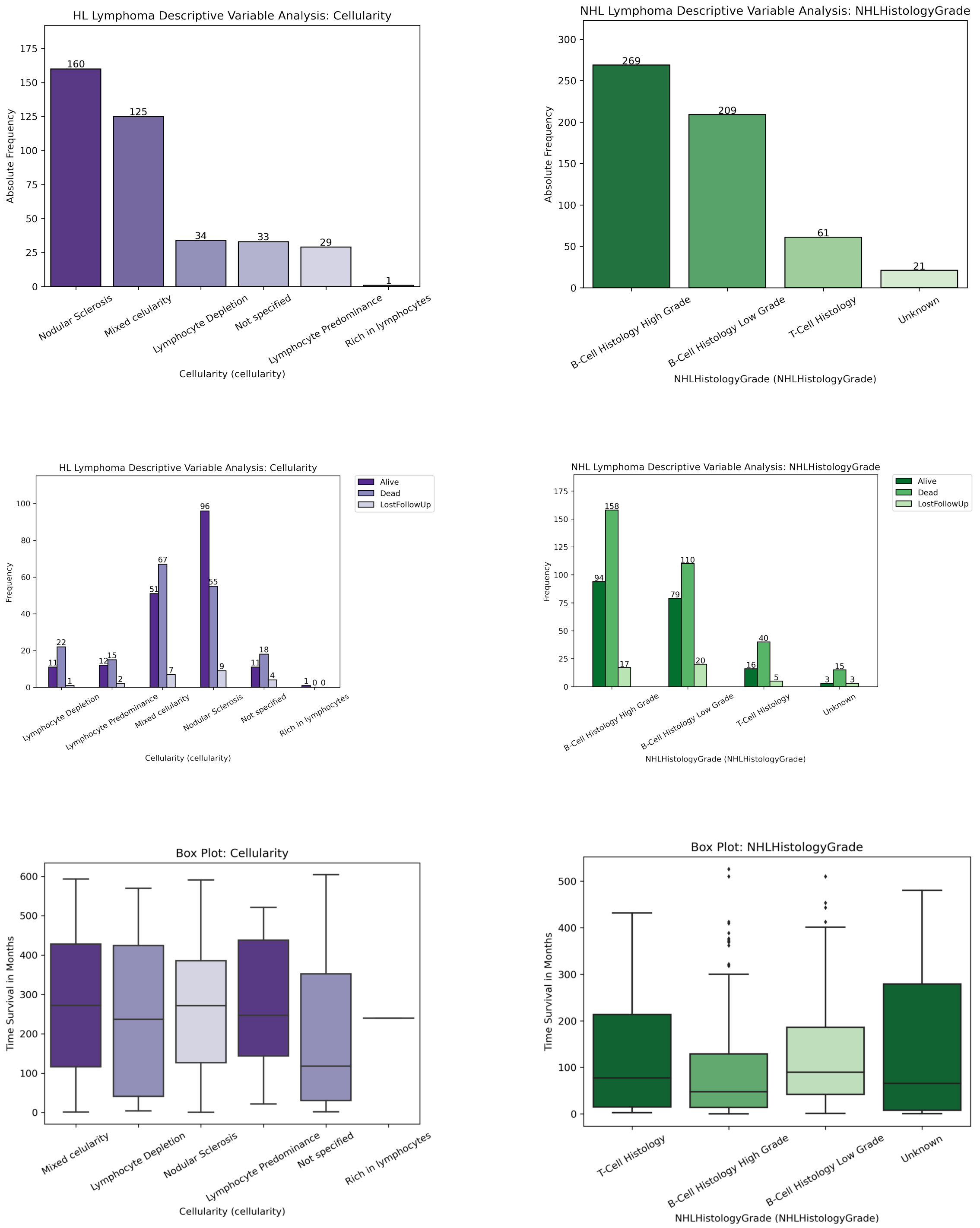


Figure 3.15: Histology descriptive graphs for both types of lymphoma

Table 3.17 confirms that "rich in lymphocytes" in HL is only composed of one patient and will be excluded when the Kaplan-Meier models are executed in the next chapter due to a lack of information. It is also clear that the patients with T-cell histologies are a minority compared to those with B-cell histologies, making up only 11% of the NHL patients.

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Histology (Cellularity)	Mixed celularity	125	32,72%	22,69	22,31	37	19	38,34
	Lymphocyte Depletion	34	8,90%	19,75	20,21	33	19	35
	Nodular Sclerosis	160	41,88%	22,61	22,23	29	23	30,43
	Lymphocyte Predominance	29	7,59%	20,57	22,75	29	20	33,93
	Not specified	33	8,64%	9,81	17,72	42	42	44,12
	Rich in lymphocytes	1	0,26%	19,99	19,99	36	36	36
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Histology (Grouped)	T-Cell Histology	61	10,89%	6,42	10,44	49	22	49,21
	B-Cell Histology High Grade	269	48,04%	3,97	7,28	63	61	61,55
	B-Cell Histology Low Grade	209	37,32%	7,47	10,54	62	79	60,32

Table 3.17: Histology distribution for Hodgkin lymphoma and non-Hodgkin lymphoma

3.3.3.7 Non-Hodgkin Lymphoma Grade

The lymphoma's grade is exclusive to the non-Hodgkin and can be classified as either very high, high or low. The results of this classification are presented below in the graphs 3.16.

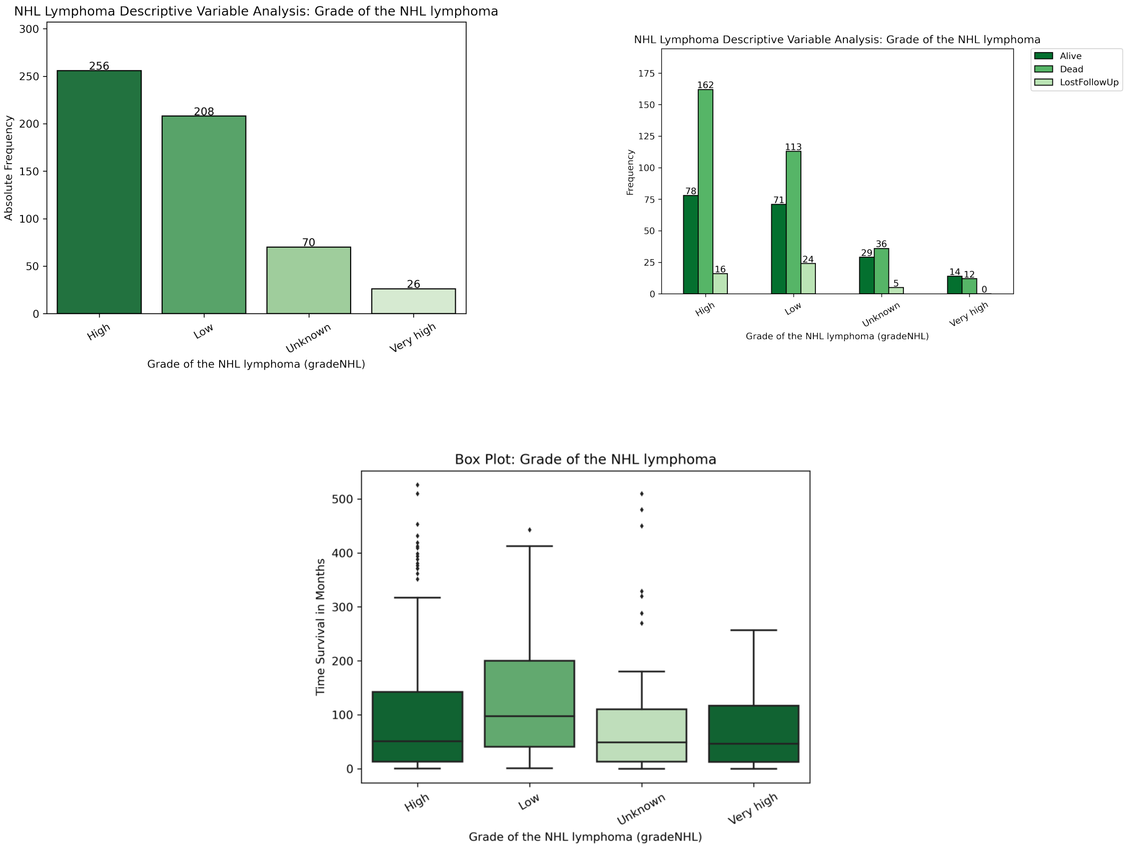


Figure 3.16: Grade of non-Hodgkin lymphoma

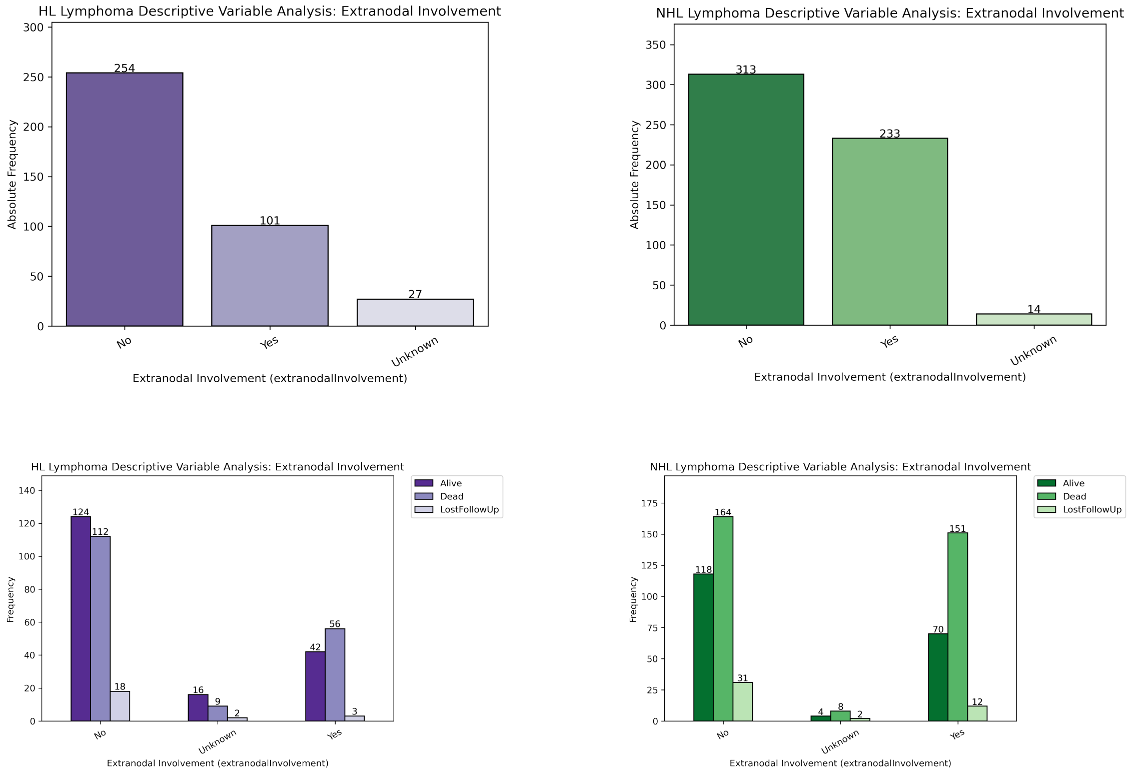
NHL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
Grade	Very high	26	4,64%	3,89	6,09	58	58	53,81
	High	256	45,71%	4,27	8,24	61,5	72	59,32
	Low	208	37,14%	8,11	10,96	61,5	63	59,71
	Unknown	70	12,50%	4,08	7,23	67	79	62,91

Table 3.18: Grade of non-Hodgkin lymphoma

As it can be observed by analysing table 3.18 the higher the grade of the lymphoma, the smaller the median survival years are. It is also necessary to note that around 46% of patients have a high-grade lymphoma.

3.3.3.8 Extranodal Involvement

The extranodal involvement variable describes whether or not the lymphoma at hand has extranodal involvement, being classified either by yes or no. The following plots 3.17 demonstrate the dispersion of the extranodal involvement for both lymphomas.



In both lymphomas, a majority percentage of patients have no extranodal involvement, 66% in HL and 56% in NHL, as seen in table presented below.

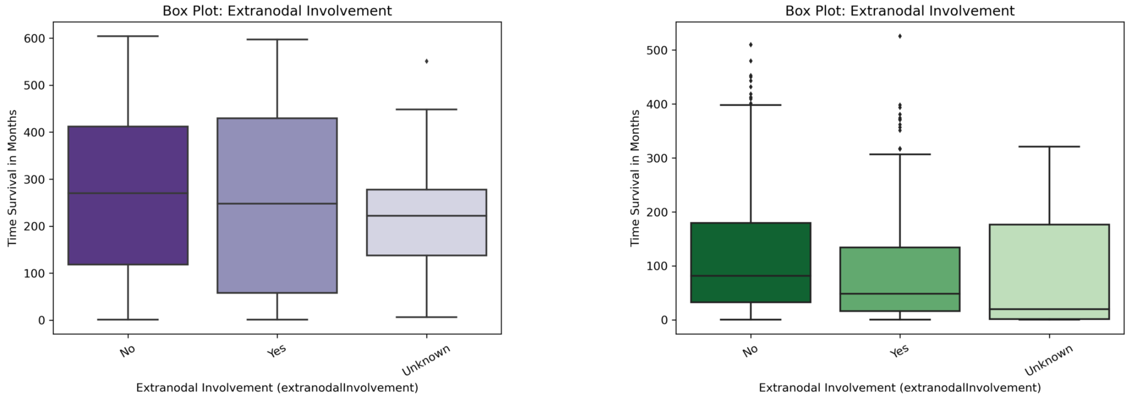


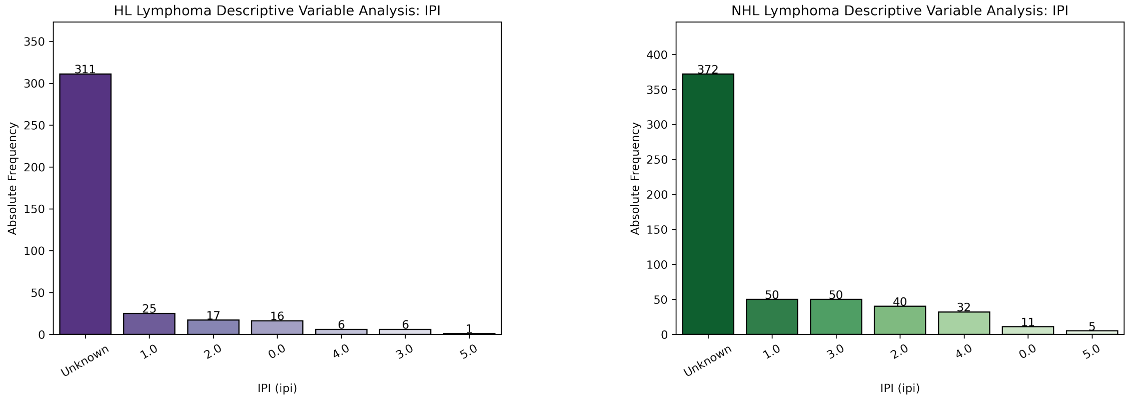
Figure 3.17: Extranodal involvement descriptive graphs for both types of lymphoma

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Extranodal Involvement	No	254	66,49%	22,49	22,21	31	30	34,33
	Yes	101	26,44%	20,66	21,35	33	29	36,81
	Unknown	27	7,07%	18,52	18,54	27	27	32,96
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Extranodal Involvement	No	313	55,89%	6,79	10,26	61	58	59,52
	Yes	233	41,61%	4,04	7,43	62	63	59,35
	Unknown	14	2,50%	1,66	7,66	73	24	67,79

Table 3.19: Extranodal Involvement for Hodgkin lymphoma and non-Hodgkin lymphoma

3.3.3.9 International Prognostic Index

As stated in 3.18, the international prognostic index or IPI can take a value between zero and five. Underneath are the plots describing the distribution of the IPI for both lymphomas.



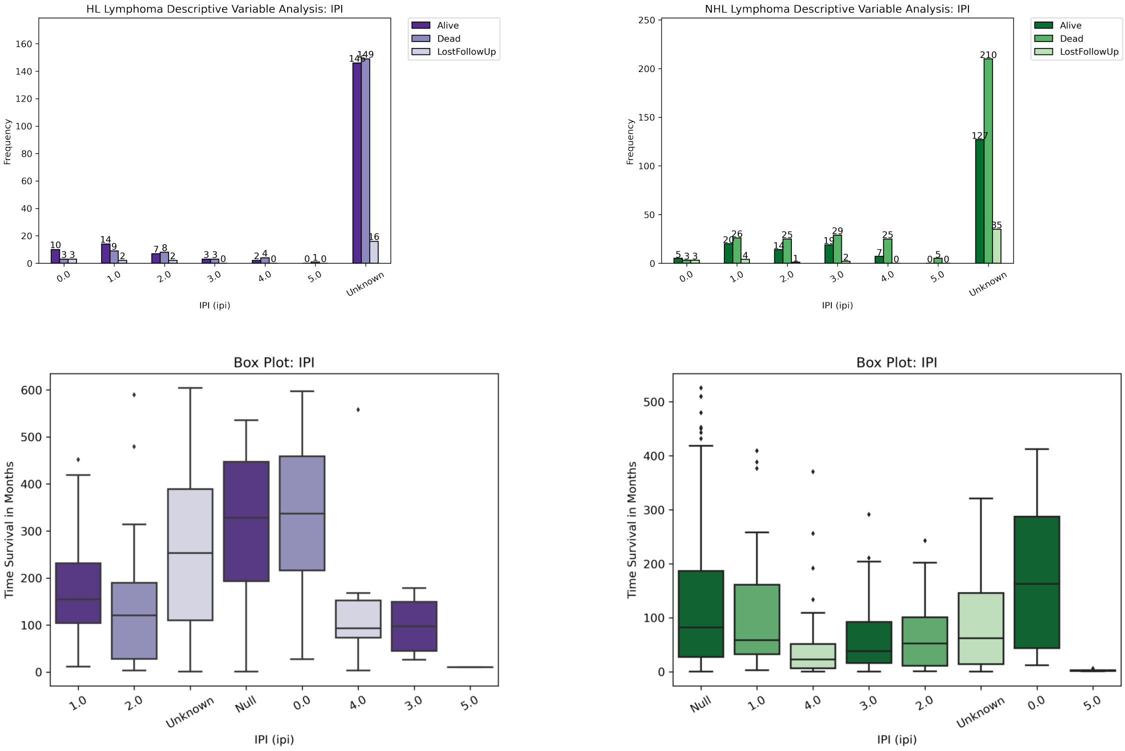


Figure 3.18: IPI descriptive graphs for both types of lymphoma

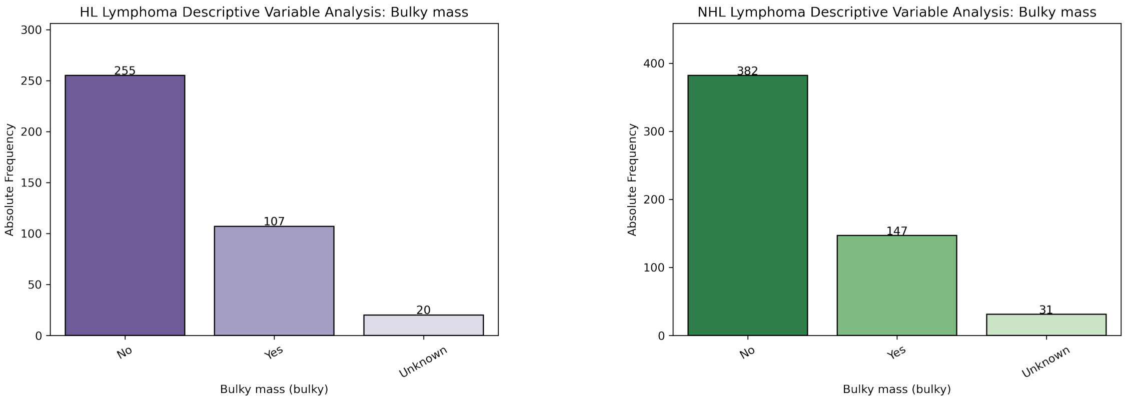
The majority of both patients in both lymphomas are classified as unknown, making the information presented only a small percentage of the original dataset, as seen in table below.

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
IPI	0.0	16	4,19%	28,06	27,53	27	30	25,69
	1.0	25	6,54%	12,89	15,38	30	29	31,16
	2.0	17	4,45%	10,04	13,15	38	9	42,18
	3.0	6	1,57%	8,08	8,24	54,5	40	55,17
	4.0	6	1,57%	7,77	13,71	51,5	12	48,17
	5.0	1	0,26%	0,86	0,86	40	40	40
	Unknown	311	81,41%	23,42	22,88	31	22	34,60
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
IPI	0.0	11	1,96%	13,56	14,16	47	58	48,18
	1.0	50	8,93%	4,88	8,76	60	54	60,72
	2.0	40	7,14%	4,35	5,74	61,5	44	64,075
	3.0	50	8,93%	3,18	5,09	69	69	68,04
	4.0	32	5,71%	1,90	4,42	70	63	68,84
	5.0	5	0,89%	0,17	0,21	73	64	73,6
	Unknown	372	66,43%	6,78	10,30	59	79	57,28

Table 3.20: International Prognostic Index for Hodgkin lymphoma and non-Hodgkin lymphoma

3.3.3.10 Bulky Mass

The distribution of whether or not the patient's lymphoma contains a bulky mass is present underneath in plots of figure 3.19.



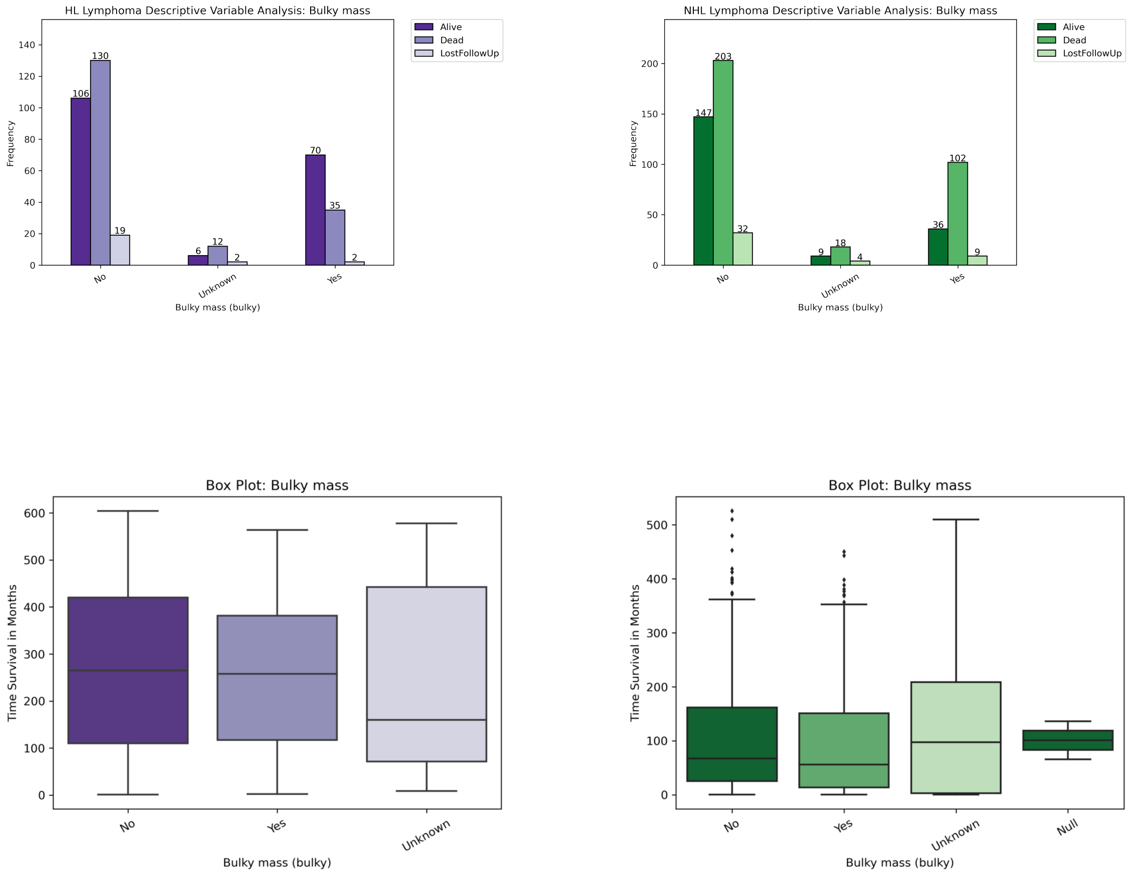


Figure 3.19: Bulky mass descriptive graphs for both types of lymphoma

HL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
Bulky Mass	No	255	66,75%	22,07	22,12	32	22	35,91
	Yes	107	28,01%	21,49	21,19	29	23	31,73
	Unknown	20	5,24%	13,33	19,47	40,5	22	38,7
NHL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
Bulky Mass	No	382	68,21%	5,60	9,07	61	72	59,57
	Yes	147	26,25%	4,66	8,52	63	79	60,22
	Unknown	31	5,54%	8,09	10,76	61	68	58,06

Table 3.21: Bulky Mass for Hodgkin lymphoma and non-Hodgkin lymphoma

Around 28% of Hodgkin lymphoma patients manifest bulky mass during their Hodgkin lymphoma; on non-Hodgkin lymphomas, only 26% of patients manifest it, as shown in table 3.21.

3.3.3.11 B-symptoms

The variable B-symptoms is also self-explanatory since it expresses in a binary way whether or not the patient has manifested any B symptoms. The distribution of this variable is visible underneath in the graphs 3.20.

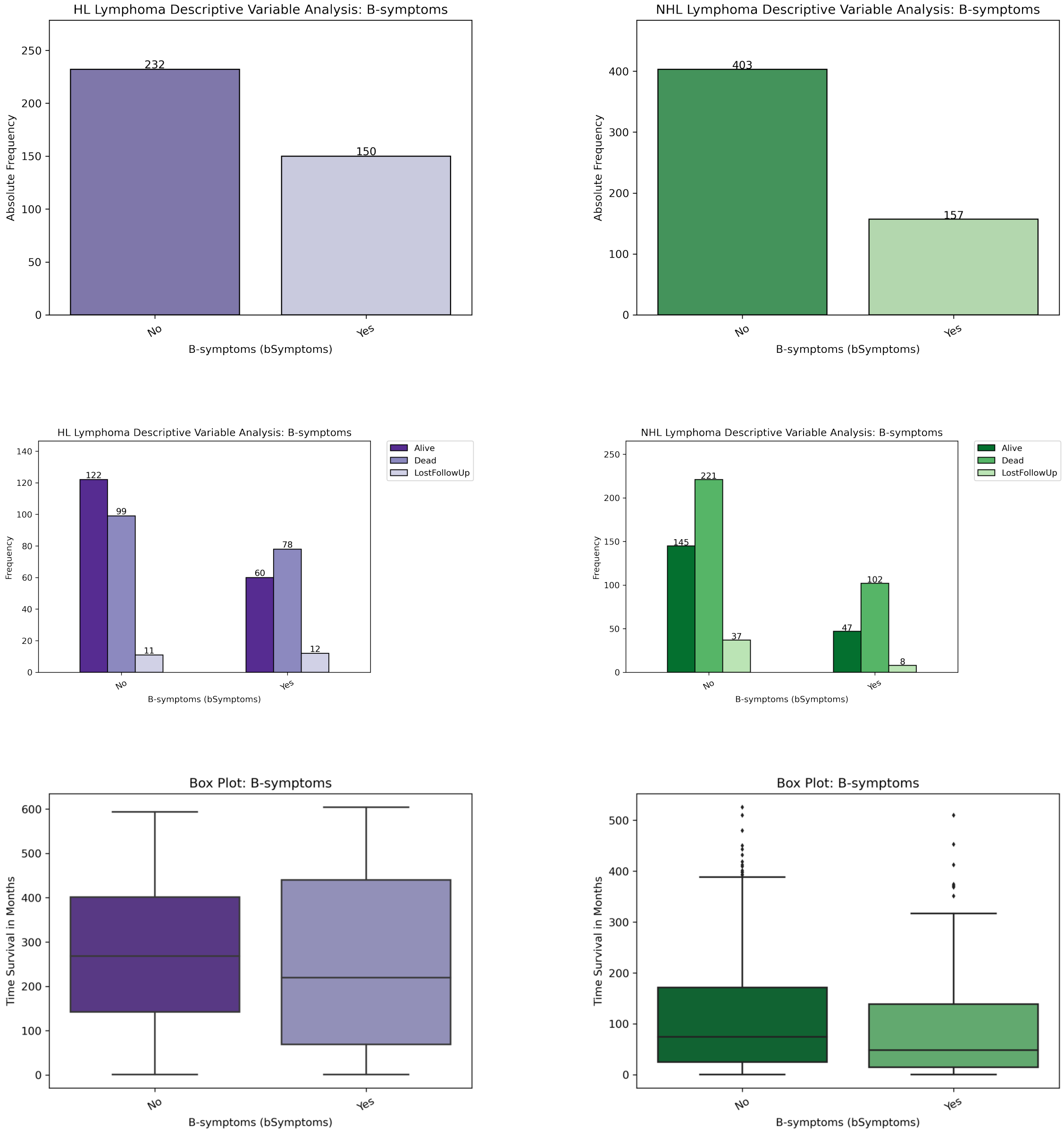


Figure 3.20: Histology descriptive graphs for both types of lymphoma

These symptoms are present in approximately 39% of HL patients and 28% of non-Hodgkin lymphoma, as shown in table 3.22 below.

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
B-symptoms	No	232	60,73%	22,38	22,36	30	30	34,21
	Yes	150	39,27%	18,32	20,72	33	19	35,94
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
B-symptoms	No	403	71,96%	6,2	9,56	61	58	60,26
	Yes	157	28,04%	4,02	7,63	62	63	58,13

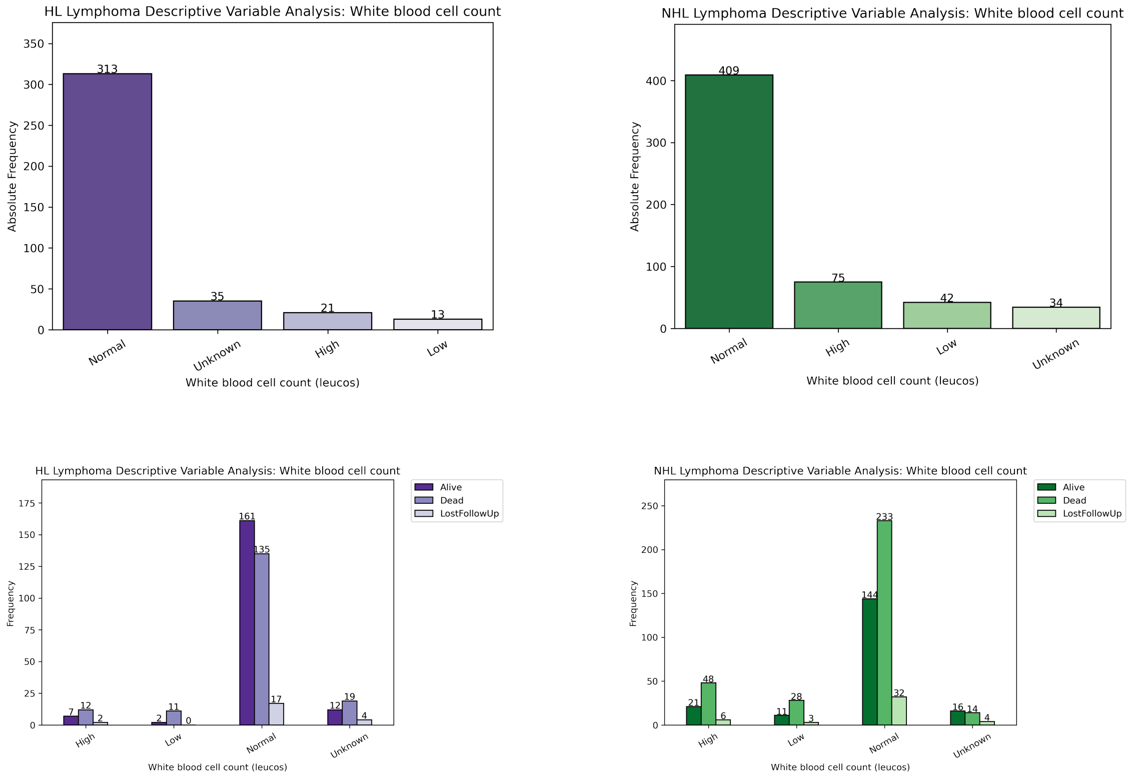
Table 3.22: B-symptoms for Hodgkin lymphoma and non-Hodgkin lymphoma

3.3.3.12 Diagnostic Analytics

The Diagnostic Analytics results in medical analysis that the patient underwent when diagnosed. Each variable expresses the results of the performed tests.

