Survival Analysis of Cancer Patients in Portugal following the Reference Centre Model Implementation

Manuel Martins de Melo Rodrigues Mateus

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Maria Margarida Martelo Catalão Lopes de Oliveira Pires Pina

Dr. Rui Gentil de Portugal e Vasconcelos Fernandes

Examination Committee:

Chairperson: Prof. Mónica Duarte Correia de Oliveira

Supervisor: Prof. Maria Margarida Martelo Catalão Lopes de Oliveira Pires Pina

Member of the Committee: Prof. António Vaz Carneiro

September 2020

Preface

The work presented in this thesis was performed at the Centro de Estudos de Gestão of Instituto Superior Técnico (Lisbon, Portugal), during the period of September 2019 to July 2020, being supervised by Professor Maria Margarida Martelo Catalão Lopes de Oliveira Pires Pina and co-supervised by Doctor Rui Gentil de Portugal e Vasconcelos Fernandes.

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgements

I would first like to thank Professor Margarida Catalão Lopes for all the availability, support and guidance throughout all the phases of this Dissertation.

I would also like to thank Doctor Rui Portugal for the guidance and support, in particular on the identification and framing of the medical context for this Dissertation.

I also extend my acknowledgment to ACSS, in particular Doctor Cláudia Medeiros Borges, for the availability and support on making available the data which served as the basis for the research conducted in this Dissertation.

I would also like to remember all my friends which were part of my path throughout my time in Técnico Lisboa, in particular Miguel Amador and Pedro Afonso for their camaraderie, but also their support and incentive for finishing this Dissertation.

A special thanks to all my friends from Ponte de Lima for their friendship, in particular Vitor Hugo Silva and Pedro Ferreira, who have been my loyal friends, ever since childhood.

Finally, I would like to thank all my family, in particular my brothers, and of course my parents, who have always been by my side and to whom I dedicate this Dissertation.

Resumo

A doença oncológica tem uma elevada incidência global, tendo afetado cerca de dezoito milhões de pessoas em todo o mundo em 2018. De acordo com a Organização Mundial de Saúde, é esperado que este número aumente para perto do dobro até 2040. Em Portugal, a doença oncológica foi diagnosticada a sessenta mil pessoas em 2018, tendo sido também a segunda principal causa de morte nesse mesmo ano, associada a aproximadamente uma em cada quatro mortes.

No seguimento da publicação da Diretiva Europeia 2011/24/EU, o Ministério da Saúde de Portugal criou em 2013 um Grupo de Trabalho que definiu o conceito de Centro de Referência (CR) como uma unidade que disponibiliza cuidados de saúde altamente especializados e de elevada qualidade para determinadas condições clínicas que, devido à sua baixa prevalência, alta complexidade ou custos associados, necessitam de uma abordagem estratégica. Uma das áreas estabelecida como prioritária para a criação de CR foi precisamente a área da oncologia.

Desde a implementação do primeiro CR para oncologia em Portugal, em 2015, não foi ainda feito um estudo do impacto na sobrevivência dos pacientes oncológicos acompanhados por CRs, isto é, pacientes que tiveram pelo menos um ou todos os episódios hospitalares num CR, em comparação com pacientes que não tiveram nenhum episódio num CR.

Nesta Dissertação de Mestrado faz-se uma análise de sobrevivência com base em dados referentes a seis tipos de cancro para os quais houve a criação de CRs de oncologia em Portugal: hepatobiliar, pancreático, sarcomas, esofágico, onco-oftalmologia e testicular. O objetivo é aferir o impacto da criação de CRs na sobrevivência destes pacientes. São considerados pacientes com episódios com alta hospitalar entre 2010 e 2019, registados na base de dados portuguesa de Grupos de Diagnósticos Homogéneos. Cada grupo de pacientes, por tipo de cancro, é analisado seguindo uma metodologia de acordo com as melhores práticas da literatura: procede-se à análise descritiva dos dados, estimam-se curvas de sobrevivência com recurso ao método de Kaplan-Meier e taxas de risco para diferentes covariáveis usando modelos multivariável estendidos de Cox.

Os resultados obtidos suportam a implementação e incentivam o aprofundar do modelo de CRs para oncologia em Portugal, já que os pacientes oncológicos acompanhados por CRs têm, em geral, uma maior probabilidade de sobrevivência a cada momento, quando comparados com pacientes que não têm nenhum episódio num CR. Estes resultados são mais evidentes para os casos dos cancros hepatobiliar e pancreático, mas também são visíveis nos casos dos sarcomas e cancro esofágico. Para a onco-oftalmologia e para o cancro testicular, devido ao relativamente pequeno número de pacientes e mortalidade registados, não foi possível obter resultados conclusivos.

Palavras-chave: Centros de Referência, Análise de Sobrevivência em Oncologia, Método de Kaplan-Meier, Modelos Estendidos de Cox

Abstract

Cancer has a globally high incidence, having affected around eighteen million people all over the world in 2018. According to the World Health Organization, this figure is expected to nearly double by 2040. In Portugal, cancer has been diagnosed in sixty thousand individuals during 2018, the second leading cause of death in that year and being associated with one in each four deaths.

Following the publication of the European Directive 2011/24/EU, the Portuguese Ministry of Health created a Work Group in 2013 which defined the concept of Reference Centre (RC) has a highly specialized healthcare providing unit, with particular knowledge and expertise to provide high-quality care to patients with a certain clinical condition, which due to its low prevalence, high complexity or costs associated, requires a strategic approach. One of the priority areas defined for the RC creation was oncology.

Ever since the implementation of the first oncology RC in Portugal, in 2015, there has not been a study of the impact on survival for patients who have been followed by RCs, that is, patients who had at least one or all hospital episodes in a RC, when compared to patients who had no hospital episode in a RC.

This Master Dissertation performs a survival analysis on data from a set of six cancer types which have seen the creation of oncology RCs in Portugal: hepatobiliary, pancreatic, sarcomas, oesophageal, onco-ophthalmology and testicular. The aim is to assess the impact of RCs on the survival probability of these patients. Patients with hospital discharges between 2010 and 2019 are considered, registered in the Portuguese database of Diagnosis-Related Groups. Each group of patients, per cancer type, is subject to a methodology developed according to the best practices described in the literature: a descriptive analysis is made, survival curves are estimated using the Kaplan-Meier methodology and hazard ratios are also estimated for different covariates, using multivariate Extended Cox models.

The results obtained support the implementation and encourage further extension of the RC model for oncology in Portugal, as the cancer patients followed by RCs, overall, have a better survival probability at each moment, when compared to patients who had no episode in a RC. These results are clearer for hepatobiliary and pancreatic cancer but are also visible for sarcomas and oesophageal cancer. Regarding onco-ophthalmology and testicular cancer, due to the relative low number of patients and deaths registered, no conclusive results were obtained.

Keywords: Reference Centre, Cancer Survival Analysis, Kaplan-Meier Method, Extended Cox Models

Contents

Figure Index

Table Index

List of abbreviations

ACSS	Central Administration of the Health System, I.P.
CI	Confidence Interval
Cox PH	Cox Proportional Hazards
DGS	General Directorate of Health
DRG	Diagnosis-Related Group
ERN	European Reference Network
HR	Hazard Ratio
ICD	International Classification of Diseases
KM	Kaplan-Meier
RC	Reference Centre
RMST	Restricted Mean Survival Time
SNS	National Health Service
WHO	World Health Organization

Conventional Signs

e.g.	For example
i.e.	That is to say
n.a.	Not applicable

1 Introduction

Over the last years, Portugal has seen a remarkable improvement in terms of health care services provided to patients, reflected on overall health indicators such as life expectancy at birth and at age 65 years [1]. Overall, as of 2019, health expenditure represents an important fraction of the country Gross Domestic Product (GDP), accounting to approximately 9% [2].

The Portuguese health system is, however, increasingly facing new challenges, one being the increase of life expectancy that Portugal has observed over the past few decades [2]. With the rapid ageing of the population, the prevalence of chronic conditions creates new demand for medical care [1]. This challenge is not unique to Portugal: it is common among several European countries, which are also observing its population ageing. This phenomenon is driving an increase in costs, promoted by higher demand for treatment for older patients. Thus, an efficiency improvement focused on outcomes is needed, through a holistic view, leveraging cooperation between different professionals and healthcare units which are distributed among different layers of the health care systems, at a country and European level [3].

Recognising the shared principles of health systems across Europe and the benefits of cooperation, the European Parliament and the European Council have published a European Directive promoting patients' rights in cross-border healthcare [3]. This Directive recognises the importance of highly specialised healthcare treatment for particular conditions, introducing the concept of highly specialized units for health care for specific conditions, framed in European Reference Networks (ERN). The introduction of these concepts aims to contribute to the free movement of patients across Europe, but also for specialization and benefiting from economies of scale [3]. Portugal, as a European member, began the transposition of this European Directive, establishing the criteria for identification and characterization of Reference Centres (RCs), as highly specialized healthcare providing units, with particular knowledge and expertise to provide high-quality care to patients with a certain clinical condition [3]. RC implementation was a priority for the Portuguese Government to promote specialization and modernization of the Portuguese healthcare units [3].