White blood cell count

The count of white blood cells or leucocytes presents its results for both lymphomas in the plots of figure 3.21 presented ayont.



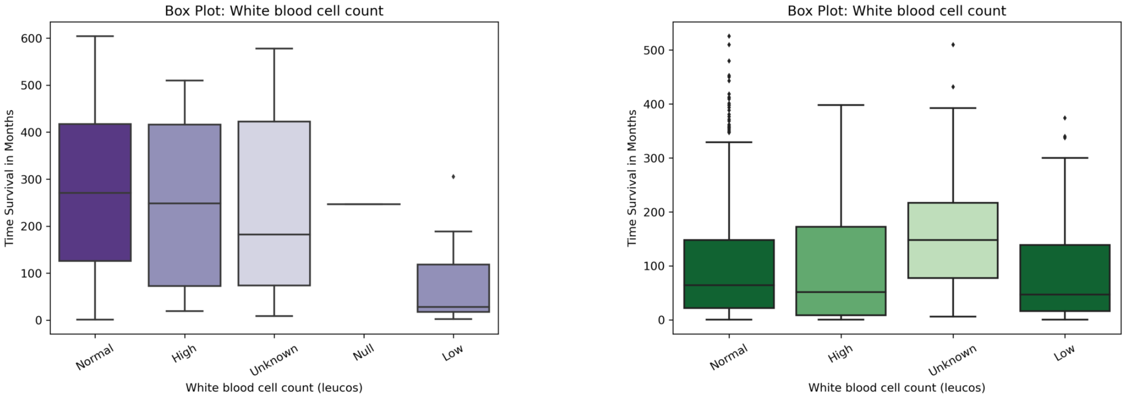


Figure 3.21: Initial Stage descriptive graphs for both types of lymphoma

HL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
White blood cell count	Low	13	3,40%	2,32	6,70	44	19	43,62
	Normal	313	81,94%	22,56	22,67	32	30	35,01
	High	21	5,50%	20,69	19,83	30	18	32,29
	Unknown	35	9,16%	18,39	19,90	27	22	32,09
NHL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
White blood cell count	Low	42	7,50%	3,90	7,69	64	79	60,95
	Normal	409	73,04%	5,35	8,88	62	63	60,44
	High	75	13,39%	4,28	8,33	57	31	56,45
	Unknown	34	6,07%	12,31	13,82	57	52	55,76

Table 3.23: White blood cell count for Hodgkin lymphoma and non-Hodgkin lymphoma

As demonstrated in table 3.24, most patients in both lymphomas have normal levels of white blood cells, and these patients are also the ones who have a higher median survival time.

Lymphocyte count

The lymphocyte count expresses the level of lymphocytes of each patient, as plots 3.22 below show.

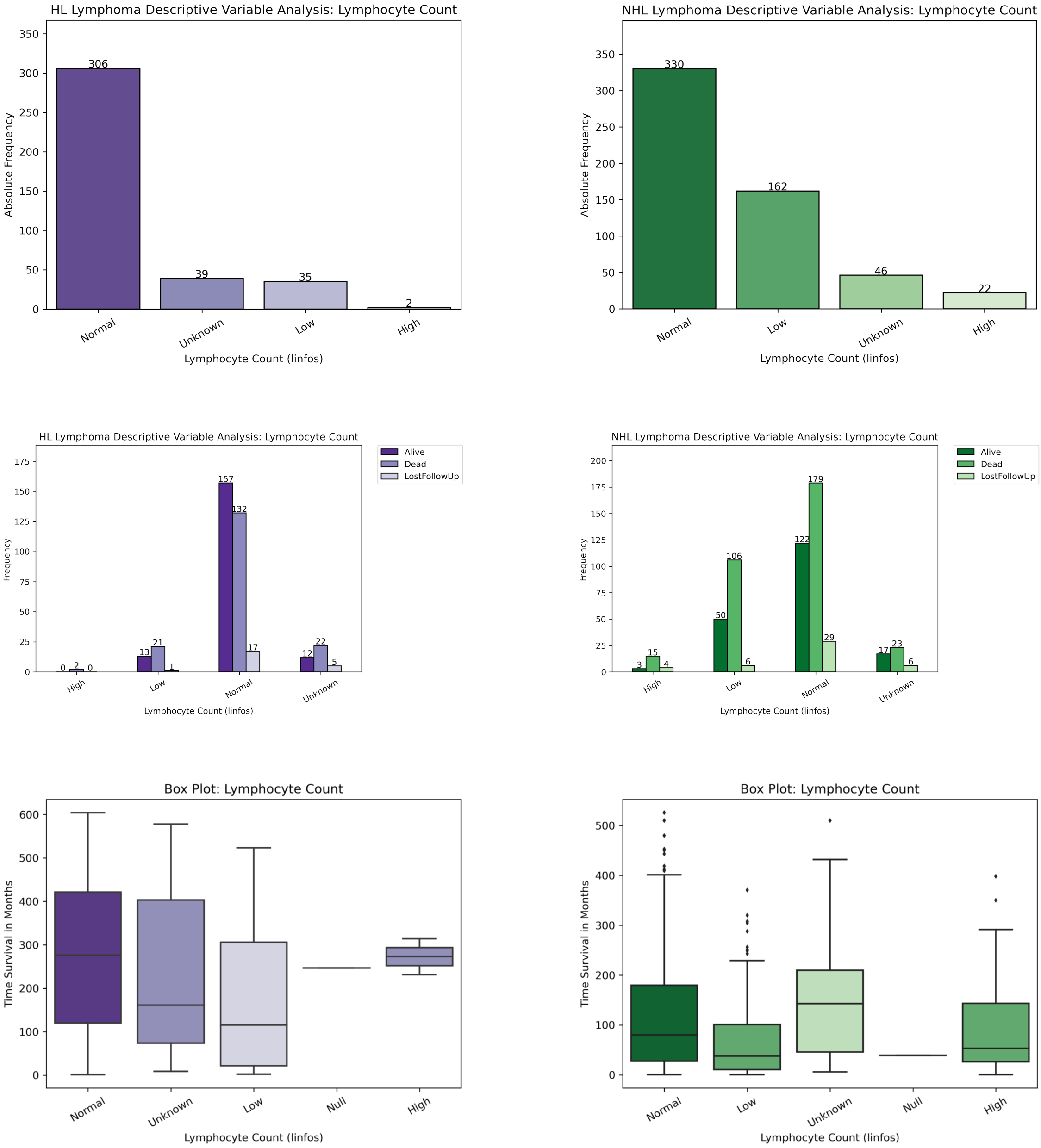


Figure 3.22: Initial Stage descriptive graphs for both types of lymphoma

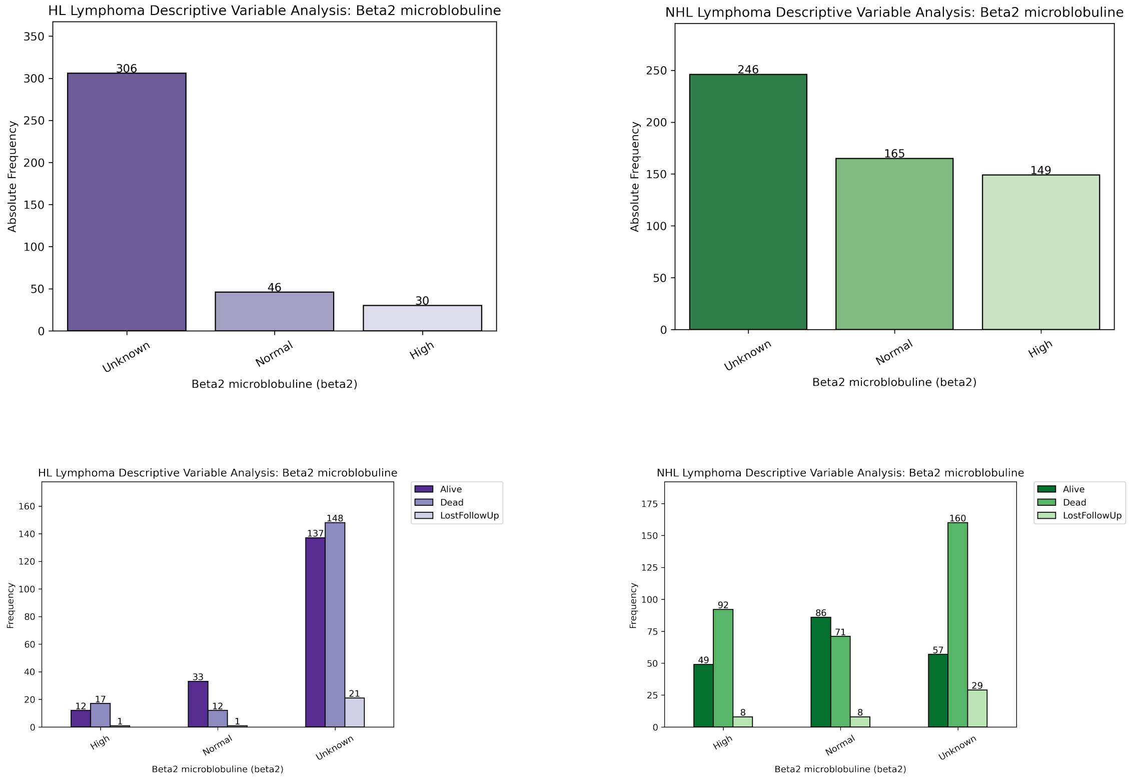
Similar to the previous variable, the lymphocyte levels in both HL and NHL are predominantly normal, as table 3.24 underneath indicates.

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Lymphocyte count	Low	35	9,16%	9,59	14,27	41	19	44,51
	Normal	306	80,10%	23	22,91	31	30	34,08
	High	2	0,52%	22,73	22,73	19,5	14	19,5
	Unknown	39	10,21%	14,86	19,04	27	22	33,38
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Lymphocyte count	Low	162	28,93%	3,15	5,77	63	64	61,73
	Normal	330	58,93%	6,68	10,16	62	63	59,54
	High	22	3,93%	4,40	9,13	60	31	60,09
	Unknown	46	8,21%	11,14	12,29	53	49	53,02

Table 3.24: Lymphocyte count for Hodgkin lymphoma and non-Hodgkin lymphoma

Beta2 microblobuline

The Beta2 microblobuline, or simply Beta2, has the same possible values of either high or normal, and the descriptive analysis is presented in the following plots 3.23.



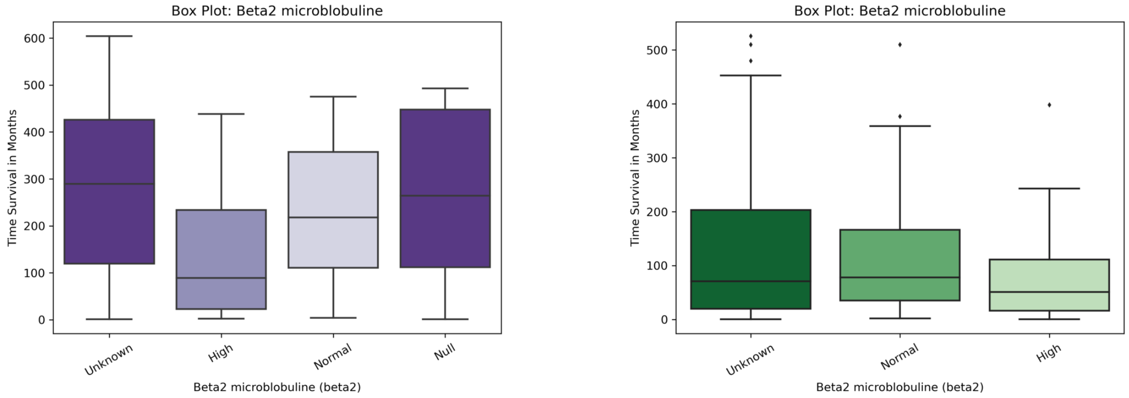


Figure 3.23: Initial Stage descriptive graphs for both types of lymphoma

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Beta2	Normal	46	12,04%	18,15	19,19	29	38	33,87
	High	30	7,85%	7,39	10,91	43	30	46,93
	Unknown	306	80,10%	23,98	23,16	31	22	33,86
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Beta2	Normal	165	29,46%	6,48	9,25	58	49	58,13
	High	149	26,61%	4,26	6,029	66	79	64,72
	Unknown	246	43,93%	5,88	10,68	60	58	57,62

Table 3.25: Beta2 microblobuline for Hodgkin lymphoma and non-Hodgkin lymphoma

Table 3.25 above demonstrates from the patients that the discrepancy between high and normal is quite slim. Despite that, both lymphomas have more patients with levels of Beta2 normal than high.

Lactate dehydrogenase - LDH

The lactate dehydrogenase or LDH has the same possible values as Beta2, and the results of the analyses are presented below in plots 3.24.

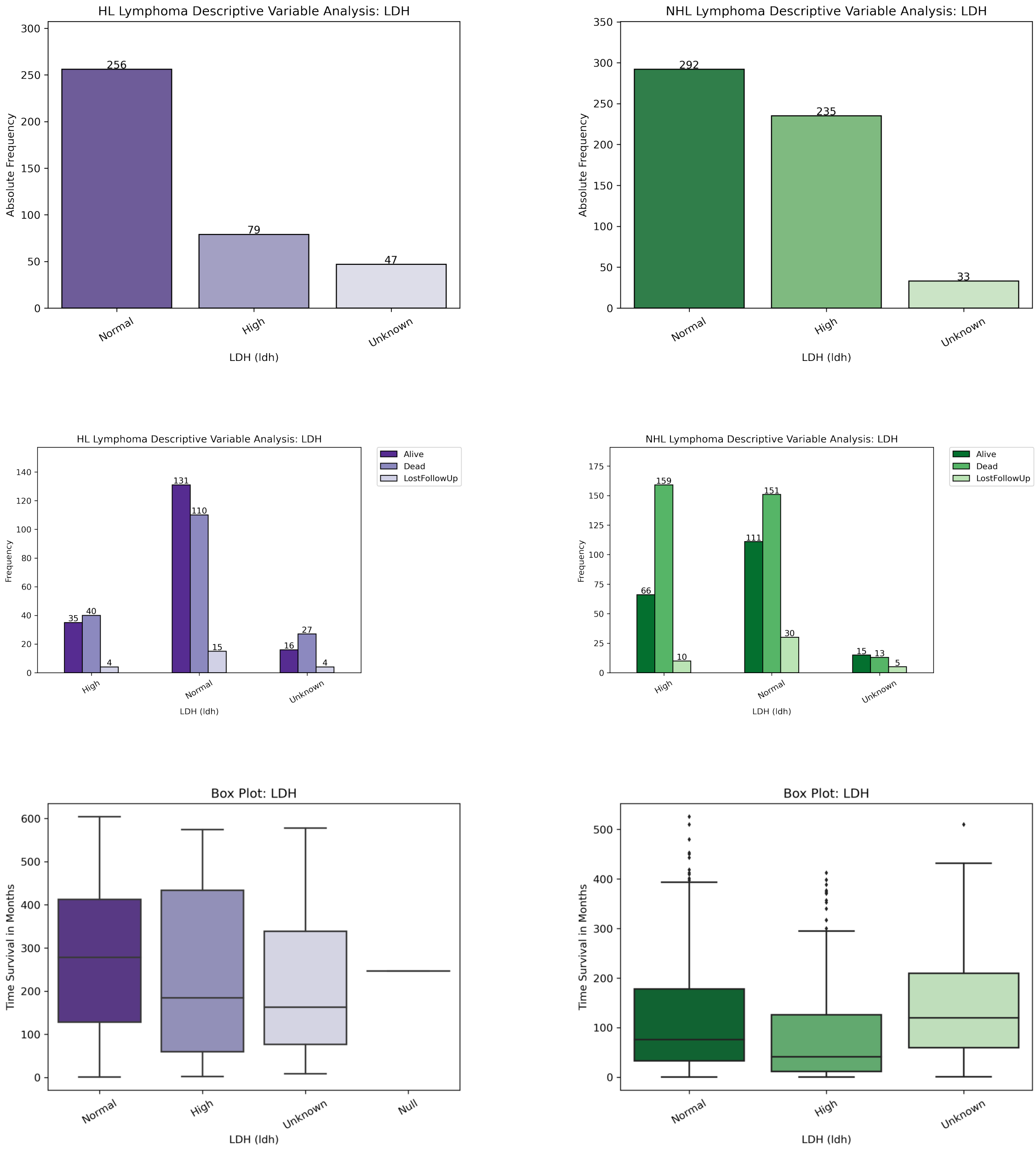


Figure 3.24: Initial Stage descriptive graphs for both types of lymphoma

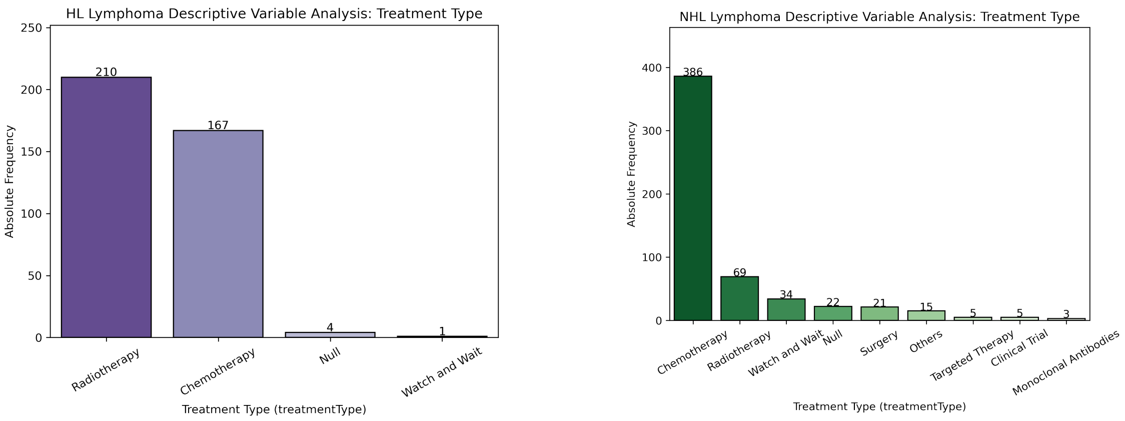
HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
LDH	Normal	256	67,02%	23,21	22,92	30	30	34,39
	High	79	20,68%	15,39	19,84	33	31	36,89
	Unknown	47	12,30%	15,15	18,33	32	22	34,21
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
LDH	Normal	292	52,14%	6,31	10,15	61	79	60,22
	High	235	41,96%	3,41	7,10	63	64	59,76
	Unknown	33	5,89%	9,97	12,73	54	45	54,03

Table 3.26: Lactate dehydrogenase for Hodgkin lymphoma and non-Hodgkin lymphoma

As stated in table 3.26, both lymphomas are predominantly normal LDH (other than unknown), despite a more significant discrepancy between normal and higher in HL with 67% normal and 21% high, than NHL with 52% normal and 42% high.

3.3.3.13 Treatment Type

The treatment of patients differs between the typology of lymphoma because each lymphoma requires a specific treatment. Presented below, in both graphs 3.25 and table 3.27, are the types of treatment that patients with both types of lymphomas have undergone.



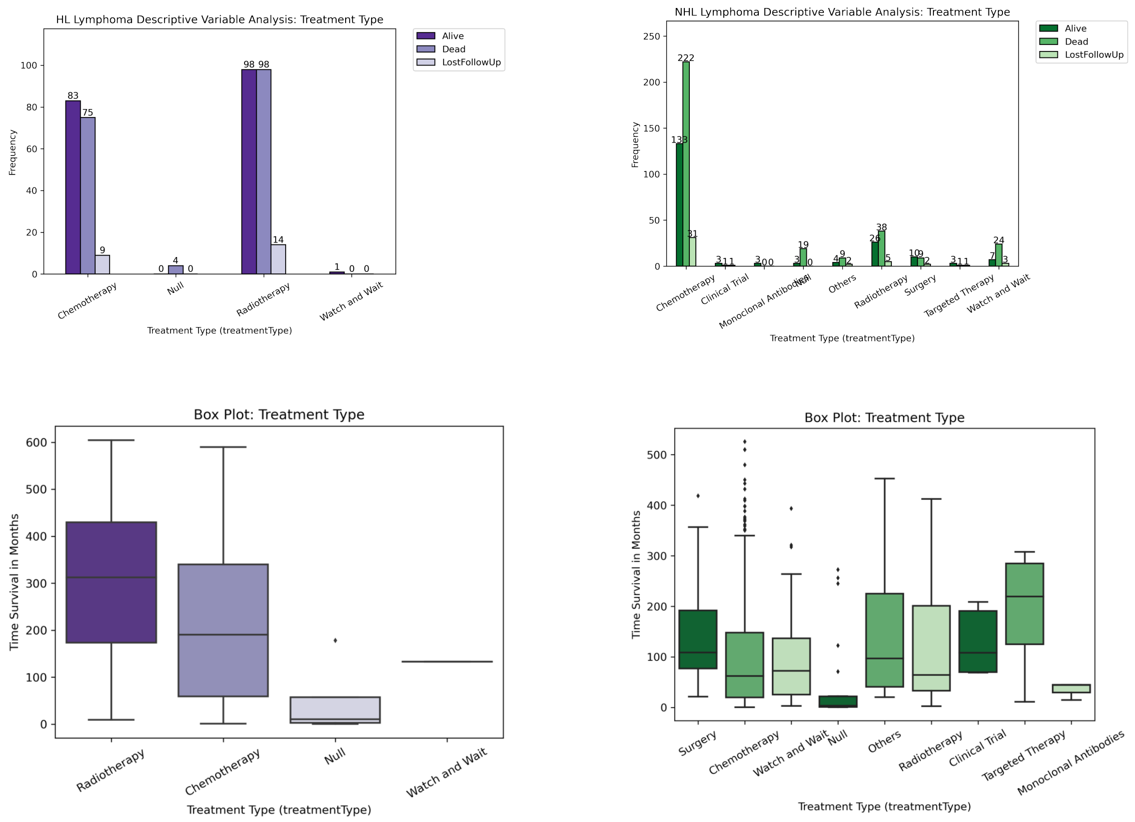


Figure 3.25: Treatment type descriptive graphs for both types of lymphoma

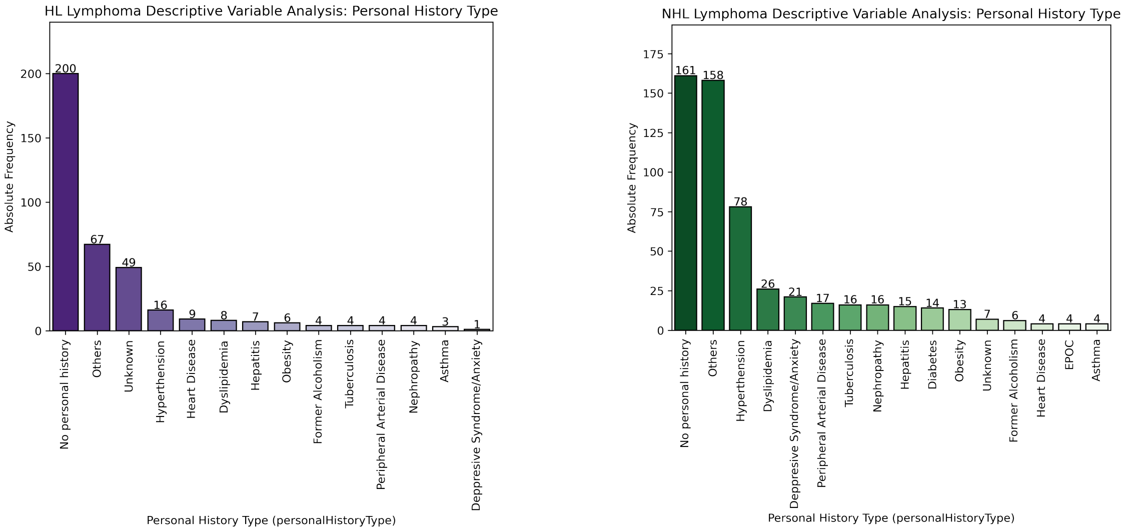
HL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
Treatment Type	Radiotherapy	210	54,97%	26,03	25	29,5	30	31,87
	Chemotherapy	167	43,72%	15,89	18,08	34	29	38,19
	Watch and Wait	1	0,26%	11,07	11,07	48	48	48
	Unknown	4	1,05%	0,86	4,17	48,5	25	52,25
NHL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
Treatment Type	Surgery	21	3,75%	9,05	13,09	59	51	60,48
	Chemotherapy	386	68,93%	5,18	8,69	60	79	58,14
	Radiotherapy	69	12,32%	5,34	10,21	61	63	59,59
	Clinical Trial	5	0,89%	9	10,76	49	11	41,4
	Targeted Therapy	5	0,89%	18,29	15,80	62	45	63,8
	Monoclonal Antibodies	3	0,54%	3,69	2,88	64	51	63,67
	Watch and Wait	34	6,07%	6,01	8,74	66,5	63	64,38
	Others	15	2,68%	8,09	12,84	60	60	60,73
	Unknown	22	3,93%	0,26	3,95	83,5	74	80,36

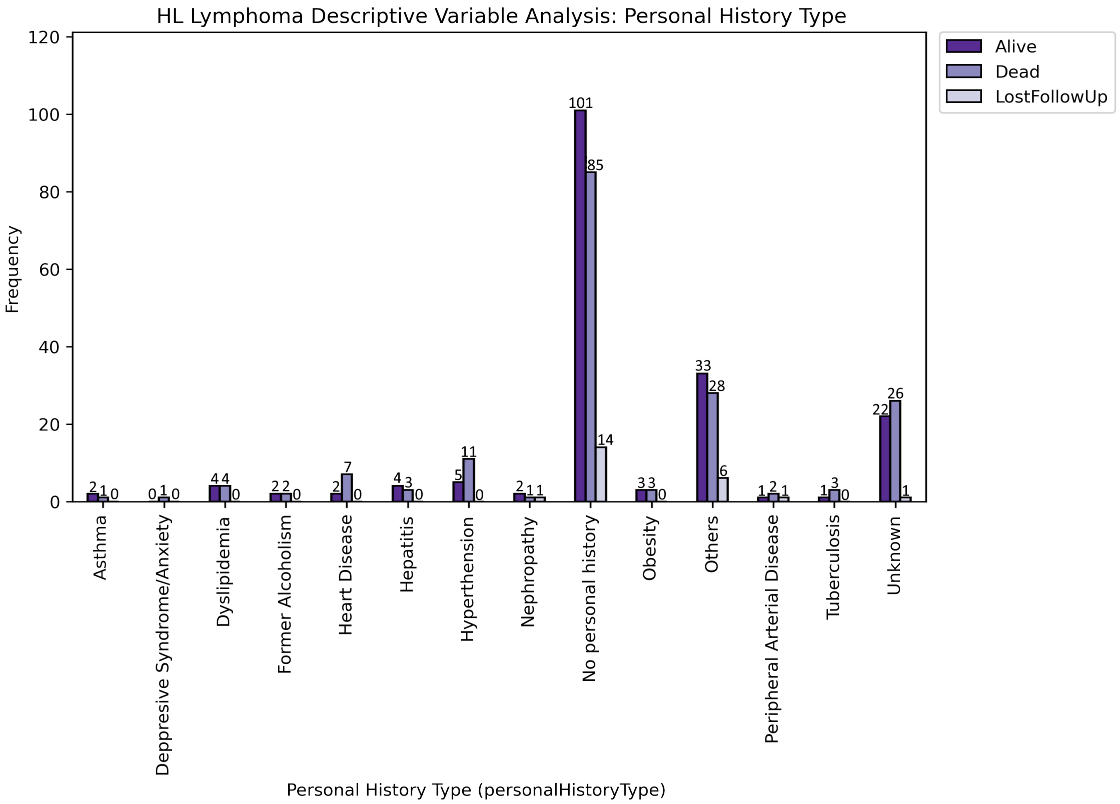
Table 3.27: Treatment type for Hodgkin lymphoma and non-Hodgkin lymphoma

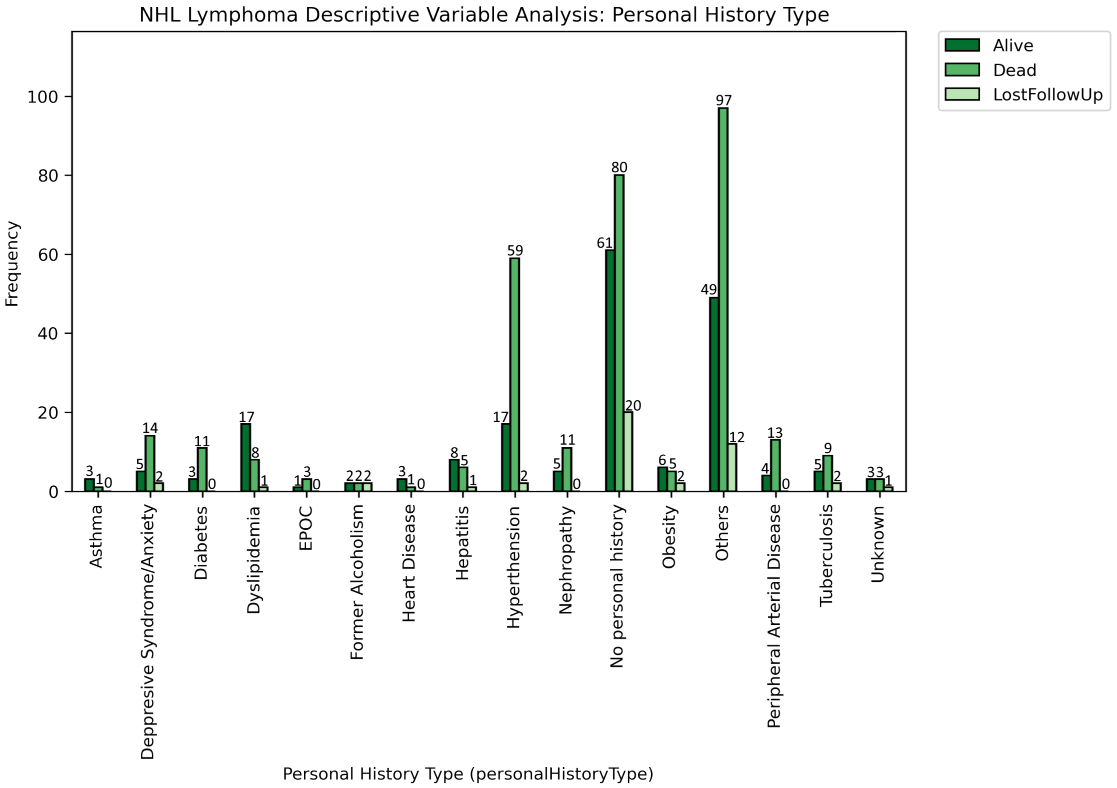
As stated in table 3.27, mentioned above, the majority of HL treatments are radiotherapy, with 55% of patients, followed by chemotherapy with 44% of patients. On the other hand, in non-Hodgkin lymphoma, the primary treatment is chemotherapy with 69% of treatments, loosely followed by radiotherapy with 12%.

3.3.3.14 Personal History Type

This variable describes patients' personal history in both datasets, which is very diverse. The plots 3.26 presented below represent the distribution of this variable for both types of lymphoma.