Population ageing in Portugal, as is also happening in Europe, has been driving a constant increase of cancer cases. According to the report "Programa Nacional para as Doenças Oncológicas de 2017" (the most recent), this increase reached a year-over-year growth rate of approximately 3% in 2015 [4]. Cancer is the leading cause of death in Portugal before the age of 70 and the second cause of death for all ages nationwide [3]. The cancer diagnosis and treatment are complex and have significant costs associated. In particular for Portugal, Lopes et al. [5] estimated cancer treatment to account for an annual cost of 867 million euro in 2017, associated with direct medical costs, representing 5.5% of Portuguese total health expenditure. Lopes et al. [5] also refer an increase of

300 million euros in the annual direct medical costs associated with cancer treatment, from 2006 until 2017, "*which may be explained by an increase in incidence and the rising cost of* drugs" [5, p. 8].

The diagnosis and treatment of cancer patients require the participation of multi-disciplinary teams, as well as technological equipment, demanding specialized resources to operate them [3]. Taking these facts into account, both to promote the concentration of resources, either in terms of human resources and technology, but also to provide the best possible treatment for cancer patients, oncology has been one of the priority areas defined for RC creation [3].

The first oncology RC was formally recognized in Portugal in 2015, for onco-ophthalmology – the *Centro Hospitalar e Universitário de Coimbra* (CHUC). Ever since, several other oncology RCs have been officially recognized, up to a total of fifty as of today for several cancer types.

Cancer represents a significant burden in Portugal, mainly due to the incidence and mortality associated [5]. According to the International Agency for Research on Cancer, there have been approximately 58000 new cases and 29000 deaths from cancer in Portugal, during the year 2018 [6]. It is therefore pertinent to study the impact of RCs in terms of patients' survival, when compared to patients treated in other healthcare units (i.e. non-RCs).

Survival analysis is a longitudinal statistical method used to study the occurrence and timing of a specific event [7]. It has been widely used to study general life events occurrence and timing, in particular in cancer research [7][8]. The event is usually referred as failure because it is often related to a negative result, such as death, as is often the case of cancer survival analysis studies [8].

As such, this Dissertation aims to study the benefits, in terms of survival, for cancer patients in Portugal, following the RC model implementation. The raw data was obtained from ACSS and included patients who were primarily diagnosed with one of the following six cancer types: hepatobiliary, pancreatic, sarcomas, oesophagus, testicular and onco-ophthalmology. The data included information about hospital episodes these patients had, with discharge date registered between the 1st of January of 2010 and 25th of November 2019 (the moment the data was collected and made available by ACSS). Several explanatory variables are taken into account, such as age, sex, number of surgeries, number of infections and the information about whether the patient had all episodes or some of them in an oncology RC.

Chapter 1 corresponds to this introduction, which lays out the motivation and objectives of the Dissertation.

Chapter 2 includes a description of the context of this Dissertation, namely the RC context, the steps taken for implementation in Portugal, as well as the criteria for identification and official recognition. A particular emphasis is given to oncology RCs, as those will be the focus of this Dissertation. This chapter also includes an overview of the six cancer types, which will be subject to the survival

analysis. Finally, since the data was obtained from the Portuguese Database of Diagnosis-Related Groups (DRGs), a description of the origin of the DRGs is made, in particular for Portugal, as well as how the variables are coded in that Database.

Chapter 3 corresponds to the literature review and includes an overview of the main published articles, which can be related to the topics in this Dissertation, such as oncology RCs, as well as the methods available for survival analysis and its application in a research context. Finally, some examples are provided from the literature of survival analysis that was developed for the six cancer types studied.

Chapter 4 describes the materials and methodology used. First, the raw data sets which were obtained from ACSS are briefly described, as well as the cleansing processes which were applied to the data to prepare it for analysis. This chapter also presents the methodology and the steps which were used to produce the survival analysis for each of the six cancer types.

Chapter 5 presents the results obtained from applying the methodology described in chapter 4. The results for each cancer type include a descriptive analysis of the data, survival curves obtained through the application of Kaplan-Meier method, but also hazard ratios for different covariates obtained from Extended Cox models.

Chapter 6 includes a discussion of the main findings, analysing them for each cancer type and from the different models. This chapter also includes a view of what implications shall be taken in consideration for Public Health Policy, based on the results obtained.

Finally, chapter 7 presents the conclusions from this Dissertation. This chapter also includes suggestions for future research related with Reference Centres and oncology, to enhance the benefits for patients.

2 Context

In this chapter an initial description is made of the RC concept, its origin, as well as the European Reference networks. Afterwards, the experience of the RC implementation in Portugal is described. A particular emphasis is put on the oncology RCs in Portugal, as the Dissertation is focused on the benefits for cancer patients who are treated in these centres.

Then an overview of the six cancer types which will be subject to analysis (hepatobiliary, pancreatic, sarcomas, oesophagus, testicular and onco-ophthalmology) is presented.

Finally, and taking into account the fact that data for these cancer patients was obtained from the Portuguese Diagnosis-Related Groups (DRG) database, a presentation of the DRG's is made, as well as an overview of its implementation in Portugal.

2.1 Reference Centres

In 9th March of 2011, a European Directive was approved by the European Parliament and the European Council (Directive 2011/24/EU), with a focus on empowering patients' rights in cross-border healthcare across Europe. This Directive mentions that the European Commission must support the development of an European Reference Network (ERN), composed by RCs, in order to promote the concentration of resources and expertise, to provide highly-differentiated medical care for specific health conditions [3]. Another objective for the ERN is to foster the cooperation between Member States, enhancing accessibility to more effective treatments [3].

Reference Centres (RCs), also referred to as Centres of Excellence or Highly Specialized Centres, are specialized centres of healthcare delivery, with an increased concentration and specialization of services for specific clinical conditions [3][9]. These centres have also as priority the continuous improvement of clinical expertise, as well as progressing research for new diagnostics and treatments [3]. According to the existing literature, a RC shall include "*healthcare practitioners, with certified technical knowledge to diagnose and provide high quality healthcare practices to patients with certain clinical conditions that require special concentration of resources and expertise due to the low frequency or complexity or high costs of those same conditions*" [3, p. 58].

In a paradigm of exponential technological evolution and increasing accountability, healthcare institutions must focus on delivering first-class medical care to their patients, striving for achieving better outcomes, while managing existing financial constraints. Therefore, there is an incentive for specialization, through the official recognition as a RC [9].

There are different areas of specialization through which RCs have been implemented, such as cardiology, oncology, ophthalmology or neurology [9].

Following the publication of the European Directive on Patient's Rights, its transposition to national law by 2013 in each EU Member state triggered the creation of 24 ERNs in 2017, supported by the EU Health Programme [10]. The main objective for the ERNs is to foster cooperation between different European healthcare units, to provide the best medical care for complex and/or rare diseases, which require highly specialised treatment, as well as a significant concentration of knowledge and resources [10]. Each European Member state is responsible for recognising RCs at a national level to be proposed to integrate a ERN; the approval of that centre to the ERN is made by a Board of European Member States [10]. The ERNs can be seen as a way to foster European states cooperation, promoting economies of scale, more efficient usage of resources, to enhance the search for better medical care for patients [10]. Currently, among the 24 existing ERNs there is one related with adult cancer, the ERN on adult cancers (solid tumours) or ERN EURACAN [10]. This ERN currently integrates three Portuguese RCs for cancer treatment - *Centro Hospitalar do Porto, Centro Hospitalar e Universitário de Coimbra* and *Instituto Português de Oncologia de Lisboa Francisco Gentil*. The ERN EURACAN has the objective of reaching all EU countries until 2022, to develop a referral system which ensures that up to 75% cancer patients are treated in an EURACAN centre [10].

2.1.1 Reference Centres in Portugal

Following the publication of the European Directive in 2011, the Portuguese Government published, in 2013, an Order of the Deputy Secretary of State of the Ministry of Health (nº 4319/2013), in the Portuguese Official Journal (series II, nº 59, of 25th of March), describing the objective to push forward the definition of Reference Centre (RC), as well as the criteria for the Ministry of Health to recognize and characterize them. This Order also stated the objective of establishing a model for RC implementation and funding [3]. The Portuguese Government was also looking for a plan to integrate the RCs within the Portuguese Hospital Network, as well as the European Reference Network [3]. To achieve these objectives, the published Order gave origin to a Work Group to study the implementation of the RCs in Portugal.

This Work Group published a report called "Final Report on Reference Centres", which defines a RC "*as a unit providing healthcare, with verified technical knowledge on the administration of high quality health care to patients in certain clinical situations, which require resources on a large scale, as well as knowledge and expertise, due to the low prevalence rate of a condition, and how complex the diagnostic or treatment procedures are and the high costs of these same situations*" [3, p. 5]. Thus, one of the objectives of the creation of RCs in Portugal, as in the European context, is to provide a centralization and specialization of healthcare treatment for specific conditions, in order to improve the expertise and clinical outcomes [3].