HL Lymphoma Descriptive Variable Analysis: Personal History Type
Alive
Dead
LostFollowUp
Frequency
0
20
40
60
80
100
120
2 1 0
0 1 0
4 4 0
2 2 0
2 7 0
4 3 0
5 11 0
2 1 1
101 85 14
3 3 0
33 28 6
1 2 1
1 3 0
22 26 1
Asthma
Deppresive Syndrome/Anxiety
Dyslipidemia
Former Alcoholism
Heart Disease
Hepatitis
Hyperthension
Nephropathy
No personal history
Obesity
Others
Peripheral Arterial Disease
Tuberculosis
Unknown
Personal History Type (personalHistoryType)


NHL Lymphoma Descriptive Variable Analysis: Personal History Type
Alive
Dead
LostFollowUp
Frequency
0
20
40
60
80
100
3 1 0
5 14 2
3 11 0
17 8 1
1 3 0
2 2 2
3 1 0
8 5 1
17 59 2
5 11 0
61 80 20
6 5 2
49 97 12
4 13 0
5 9 2
3 3 1
Asthma
Deppresive Syndrome/Anxiety
Diabetes
Dyslipidemia
EPOC
Former Alcoholism
Heart Disease
Hepatitis
Hyperthension
Nephropathy
No personal history
Obesity
Others
Peripheral Arterial Disease
Tuberculosis
Unknown
Personal History Type (personalHistoryType)

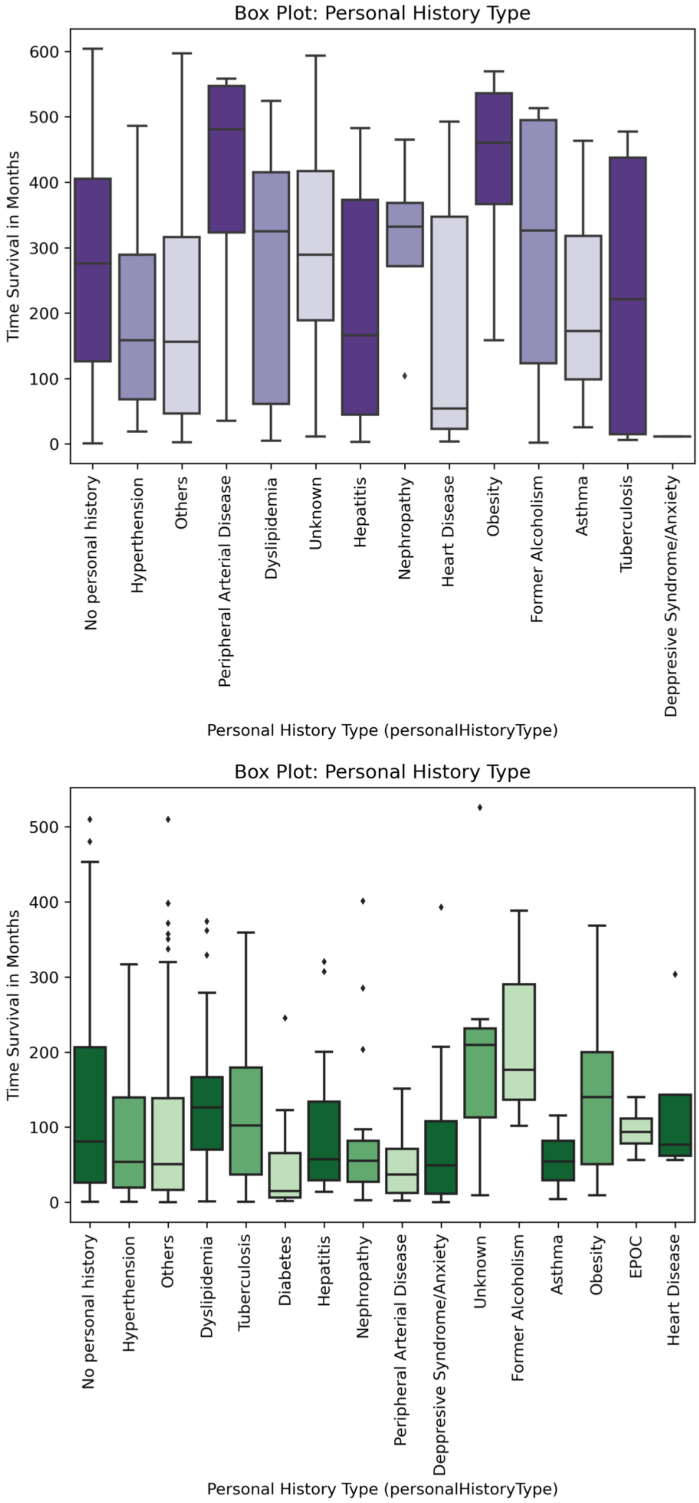


Figure 3.26: Personal history type descriptive graphs for both types of lymphoma

Observing table 3.28 presented below, it is clear that there is a large number of values that this variable can contain. The most common specified condition in both lymphomas is hypertension, with 4% and 14% in HL and NHL, respectively.

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Personal History	No personal history	200	52,36%	23	22,87	28	18	30,05
	Hyperthension	16	4,19%	13,2	17,27	55,5	71	56,06
	Others	67	17,54%	12,99	17,17	34	29	39
	Peripheral Arterial Disease	4	1,05%	40,09	32,42	39,5	28	46,75
	Dyslipidemia	8	2,09%	27,08	22,64	42,5	15	39,13
	Unknown	49	12,83%	24,11	24,19	30	19	34,55
	Hepatitis	7	1,83%	13,84	17,72	35	17	43,57
	Nephropathy	4	1,05%	27,66	25,68	39,5	30	39,5
	Heart Disease	9	2,36%	4,49	14,48	51	51	47
	Obesity	6	1,57%	38,37	35,36	40	12	37,67
	Former Alcoholism	4	1,05%	27,18	24,32	47	35	48,75
	Asthma	3	0,79%	14,37	18,35	32	14	31
	Tuberculosis	4	1,05%	18,42	19,28	40,5	32	46,25
	Deppresive Syndrome/Anxiety	1	0,26%	0,94	0,94	30	30	30
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Personal History	Dyslipidemia	26	4,64%	10,53	11,56	57	43	58,38
	Tuberculosis	16	2,86%	8,52	10,07	58	58	57,75
	Diabetes	14	2,50%	1,27	4,25	74	70	72,07
	Hepatitis	15	2,68%	4,79	8,45	55	49	56,73
	Nephropathy	16	2,86%	4,60	7,58	72,5	45	71,06
	Peripheral Arterial Disease	17	3,04%	3,06	4,33	73	64	71,94
	Deppresive Syndrome/Anxiety	21	3,75%	4,12	6,03	61	56	62,38
	Former Alcoholism	6	1,07%	14,67	17,96	56	55	52,17
	Asthma	4	0,71%	4,52	4,75	63	58	64
	Obesity	13	2,32%	11,67	12,84	63	63	64,54
	EPOC	4	0,71%	7,81	7,99	69	62	69,75
	Heart Disease	4	0,71%	6,38	10,68	53,5	25	53
	Hyperthension	78	13,93%	4,49	7,14	70	54	69,40
	No personal history	161	28,75%	6,74	11,10	49	36	48,12
	Others	158	28,21%	4,23	7,91	65,5	59	63,29
	Unknown	7	1,25%	17,46	17,07	54	31	54

Table 3.28: Personal history for Hodgkin lymphoma and non-Hodgkin lymphoma

3.3.3.15 Rheumatic Disease Type

Another type of disease deemed valuable to develop the rheumatic disease. Below is presented the descriptive analysis for HL and NHL in graphs 3.27.

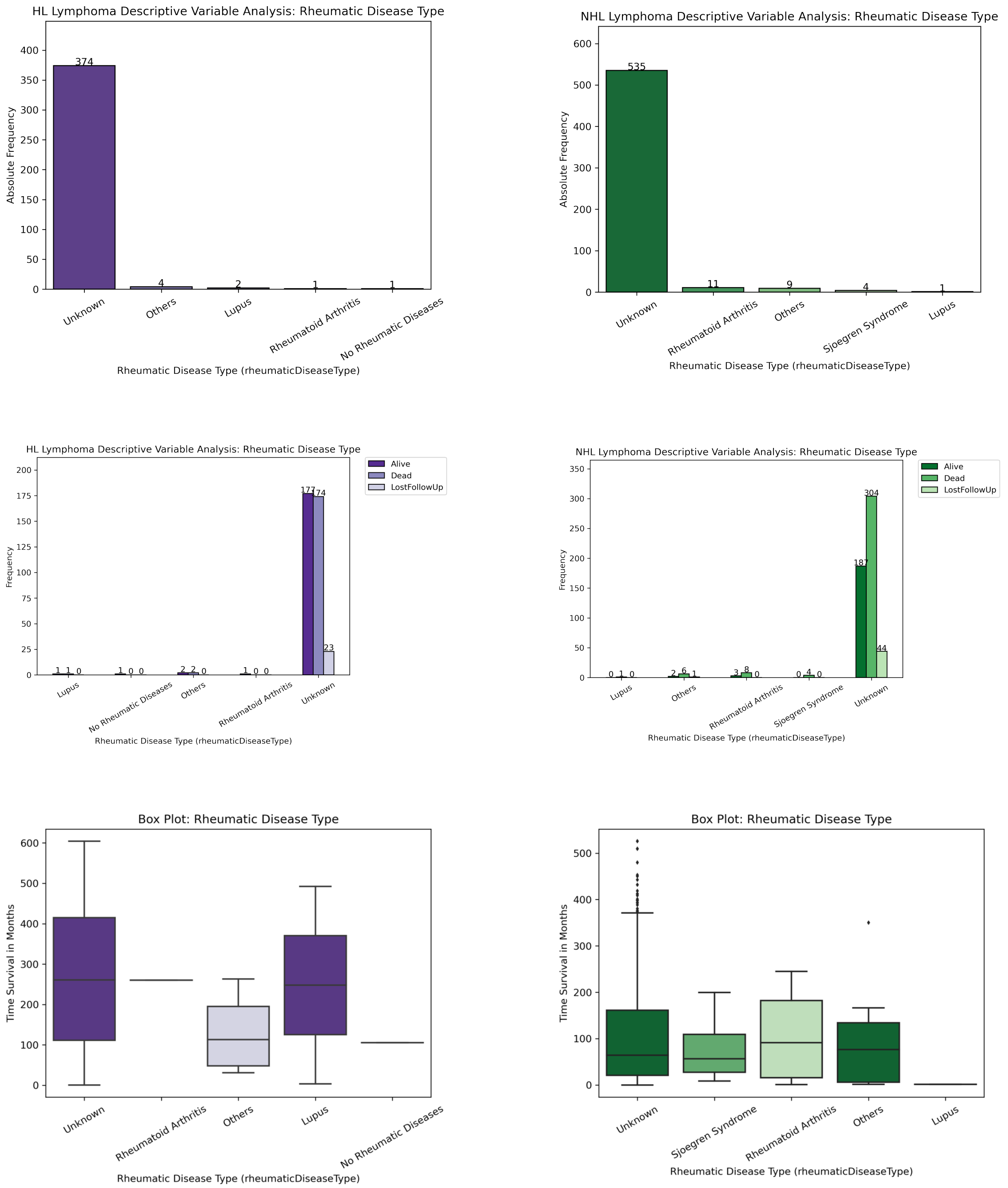


Figure 3.27: Rheumatic history type descriptive graphs for both types of lymphoma

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Rheumatic Disease	Lupus	2	0,52%	20,66	20,66	56,5	26	56,5
	Rheumatoid Arthritis	1	0,26%	21,68	21,68	37	37	37
	Others	4	1,05%	9,43	10,84	38,5	24	44,25
	No Rheumatic Diseases	1	0,26%	8,77	8,77	66	66	66
	Unknown	374	97,91%	21,74	21,88	31	22	34,58
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Rheumatic Disease	Lupus	1	0,18%	0,12	0,12	63	63	63
	Sjoegren Syndrome	4	0,71%	4,71	6,68	70	57	69
	Rheumatoid Arthritis	11	1,96%	7,63	8,84	67	60	67,55
	Others	9	1,61%	6,38	8,26	59	76	60,78
	Unknown	535	95,54%	5,35	9,07	61	79	59,40

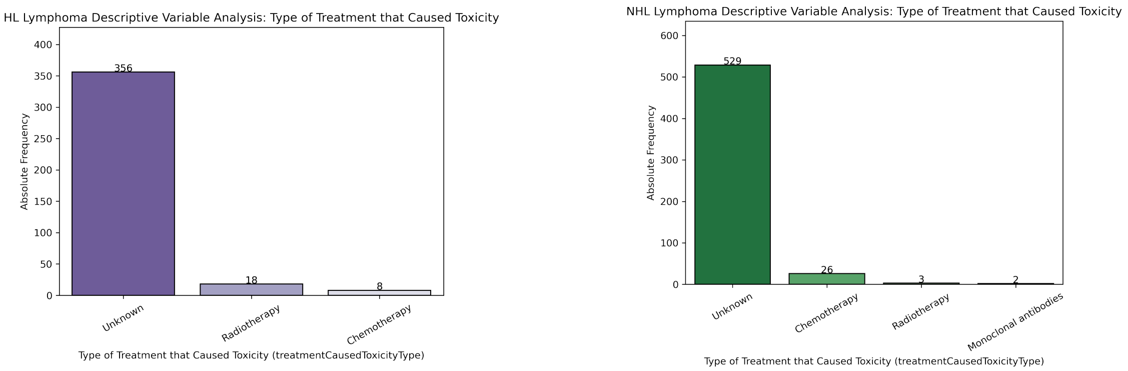
Table 3.29: Rheumatic disease type for Hodgkin lymphoma and non-Hodgkin lymphoma

As it can be concluded by analysing table , very few patients manifested any rheumatic disease. Despite being one of the recommended variables to analyse to determine the probability of survival, the lack of data makes this variable unsuitable for extracting any conclusions in future chapters.

3.3.3.16 Treatment-caused Toxicity Type

As one of the initial objectives was to study the effects of treatment-caused toxicities, the analysis of this variable was pertinent. The variable studied in this subsection describes the toxicities that were formed due to treatments administered to the patients.

The graphs 3.28 below express the possible values the variable can have in both types of lymphoma.



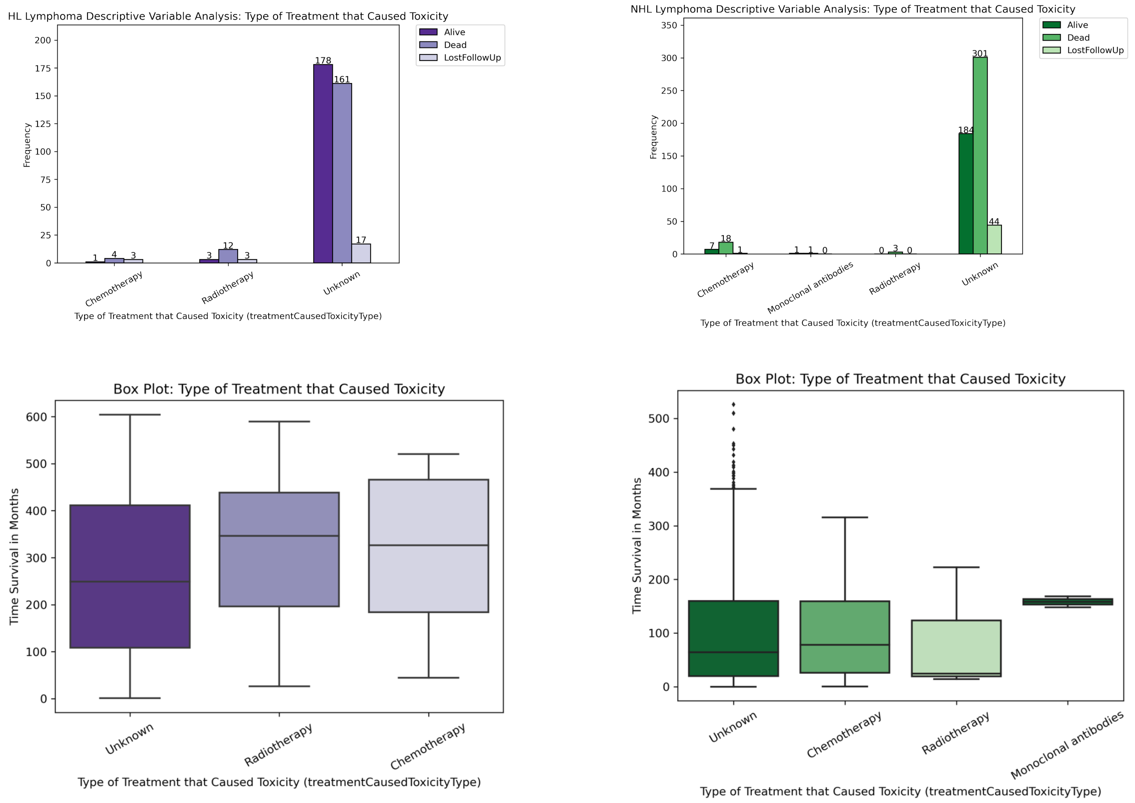


Figure 3.28: Type of treatment that caused toxicity descriptive graphs for both types of lymphoma

Observing the above plots and combining that information with the tables below, 3.30, it can be noted that in both lymphomas, the values are predominantly unknown, nearly 93% in HL and 94% in NHL. Due to the lack of classified data, this variable should not be used to extract any knowledge from the datasets.

HL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Treatment Caused Toxicity Type	Radiotherapy	18	4,71%	28,85	27,04	32	21	31,33
	Chemotherapy	8	2,09%	27,18	25,46	26,5	24	29,25
	Unknown	356	93,19%	20,75	21,37	32	22	35,19
NHL	Variable	Absolute	Relative	Survival Years		Age Diagnosis		
	Value	Count	Count	Median	Mean	Median	Mode	Mean
Treatment Caused Toxicity Type	Radiotherapy	3	9,68%	2,05	7,29	48	41	49,33
	Chemotherapy	26	83,87%	6,50	8,19	59,5	49	60,31
	Monoclonal antibodies	2	6,45%	13,2	13,2	71	70	71
	Unknown	529	94,46%	5,35	9,06	62	79	59,64

Table 3.30: Treatment caused toxicity type for Hodgkin lymphoma and non-Hodgkin lymphoma

3.3.3.17 Treatment Related Diseases

Alongside the previous subsection, the analysis of the treatment-related diseases was one of the initial objectives of the overall analysis. This variable describes the type of disease originating from the provided treatment. The plots 3.29 below show the different body zone affected.

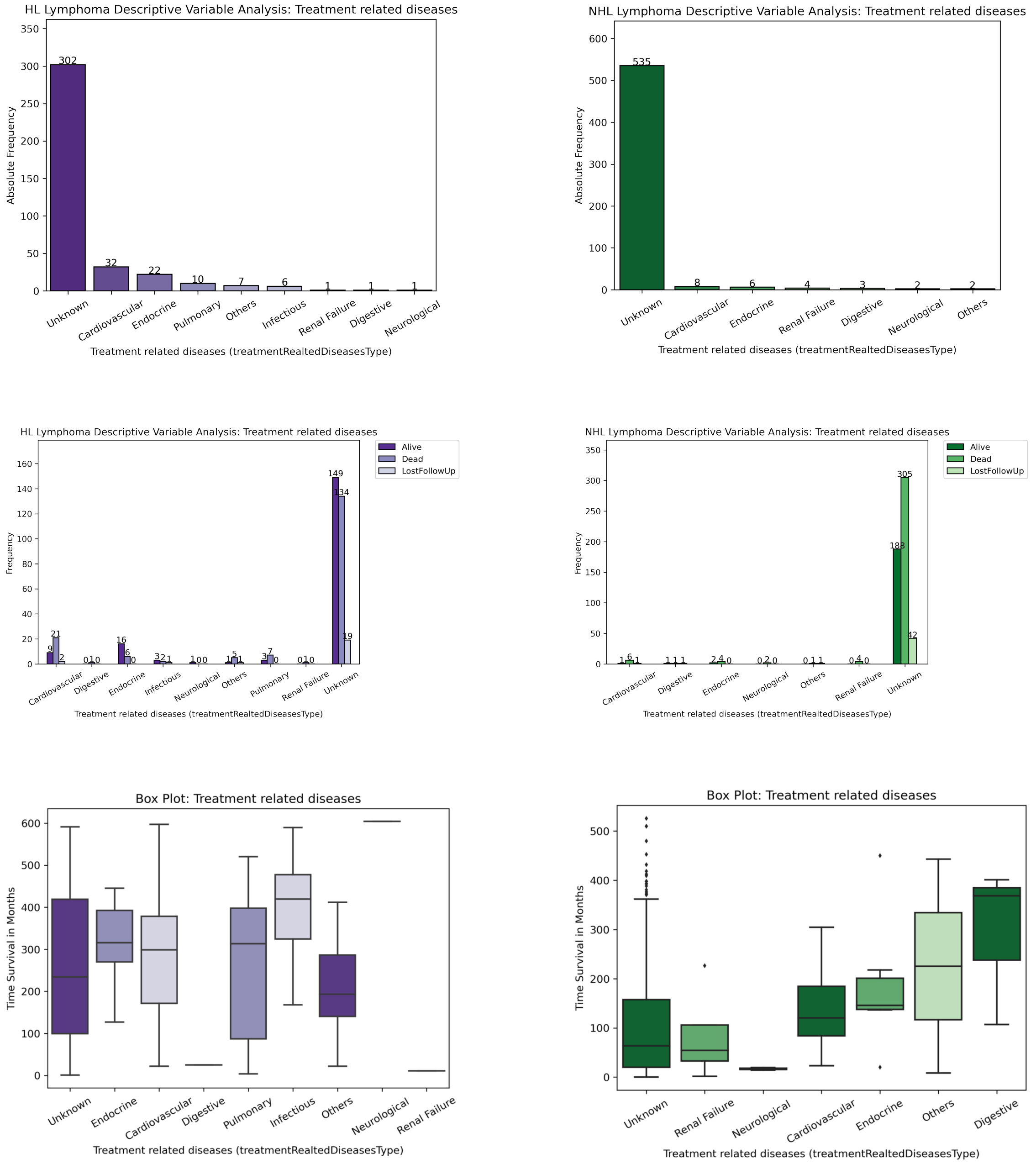


Figure 3.29: Type of treatment that caused toxicity descriptive graphs for both types of lymphoma

HL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
Treatment Related Diseases	Renal Failure	1	0,26%	0,93	0,93	75	75	75
	Neurological	1	0,26%	50,36	50,36	52	52	52
	Cardiovascular	32	8,38%	24,89	24,38	34,5	19	37,75
	Endocrine	22	5,76%	26,32	26,19	26	23	27,05
	Digestive	1	0,26%	2,08	2,08	29	29	29
	Pulmonary	10	2,62%	26,13	22	21,5	19	32,6
	Infectious	6	1,57%	34,95	33,14	33	19	31,83
	Others	7	1,83%	16,09	17,63	33	21	34,71
	Unknown	302	79,06%	19,53	21,01	32	22	35,13
NHL	Variable Value	Absolute Count	Relative Count	Survival Years		Age Diagnosis		
				Median	Mean	Median	Mode	Mean
Treatment Related Diseases	Renal Failure	4	0,71%	4,53	7,01	71	56	69,25
	Neurological	2	0,36%	1,40	1,40	59,5	59	59,5
	Cardiovascular	8	1,43%	10,03	11,63	50	14	46,75
	Endocrine	6	1,07%	12,17	15,51	46	22	47,83
	Digestive	3	0,54%	30,71	24,34	62	52	62
	Others	2	0,36%	18,79	18,79	35	17	35
	Unknown	535	95,54%	5,30	8,83	62	63	59,99

Table 3.31: Treatment related diseases for Hodgkin lymphoma and non-Hodgkin lymphoma

Similarly to the previous section, the treatment-related diseases have hardly any representation in the data received. The tables underneath present the quantitative results, and as it can be observed in 3.31, approximately 76% of HL patients do not have any associated disease. In comparison, nearly 96% of NHL patients are in the same conditions. Therefore like the previously analysed variable, the number of patients to study if the phenomenon is too low and will not be executed for the received dataset.

3.3.4 Document Generator Tool

An additional tool was developed to aid in the creation of reports and gatherings of the initial data analysis to optimise the validation and result presentation to the medical specialist's team.

This tool was developed using the previously presented docx library. The ability to integrate the previously saved graphs and tables that describe the data distribution, with and without unknown values, as per project requisite, was done beforehand using the ability to save matplotlib graphs as images and using xlsxWriter to create the tables.

The choice to save the graphs as images beforehand was deliberate to keep the image quality to the max and avoid distortion and data visualisation problems. The standard dpi (dots per inch) used across the board was 400 to keep the crispness of the image.

Likewise, making the creation of the tables an independent asset instead of calculating it within the document creation process was deliberate, as this separation and modulation of features allow the independent use of the created tables. The modularity allows for exporting all the tables corresponding to a specific type of lymphoma without any extra problems and, consequently, a better and more objective communication line and data sharing within the project.

3.3.5 Initial Statistical Analysis Results Ponderations

Due to the lack of proportion in the second and third data divisions of HL structured time since diagnosis, the choice to agglomerate the partitions was made. Consequently, when analysing the time since diagnosis in HL, there will only be two structured intervals less than 45 years of age and greater or equal to 45 years of age since after this age, there is no significant distinction of a possible regression. This decision was reinforced by the medical team's professional opinion regarding the medical analysis in these same age groups.

The choice to present the initial structure is to show the initial assumption and what the medical assumption expressed when analysing the initial data analysis. It was also important to divulge the default separation since reanalysing the age intervals is crucial if or when the data is updated.

4

Kaplan-Meier survival analysis

This chapter will focus on the Kaplan-Meier estimator analysis. As stated in subsection 2.4.1, the Kaplan-Meier estimator is a univariate analysis for survival analysis. The event in this study will be the death of a patient, and the time variable will be presented in years to facilitate interpretation

The log-rank test presented in subsection 2.4.2 will be used to help discern whether or not there are significant differences in the survival of different groups within the same variable. The p-value obtained in the log-rank test determines whether or not the survival is statistically different.

After the Kaplan-Meier analysis, some conclusions regarding the results will be presented regarding the data and the following procedures.

4.1 Kaplan-Meier Ponderations

In addition to the following analysed and shown variables presented in section 4.2, the remnant variables were analysed but were deemed as not relevant partially due to the lack of valuable data that characterises them.

The reached results are not very helpful in the continuation of this analysis, and this is due to the data received. Despite that, all the processes and the previously presented analysis were created in modules, so any time there is an updated database, the process is plug-and-play without any additional problems.

Since the data does not provide enough content to either proceed to a Cox analysis or deepen the Kaplan-Meier analysis, services and considering that all the created modules are plug and play, the elaboration of a new tool to aid the overall project analysis was conceived.

The tool in question aims to be a widely applicable tool to perform the Kaplan-Meier analysis for any given variable of various cancer types contained within the project.

Although this does not provide a full survival analysis for the received data, that analysis would not provide any additional information since the data used would not be of good enough quality following the data science proverb of "garbage in garbage out"[109].

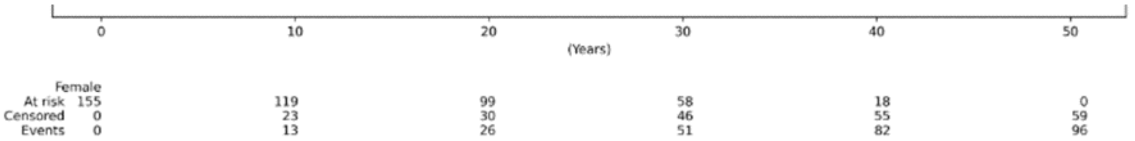
Despite the lack of valuable conclusions, the following presented Kaplan-Meier estimates will present the survival knowledge that could be extracted from the analysed data.

4.2 Kaplan-Meier Estimator

The Kaplan-Meier estimates presented in this section will mostly coincide with an analysis of the previously used variables for the initial descriptive analysis. Some of the variables from that analysis will be omitted since there is not enough data to elaborate a significant Kaplan-Meier analysis.

Taking this into consideration and despite not being presented in the dissertation due to the formerly present argument, all Kaplan-Meier's analyses were done to confirm the poor results and if there was any result worth presenting.

Underneath each presented plot, there will be a table with values accompanying the x-axis like the one presented below in figure 4.1



	0	10	20	30	40	50
Female						
At risk	155	119	99	58	18	0
Censored	0	23	30	46	55	59
Events	0	13	26	51	82	96

Figure 4.1: Example of the table presented below each Kaplan-Meier estimator plot

Figure 4.1 presents an example of some results using the Kaplan-Meier estimator. Every value of the variable analysed will present an identical table to the abovementioned one. Each table will contain three rows, the "at risk" row, the "censored" row and the "events" row.

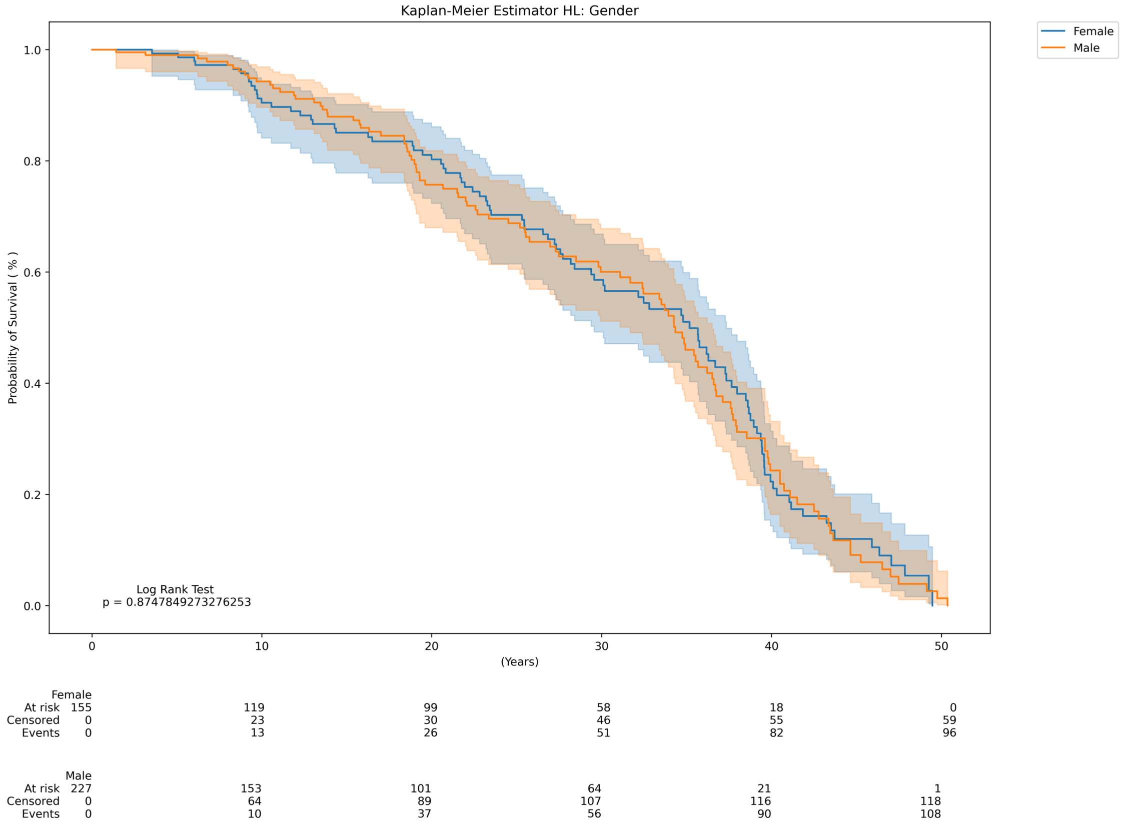
The "at risk" presents the number of patients that, in that instant of time, are eligible for the study and are alive. The second row, "censored", is the number of patients that have not yet survived the needed amount of time represented in the x-axis and, therefore, are censored and do not contribute to further estimates in the timeline . Lastly, the row, "events" presents the number of patients in which the studied event has occurred which, in the case of this analysis, is the death of the studied patient.

These tables complement the plots they accompany, making it easier to discern the substantial amount of patients that have died or have not survived and the sufficient amount of time to be used forward in the rest of the Kaplan-Meier analysis.