For the establishment of a RC in Portugal, the following components are considered essential to be considered for evaluation [3]:

- Experienced and highly qualified teams, with different competences incorporated;
- Highly specialized structures and equipment;
- Healthcare delivery with the highest standards of quality;
- Capability to promote training, education and research.

With the implementation of the RC Model, Portugal aimed to achieve benefits such as economies of scale, best practices application, efficiency maximization, clinical innovation, as well as cost-effectiveness for the different treatments [3].

Regarding the areas for RC creation, the Work Group has followed the European approach to focus on highly complex and high-cost areas of health care [3]. The Work group defined the following priority areas to be initially considered for RC creation: transplantation of solid organs, oncology, inherited metabolic diseases, haemophilia and haemodynamic and intervention cardiology [3].

2.1.2 Criteria for identification of Reference Centres in Portugal

The Work Group created by the Order of the Ministry of Health of Portugal defined a set of necessary conditions and criteria which health care providers must fulfil in order to be considered for identification as RC [3]. These conditions include factors such as clear accountability for the patients, information systems, quality control or the potential for research and training [3]. In terms of criteria, the Work Group describes general and specific criteria which healthcare providers must satisfy to be considered as RC. Regarding general criteria, healthcare providers must meet cumulatively criteria grouped over the following topics [3]:

- Mutual accountability and patient-focused healthcare;
- Quality, safety and good practices;
- Organization and Management;
- Capacity for research and training;
- Information systems.

Regarding specific criteria, healthcare providers must document and show evidence of satisfying the following conditions [3]:

- Have competence, expertise and recorded practices;
- Show casuistic (minimum and ideal numbers, ratio of patients per year);
- Have good clinical results, according to the available evidence;
- Type, number, qualification and competencies of its human resources;

- Meet organizational and functional prerequisites;
- Have access to specific equipment, inside or outside the RC, as well as e-health tools;
- Describe evidence of multidisciplinary approaches for clinical conditions.

There are healthcare providers which may not fulfil all those demanding conditions and criteria required to be officially recognised as a RC. Nevertheless, based on the knowledge and expertise in providing clinical care for a specific condition, these units can be recognized by the Ministry of Health of Portugal as an Affiliated Centre. The Affiliated Centres can then be related and interact with a RC, on the same area of healthcare specialization, promoting cooperation between different health care providers in the system [3].

2.1.3 Oncology Reference Centres in Portugal

The diagnosis and treatment of cancer patients requires a specialized and multi-disciplinary approach [3][11]. Cancer requires healthcare providers to have specialized and multi-disciplinary teams, with access to high technological equipment, both for diagnosis and treatment, such as radiotherapy or chemotherapy.

In a report published in 2014 by the European Partnership for Action Against Cancer (EPAAC), the variations of service delivery for cancer patients across Europe, based on each country national cancer plan, account for a quarter of survival differences [11]. Thus, it is important to promote specialization and harmonization of centres across the different healthcare systems to ensure that patients receive the best therapies available to increase survival. Taking into account the complexity of treatments for the different cancer types, as well as the high costs associated, the concentration of resources is crucial to achieve a specialization and cost-effective approach for treating these patients [3].

In 2014, the Work Group for RCs implementation in Portugal identified priority areas of intervention which should be the focus for the RC implementation. Oncology has been one of the areas defined as a priority for RC implementation [3]. One of the reasons for this priority is the fact that cancer is the main cause of death before 70 and the second cause of death, among all ages, in Portugal [3]. By the time the Work Group produced the final report, in 2014, there were around 45.000 new cases of cancer every year in Portugal, with around 24.000 deaths every year [3].

As such, Portugal has established RCs for adult oncology, as well as for paediatric oncology, in order to concentrate resources and create multi-disciplinary specialized teams. The current number of oncology RCs in Portugal, per cancer type, can be seen in Table 1.

Table 1 - Distribution of oncology RCs in Portugal

Cancer Type	Number of RCs in Portugal
Adult oncology – Oesophagus Cancer	6
Adult oncology – Rectum Cancer	21
Adult oncology – Testicular Cancer	4
Adult oncology – Hepatobiliary/Pancreatic Cancer	10
Adult oncology – Sarcomas Cancer	5
Paediatric oncology	3
Onco-ophthalmology (Retinoblastoma and Ocular Melanoma)	1

The RC creation in Portugal aggregated the cases of hepatobiliary and pancreatic cancer patients in the same RCs, taking into account the similarity between organs, as described by Penedo et al. [3].