It is also important to note that each plot's confidence intervals are represented by the shaded area surrounding the curve with match colours.

4.2.1 Gender

This subsection presents the results for the Kaplan-Meier estimator for gender for both lymphomas.



Female						
At risk	155	119	99	58	18	0
Censored	0	23	30	46	55	59
Events	0	13	26	51	82	96
Male						
At risk	227	153	101	64	21	1
Censored	0	64	89	107	116	118
Events	0	10	37	56	90	108

Figure 4.2: Kaplan-Meier estimator for Hodgkin lymphoma's variable gender

Figure 4.2 above presents the Kaplan-Meier analysis of Hodgkin lymphoma results. As it can be observed with a p-value of approximately 0.87, there is significant statistical relevance to conclude that the curves for male and female genders are not sufficiently different to make gender a significative variable when determining the probability of surviving the lymphoma at hand.

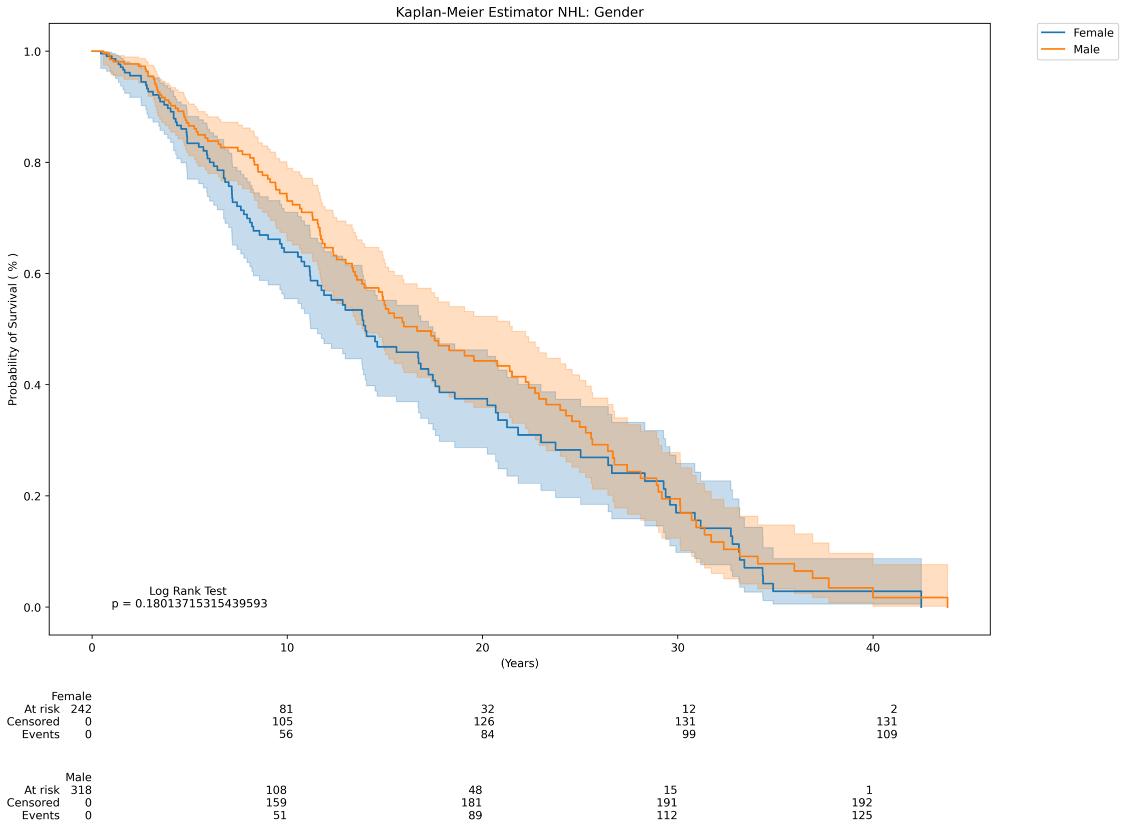
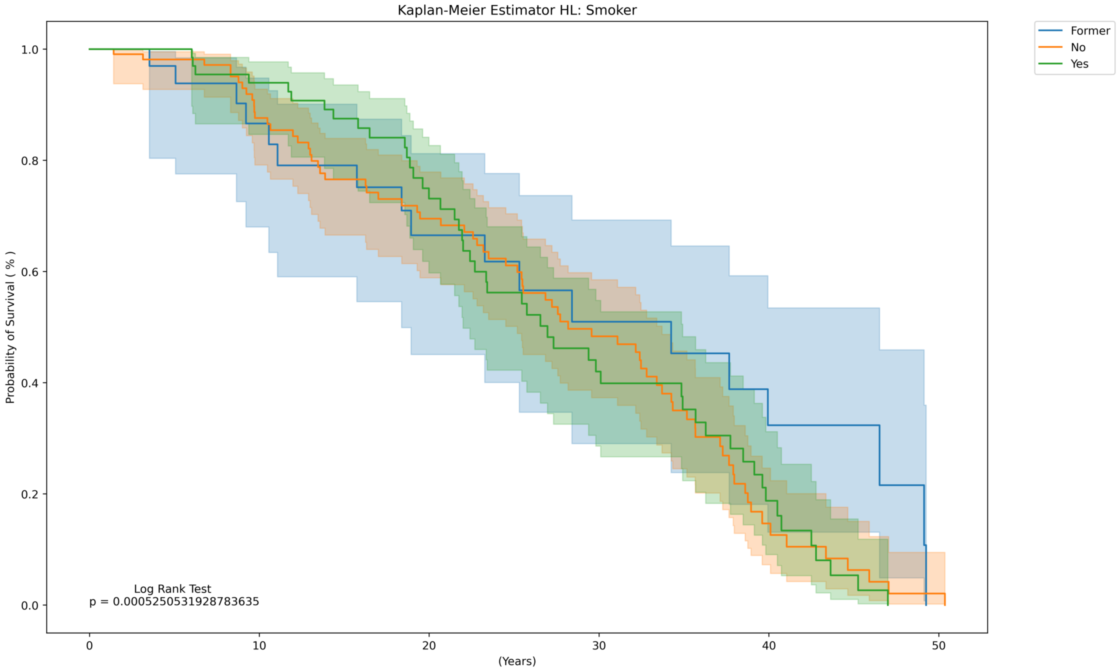


Figure 4.3: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable gender

Similarly, the p-value result of 0.18, as portrayed in figure 4.3, is still greater than the established threshold to determine a significant difference between both curves. This result can be a consequence of the curves crossing each other at around 27 years, and the crossing probably results from the lack of patients in those conditions. Otherwise, before the survival curves cross each other, both appear to be distinct between them.

4.2.2 Smoking Habits

The figures presented below represent the results for the Kaplan-Meier estimator on the smoker variable for Hodgkin lymphoma and non-Hodgkin lymphoma.



Former						
At risk	43	24	15	9	5	0
Censored	0	15	19	22	23	25
Events	0	4	9	12	15	18

No						
At risk	115	81	59	36	7	1
Censored	0	22	28	34	42	42
Events	0	12	28	45	66	72

Yes						
At risk	82	60	39	20	8	0
Censored	0	18	27	30	32	33
Events	0	4	16	32	42	49

Figure 4.4: Kaplan-Meier estimator for Hodgkin lymphoma's variable smoking habits

Analysing the results for Hodgkin lymphoma in figure 4.4, it can observe that the p-value, despite being lower than the considered threshold of 0.05 with the value approximately of 0.0005, that this test is not highly dependable since the curves cross each other, and consequently, the log-rank test has no longer the prerequisites to do be correctly performed.

It is also important to note that the sample size does not bring any assurance for the obtained estimates, which are corroborated by the 95% confidence intervals presented.

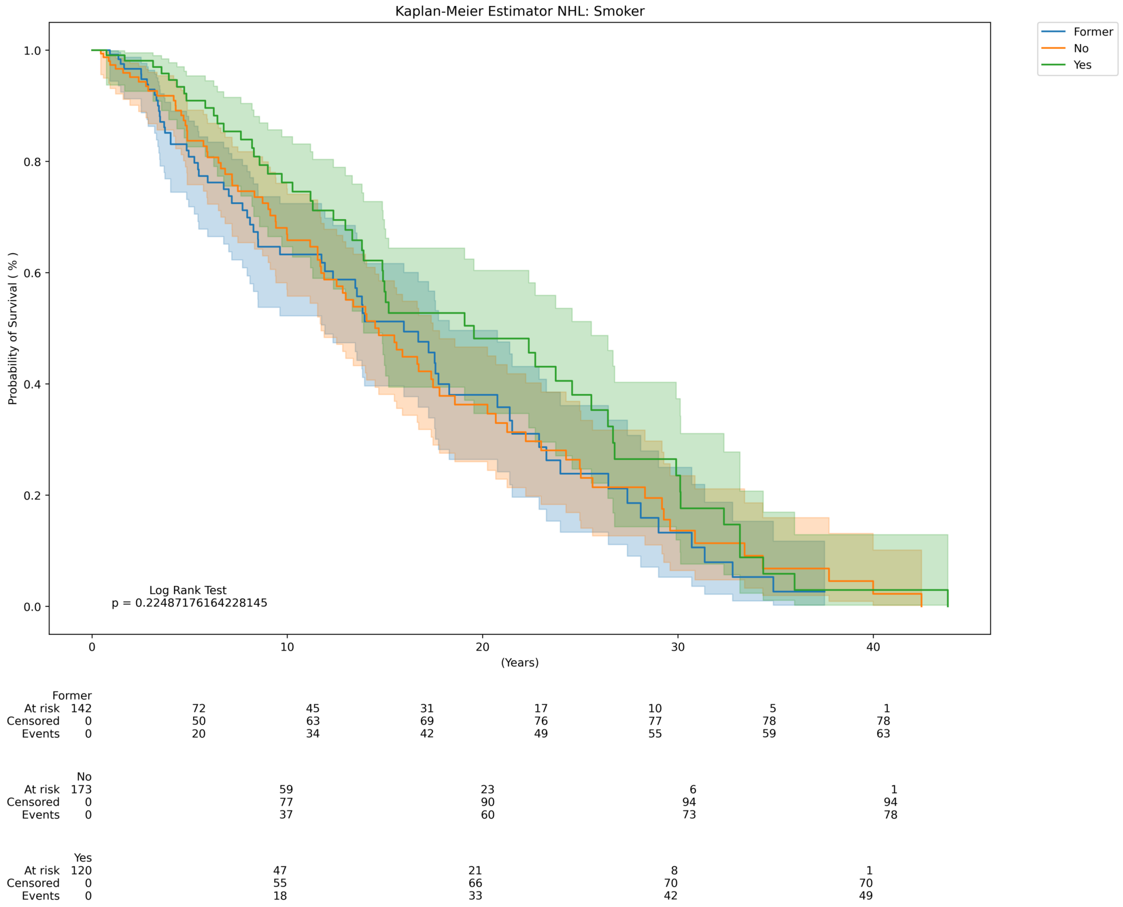
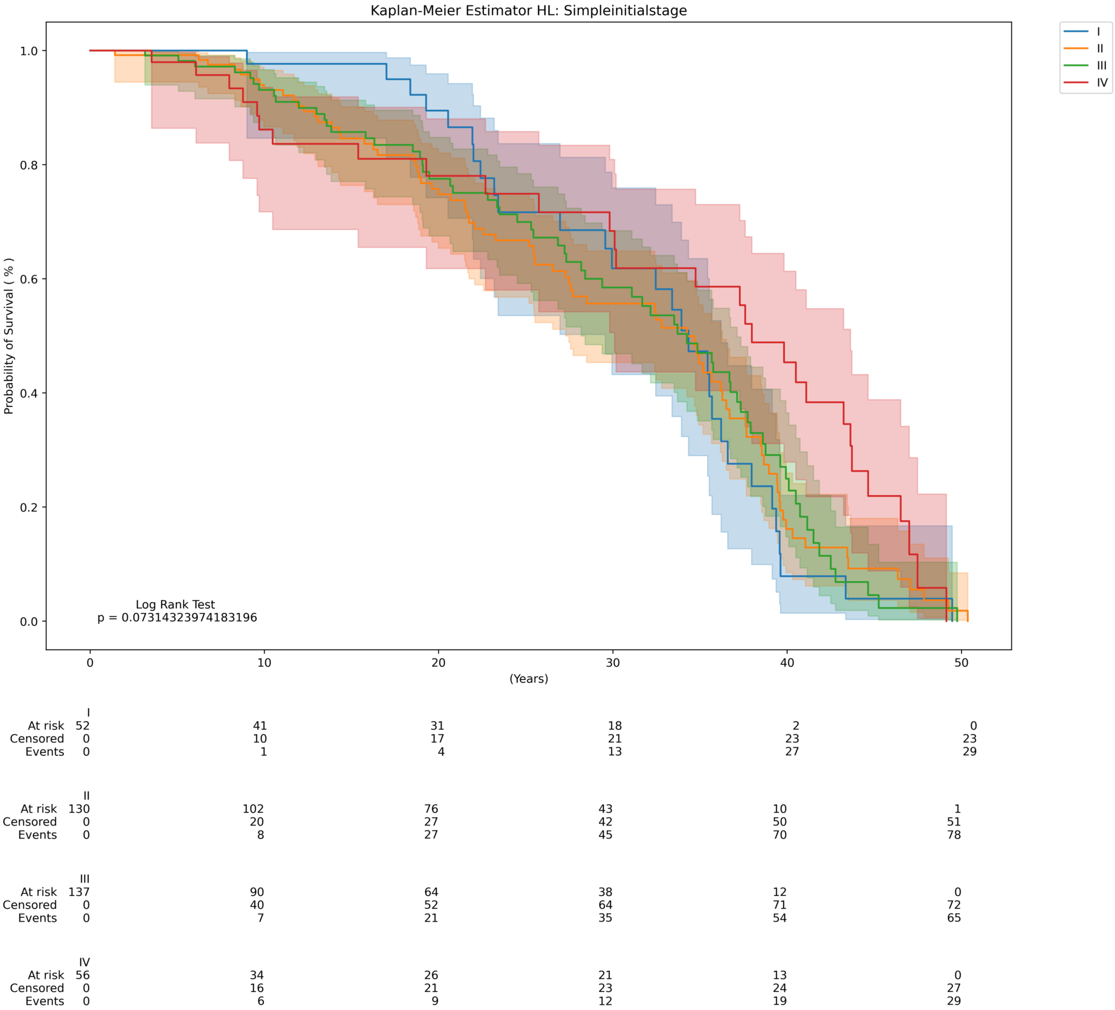


Figure 4.5: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable smoking habits

On the other hand, figure 4.5 presents the results for the analysis of the variable smoker for non-Hodgkin lymphoma. The p-value for this test is 0.22, which is larger than the limit of 0.05, making the curves not different enough to be considered statistically important. The number of censored patients furthers the inapplicability of this variable as significant.

4.2.3 Initial Stage

This subsection focus on presenting the results for the Kaplan-Meier analysis of the stage at diagnosis, also known as the initial stage.

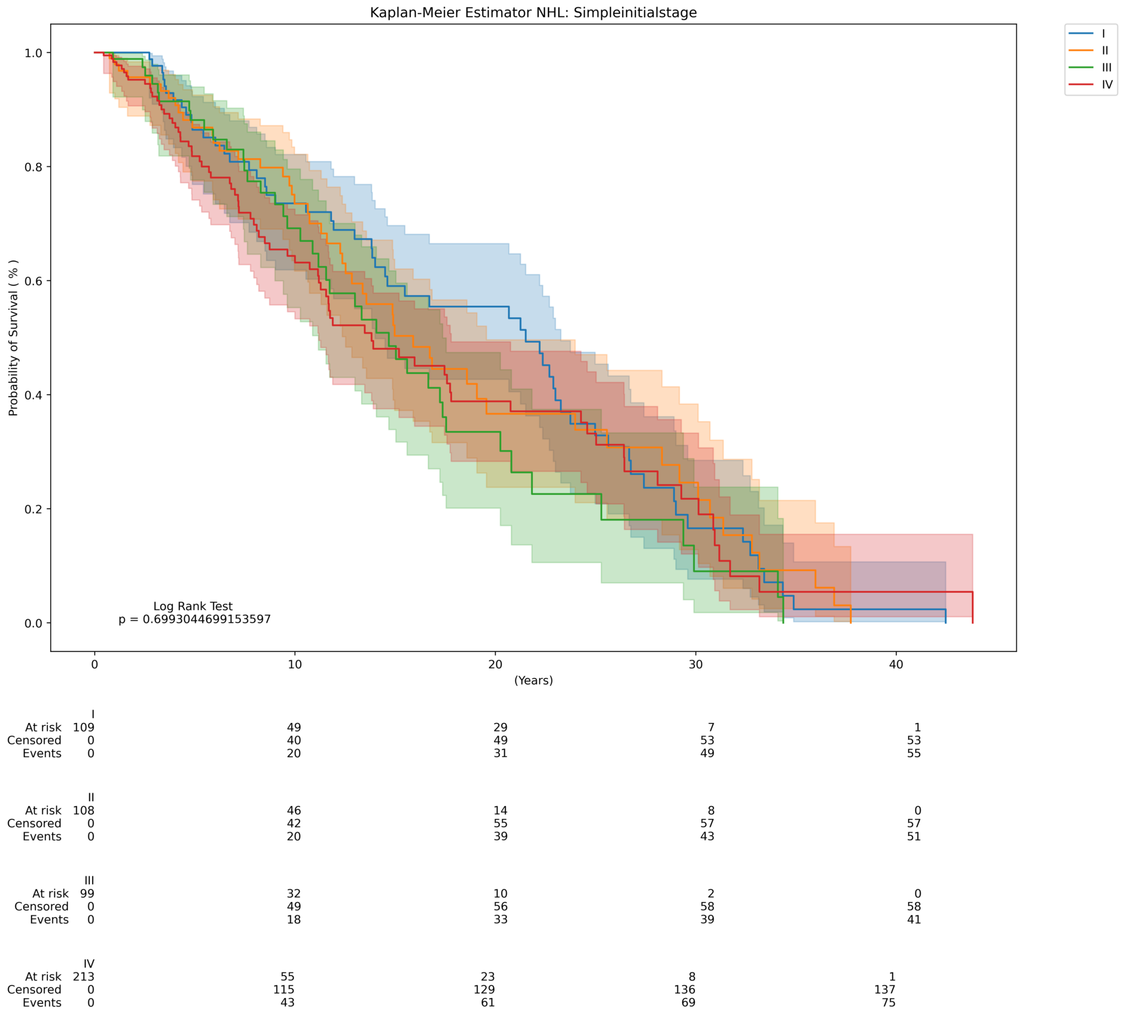


		0	10	20	30	40	50
I	At risk	52	41	31	18	2	0
	Censored	0	10	17	21	23	23
	Events	0	1	4	13	27	29
II	At risk	130	102	76	43	10	1
	Censored	0	20	27	42	50	51
	Events	0	8	27	45	70	78
III	At risk	137	90	64	38	12	0
	Censored	0	40	52	64	71	72
	Events	0	7	21	35	54	65
IV	At risk	56	34	26	21	13	0
	Censored	0	16	21	23	24	27
	Events	0	6	9	12	19	29

Figure 4.6: Kaplan-Meier estimator for Hodgkin lymphoma's variable initial stage

The Hodgkin lymphoma plot in figure 4.6 shows the achieved result for the log-rank test of a p-value equal to 0.073, which is higher than the limit of 0.05. Like the last variable in this type of lymphoma, the test is not the most appropriate since the curves cross each other.

The achieved results, similar to the previous variable, suffer from a lack of data that directly impacts the obtained results.



I	0	10	20	30	40
At risk	109	49	29	7	1
Censored	0	40	49	53	53
Events	0	20	31	49	55

II	0	10	20	30	40
At risk	108	46	14	8	0
Censored	0	42	55	57	57
Events	0	20	39	43	51

III	0	10	20	30	40
At risk	99	32	10	2	0
Censored	0	49	56	58	58
Events	0	18	33	39	41

IV	0	10	20	30	40
At risk	213	55	23	8	1
Censored	0	115	129	136	137
Events	0	43	61	69	75

Figure 4.7: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable initial stage

Figure 4.7 shows the results for non-Hodgkin lymphoma, although the curves so very slightly cross the p-value to discern whether or not they are sufficiently different is 0.69, which is vastly above the bounds of 0.05.

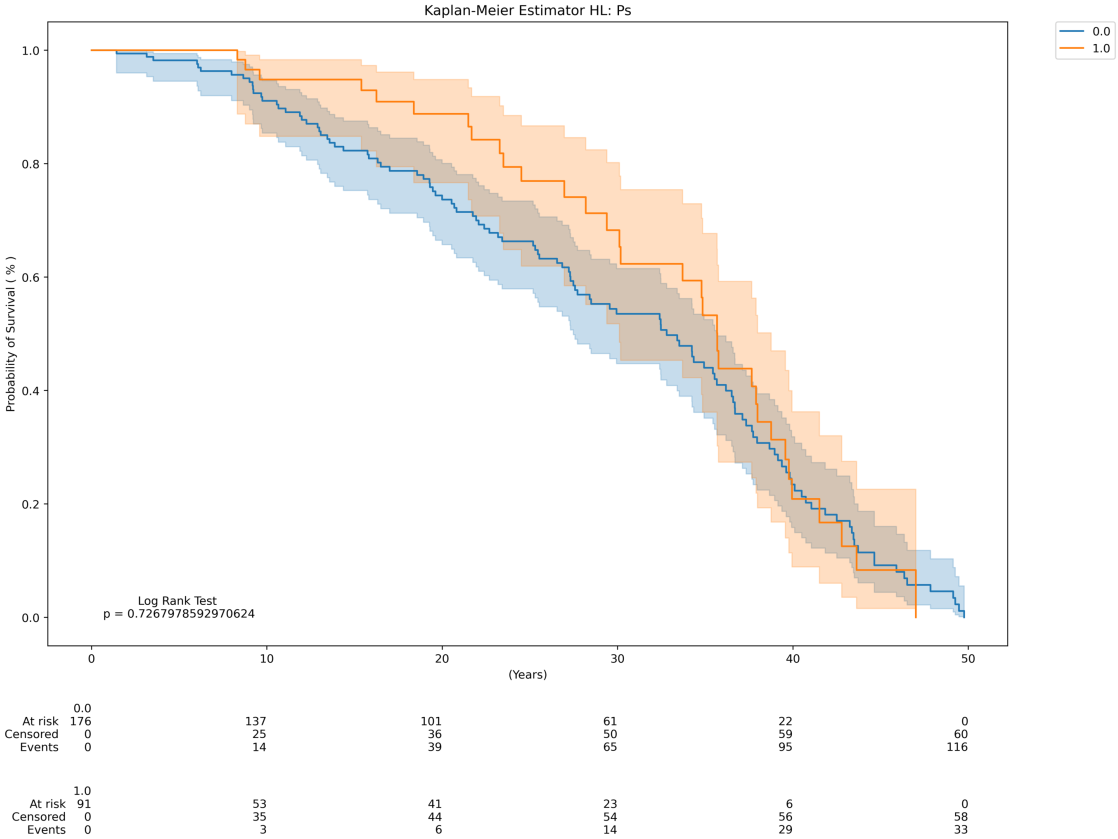
The table below the Kaplan-Meier estimate in figure 4.7 shows that nearly half of the patients have not reached a time after diagnosis high enough not to be censored after the first analysis interval.

4.2.4 ECOG Performance Status

The following plots represent the results of the Kaplan-Meier analysis of the ECOG performance status. The curves in both lymphomas that have considerable confidence intervals are the ones that contain the least number of patients, which is correlated to the uncertainty expressed by the intervals.

For the achieved estimate for HL in figure 4.8, the value "2", which is solely constituted by eight patients, was removed since it does not contribute in any aspect to a better understanding of the achieved estimate consequence of the lack of patients.

Similarly, the NHL estimate represented in figure 4.9 had both the value "3", which is composed of fourteen patients and the value "4", which is only constituted by four patients removed from the final estimate for the same reason as the HL estimate.



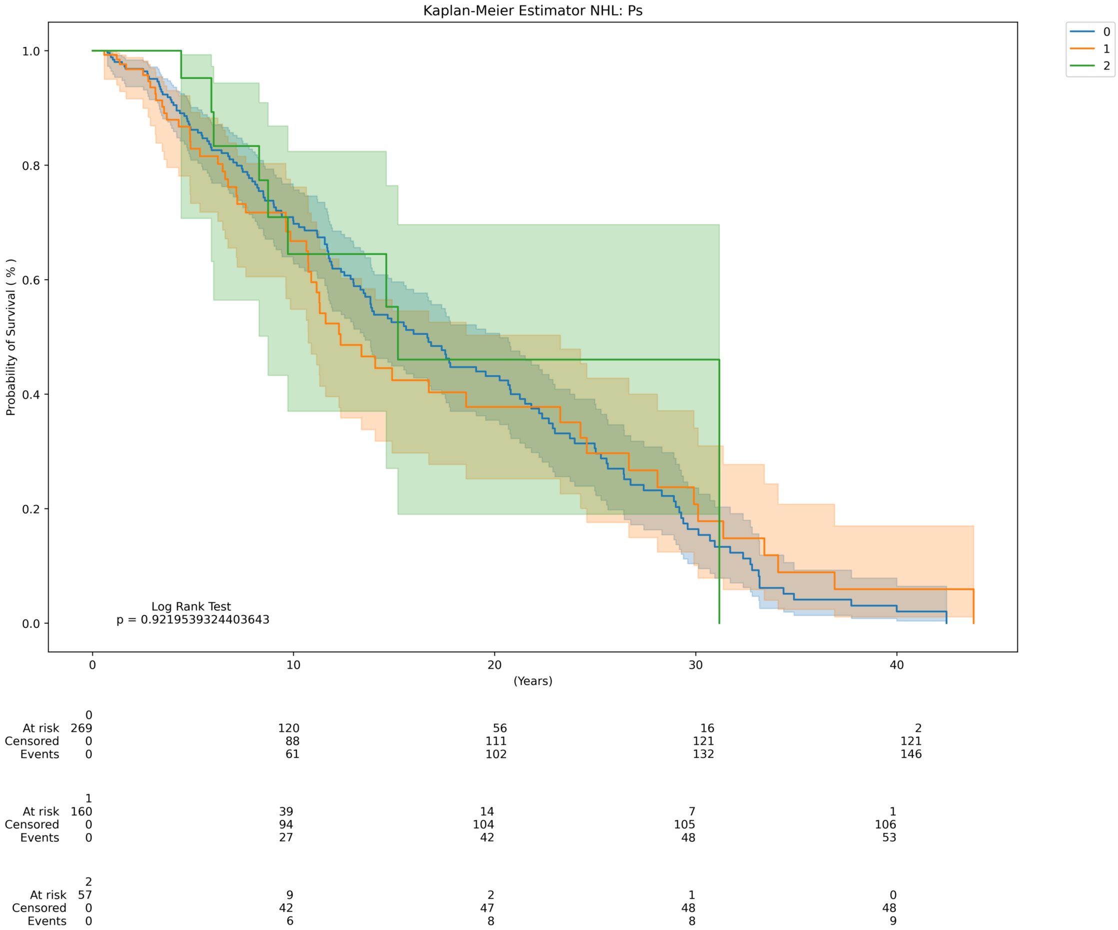
0.0						
At risk	176	137	101	61	22	0
Censored	0	25	36	50	59	60
Events	0	14	39	65	95	116

1.0						
At risk	91	53	41	23	6	0
Censored	0	35	44	54	56	58
Events	0	3	6	14	29	33

Figure 4.8: Kaplan-Meier estimator for Hodgkin lymphoma's variable performance status

The results for HL presented in figure 4.8 show a p-value of 0.72, which is greater than the threshold to be considered a valuable variable.

Again, the lack of data induces a "rough" estimation, which does not allow great confidence in the obtained results.



	0				
At risk	269	120	56	16	2
Censored	0	88	111	121	121
Events	0	61	102	132	146
	1				
At risk	160	39	14	7	1
Censored	0	94	104	105	106
Events	0	27	42	48	53
	2				
At risk	57	9	2	1	0
Censored	0	42	47	48	48
Events	0	6	8	8	9

Figure 4.9: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable performance status

Identically, figure 4.9 which represents the results of NHL, has a p-value of 0.92, confirming that the initial stage variable does not have significance for either lymphoma with the collected data.

4.2.5 Histology

The following plots showcase the results for the Kaplan-Meier estimator. Unlike other variables analysed, this variable has distinct values between both lymphomas and can not be directly compared.

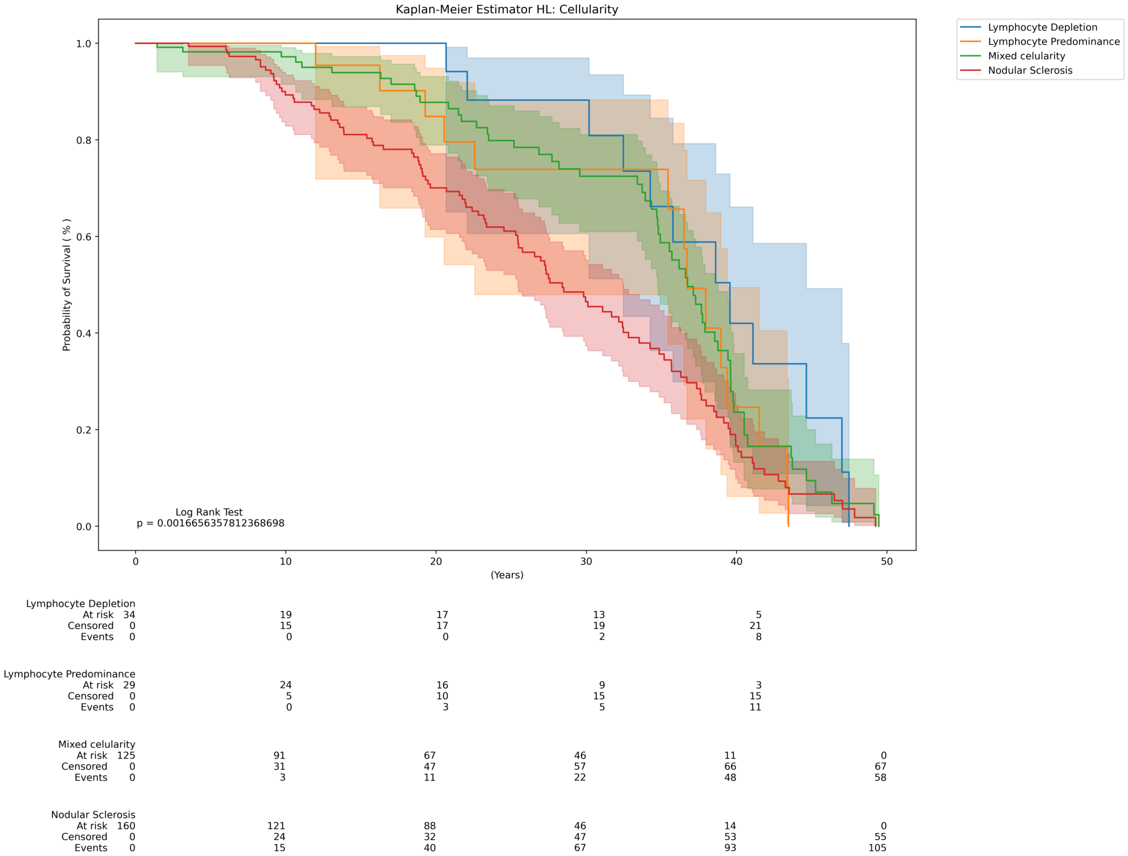


Figure 4.10: Kaplan-Meier estimator for Hodgkin lymphoma's variable histology

The above-presented figure 4.10 expresses the histology of Hodgkin lymphoma, cellularity. The obtained p-value is 0.0016, which is well within the margin, although the curves overlap. It is also required to consider that the value "Rich in lymphocytes" only has one patient and is henceforth not considered.