2.2 Cancer

Cancer affected around 18.1 million people all over the world in 2018, being responsible for 9.6 million deaths in that year [12]. According to the World Health Organization, those numbers are expected to nearly double by 2040, as shown in Figure 1 [12]. These increases of new cases and deaths can be related with the increase of life expectancy and epidemiological and demographic transitions [13]. In Portugal, cancer is the second leading cause of death and the fastest growing, over the last years [4].

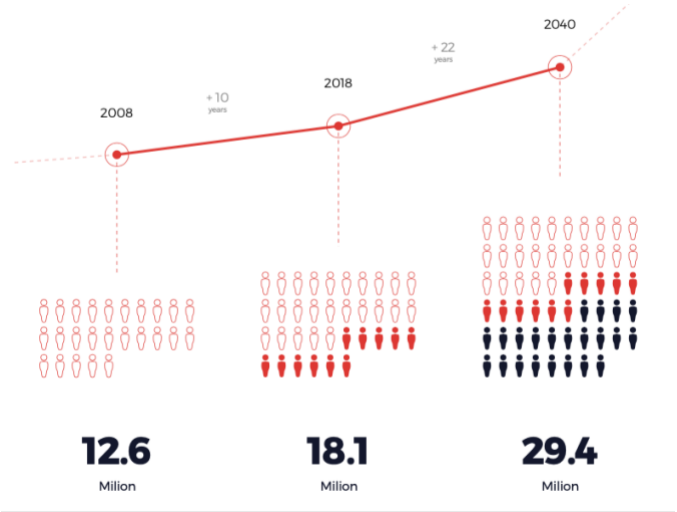

Figure 1 – Estimated evolution of global cancer cases, from 2008 to 2040 (projected) – taken from [12]

In particular for Portugal, as in the rest of Europe, there has been a steady rise of cancer incidence of about 3% per year, as can be seen in Figure 2 [4].

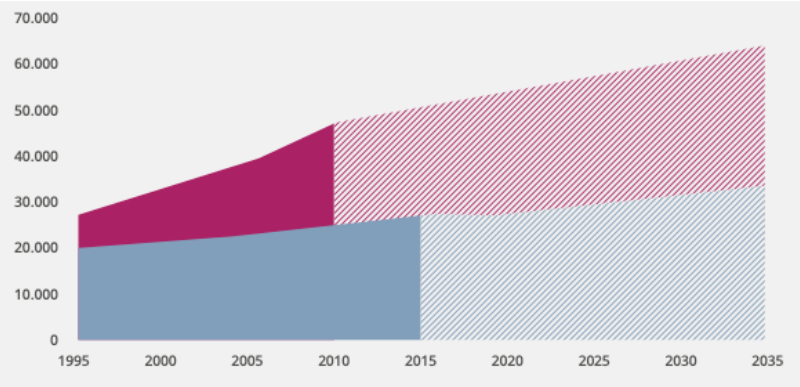

Figure 2 – Projected cancer incidence (rose) and mortality (blue) for Portugal – adapted from [4]

The projection from DGS for the evolution of incidence and mortality in Portugal also shows a steady increase for the next years, up until 2035, in line with the tendency presented by WHO. The incidence is however growing more than mortality, which shows that there is a growing success on the treatment side [4].

In the next sub-sections, the six cancer types which will be analysed in this Dissertation (hepatobiliary, pancreatic, sarcomas, oesophageal, onco-ophthalmology and testicular) are summarily described.

2.2.1 Hepatobiliary Cancer

Hepatobiliary cancer refers to a group of malignancies which affect the liver, either hepatocellular carcinoma (HCC), or in the intra and extra-hepatic biliary ductal system, which present as biliary tract carcinomas (BTC) [14]. Liver disease is considered to be responsible for approximately two million deaths per year worldwide, but is also responsible for a major source of morbidity and mortality [15]. Liver disease can be decomposed into complications from cirrhosis (close to one million deaths per year in the world), viral hepatitis and HCC [15].

Rocha et al. [16] analysed the mortality associated with hepatobiliary disease in Portugal, between 2006 and 2012, finding that, for this period, there were a total of 12279 deaths (24.5 per 100000), making hepatobiliary disease the 8th leading cause of death. The main causes of death for hepatobiliary disease were alcoholic liver disease (7.1 per 100000), unspecified cirrhosis of liver (5.5 per 100000) and hepatocellular carcinoma (4.3 per 100000) [16]. Rocha et al. [16] also found, for the same period, a 66% mortality increase associated with HCC. There are other types of malignant neoplasms of liver and intrahepatic bile ducts, such as unspecified malignant neoplasm of liver or cholangiocarcinoma.