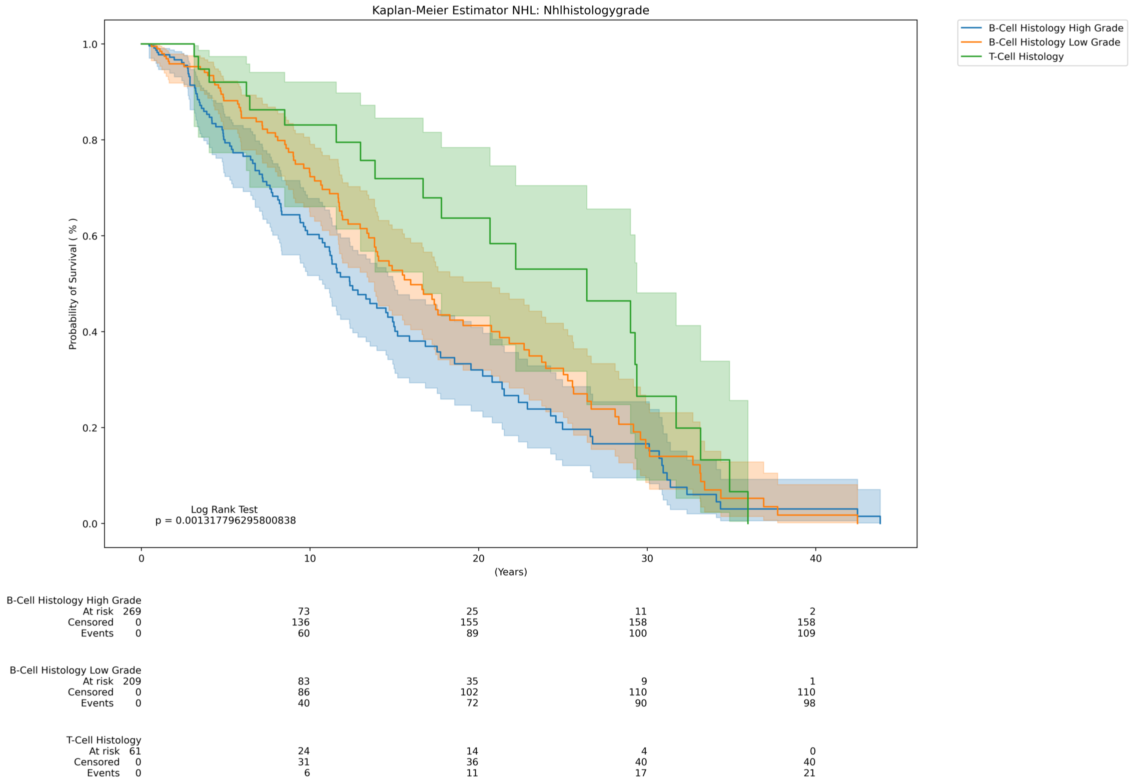


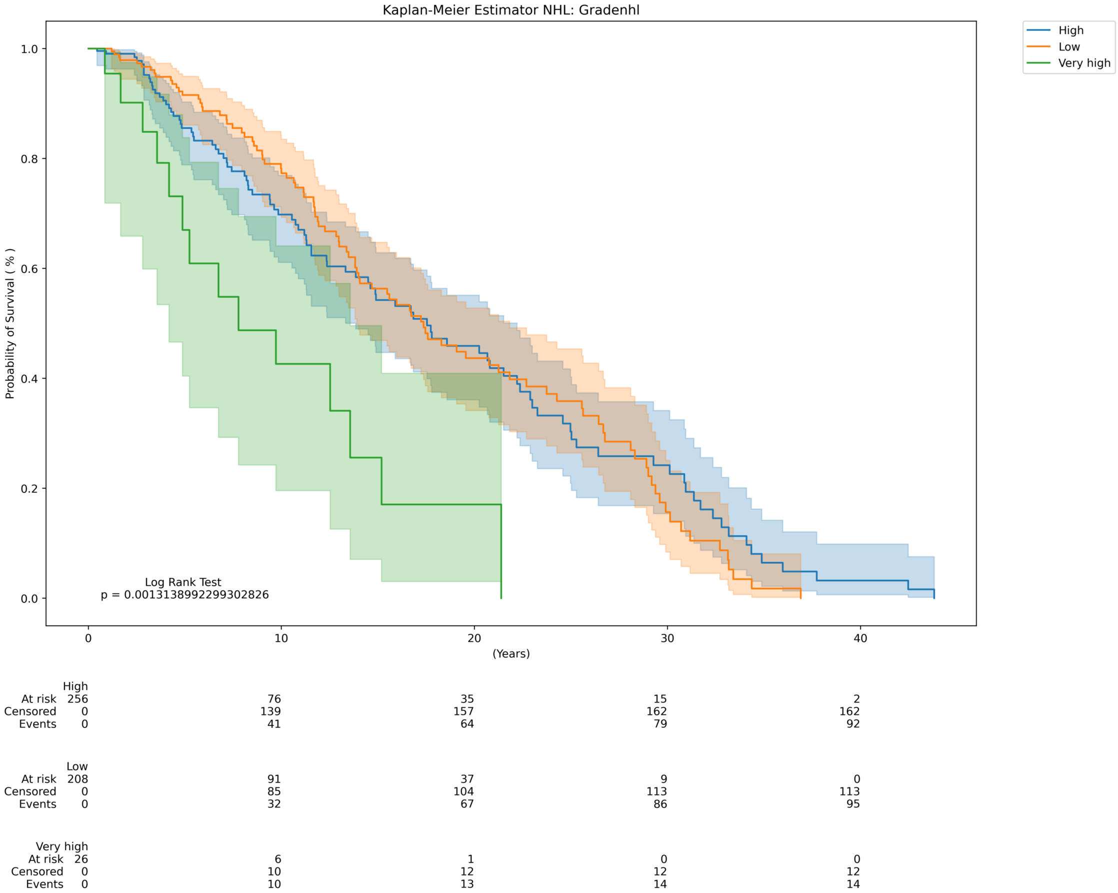
Figure 4.11: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable histology

Conversely, the non-Hodgkin results in figure 4.11 are only constituted by three different values and have a p-value of 0.0013, which is also under the margin to consider sufficiently different survival curves.

The NHL curves, not unlike the histology for HL, overlap, but this happens later when the data is scarce, so for both lymphomas, the variable can be considered a value to determine the survival of the patient.

4.2.6 Non-Hodgkin Lymphoma Grade

As the grade of lymphoma is a variable exclusive to NHL, the figure underneath displays the results for the estimator when analysing the grade of non-Hodgkin lymphoma.



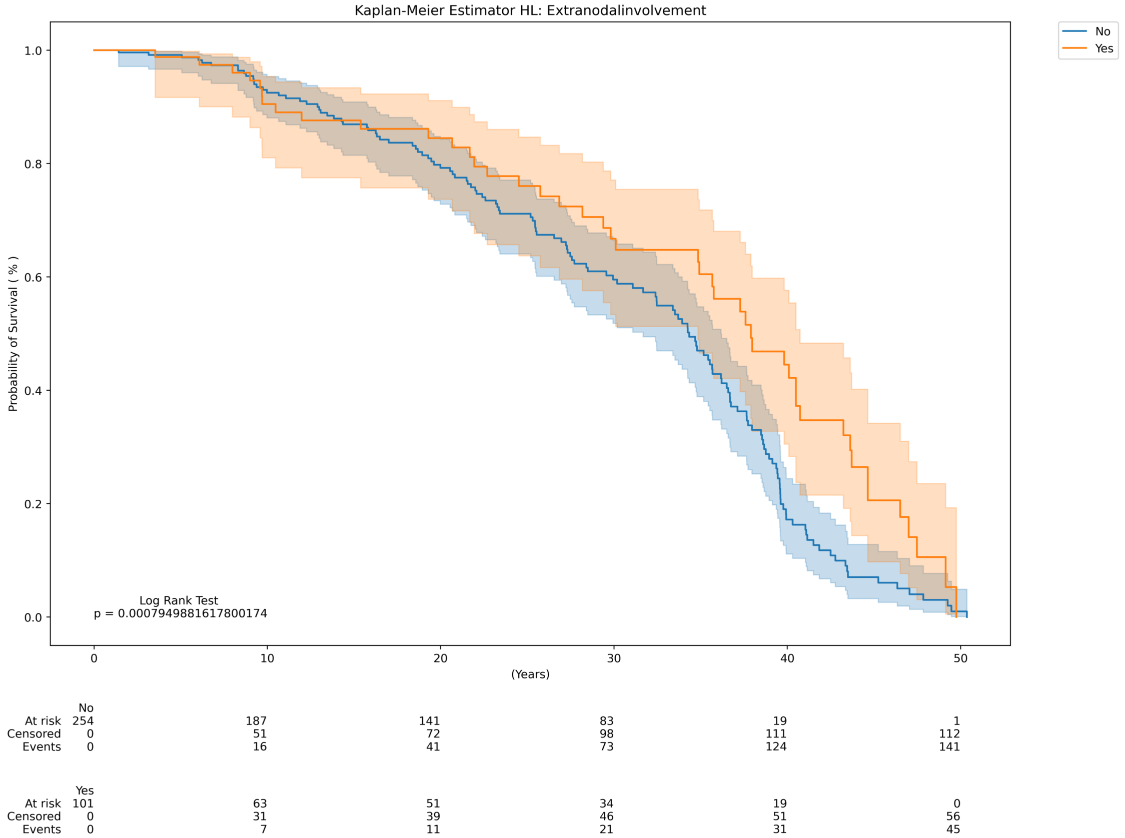
	0	10	20	30	40
High					
At risk	256	76	35	15	2
Censored	0	139	157	162	162
Events	0	41	64	79	92
Low					
At risk	208	91	37	9	0
Censored	0	85	104	113	113
Events	0	32	67	86	95
Very high					
At risk	26	6	1	0	0
Censored	0	10	12	12	12
Events	0	10	13	14	14

Figure 4.12: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable lymphoma grade

Whilst the p-value of figure 4.12 is equal to 0.0013, which is lower than the threshold for the Kaplan-Meier estimator curves to be sufficiently different, the curves cross each other, making the result of the test not very dependable. Despite that, it is clear that despite the curves corresponding to "High" and "Low" crossing each other, there is a significant difference in the curve "Very High".

4.2.7 Extranodal Involvement

This subsection exhibits the estimator's results on the following plots for the variable that determines whether or not there was extranodal involvement.



No	0	10	20	30	40	50
At risk	254	187	141	83	19	1
Censored	0	51	72	98	111	112
Events	0	16	41	73	124	141

Yes	0	10	20	30	40	50
At risk	101	63	51	34	19	0
Censored	0	31	39	46	51	56
Events	0	7	11	21	31	45

Figure 4.13: Kaplan-Meier estimator for Hodgkin lymphoma's variable extranodal involvement

The results for the HL in figure 4.13 show a large discrepancy within the number of patients with each value. Since 313 have "Normal" values, this curve is the most comprehensive, while the values "High" and "Very High" are only constituted by 21 and 13 patients, respectively, making both of these curves probably not indicative of the phenomenon they are representing.

The p-value shows no distinct curves due to its value of 0.69.This p-value can be misleading by crossing the survival lines at the beginning of the analysed timeline. As observed in the same figure 4.13, it is clear that there is a visual difference between both survival curves for patients that have survived more than 18 years.

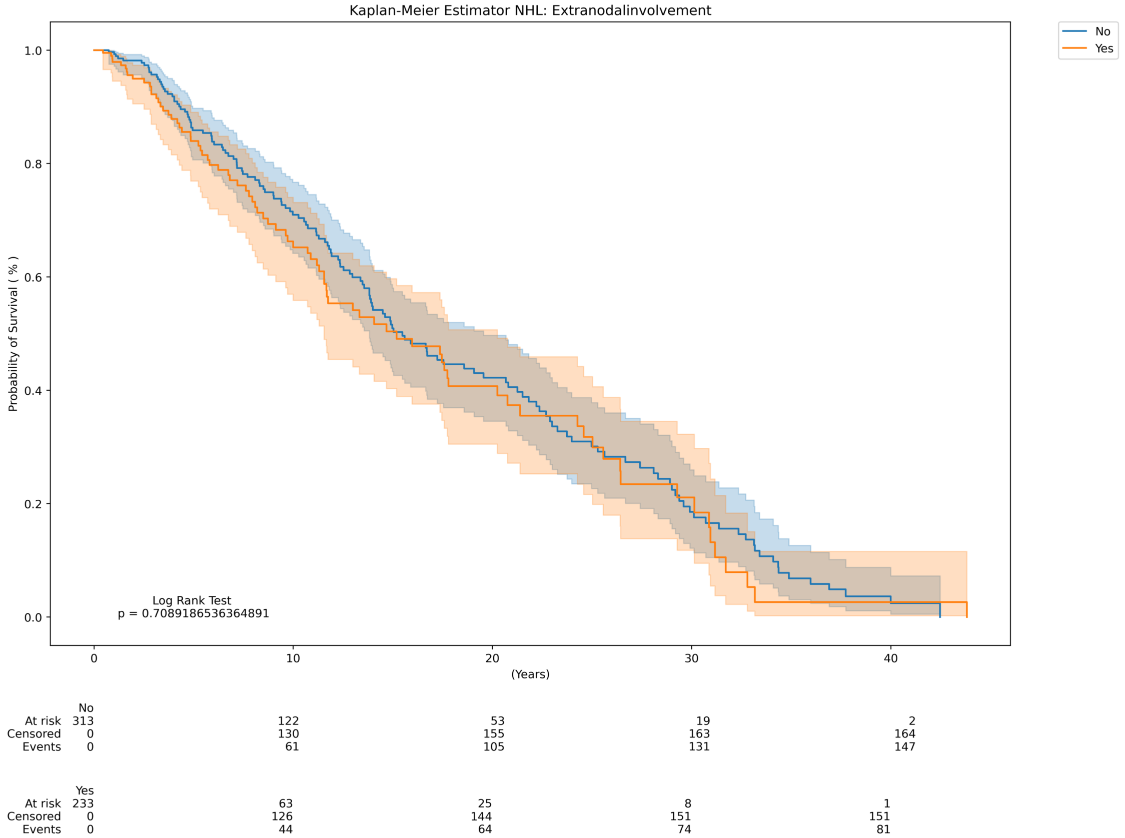
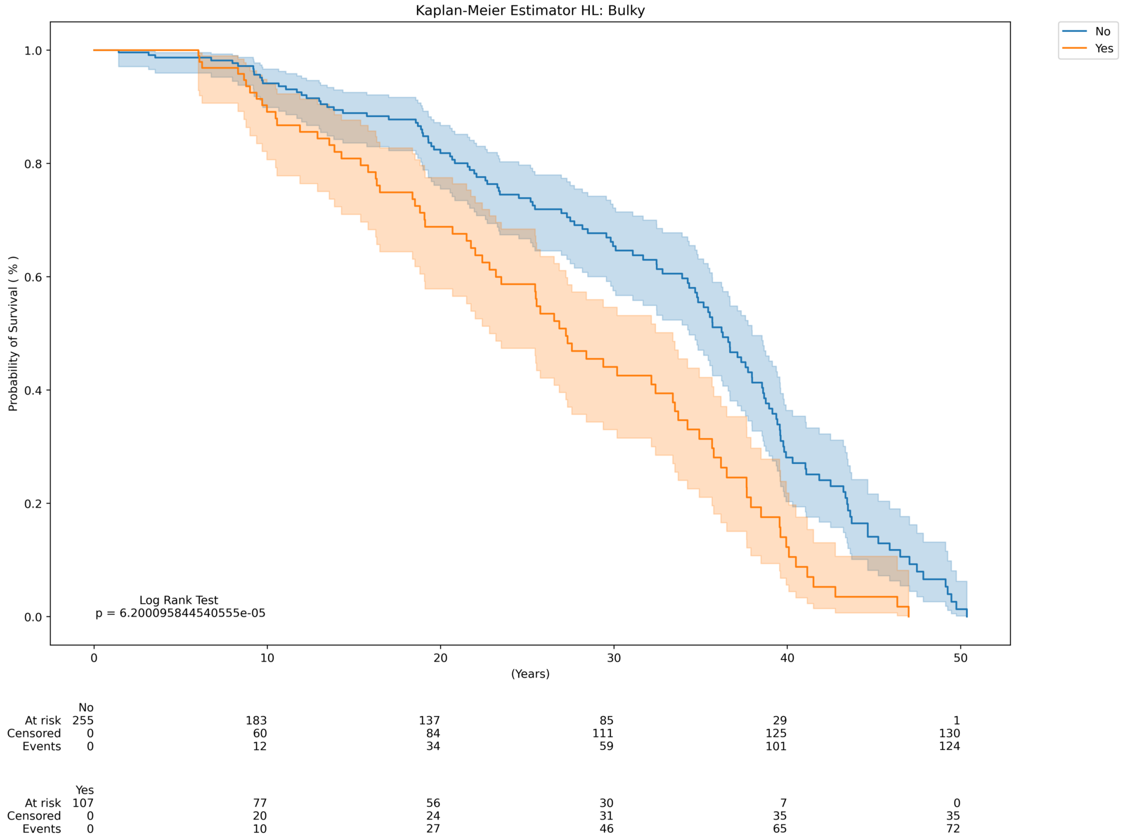


Figure 4.14: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable extranodal involvement

Despite the similar discrepancy of numbers between the variable's values, the NHL results in figure 4.14 contain a more considerable number of patients in each value, making the curves more representative of that value. The p-value that gives the relation between curves is also higher with a value of 0.81, making the extranodal involvement variable not relevant to the survival study in both lymphomas.

4.2.8 Bulky Mass

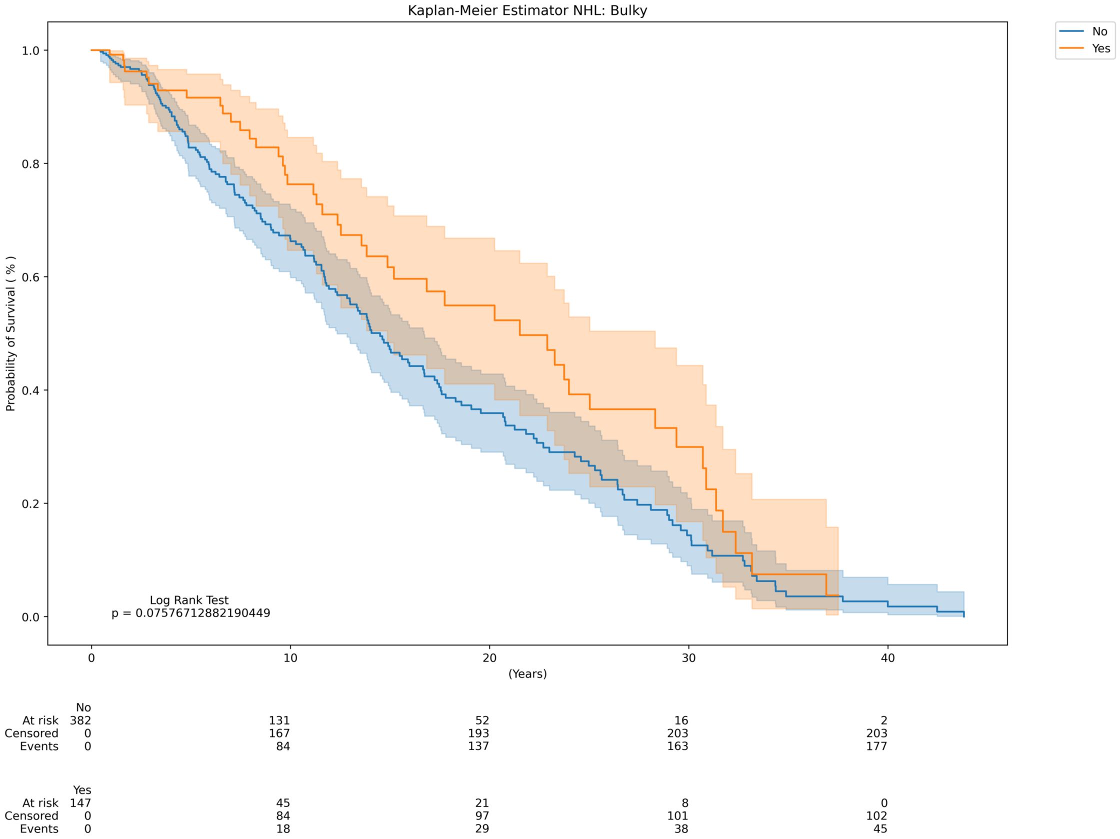
The following plots exhibit the results for the variable that describes the presence of a bulky mass in the patients.



	0	10	20	30	40	50
No						
At risk	255	183	137	85	29	1
Censored	0	60	84	111	125	130
Events	0	12	34	59	101	124
Yes						
At risk	107	77	56	30	7	0
Censored	0	20	24	31	35	35
Events	0	10	27	46	65	72

Figure 4.15: Kaplan-Meier estimator for Hodgkin lymphoma's variable bulky mass

Figure 4.15 provides the results for Hodgkin lymphoma, in which there is a significant difference between the survival of both groups, furthered by the obtained p-value of $6.2 \cdot 10^{-5}$.



No					
At risk	382	131	52	16	2
Censored	0	167	193	203	203
Events	0	84	137	163	177

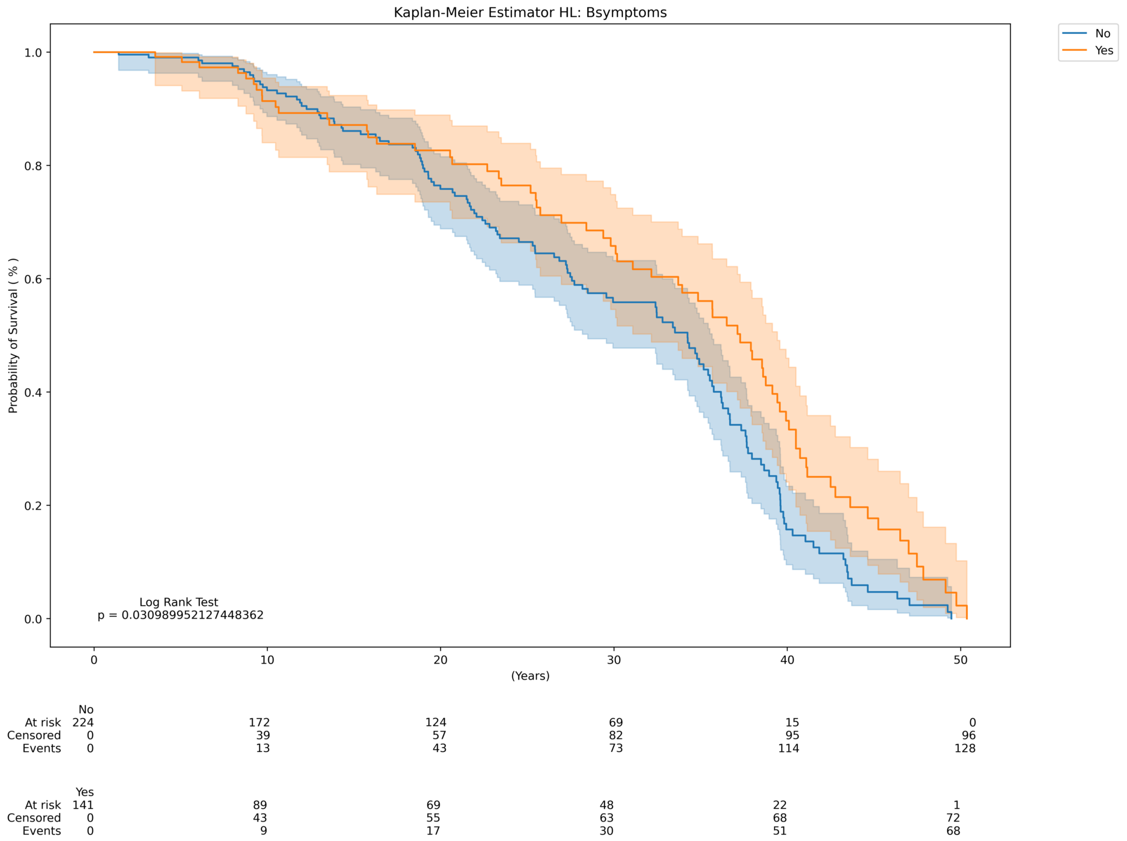
Yes					
At risk	147	45	21	8	0
Censored	0	84	97	101	102
Events	0	18	29	38	45

Figure 4.16: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable bulky mass

In contrast, plot 4.16 which presents the results for NHL shows a different conclusion regarding the significance of the studied variable since the p-value is 0.075. The proximity to the threshold limit, as well as the two times the survival curves cross (in the beginning and at the end when there is fewer patient data), which can make the p-value not reliable, make this variable a possibility worth studying in case of a lack of other significant variables that describe the survivability in NHL.

4.2.9 B-symptoms

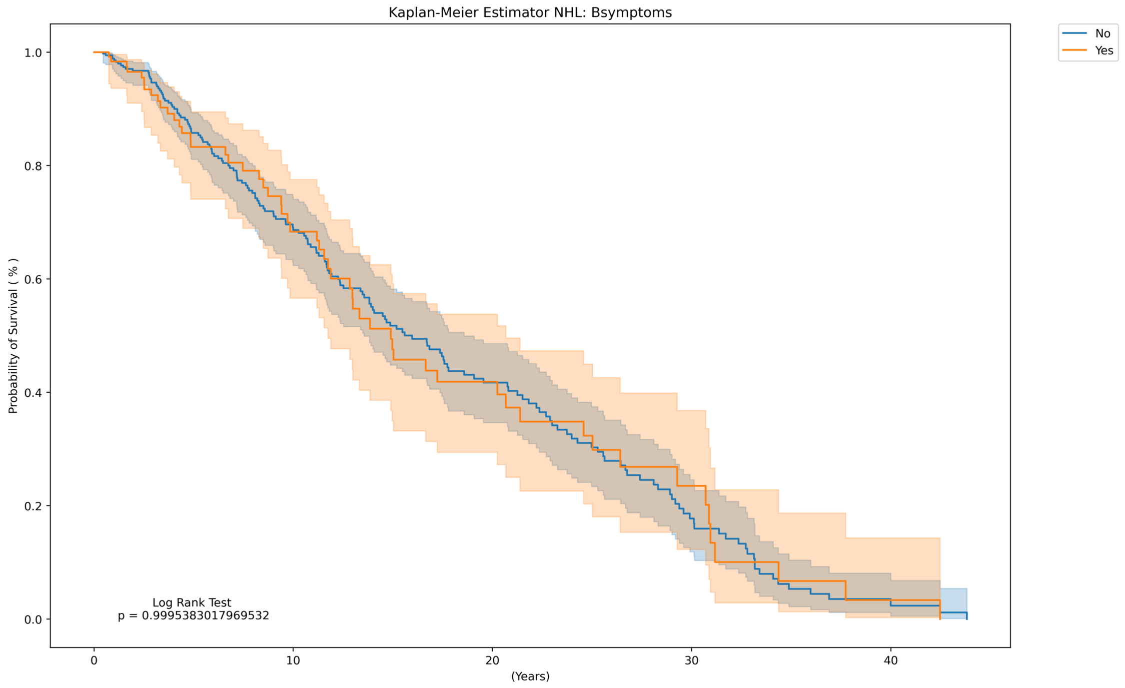
This subsection bares the results for the Kaplan-Meier analysis of the variable that depicts the B-symptoms of the patients for both Hodgkin and non-Hodgkin lymphomas.



	No					
At risk	224	172	124	69	15	0
Censored	0	39	57	82	95	96
Events	0	13	43	73	114	128
	Yes					
At risk	141	89	69	48	22	1
Censored	0	43	55	63	68	72
Events	0	9	17	30	51	68

Figure 4.17: Kaplan-Meier estimator for Hodgkin lymphoma's variable B-symptoms

Despite an initial crossing of the curves, the results in figure 4.17 for Hodgkin lymphoma show a significant difference between both curves for values above 20 years, which makes the curves seemingly sufficiently different from one another. The achieved p-value is 0.03, but the crossing of the curves heavily influences this value and, therefore, can not be used for any further conclusions regarding the statistical relevance of the variable at hand.



No					
At risk	394	141	60	20	2
Censored	0	173	206	216	217
Events	0	80	128	158	175

Yes					
At risk	155	43	19	7	1
Censored	0	86	95	100	100
Events	0	26	41	48	54

Figure 4.18: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable B-symptoms

Diversely the results for non-Hodgkin lymphoma in figure 4.18 show both curves almost overlapping each other, which is confirmed by the large p-value of approximately 0.99, confirming the similarity of the curves.

4.2.10 Diagnostic Analytics

Similarly to the previous chapter, the Diagnostic Analytics results in the medical analysis that the patient underwent.

4.2.10.1 White blood cell count

This subsection focuses on the variable that studies the number of white blood cells. The following plots present the results of the Kaplan-Meier analysis.

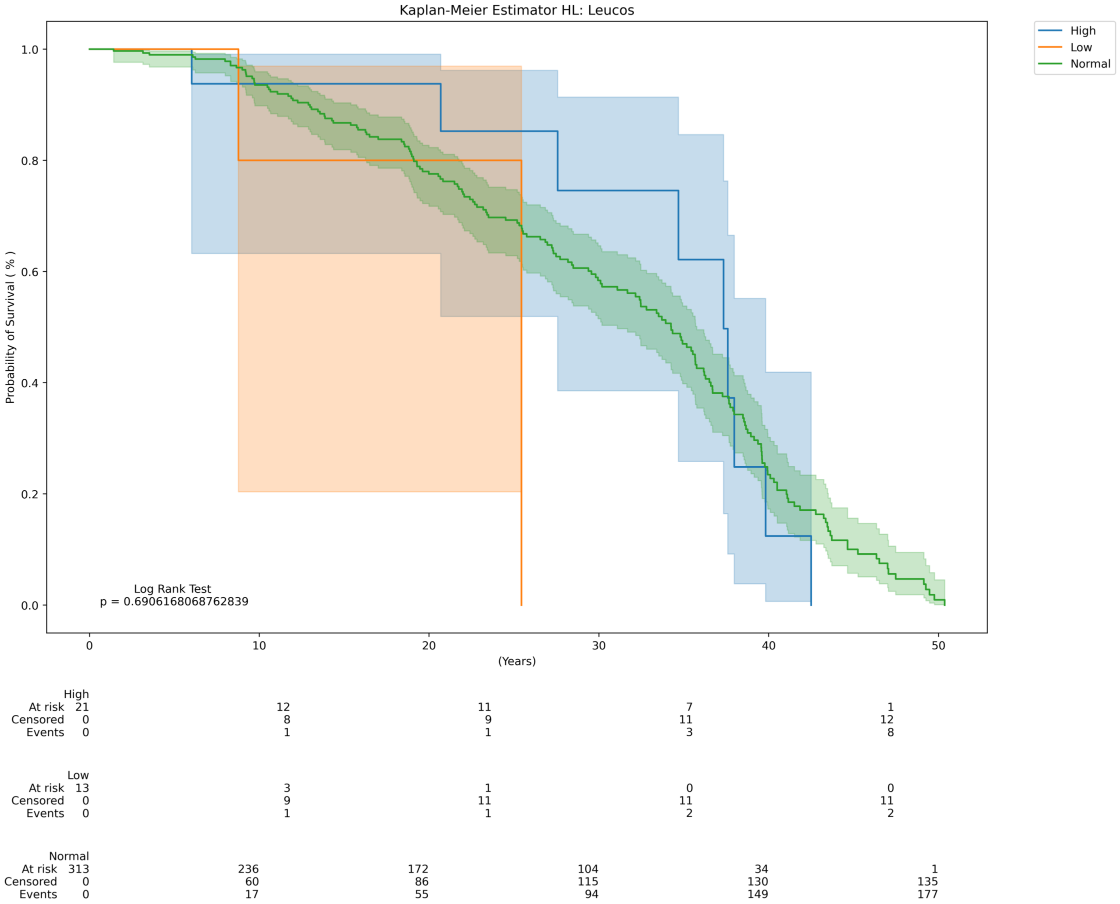


Figure 4.19: Kaplan-Meier estimator for Hodgkin lymphoma's variable that counts white blood cells

Observing plot 4.19, which represents the results for HL, there is a clear dominance for the value "Normal" this makes the representation of the two other values, "Low" and "High", with 21 and 13 patients, respectively, not significant. The calculated p-value is approximately 0.69 making the curves not sufficiently different from one another.

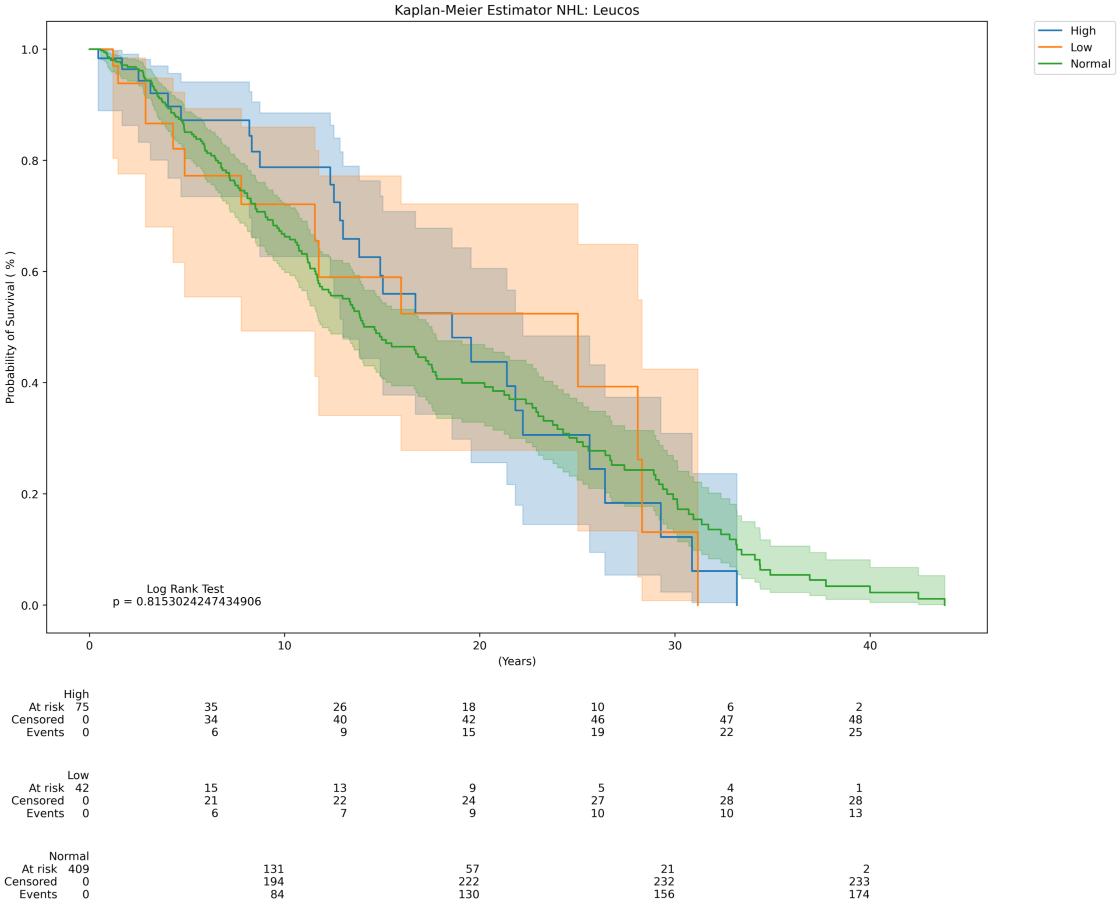
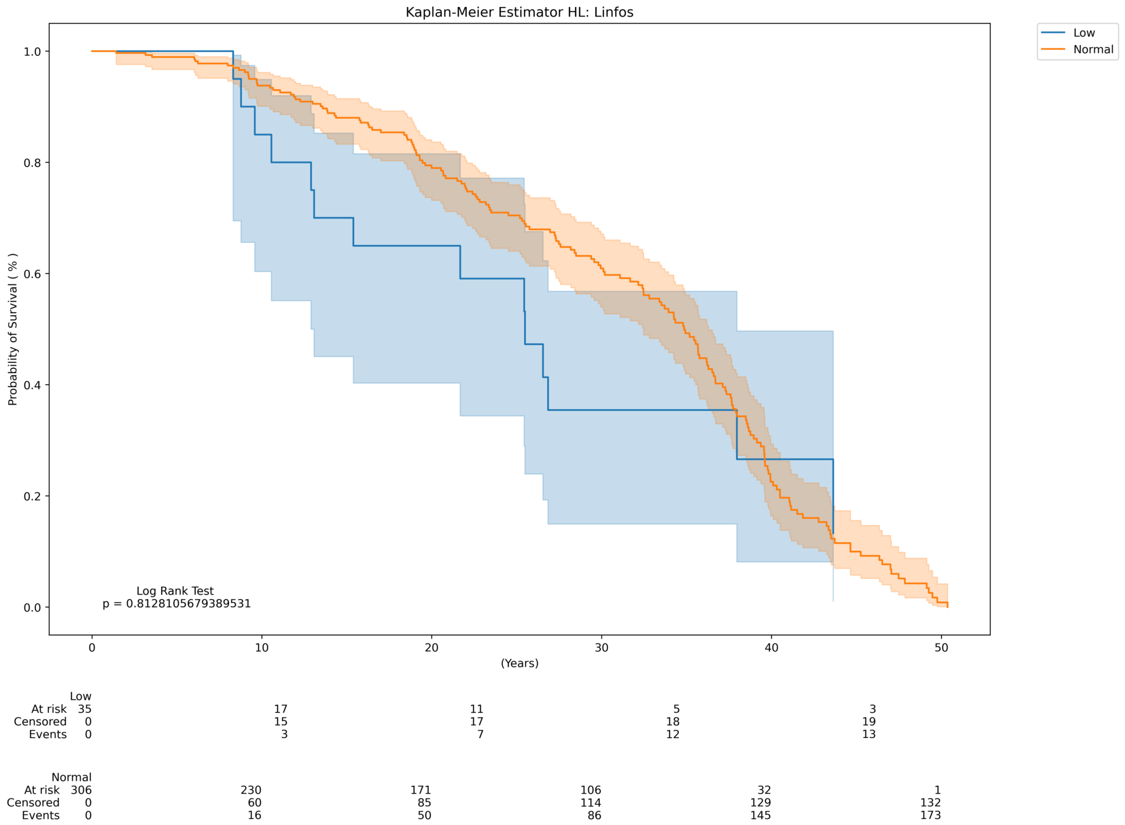


Figure 4.20: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable that counts white blood cells

The results for NHL in figure 4.20 follow the HL's conclusion, and there is a large discrepancy in the distribution. In the case of the NHL results, the less represented variables still constitute a considerable number of patients, 75 patients classified as "High" and 42 patients classified as "Low", despite the obtained p-value of 0.81 being far from the established threshold.

4.2.10.2 Lymphocyte count

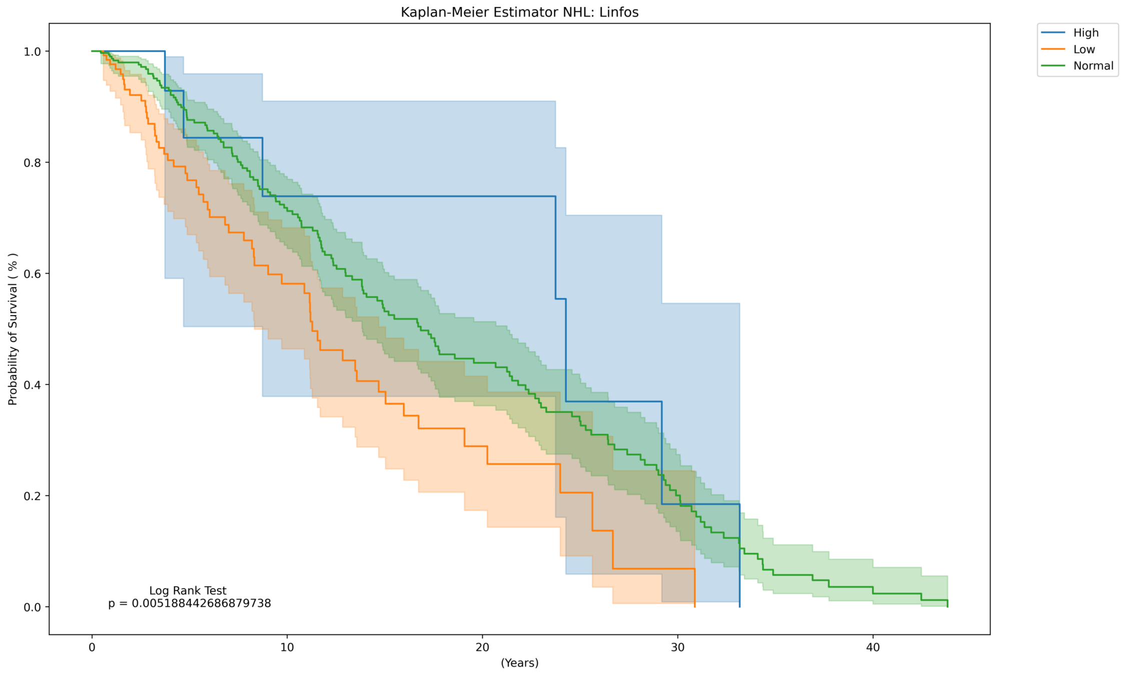
The lymphocyte count results are presented for both lymphomas in the consequent plots.



Low						
At risk	35	17	11	5	3	
Censored	0	15	17	18	19	
Events	0	3	7	12	13	
Normal						
At risk	306	230	171	106	32	1
Censored	0	60	85	114	129	132
Events	0	16	50	86	145	173

Figure 4.21: Kaplan-Meier estimator for Hodgkin lymphoma's variable that counts lymphocytes

Analysing the results obtained in 4.21 for HL lymphoma, the same effect of the previous variable, a large number of patients have the classification of "Normal" whilst a minority are categorised as "Low" with 35 patients. The patients with the value "High" (2 patients total) were removed due to the reduced frequency of this value.



High							
At risk	22	9	6	5	5	2	1
Censored	0	11	13	14	14	15	15
Events	0	2	3	3	3	5	6

Low							
At risk	162	62	35	18	9	4	1
Censored	0	77	91	97	102	105	106
Events	0	23	36	47	51	53	55

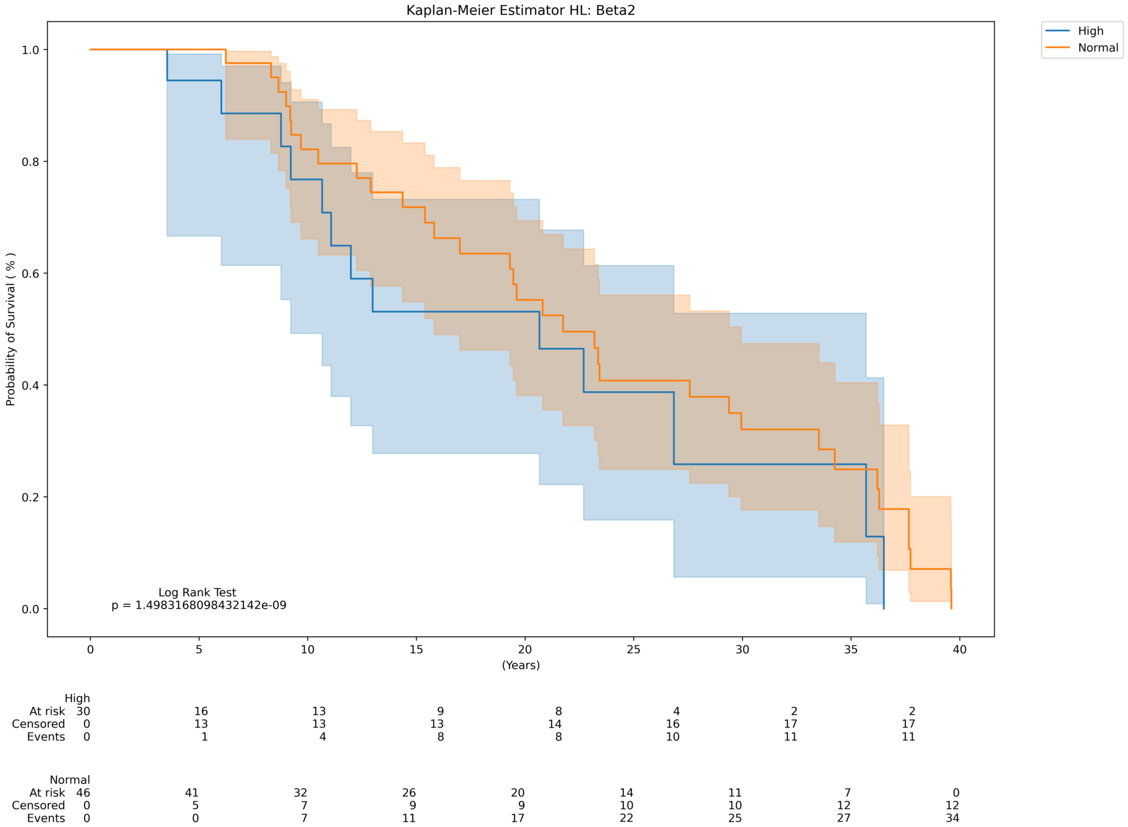
Normal					
At risk	330	124	57	21	2
Censored	0	145	170	178	179
Events	0	61	103	131	149

Figure 4.22: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable that counts lymphocytes

On the other hand, the NHL results present only one of the values, "High", as a low number of patients, with only 22 registered patients. The two other values constitute enough patients to consider the results. The achieved p-value is 0.005, which is well within the threshold and makes the curves represented in figure 4.22 distinct enough to consider in survival analysis.

4.2.10.3 Beta2 microblobuline

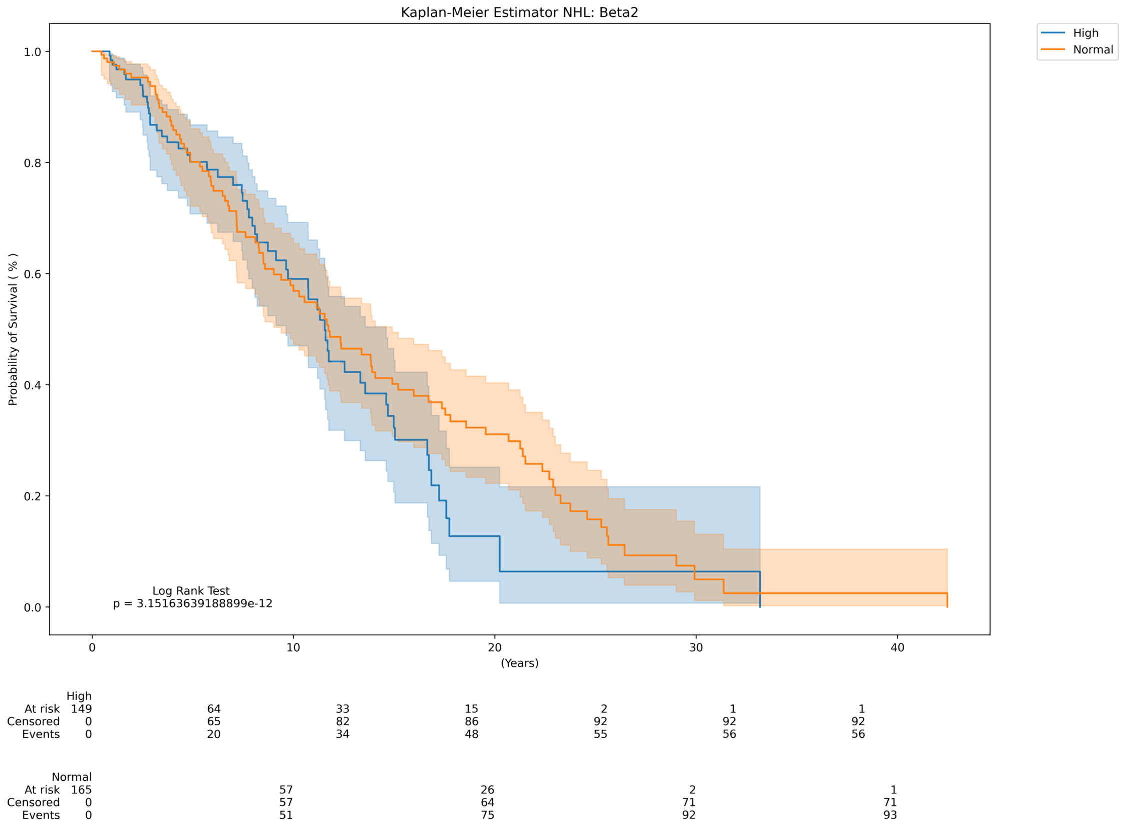
This subsection presents in the plots below the achieved results for the variable that defines the presence of a bulky mass in the patients from either lymphoma.



High									
At risk	30	16	13	9	8	4	2	2	
Censored	0	13	13	13	14	16	17	17	
Events	0	1	4	8	8	10	11	11	
Normal									
At risk	46	41	32	26	20	14	11	7	0
Censored	0	5	7	9	9	10	10	12	12
Events	0	0	7	11	17	22	25	27	34

Figure 4.23: Kaplan-Meier estimator for Hodgkin lymphoma's variable beta2

Figure 4.23 above represents the results obtained for the values "Normal" and "High" of the studied variable. Both curves are sufficiently distinct from one another as comprised by the p-value of $1.49 \cdot 10^{-9}$.



High							
At risk	149	64	33	15	2	1	1
Censored	0	65	82	86	92	92	92
Events	0	20	34	48	55	56	56

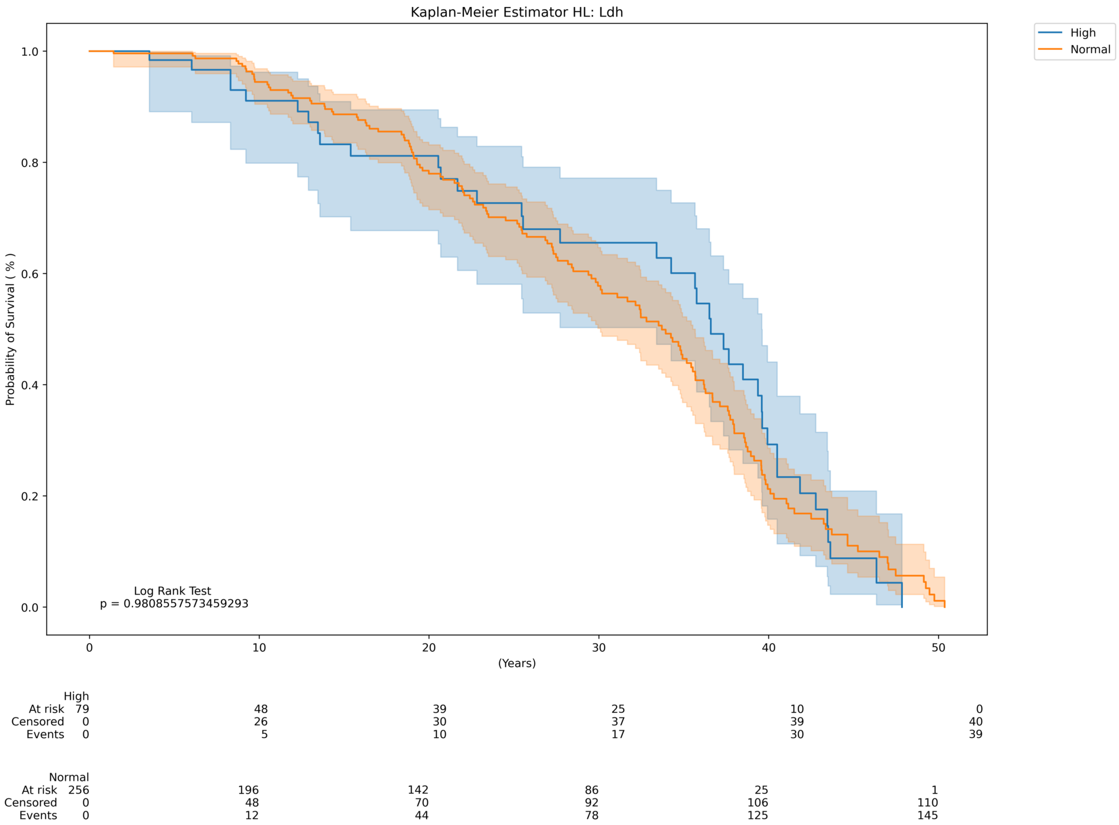
Normal					
At risk	165	57	26	2	1
Censored	0	57	64	71	71
Events	0	51	75	92	93

Figure 4.24: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable beta2

Similarly to the HL curves, the results from NHL in figure 4.24 present a very residual p-value. It is important to note that both curves cross each other on several occasions, directly impacting the obtained p-value.

4.2.10.4 Lactate dehydrogenase - LDH

The results for the Non-Hodgkin lymphoma Kaplan-Meier analysis are shown in the following figures.

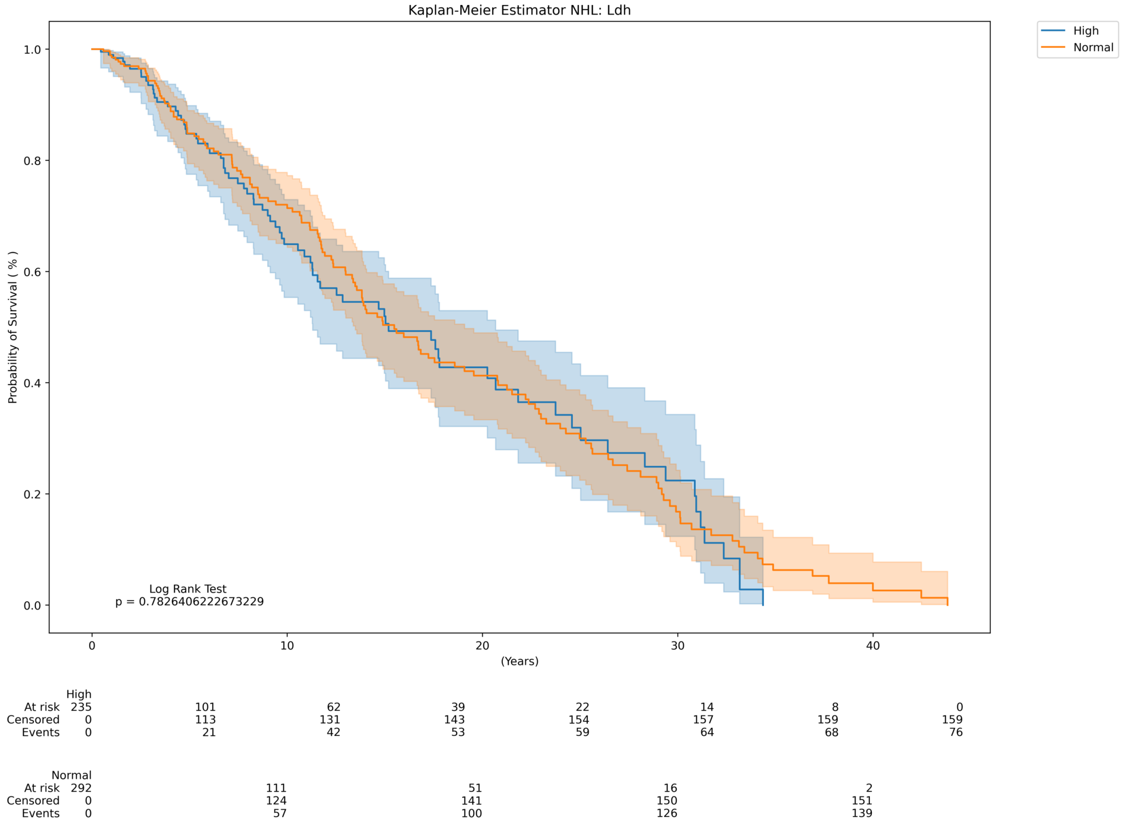


High						
At risk	79	48	39	25	10	0
Censored	0	26	30	37	39	40
Events	0	5	10	17	30	39

Normal						
At risk	256	196	142	86	25	1
Censored	0	48	70	92	106	110
Events	0	12	44	78	125	145

Figure 4.25: Kaplan-Meier estimator for Hodgkin lymphoma's variable LDH

As shown in figure 4.25, the HL results have two very similar curves. Using the obtained p-value, it can be confirmed that with a p-value of 0.98, both curves are not sufficiently different from being considered further.



High								
At risk	235	101	62	39	22	14	8	0
Censored	0	113	131	143	154	157	159	159
Events	0	21	42	53	59	64	68	76

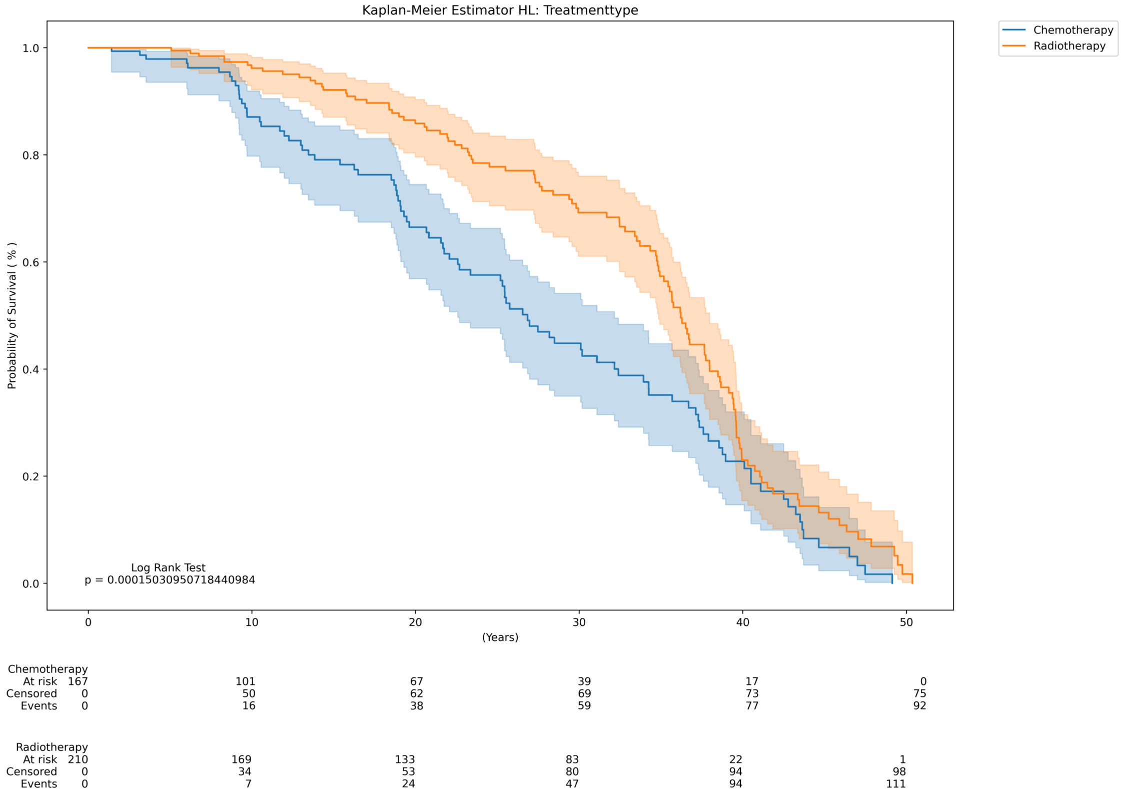
Normal					
At risk	292	111	51	16	2
Censored	0	124	141	150	151
Events	0	57	100	126	139

Figure 4.26: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable LDH

The same phenomenon occurs in the results of the NHL in figure 4.26. The curves achieved are very similar, with a p-value of 0.78, making the result curves not sufficiently different.

4.2.11 Treatment Type

The last analysed variable is the variable that describes the type of treatment each patient received.



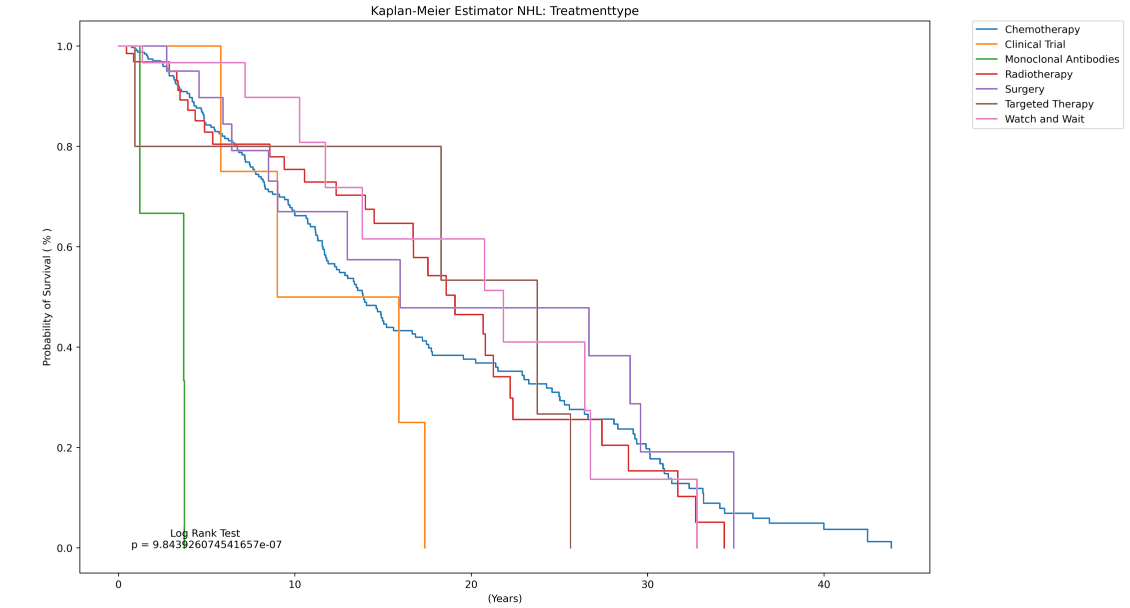
Chemotherapy	0	10	20	30	40	50
At risk	167	101	67	39	17	0
Censored	0	50	62	69	73	75
Events	0	16	38	59	77	92

Radiotherapy	0	10	20	30	40	50
At risk	210	169	133	83	22	1
Censored	0	34	53	80	94	98
Events	0	7	24	47	94	111

Figure 4.27: Kaplan-Meier estimator for Hodgkin lymphoma's variable treatment type

Figure 4.27 presents the HL Kaplan-Meier analysis results, showing that two of the curves are visually different. The p-value that testifies to this conclusion has a value of 0.0001

"Watch and Wait" was removed from the analysis because it was only composed of one patient and would not bring any additional information to the resulting estimate



Chemotherapy					
At risk	386	124	49	20	3
Censored	0	184	212	221	222
Events	0	78	125	145	161
Clinical Trial					
At risk	5	2	0	0	0
Censored	0	1	1	1	1
Events	0	2	4	4	4
Monoclonal Antibodies					
At risk	3	0	0	0	0
Censored	0	0	0	0	0
Events	0	3	3	3	3
Radiotherapy					
At risk	69	30	12	3	0
Censored	0	27	36	38	38
Events	0	12	21	28	31
Surgery					
At risk	21	8	5	1	0
Censored	0	7	8	9	9
Events	0	6	8	11	12
Targeted Therapy					
At risk	5	4	2	0	0
Censored	0	0	1	1	1
Events	0	1	2	4	4
Watch and Wait					
At risk	34	10	6	1	0
Censored	0	22	23	24	24
Events	0	2	5	9	10

Figure 4.28: Kaplan-Meier estimator for non-Hodgkin lymphoma's variable treatment type

Unlike the majority of plots presented, figure 4.28 does not contain the respective confidence intervals to facilitate the survival curves' visualisation.

In contrast to the previous plot, plot 4.28 is composed of seven different curves. Three out of the seven curves have only a few patients and are therefore not helpful to the overall analyses. The curves that have a residual amount of patients are "Clinical Trial", "Monoclonal Antibodies", and "Targeted Therapy".The obtained p-value result is clouded by the number of different curves with hardly any patients.

5

Implementation of Kaplan-Meier Surviving Tool

This chapter focus on the developed Kaplan-Meier survival tool, which entails not only the explanation of the used mechanics and project design but also the achieved results and future expansions.