2.2.2 Pancreatic Cancer

Pancreatic cancer is one of the most deadly cancers all over the world, being ranked as the fourth leading cause of death for cancer patients in western societies [17][18]. Costa et al. [19] analysed the annual mortality associated with pancreatic cancer in Portugal, having found that it doubled from 701 deaths in 1991 to 1415 deaths in 2015. These authors [19] expected pancreatic cancer to have an increase of around 50% in the number of deaths, over the next two decades. The prognosis for pancreatic cancer is not positive, as the 5-year survival rate accounts to a bare 7% [18]. One of the reasons for the prognosis to be so poor is related with the often late diagnosis, usually already in presence of metastatic cancer [17][20]. This can be due to lack of effective early screening methods, which leads to mainly palliative treatment when pancreatic cancer is found [17][20]. Some of the patients (approximately one in five) who are diagnosed with pancreatic cancer may be eligible for surgical resection, depending on their cancer stage [17]. These patients have a much better survival prognosis than those who do not undergo surgical resection [17].

2.2.3 Sarcomas

The sarcoma of bone and soft issue is a type of rare (1% of all tumours) heterogeneous group of malignant tumours deriving from mesenchymal progenitor cells, often differentiating to different mesenchymal cells (such as fibrous tissue, adipose tissue or muscle) [21][22][23]. The sarcomas are a type of tumours with a great histological variety (more than seventy different histological types), which can occur in any part of the body [22][23]. There are two main types of sarcomas: the osteosarcoma, which develops from bone, and the soft tissue sarcoma, which affects soft tissue, such as muscle or adipose tissue (STS) [23]. The sarcoma is a specific type of infrequent cancer that, due to its rare aspect, is often not correctly diagnosed at an initial stage [24]. The incorrect diagnosis at an initial stage can also be influenced by the fact that sarcomas can display a wide difference in terms of clinical presentations [24]. Taking into account the complexity of the sarcomas, according to practice guidelines and literature, patients with sarcomas shall be followed by multidisciplinary teams, including both pathologists and surgeons [24].

2.2.4 Oesophageal Cancer

The oesophageal cancer is a type of cancer which occurs in the oesophagus, a muscular tube organ, part of the digestive tract, which extends from the sixth cervical vertebra to the eleventh thoracic vertebra [25]. The oesophageal cancer usually affects more men than women, mostly because this type of cancer can be associated with tobacco and alcohol use, as well as body weight excess, factors which can be more common in men than women [26].

Oesophageal cancer is considered to be one of the main leading causes of cancer-death recorded in the whole world, being responsible for more than 400000 deaths every year [27]. There has been an increase of the overall survival for oesophageal cancer patients over the last years – in Europe overall survival increased from 10% in 1999-2001 to 13% in 2005-2007 [28][29]. Nevertheless, as an example, the 5-year relative survival rate was around 17% in 2014 in the United States, which is particularly poor [28].

2.2.5 Onco-ophthalmology

There are several malignancies which can develop in the eye and which can spread to other parts of the body. The retinoblastoma and the ocular melanoma, or the primary intraocular lymphoma are among the main ones.

The retinoblastoma is the most common eye neoplasm affecting children, accounting for 2-4% of all childhood cancers [30]. This cancer has particular incidence among the younger children: 66% affect children younger than 2 years and 95% occur in children younger than 5 years, being uncommon for patients older than 10 years [31][30]. As such, it is particularly important to effectively manage its treatment and cure, preserving the vision of the patients and avoiding, as much as possible, any side-effects on the long term [31]. The successful treatment of retinoblastoma highly depends on early detection, while the disease is still intra-ocular, to contribute to ocular preservation [31]. According to the literature, up to 90% of these patients survive this disease [31]. Patients in which the disease progresses to extra-ocular body parts have worse outcome, often associated with metastases [31]. As such, early diagnosis is of prime importance to enhance the probability of a cure [31].

The melanoma is a type of tumour which can affect different anatomic sites, such as skin, mucous membrane or ocular region [32]. As for the ocular melanomas, it affects the eye of the patient, in particular the uvea, being the uveal melanoma the most frequent (more than 80%). Uveal melanoma is the most common primary intraocular tumour in adults [32][33]. The incidence for uveal melanoma is around 1.3 to 8.6 cases per million, in Europe [32]. Unlike retinoblastoma, uveal melanoma affects mainly older people, with the peak incidence around 70 years [32]. The uvea is the vascular middle layer of the eye, composed by the iris, ciliary body and choroid. The most common uveal melanomas are located in the choroid (around 90%) [32]. The most common first-line treatment options for uveal melanoma are the resection, radiation therapy and enucleation [32]. As in the case of retinoblastoma, long-term survival outcomes for uveal melanoma are associated with early diagnosis and treatment [32]. The early diagnosis for uveal melanoma is particularly important because, according to the literature, more than half of these tumours metastasize, contributing for a median overall survival of less than six months [33].