The tool's objective is to incorporate an easy-to-use, quick-to-calculate Kaplan-Meier estimator so it can be consulted on demand according to the prevailing database. This objective entails the need to create a versatile and effective tool that could be used for any different type of cancer that the CLARIFY platform supports.

As this dissertation's objective is to analyse both types of lymphomas, Hodgkin lymphoma and non-Hodgkin lymphoma and the results will reflect both lymphomas studied.

In addition, and as a proof of concept for both the optimisation of the tool and the range in which it can be used, lung cancer was included since this type of cancer already constituted part of the CLARIFY project [105].

5.1 Tool Implementation

This subchapter presents the development and the decisions made to create the Kaplan-Meier survival tool.

It will present both the backend with the integration of new services within the CLARIFY project using cloud functions and the User Interface programmed using React, along with the connecting layer between the backend and the frontend that includes the User Interface.

5.1.1 Google Cloud Functions

This section aims to explain the services implemented through Google cloud functions using the Google cloud platform function-as-a-service (Faas). As the overall project follows a microservice approach, all the implemented Google cloud functions that inhibit the use of the tool follow the same postulation.

As mentioned in subsection 2.3.8, all communication within cloud functions is based on JSON, and the expected cloud functions will follow that format.

The implemented cloud functions to make the survival tool possible were:

- **kmSearchVariable** – This function is responsible for searching every implemented type of cancer and pre-approved variable so the user can select which one to analyse;

- **kaplan_meier** – This function is responsible for computing the Kaplan-Meier estimator, plot the survival tools and the correspondent 95% confidence intervals.

5.1.1.1 Function “kmSearchVariable”

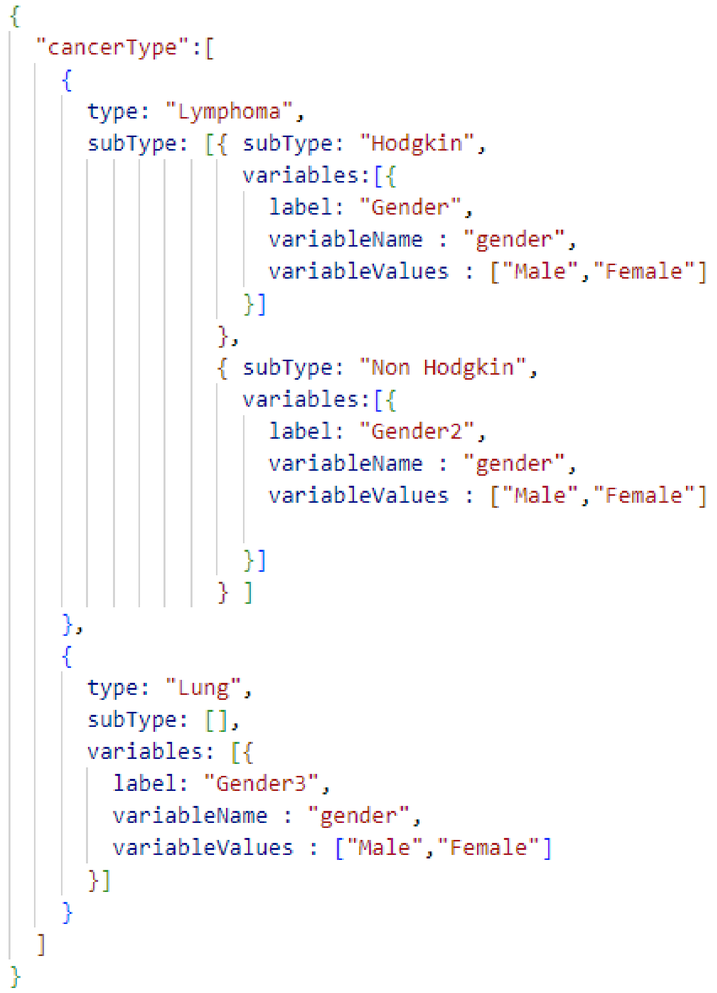
The kmSearchVariable functions’ objective is to retrieve all the different types of cancer in a pre-approved list along with the correspondent explanatory variables and respective possible values. Using a service to complete this task instead of making a static list allows the result to be dynamic and automatic instead of hard coding each type of cancer and the dozens of variables, each with many possible values. This approach makes the developed tool scalable and non-uniform.

This function does not require any parameters since the function always returns the integral search of possible results.

The function uses a medical chosen, pre-approved list of variables instead of the integral list of variables the database has to help filter the quality of information while keeping the tool simple and easy to use instead of making it a counter-intuitive extensive list with all the possibilities. This choice helps the tool to be intuitive and easy to use by any user.

A side effect of restricting the list of variables is that the computing time is primarily reduced by the fact that the Google cloud function does not need to return hundreds to thousands of variables and the respective values, which is an excellent optimisation to speed up the function run-time. This list of variables can be quickly edited without any dependencies making this function even more modular and future-proof.

The results as standard per Google cloud functions are returned in JSON formatted as shown in the example below.



```
{
  "cancerType":[
    {
      type: "Lymphoma",
      subType: [{ subType: "Hodgkin",
                  variables:[{
                    label: "Gender",
                    variableName : "gender",
                    variableValues : ["Male","Female"]
                  }]
                },
                { subType: "Non Hodgkin",
                  variables:[{
                    label: "Gender2",
                    variableName : "gender",
                    variableValues : ["Male","Female"]

                  }]
                } ]
    },
    {
      type: "Lung",
      subType: [],
      variables: [{
        label: "Gender3",
        variableName : "gender",
        variableValues : ["Male","Female"]
      }]
    }
  ]
}
```

Figure 5.1: Example of received dictionary for the kmSearchVariable cloud function

As depicted in figure 5.1 above, the function returns a dictionary, "cancerType", that contains the formatted function response. This dictionary has as its value for the previously stated key a list of dictionaries whose keys are the type of cancer they represent. Each cancer only has one "type", but it can have various "subtypes", as figure 5.1 shows for the lymphoma cancer type.

The cancer subtype is a list of dictionaries containing the specified subtype along with all the variables' names in the database and their corresponding values. If studied type of cancer does not have any subtypes, the variables will be in the original dictionary instead of a list of subtypes dictionaries since this field will be empty, as is the case for lung cancer in figure 5.1.

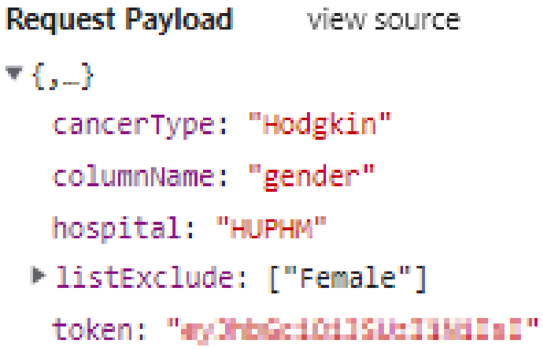
5.1.1.2 Function "kaplan_meier"

The other function is "kaplan_meier", which aims to return any needed plots obtained from the Kaplan-Meier estimator analysis to the User Interface.

Since the cloud function is not involved in the project's front end and can only transmit data through a JSON format, the previously shown plots presented in the previous chapter could not be simply transposed since there was no way to send them to the front end. In addition, these plots were not interactive, and the tool user could not interact with the point that delineates the resultant curve.

To mitigate this problem, the function returns the curves as a list of points within the JSON response that will be used to create a responsive chart in the front end.

Since this cloud function requires arguments, it is essential to specify the function of each one and its significance in making the overall function more reusable and modular as possible.

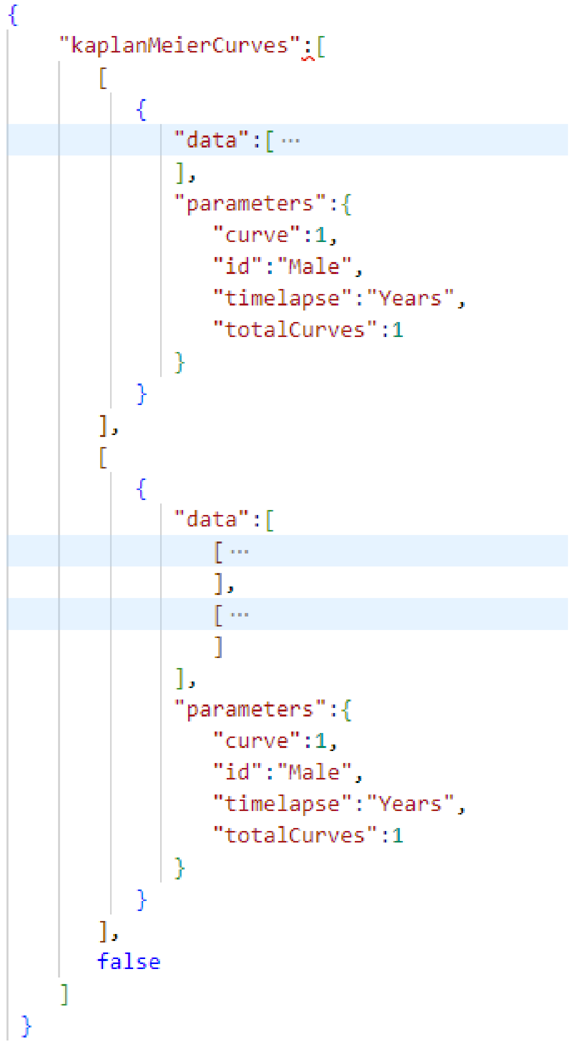


Request Payload view source

```
▾{,…}
   cancerType: "Hodgkin"
   columnName: "gender"
   hospital: "HUPHM"
  ▸listExclude: ["Female"]
   token: "[redacted]"
```

Figure 5.2: Example of sent payload parameters

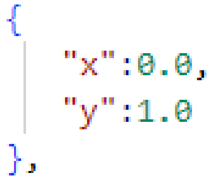
Figure 5.2 above represents the cloud functions payload with all the included arguments used as an example. The presented example to explain the implementation of the overall cloud function is the Kaplan-Meier estimator for Hodgkin lymphoma, where the analysed variable is gender, and the value "Female" is excluded to include every possible argument the function handles. For confidentiality reasons, the token has been omitted. This token is used to authenticate whether or not the cloud function should execute the request to access the database as a security measure. As the example only requires one curve with the respective intervals, the response is easily interpreted and is shown partially in figure 5.3 below.



```
{
    "kaplanMeierCurves":[
       [
          {
             "data":[…
             ],
             "parameters":{
                "curve":1,
                "id":"Male",
                "timelapse":"Years",
                "totalCurves":1
             }
          }
       ],
       [
          {
             "data":[
                […
                ],
                […
                ]
             ],
             "parameters":{
                "curve":1,
                "id":"Male",
                "timelapse":"Years",
                "totalCurves":1
             }
          }
       ],
       false
    ]
 }
```

Figure 5.3: Example of received dictionary for the kaplan_meier cloud function

The response dictionary, "kaplanMeierCurves", is composed of three arguments, as seen in figure 5.3, two lists and one flag. The first list contains the different Kaplan-Meier survival curves since, in the presented example, only one survival curve is returned, and consequently, there is only one dictionary. This sub-dictionary contains two further dictionaries, one containing the data of the curve itself, as demonstrated in figure 5.4 below, and the other containing the parameters that characterise the survival curve.



```
{
   "x":0.0,
   "y":1.0
},
```

Figure 5.4: Example of a list of coordinates present in the data dictionary

The second list is likewise a list of dictionaries. Similar to the first list, these dictionaries contain two sub-dictionaries of their own; the first is the data that contains a list of two lists, the first the upper limit of the confidence interval and the second the lower limit of the same confidence interval. The confidence intervals are static and always correspond to the 95% confidence intervals. The second sub-dictionary contains the different coordinates parameters that characterise the confidence intervals, as seen in figure 5.4.

The third and final argument is a flag that can take the value of true or false depending if the margin within both confidence intervals exceeds a designated threshold. This optimisation is a quality of life upgrade so the user can be warned whether the conclusions retrieved from the graphs are sufficiently good enough to be done.

The threshold is easily changeable at any given time to further maximise the possibilities and uses of the cloud function. This flag will be responsible for the appearance of a warning banner in the User Interface in the following subsection.

5.1.2 User Interface

This section focuses on the implementation of the front-end portion of the developed tool using React. As stated in section 2.3.9, the React platform is a single-page app, and as such, the CLARIFY project needs to be integrated and interconnected with the existing logic and page landing layout. Considering the Kaplan-Meier tool is used to analyse populations, it will be integrated into the section for population analysis as shown in the results section 5.2.

One of the main objectives of the tool is the simplicity of use without a simple and smooth learning curve so the medical staff can easily and readily access and use the tool to its fullest.

The following diagram in figure 5.5 idealises the layout of the tool's UI.

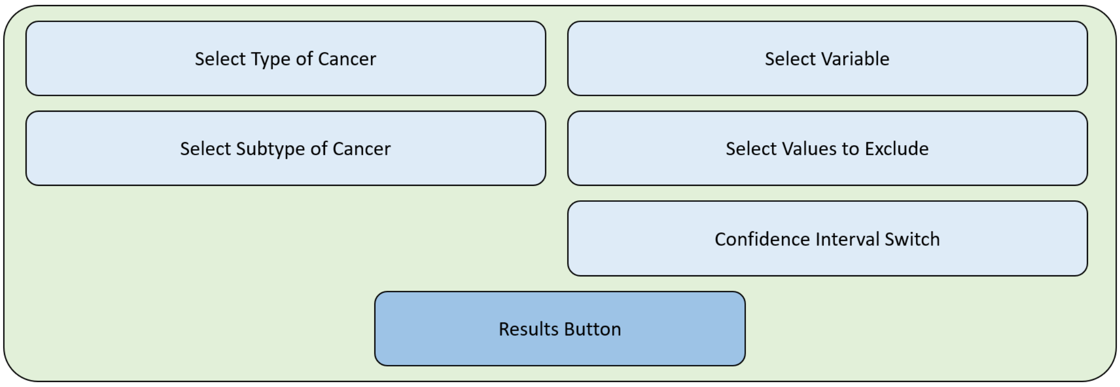


Figure 5.5: Idealisation for the tool's layout

To help facilitate usability, the fields presented in figure 5.5 appear necessary for the user to complete. Consequently, not all fields appear depending on the typology of cancer;

for instance, lung cancer does not have any subtypes and, as such, does not require a field to specify the inexistence of a subtype and therefore is not present.

Four of the six represented elements, "Select Type of Cancer", "Select Subtype of Cancer", "Select Variable", and "Select Values to Exclude", are autocomplete fields, as shown by the example in the following figure 5.6. The autocomplete fields can either be chosen as a selector dropdown menu like figure 5.6 ilustrates, or the user can write and autocomplete the field by hand as the element's name entails. The autocomplete for "Select Values to Exclude" is different since more than one variable's value can be selected, and as such, a multiple autocomplete element is used.

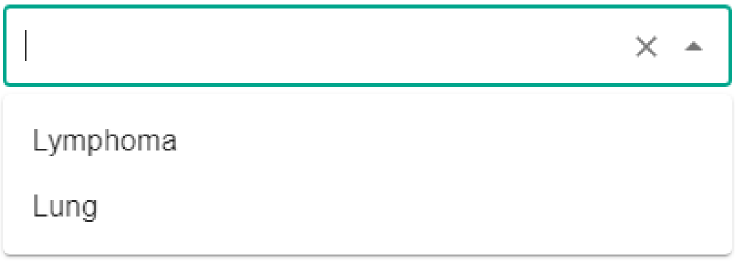


Figure 5.6: Example of a drop down selector menu

The other two elements presented in the schematic for the tool are a toggle switch to turn on or off the 95% confidence terminals and a clickable button to call the previously presented cloud function that returns the values to the plot.

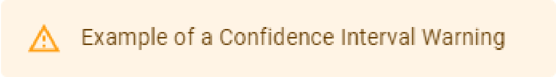
The change of state of the toggle switch does not require a recalculation of the estimator since the cloud function returns the confidence level intervals in both states. The switch, therefore, controls the visualisation of these intervals. The choice to use the switch as a visual flag instead of a cloud function payload flag is to minimise the waiting time if any user wants to take a quick peak of the overall Kaplan-Meier estimates without needing loads of lines confidence levels cause.

The clickable button calls the cloud function is enabled since the choice of the variable is made, except if the user selects all the possible values to exclude, since if the user wants to exclude all the possible values, there would be no curves to be represented, and therefore the button is disabled at that moment.

The last element designed, a responsive line chart, is created after the JSON response since it requires data to portray the result. The responsive line chart used is part of the Nivo library that provide robust data visualisation models that are appellative, so the appearance of the resulting data is more eye-catching and less daunting for the regular user.

The responsive line chart used has the capability of overlapping curves and, as the name indicates, is responsive and interactive in that the cursor overlaps the mapped points.

This feature by design is only permitted for the Kaplan-Meier estimator curves and not the confidence intervals since the latter only serve as visual aids to the interpretation and analysis of the estimator results.



⚠ Example of a Confidence Interval Warning

Figure 5.7: Example of a warning pop-up

Depending on the flag that determines whether or not a warning similar to figure 5.7 is posted to the user to visualise, an extra element may appear above the responsive line chart. This warning alert's only purpose is to inform the user of a possible large confidence interval that could entail a not-so-precise estimator and, consequently, a poorly responsive line chart to make further conclusions, due to the high standard error of the estimates.

5.2 Results

In this section is responsible for presenting the final results of the developed Kaplan-Meier estimator tool are presented. The tool is already implemented in the dashboard of the CLARIFY project, providing results on the survival of cancer patients, using Kaplan-Meier estimates for both lymphomas datasets received (Hodgkin and non-Hodgkin) and for non-small cell lung cancer, for data collected from the Hospital Universitario Puerta de Hierro Majadahonda (HUPHM) and Spanish Lung Cancer Group (SLCG)

The chosen characteristics for all the following images are merely exemplary and do not encompass all the different possibilities and variables the tool enables the user to create.

As previously mentioned, the tool is integrated into the already defined existing platform and, as such, is integrated within the CLARIFY project. The tool is located within the path demonstrated in figure 5.8 within the subsection of figure 5.9 below.

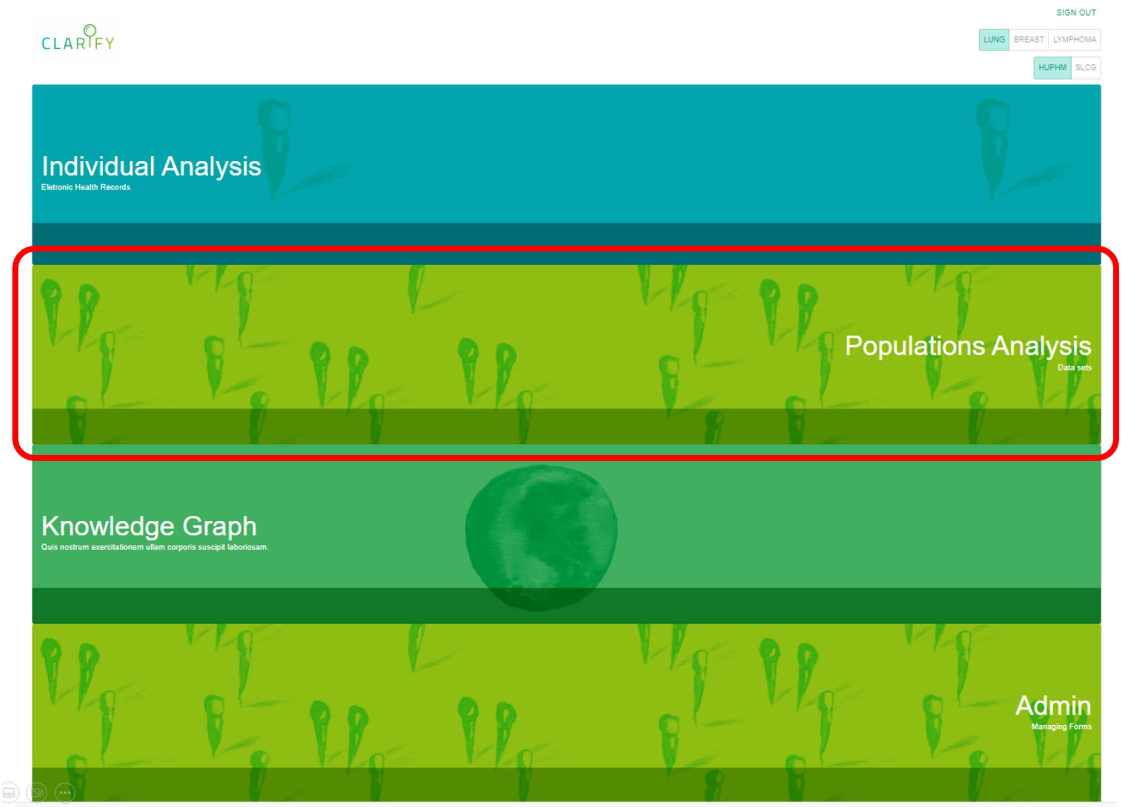
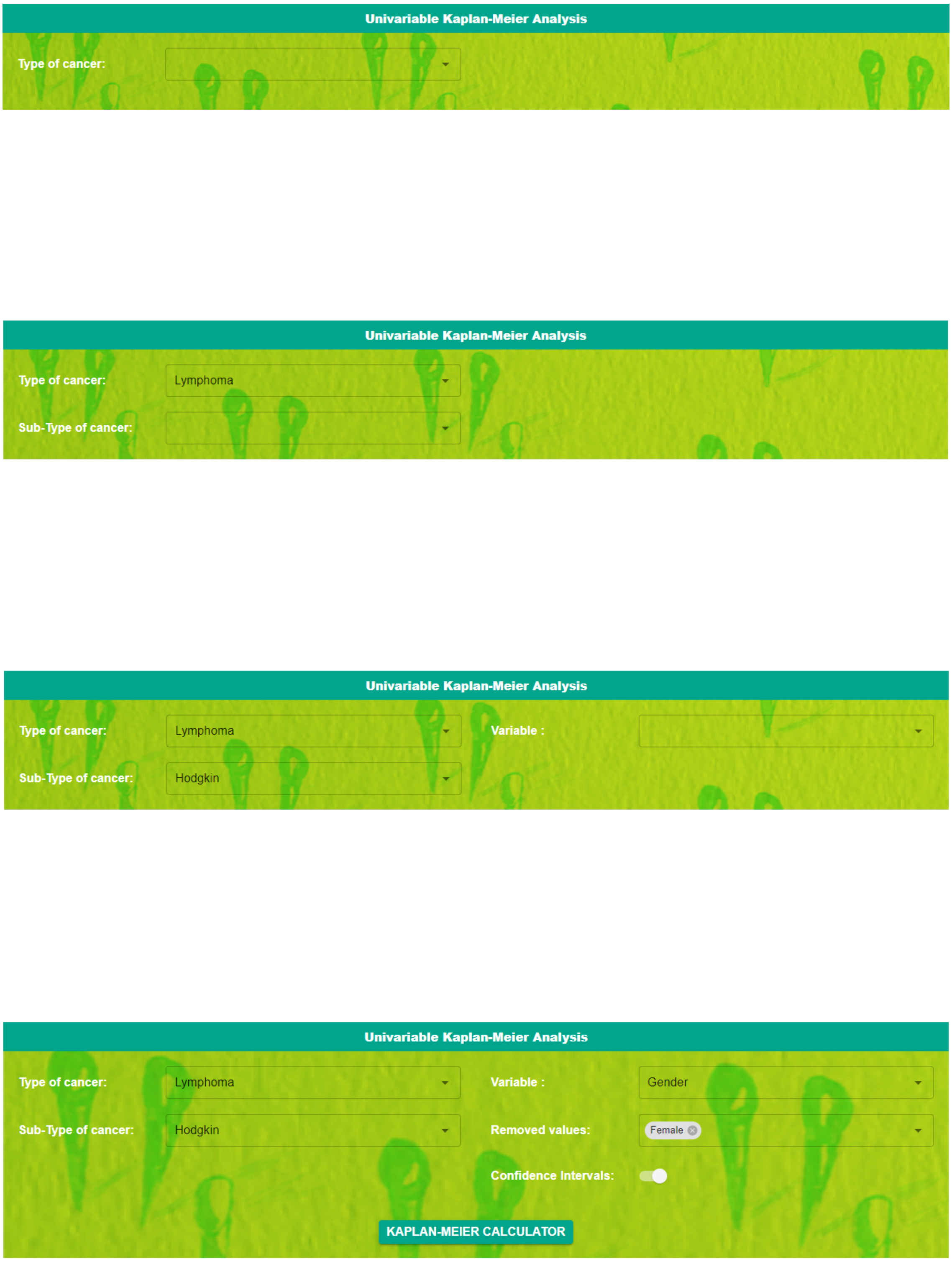


Figure 5.8: Initial menu location of the implemented tool

Figure 5.9: Implemented tool's subsection within the Population Analysis tab

The integrated results for the tool with the standard layout and aesthetic are shown in figure 5.10. In the presented figures, the revelling autocomplete fields are also shown as intended with a need-to-use basis, as stated in the previous section.



Univariable Kaplan-Meier Analysis

Type of cancer:

Univariable Kaplan-Meier Analysis

Type of cancer: Lymphoma

Sub-Type of cancer:

Univariable Kaplan-Meier Analysis

Type of cancer: Lymphoma

Sub-Type of cancer: Hodgkin

Variable :

Univariable Kaplan-Meier Analysis

Type of cancer: Lymphoma

Sub-Type of cancer: Hodgkin

Variable : Gender

Removed values: Female

Confidence Intervals:

Figure 5.10: Different stages of the Kaplan-Meier tool UI

As ilustrated in the sequence of images in figure 5.10, the tool is interactive, allowing the user to choose the components which avoids unnecessary complexity.

The entirety of the tool is shown in the last division of the figure 5.10. If any choices are changed, the options ahead of the changed one are cleared.

The option to show the confidence intervals with the switch on the down right-hand side of the tools box is the only feature that does not require recalculation when enabling or disabling this option.

As a consequence of the use of the tool, a plot is presented that expresses the results for the requirement of the request payload. An example of the two formats in which the data can be presented, either with or without a 95% confidence intervals are presented in figures 5.11 and 5.12, respectively.

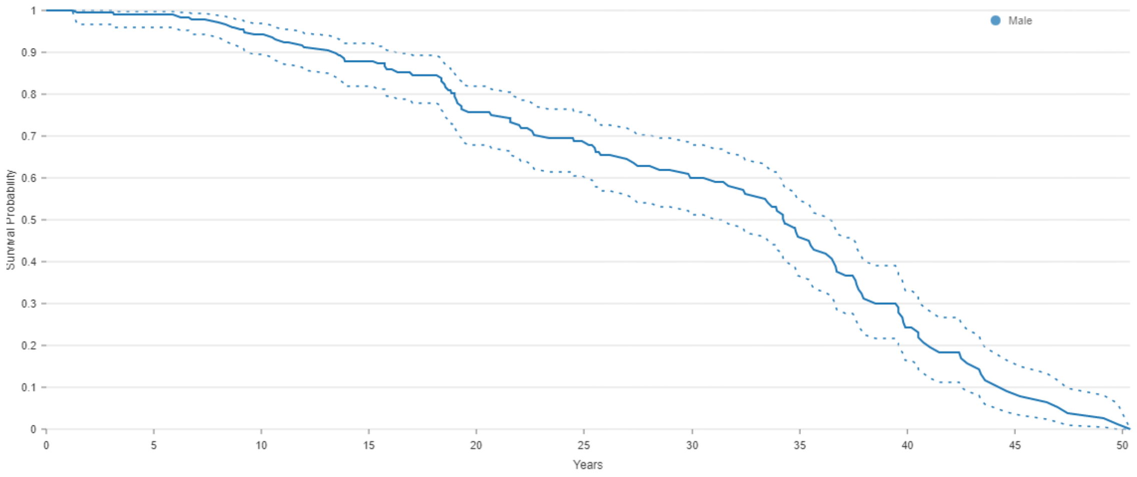


Figure 5.11: Example of a responsive line chart plot with confidence intervals

Figure 5.11 shows the general appearance of the Kaplan-Meier survival curve using the responsive line element with the 95% confidence intervals, which is the default option for any estimator calculation request.

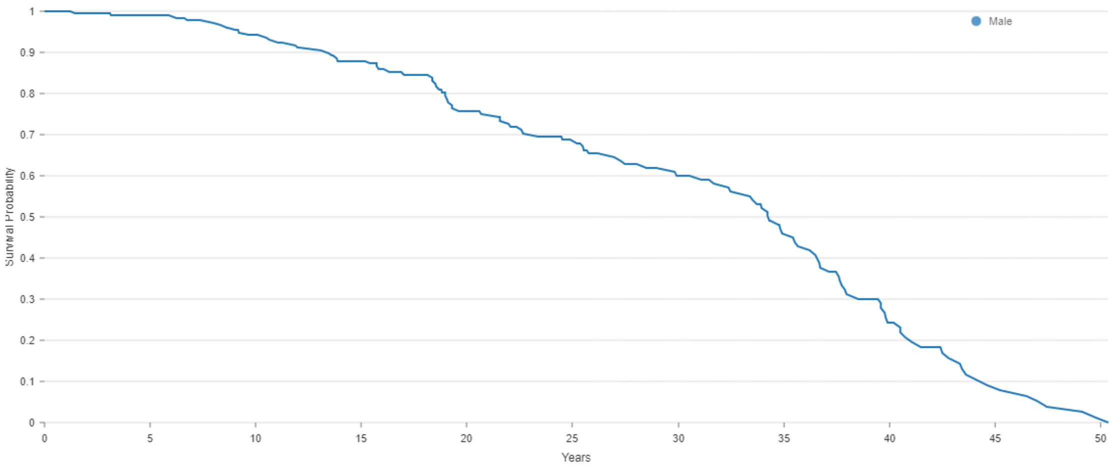


Figure 5.12: Example of a responsive line chart plot without confidence intervals

Alternatively, figure 5.12 shows the optional view of the survival probability estimated (y) for each survival time (x). The choice to only display the survival curve's coordinates instead of the 95% confidence levels coordinates is to make the resulting plot less cluttered and easier to interpret despite only one curve in the figure.

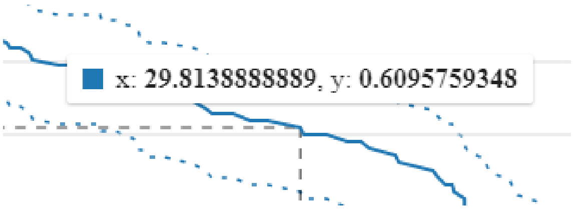


Figure 5.13: Zoomed example responsive capability of the responsive line chart

In both cases, whether the confidence intervals are displayed or not, if the user overlays the mouse's cursor over the line, the coordinates of each point are presented, where x represents the survival time (in years) and y the probability of a patient with the chosen characteristics (type of cancer and associated variables) to survive up to that time according to the Kaplan-Meier estimates. This feature is represented in the zoomed result in figure 5.13 above.

The overall implemented tool represents a more straightforward method of estimating the Kaplan-Meier survival curves and a more quick and interactive system that helps the user to achieve the wanted results more accessible. The added accessibility makes the tool ideal for uses such as diagnosis and progression checks for any patient.

This tool also entails a significant progression line in the overall project since it can be expanded to create modules that do the same procedures for further survival analysis, such as the Cox regression and even multivariable regression for instances.

6

Conclusions and Future Work

This chapter will focus on a comprehensive view of the developed work, taking into consideration all the difficulties and workarounds to implement the initial objective as well as considerations regarding future work-related and derived from the achieved work in this dissertation.

6.1 Conclusions

The conclusion of the work will present an overview of the developments performed within the dissertation and further comments on the overall optimisation and remarks to be taken into consideration. Assessing the produced work regarding the original objectives, it is clear to see an evolutionary line adopted because of the available data.

When working on data in any capacity, it is essential to have at least an overall understanding of the field so that any decisions backed up with conclusions from any statistical analysis do not collide with the world where the data originated. As such, the first step of the work was acquiring a better understatement of not only the technology, algorithms and statistics behind the theory the project entailed but also the subject where these computations and analyses were applied. The study of both Hodgkin and non-Hodgkin lymphoma was paramount to a better understanding of the results and the possible shortcomings of the accoutred data.

Secondly, an initial descriptive analysis was elaborated in order to evaluate the dispersion of values within the entirety of the variables that constituted both datasets. This initial analysis permitted the elimination of non-essential variables and helped to focus the analysis.

Thirdly the Kaplan-Meier estimator analysis was performed to study the survival of both lymphoma patients. The results from this analysis were not ideal, making the following steps of the initial natural progression not viable.

After analysing the initial descriptive analysis and the Kaplan-Meier estimator survival analysis, it was clear that providing a Cox regression would not be feasible since the data used was not good enough to process with that analysis.

Despite the quality of the data, the decision to develop a modular tool to integrate the Kaplan-Meier estimates in an easy-to-use platform within the dashboard was not hindered since the lymphomas' data could be retroactively fitted.

The modularisation for the inclusion of different cancers broadens the horizons for the possibly included types of cancer. This future expansion possibility for any type of cancer that the data provided will be a straightforward implementation without the need for significant resources.

Taking that into consideration, the next phase of the dissertation was developing the Kaplan-Meier estimator tool that would include the grounds for future expansion and the entirety of the Kaplan-Meier analysis realised for the twosome of Hodgkin and non-Hodgkin lymphomas.

In conclusion, despite the not very promising results obtained in the Kaplan-Meier analysis, all the structuring and groundwork for when the updated dataset arrives is as much of a plug-and-play procedure. These results did not significantly impact the overall developed tool due to the robust and adaptive module architecture design for the provided data.

Furthermore, the ability to expand the analysis for any given variables in the dataset is integrated so the received data can be thoroughly analysed.

In addition, creating an all-encompassing Kaplan-Meier estimator tool reduces the development time for each type of cancer and optimises the resources used for the CLARIFY project.

6.2 Future Work

Expanding on the objectives of this dissertation, some steps would further improve upon the development and contribution made towards a better understanding of the received data within the context analysed along with the CLARIFY dashboard.

Following the conclusion of the dissertation, there is a clear progression in integrating the other oncoming cancers into the clarify platform, for instance, breast cancer and future integrated cancers. This natural progression and expansion of the existing Kaplan-Meier curves and respective analysis offer a dynamic evolution of the usefulness of the created tool without a need for further major adaption due to the tool's modularity.

On the occasion of an updated Lymphoma dataset, the reconsideration of developing the survival analysis of both types of lymphoma using a Cox regression is an appropriate step to comprehend better and analyse the tumours in question.

Independently a further elaboration in line with the modularisation and interchangeability of the platforms used tools is to create a new tool in line with the one presented for the Kaplan-Meier analysis for the Cox regression or any number of new applicable survival analysis regression independent of cancer. This approach requires that the stored data follows a generic structure independent from the cancer type, so the adjustments within each type of cancer are not tenuous.

The applicability of neural networks for the survival analysis of patients with lymphoma cancer would also be of possible interest to aid the decision-making by the medical professionals, giving them more information to assist the analysis of the overall problem.

All these alterations and models would be available in the CLARIFY dashboard as well as other possible future hospitals with different data and, therefore, different conclusions.

Bibliography

[1] "Non communicable diseases Internet". en. In: *Who.int* (2022). cited 29 January 2022. Available from: URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (cit. on p. 1).

[2] F. Bray et al. "The ever-increasing importance of cancer as a leading cause of premature death worldwide". In: *Cancer* 127 (16 Aug. 2021), pp. 3029–3030. ISSN: 0008-543X. DOI: 10.1002/cncr.33587 (cit. on p. 1).

[3] T. Hofmarcher et al. *Comparator Report on Cancer in Europe 2019 – Disease Burden, Costs and Access to Medicines*. en. IHE Report 2019:7. Lund, Sweden: IHE (cit. on p. 1).

[4] C. Statistics. *National Cancer Institute*. en. cited 31 January 2022. Available from: 2022. URL: https://www.cancer.gov/about-cancer/understanding/statistics (cit. on p. 1).

[5] "Side effects of chemotherapy | Cancer Council Victoria [Internet". en. In: *Cancervic.org.au* (2022). cited 30 January 2022]. Available from: URL: https://www.cancervic.org.au/cancer-information/treatments/treatments-types/chemotherapy/side_effects_of_chemotherapy.html (cit. on p. 1).

[6] S. D. Fosså, A. A. Dahl, and J. H. Loge. "Fatigue, Anxiety, and Depression in Long-Term Survivors of Testicular Cancer". In: *Journal of Clinical Oncology* 21 (7 Apr. 2003), pp. 1249–1254. ISSN: 0732-183X. DOI: 10.1200/JCO.2003.08.163 (cit. on p. 1).

[7] M. van den Beuken-van Everdingen et al. "Prevalence of pain in patients with cancer: a systematic review of the past 40 years". In: *Annals of Oncology* 18 (9 Sept. 2007), pp. 1437–1449. ISSN: 09237534. DOI: 10.1093/annonc/mdm056 (cit. on p. 2).

[8] Y. Y. Usta. "Importance of Social Support in Cancer Patients". In: *Asian Pacific Journal of Cancer Prevention* 13 (8 Aug. 2012), pp. 3569–3572. ISSN: 1513-7368. DOI: 10.7314/APJCP.2012.13.8.3569 (cit. on p. 2).

[9] K. W. Brown et al. "Psychological Distress and Cancer Survival". In: *Psychosomatic Medicine* 65 (4 July 2003), pp. 636–643. ISSN: 0033-3174. DOI: 10.1097/01.PSY.0000077503.96903.A6 (cit. on p. 2).

[10] E. P. Wright et al. "Social problems in oncology". In: *British Journal of Cancer* 87 (10 Nov. 2002), pp. 1099–1104. ISSN: 0007-0920. DOI: 10.1038/sj.bjc.6600642 (cit. on p. 2).

[11] M. Jeffery et al. "Follow-up strategies for patients treated for non-metastatic colorectal cancer". In: *Cochrane Database of Systematic Reviews* (Nov. 2016). ISSN: 14651858. DOI: 10.1002/14651858.CD002200.pub3 (cit. on p. 2).

[12] R. A. Lewis et al. "Follow-up of cancer in primary care versus secondary care: systematic review." In: *The British journal of general practice : the journal of the Royal College of General Practitioners* 59 (564 July 2009), e234–47. ISSN: 1478-5242. DOI: 10.3399/bjgp09X453567 (cit. on p. 2).

[13] C. M. Alfano et al. "Building Personalized Cancer Follow-up Care Pathways in the United States: Lessons Learned From Implementation in England, Northern Ireland, and Australia". In: *American Society of Clinical Oncology Educational Book* (39 May 2019), pp. 625–639. ISSN: 1548-8748. DOI: 10.1200/EDBK_238267 (cit. on p. 2).

[14] C. Janeway. *Immunobiology 5 : the immune system in health and disease*. New York: Garland Pub, 2001. ISBN: 0-8153-3642-X (cit. on p. 4).

[15] "Lymphoma Action | The immune system". en. In: *Lymphoma Action* (2022). [Internet] cited 4 February 2022. URL: https://lymphoma-action.org.uk/about-lymphoma-what-lymphoma/immune-system#lymphocytes (cit. on p. 4).

[16] "Natural Killer Cells | British Society for Immunology ,Internet". en. In: *Immunology.org* (2022). cited 5 February 2022. URL: https://www.immunology.org/public-information/bitesized-immunology/cells/natural-killer-cells (cit. on p. 4).

[17] "Lymphoma Action | What is lymphoma?" en. In: *Lymphoma Action* (2022). cited 5 February 2022. Available from: URL: https://lymphoma-action.org.uk/about-lymphoma/what-lymphoma (cit. on pp. 4, 5, 7).

[18] "Lymphoma Survival Rate | Blood Cancer Survival Rates | LLS". en. In: *Lls.org* (2022). [Internet] cited 5 February 2022. URL: https://www.lls.org/facts-and-statistics/facts-and-statistics-overview#General%20Blood%20Cancers (cit. on p. 4).

[19] W. I. H. Lymphoma? "Define Hodgkin Lymphoma". en. In: *Cancer.org* (2022). [Internet] cited 5 February 2022. Available from: URL: https://www.cancer.org/cancer/hodgkin-lymphoma/about/what-is-hodgkin-disease.html (cit. on p. 5).

[20] P. Lanzkowsky. *Lanzkowsky's manual of pediatric hematology and oncology*. London: Academic Press is an imprint of Elsevier, 2016. ISBN: 9780128013687 (cit. on p. 5).

[21] C. Lees et al. "Biology and therapy of primary mediastinal B-cell lymphoma: current status and future directions". en. In: *Br J Haematol* (2019) (cit. on p. 5).

[22] M. Metzger and C. Mauz-Körholz. "Epidemiology, outcome, targeted agents and immunotherapy in adolescent and young adult non-Hodgkin and Hodgkin lymphoma". en. In: *Br J Haematol* (June 6, 2019) (cit. on p. 5).

[23] D. A. Eichenauer and A. Engert. "Nodular lymphocyte-predominant Hodgkin lymphoma: a unique disease deserving unique management". In: *Hematology* 2017 (1 Dec. 2017), pp. 324–328. ISSN: 1520-4391. DOI: 10.1182/asheducation-2017.1.324 (cit. on p. 6).

[24] D. A. Eichenauer et al. "Long-Term Follow-Up of Patients With Nodular Lymphocyte-Predominant Hodgkin Lymphoma Treated in the HD7 to HD15 Trials: A Report From the German Hodgkin Study Group". In: *Journal of Clinical Oncology* 38 (7 Mar. 2020), pp. 698–705. ISSN: 0732-183X. DOI: 10.1200/JCO.19.00986 (cit. on p. 6).

[25] H. Kaseb and B. H. Lymphoma. en. In: StatPearls [Internet]. Treasure Island (FL): Mar. 19, 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK499969/ (cit. on pp. 6, 11, 12).

[26] "Lymphoma Action | Hodgkin lymphoma [Internet". en. In: *Lymphoma Action* (Feb. 5, 2022). Available from: URL: https://lymphoma-action.org.uk/types-lymphoma/hodgkin-lymphoma (cit. on p. 6).

[27] S. H. Swerdlow et al. "The 2016 revision of the World Health Organization classification of lymphoid neoplasms." In: *Blood* 127 (20 2016), pp. 2375–90. ISSN: 1528-0020. DOI: 10.1182/blood-2016-01-643569 (cit. on p. 6).

[28] "How doctors group non-Hodgkin lymphoma | non-Hodgkin lymphoma | Cancer Research UK [Internet". en. In: *Cancerresearchuk.org* (2022). cited 6 February 2022]. Available from: URL: https://www.cancerresearchuk.org/about-cancer/non-hodgkin-lymphoma/types/group (cit. on p. 6).

[29] *Cancer.org., What Is Non-Hodgkin Lymphoma? [Internet]*. en. Feb. 6, 2022. URL: https://www.cancer.org/cancer/non-hodgkin-lymphoma/about/what-is-non-hodgkin-lymphoma.html (cit. on pp. 6, 7).

[30] J. O. Armitage and D. D. Weisenburger. "New approach to classifying non-Hodgkin's lymphomas: clinical features of the major histologic subtypes. Non-Hodgkin's Lymphoma Classification Project." In: *Journal of Clinical Oncology* 16 (8 Aug. 1998), pp. 2780–2795. ISSN: 0732-183X. DOI: 10.1200/JCO.1998.16.8.2780 (cit. on pp. 7, 11).

[31] G. Lenz and L. M. Staudt. "Aggressive Lymphomas". In: *New England Journal of Medicine* 362 (15 Apr. 2010), pp. 1417–1429. ISSN: 0028-4793. DOI: 10.1056/NEJMra0807082 (cit. on p. 7).

[32] S. Sapkota and S. H.-H. Lymphoma. en. In: StatPearls [Internet]. Treasure Island (FL): May 1, 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK559328/ (cit. on pp. 7, 9, 11, 12, 15).

[33] A. Ciobanu et al. "Indolent lymphoma: diagnosis and prognosis in medical practice." In: *Maedica* 8 (4 Sept. 2013), pp. 338–42. ISSN: 1841-9038 (cit. on p. 7).

[34] P. T. E. Board. "Adult Non-Hodgkin Lymphoma Treatment (PDQ®): Health Professional Version". en. In: *PDQ Cancer Information Summaries [Internet*. Table], Table 2. Lugano Classification for Hodgkin and Non-Hodgkin Lymphomaa. Available from: Bethesda (MD: National Cancer Institute (US, Jan. 18, 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK66057/table/CDR0000062707__1075/ (cit. on p. 8).

[35] "Recommendations for Initial Evaluation, Staging, and Response Assessment of Hodgkin and Non-Hodgkin Lymphoma: The Lugano Classification". In: *Journal of Clinical Oncology* 32 (27 Sept. 2014), pp. 3059–3067. ISSN: 0732-183X. DOI: 10.1200/JCO.2013.54.8800 (cit. on p. 8).

[36] "Lugano classification for staging of lymphomas [Internet". en. In: *Uptodate.com* (2022). cited 7 February 2022]. Available from: URL: https://www.uptodate.com/contents/image?imageKey=HEME%2F66651 (cit. on p. 8).

[37] M. Matasar. "Late mortality and morbidity of patients with Hodgkin lymphoma treated in adulthood". en. In: *J Clin Oncol* 27.15s (2009), p. 8547 (cit. on p. 9).

[38] M. Provencio. "Late relapses in Hodgkin lymphoma: a clinical and immunohistochemistry study". pt. In: *Leuk Lymphoma* 51 (2010), pp. 1686–91 (cit. on p. 9).

[39] S. Hess. "Adult survivors of childhood malignant lymphoma are not aware of their risk of late effects". en. In: *Acta Oncol* 50 (2011), pp. 653–9 (cit. on p. 9).

[40] L. Anderson et al. "Population-based study of autoimmune conditions and the risk of specific lymphoid malignancies". en. In: *Int J Cancer. 2009 Jul* 15;125(2):398-405 () (cit. on pp. 9, 13).

[41] A. Smith et al. "Incidence of haematological malignancy by sub-type: a report from the Haematological Malignancy Research Network". en. In: *Br J Cancer. 2011 Nov* 22;105(11):1684-92 (). PMC free article] [PubMed (cit. on pp. 9, 10).

[42] M. Provencio. "Analysis of competing risks of causes of death and their variation over different time periods in Hodgkin's disease". en. In: *Clin Cancer Res* 14 (2008), pp. 5300–5 (cit. on p. 9).

[43] D. Hodgson. "Long-term solid cancer risk among 5-year survivors of Hodgkin's lymphoma". en. In: *J Clin Oncol* 25 (2007), pp. 1489–97 (cit. on p. 10).

[44] T. Best. "Variants at 6q21 implicate PRDM1 in the etiology of therapy-induced second malignancies after Hodgkin's lymphoma". en. In: *Nat Med* 17 (2011), pp. 941–3 (cit. on p. 10).

[45] P. Barbaro. "Reduced incidence of second solid tumors in survivors of childhood Hodgkin's lymphoma treated without radiation therapy". en. In: *Ann Oncol* 22 (2011), pp. 2569–74 (cit. on p. 10).

[46] Lymphoma.org. "Lymphoma, Hodgkin Diagnosis | Lymphoma Research Foundation [Internet]". en. In: *Lymphoma Research Foundation* (2022). [cited 7 February 2022]. Available from: URL: https://lymphoma.org/aboutlymphoma/hl/hldiagnosis/ (cit. on p. 10).

[47] Lymphoma.org. "Non-Hodgkin Lymphoma Diagnosis - Lymphoma Research Foundation [Internet]". en. In: *Lymphoma Research Foundation* (2022). [cited 7 February 2022]. Available from: URL: https://lymphoma.org/aboutlymphoma/nhl/nhldiagnosis/ (cit. on p. 10).

[48] [. Next.amboss.com. en. [cited 7 February 2022]. Feb. 7, 2022. URL: https://next.amboss.com/us/article/mT0Vr2?q=hodgkin%20lymphoma#Z419133b57ea628b183f73a28c5ad7a8c (cit. on p. 10).

[49] L. Goldman. *Goldman-Cecil medicine*. Philadelphia, PA: Elsevier, 2020. ISBN: 9780323532662 (cit. on p. 10).

[50] I. Next.amboss.com. en. [cited 7 February 2022]. Feb. 7, 2022. URL: https://next.amboss.com/us/article/NT0-I2?q=non-hodgkin%20lymphomas#Z46cc87b0bb6d0dccc4fedce8fd20de06 (cit. on p. 11).

[51] A. M. Evens, M. Hutchings, and V. Diehl. "Treatment of Hodgkin lymphoma: the past, present, and future". In: *Nature Clinical Practice Oncology* 5 (9 Sept. 2008), pp. 543–556. ISSN: 1743-4254. DOI: 10.1038/ncponc1186 (cit. on p. 12).

[52] S. M. Ansell. "Hodgkin Lymphoma: Diagnosis and Treatment". In: *Mayo Clinic Proceedings* 90 (11 Nov. 2015), pp. 1574–1583. ISSN: 00256196. DOI: 10.1016/j.mayocp.2015.07.005 (cit. on pp. 12, 13).

[53] K. Ardeshna et al. "Long-term effect of a watch and wait policy versus immediate systemic treatment for asymptomatic advanced-stage non-Hodgkin lymphoma: a randomised controlled trial". In: *The Lancet* 362 (9383 Aug. 2003), pp. 516–522. ISSN: 01406736. DOI: 10.1016/S0140-6736(03)14110-4 (cit. on p. 13).

[54] J. L. Jameson et al. *Harrison's Principles of Internal Medicine*. Twentieth. Vol. 1 Vol.2. McGraw-Hill Education / Medical (cit. on p. 13).

[55] G. Dores. "Second malignant neoplasms among long-term survivors of Hodgkin's disease: a population-based evaluation over 25 years". en. In: *J Clin Oncol* 20 (2002), pp. 3484–94 (cit. on p. 13).

[56] L. Jiang and N. Li. “B-cell non-Hodgkin lymphoma: importance of angiogenesis and antiangiogenic therapy”. en. In: *Angiogenesis* (2020) (cit. on p. 15).

[57] B. von Tresckow and C. H. Moskowitz. “Treatment of relapsed and refractory Hodgkin Lymphoma”. In: *Seminars in Hematology* 53 (3 July 2016), pp. 180–185. ISSN: 00371963. DOI: 10.1053/j.seminhematol.2016.05.010 (cit. on p. 16).

[58] M. Liedtke et al. “Surveillance imaging during remission identifies a group of patients with more favorable aggressive NHL at time of relapse: a retrospective analysis of a uniformly-treated patient population”. In: *Annals of Oncology* 17 (6 June 2006), pp. 909–913. ISSN: 09237534. DOI: 10.1093/annonc/mdl049 (cit. on p. 16).

[59] B. Dabaja, C. S. Ha, and J. D. Cox. *Chapter 34 - Leukemias and Lymphomas*. Ed. by J. D. Cox and K. K. Ang. Ninth Edition. Mosby, 2010, pp. 875–911. ISBN: 978-0-323-04971-9. URL: https://www.sciencedirect.com/science/article/pii/B9780323049719000342 (cit. on p. 16).

[60] *Diffuse Large B-Cell Lymphoma, Not Otherwise Specified*. Elsevier, 2018, pp. 370–377. DOI: 10.1016/B978-0-323-47779-6.50058-2 (cit. on p. 16).

[61] C. Buske et al. “The Follicular Lymphoma International Prognostic Index (FLIPI) separates high-risk from intermediate- or low-risk patients with advanced-stage follicular lymphoma treated front-line with rituximab and the combination of cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) with respect to treatment outcome”. In: *Blood* 108 (5 Sept. 2006), pp. 1504–1508. ISSN: 0006-4971. DOI: 10.1182/blood-2006-01-013367 (cit. on p. 16).

[62] M. M. Oken et al. “Toxicity and response criteria of the Eastern Cooperative Oncology Group.” In: *American journal of clinical oncology* 5 (6 Dec. 1982), pp. 649–55. ISSN: 0277-3732 (cit. on pp. 16, 17).

[63] *ECOG Performance Status Scale - ECOG-ACRIN Cancer Research Group [Internet*. en. Available from: June 3, 2022. URL: https://ecog-acrin.org/resources/ecog-performance-status/ (cit. on p. 16).

[64] S. Imran and I. Hyder. “Security Issues in Databases”. In: *2009 Second International Conference on Future Information Technology and Management Engineering*. 2009, pp. 541–545. DOI: 10.1109/FITME.2009.140 (cit. on pp. 17, 22).

[65] “The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with diffuse large B-cell lymphoma treated with R-CHOP.” In: *Blood* 109 (5 Mar. 2007), pp. 1857–61. ISSN: 0006-4971. DOI: 10.1182/blood-2006-08-038257 (cit. on pp. 17, 18).

[66] "A Predictive Model for Aggressive Non-Hodgkin's Lymphoma". In: *New England Journal of Medicine* 329.14 (1993). PMID: 8141877, pp. 987–994. DOI: 10.1056/NEJM199309303291402. eprint: https://doi.org/10.1056/NEJM199309303291402. URL: https://doi.org/10.1056/NEJM199309303291402 (cit. on p. 17).

[67] A. S. Ruppert et al. "International prognostic indices in diffuse large B-cell lymphoma: a comparison of IPI, R-IPI, and NCCN-IPI". In: *Blood* 135.23 (June 2020), pp. 2041–2048. ISSN: 0006-4971. DOI: 10.1182/blood.2019002729. eprint: https://ashpublications.org/blood/article-pdf/135/23/2041/1743304/bloodbld2019002729.pdf. URL: https://doi.org/10.1182/blood.2019002729 (cit. on p. 17).

[68] "International Prognostic Index for Agressive Non-Hodgkin's Lymphoma [Internet". en. In: *Oncologypro.esmo.org* (June 4, 2022). Available from: URL: https://oncologypro.esmo.org/oncology-in-practice/practice-tools/international-prognostic-index-tools-for-lymphoma/prognostic-index-non-hodgkin-s-lymphoma (cit. on p. 17).

[69] I. N.-H. L. P. F. Project. "A predictive model for aggressive non-Hodgkin's lymphoma." In: *The New England journal of medicine* 329 (14 1993), pp. 987–94. ISSN: 0028-4793. DOI: 10.1056/NEJM199309303291402 (cit. on p. 18).

[70] S. Schneeweiss. "Learning from Big Health Care Data". In: *New England Journal of Medicine* 370 (23 June 2014), pp. 2161–2163. ISSN: 0028-4793. DOI: 10.1056/NEJMp1401111 (cit. on p. 18).

[71] S. Ajami and T. BagheriTadi. "Barriers for Adopting Electronic Health Records (EHRs) by Physicians". In: *Acta Informatica Medica* 21 (2 2013), p. 129. ISSN: 0353-8109. DOI: 10.5455/aim.2013.21.129-134 (cit. on p. 18).

[72] M. A. R. "Health Privacy in the Electronic Age". In: *Journal of Legal Medicine* 28 (4 Dec. 2007), pp. 487–501. ISSN: 0194-7648. DOI: 10.1080/01947640701732148 (cit. on p. 18).

[73] *Electronic health records – standard formats [Internet]*. en. cited 26 February 2022]. Available from: 2022. URL: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/1999-Registos-de-saude-eletronicos-formatos-normalizados_pt (cit. on p. 18).

[74] J. Vora et al. "Ensuring Privacy and Security in E- Health Records". In: IEEE, July 2018, pp. 1–5. ISBN: 978-1-5386-4599-4. DOI: 10.1109/CITS.2018.8440164 (cit. on p. 18).

[75] B. M. Knoppers and A. M. Thorogood. "Ethics and Big Data in health". In: *Current Opinion in Systems Biology* 4 (Aug. 2017), pp. 53–57. ISSN: 24523100. DOI: 10.1016/j.coisb.2017.07.001 (cit. on p. 19).

[76] E. S. Dove and C. Garattini. "Expert perspectives on ethics review of international data-intensive research: Working towards mutual recognition". In: *Research Ethics* 14 (1 Jan. 2018), pp. 1–25. ISSN: 1747-0161. DOI: 10.1177/1747016117711972 (cit. on p. 19).

[77] A. Panesar. *Machine Learning and AI for Healthcare*. Apress, 2021. ISBN: 978-1-4842-6536-9. DOI: 10.1007/978-1-4842-6537-6 (cit. on pp. 19, 20, 24–26).

[78] D. Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. META Group, Feb. 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (cit. on p. 19).

[79] M. Dave and J. Kamal. "Identifying big data dimensions and structure". In: IEEE, Sept. 2017, pp. 163–168. ISBN: 978-1-5090-5838-9. DOI: 10.1109/ISPCC.2017.8269669 (cit. on p. 20).

[80] M. R. TRIFU and M. L. IVAN. "Big Data: present and future". In: *Database Systems Journal* 5 (1 2014) (cit. on p. 20).

[81] S. Sivabalan and R. I. Minu. "Heterogeneous Data Integration with ELT and Analytical MPP Database for Data Analysis Application". In: IEEE, Nov. 2021, pp. 1–5. ISBN: 978-1-6654-2691-6. DOI: 10.1109/i-PACT52855.2021.9696841 (cit. on p. 21).

[82] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos. "Conceptual modeling for ETL processes". In: ACM Press, 2002, pp. 14–21. ISBN: 1581135904. DOI: 10.1145/583890.583893 (cit. on p. 21).

[83] M. Nasution. *Relational Database Management Systems*. Mar. 2021 (cit. on p. 21).

[84] G. Blokdyk. *RDBMS Relational Database Management System a Complete Guide - 2020 Edition*. Emereo Pty Limited, 2019. ISBN: 9780655942290. URL: https://books.google.pt/books?id=0YTIywEACAAJ (cit. on p. 22).

[85] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. "Knowledge Discovery and Data Mining: Towards a Unifying Framework". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, 1996, pp. 82–88 (cit. on pp. 22, 23).

[86] T. Davenport and R. Kalakota. "The potential for artificial intelligence in healthcare". In: *Future Healthcare Journal* 6 (2 June 2019), pp. 94–98. ISSN: 2514-6645. DOI: 10.7861/futurehosp.6-2-94 (cit. on p. 25).

[87] B. Mahesh. *Machine Learning Algorithms -A Review*. Jan. 2019. DOI: 10.21275/ART20203995 (cit. on p. 25).

[88] A. Moubayed et al. "E-Learning: Challenges and Research Opportunities Using Machine Learning amp; Data Analytics". In: *IEEE Access* 6 (2018), pp. 39117–39138. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2851790 (cit. on p. 25).

[89] P. Cunningham, M. Cord, and S. J. Delany. *Supervised Learning*. Springer Berlin Heidelberg, pp. 21–49. DOI: 10.1007/978-3-540-75171-7_2 (cit. on p. 26).

[90] M. E. Celebi and K. Aydin, eds. *Unsupervised Learning Algorithms*. Springer International Publishing, 2016. ISBN: 978-3-319-24209-5. DOI: 10.1007/978-3-319-24211-8 (cit. on p. 26).

[91] X. Zhu and A. B. Goldberg. "Introduction to Semi-Supervised Learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3 (1 Jan. 2009), pp. 1–130. ISSN: 1939-4608. DOI: 10.2200/S00196ED1V01Y200906AIM006 (cit. on p. 26).

[92] T. G. Dietterich. "Ensemble Methods in Machine Learning". In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. ISBN: 978-3-540-45014-6 (cit. on p. 27).

[93] M. Chen, S. Mao, and Y. Liu. "Big Data: A Survey". In: *Mobile Networks and Applications* 19 (2 Apr. 2014), pp. 171–209. ISSN: 1383-469X. DOI: 10.1007/s11036-013-0489-0 (cit. on p. 27).

[94] S. Dash et al. "Big data in healthcare: management, analysis and future prospects". In: *Journal of Big Data* 6 (1 Dec. 2019), p. 54. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0217-0 (cit. on p. 27).

[95] J. Peng et al. "Comparison of Several Cloud Computing Platforms". In: IEEE, Dec. 2009, pp. 23–27. ISBN: 978-1-4244-6325-1. DOI: 10.1109/ISISE.2009.94 (cit. on p. 27).

[96] C. Gackenheimer. *What Is React?* 2015. DOI: 10.1007/978-1-4842-1245-5_1 (cit. on p. 27).

[97] React. "A JavaScript library for building user interfaces [Internet". en. In: *Reactjs.org* (2022). cited 7 June 2022]. Available from: URL: https://reactjs.org/ (cit. on p. 27).

[98] K. Boczkowski and B. Pańczyk. "Comparison of the performance of tools for creating a SPA application interface - React and Vue.js". In: *Journal of Computer Sciences Institute* 14 (Mar. 2020), pp. 73–77. DOI: 10.35784/jcsi.1579. URL: https://ph.pollub.pl/index.php/jcsi/article/view/1579 (cit. on p. 28).

[99] C. Abernathy. *Open source in 2015: A year of growth [Internet*. en. Engineering at Meta. 2022 [cited 7 June 2022]. Available from: URL: https://engineering.fb.com/2015/12/29/developer-tools/open-source-in-2015-a-year-of-growth/ (cit. on p. 28).

[100] J. M. Bland and D. G. Altman. "Statistics Notes: Survival probabilities (the Kaplan-Meier method)". In: *BMJ* 317 (7172 Dec. 1998), pp. 1572–1580. ISSN: 0959-8138. DOI: 10.1136/bmj.317.7172.1572 (cit. on p. 28).

[101] C. Rocha and A. Papoila. *Análise de sobrevivência*. pt. Sociedade Portuguesa de Estatística (cit. on pp. 28, 29).

[102] L. J. A. Stalpers and E. L. Kaplan. "Edward L. Kaplan and the Kaplan-Meier Survival Curve". In: *BSHM Bulletin: Journal of the British Society for the History of Mathematics* 33 (2 May 2018), pp. 109–135. ISSN: 1749-8430. DOI: 10.1080/17498430.2018.1450055 (cit. on p. 28).

[103] J. M. Bland and D. G. Altman. "The logrank test". In: *BMJ* 328 (7447 May 2004), p. 1073. ISSN: 0959-8138. DOI: 10.1136/bmj.328.7447.1073 (cit. on p. 29).

[104] D. G. Kleinbaum and M. Klein. *Survival Analysis*. Springer New York, 2012. ISBN: 978-1-4419-6645-2. DOI: 10.1007/978-1-4419-6646-9 (cit. on p. 29).

[105] M. Torrente et al. "Clinical Factors Influencing Long-Term Survival in a Real-Life Cohort of Early Stage Non-Small-Cell Lung Cancer Patients". In: (Aug. 2022). DOI: 10.21203/rs.3.rs-1788174/v2. URL: https://doi.org/10.21203/rs.3.rs-1788174/v2 (cit. on pp. 29, 116).

[106] L. E. Abrey, L. M. DeAngelis, and J. Yahalom. "Long-term survival in primary CNS lymphoma." In: *Journal of Clinical Oncology* 16.3 (1998). PMID: 9508166, pp. 859–863. DOI: 10.1200/JCO.1998.16.3.859. eprint: https://doi.org/10.1200/JCO.1998.16.3.859. URL: https://doi.org/10.1200/JCO.1998.16.3.859 (cit. on p. 30).

[107] R. Fisher et al. "Factors predicting long-term survival in diffuse mixed, histiocytic, or undifferentiated lymphoma". In: *Blood* 58 (1 July 1981), pp. 45–51. ISSN: 0006-4971. DOI: 10.1182/blood.V58.1.45.45 (cit. on p. 30).

[108] H. Schulz et al. "Rituximab in relapsed lymphocyte-predominant Hodgkin lymphoma: long-term results of a phase 2 trial by the German Hodgkin Lymphoma Study Group (GHSG)". In: *Blood* 111 (1 Jan. 2008), pp. 109–111. ISSN: 0006-4971. DOI: 10.1182/blood-2007-03-078725 (cit. on p. 30).

[109] R. S. Geiger et al. ""Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data?" In: *Quantitative Science Studies* 2 (3 Nov. 2021), pp. 795–827. ISSN: 2641-3337. DOI: 10.1162/qss_a_00144 (cit. on p. 88).