2.2.6 Testicular Cancer

Testicular cancer is a rare form of cancer, corresponding to approximately 1% of solid malignancy tumours [34]. Nevertheless, it is the most common cancer diagnosed in males between the age of 15 and 35 [34][35].

Testicular tumours can be classified based on their cellular origin: either Germ Cell Tumours (GCTs), which represent approximately 95% of testicular tumour cases and include seminomas and nonseminomas; or other testicular tumours, which account for the remaining 5%, corresponding to the sex cord stromal tumours, which include the Leydig cell tumours (LCTs) and Sertoli cell tumours (SCTs) [36][34][37]. Testicular cancer treatment has improved significantly with the approval, in 1970, of cisplatin-based chemotherapy for advanced disease [37]. Cisplatin-based chemotherapy has shown to have an effect on the long-term survival rates of testicular cancer patients [38]. Overall survival has also been improving over the last decades, since 1970, both for seminomas and nonseminomas (although more evident for nonseminomas) [37].

2.3 Diagnosis-Related Groups

The Diagnosis-Related Groups (DRGs) consist in a way of aggregating a large number of different (individual) patients into manageable, clinically meaningful and economically homogeneous groups [39]. The DRGs were firstly presented in 1970 by researchers at Yale University, led by Robert B. Fetter and John Thompson, as a way of defining hospital products and the cost for the hospital output production, but also as a way for comparison and benchmarking [40]. Later, the United States adopted DRGs with the purpose of monitoring, but also as a platform for reimbursement [40].

Ever since, DRGs have been adopted all over the world, particularly in industrialized countries [40]. The main objective for the adoption of DRGs in Europe was to make health services delivery more transparent and efficient [40]. This transparency is enhanced by making hospitals group large numbers of heterogenous patients into clinically related and economically homogenous groups of patients [40]. This also allows to compare the treatment provided to the patients, allowing to assess the quality of healthcare delivery [40]. It also allows to evaluate the efficiency of different hospitals in providing treatment to similar groups of patients, comparing resource usage [40]. This efficiency dimension of the DRGs works as an incentive for hospitals to improve efficiency, pushing them to focus on the resource usage per patient, as well as on treating more patients [40].

2.3.1 Diagnosis-Related Groups in Portugal

Overall, the adoption of DRG systems in Europe occurred in the 1990s, but Portugal stands out as front-runner by starting the adoption in the early 1980s [40][39]. The original motivations for DRG introduction in Portugal were to improve resource allocation and increase transparency [39].

Many European countries have adopted DRG systems and adapted those to their country-specific characteristics. Portugal, Ireland and Spain have originally imported their DRG system either from the All Patient (AP) DRG, originally from the United States (in the case of Portugal and Spain), or from Australia (Australian Refined (AR)-DRGs). The Nordic countries, such as Finland, Sweden and Estonia, have collaborated and developed their own DRG, the NordDRG system [39].

By 2006, a non-modified version of All Patient DRG (AP-DRG) was implemented in Portugal and was applicable to all patients (inpatient and ambulatory) treated in public hospitals [39]. This system has 669 DRGs, grouped into 25 Major Diagnostic Categories (MDCs), each of them related with a specific organ or physiological system [39]. There is also a Pre-MDC which is used for coding specific high-cost cases, such as transplantation [39].

The Portuguese DRG System is managed by ACSS ("*Administração Central do Sistema de Saúde*"), a Portuguese health institution within the Portuguese Ministry of Health.

2.3.2 Diagnosis-Related Groups – Portuguese Database

The DRG, as a Patient Classification System (PCS), has at its core the routine data collection on patient discharge, to classify patients into manageable groups, clinically meaningful and economically homogeneous [39]. This has made available a database, the Portuguese DRG Database, managed by ACSS, which includes the information for patients treated in public health care providers in Portugal, grouped by DRGs.

The classification of patients in their DRGs in Portugal is made by physicians, within hospitals, who have specific training on codification [39]. Their activity is essential, since the classification and invoicing, as well as registry for data analysis and statistics, vastly depends on this work. Taking into account the importance of coding, the voluntary physicians receive specific training [39].

The coding of patients into AP-DRGs is made based on principal diagnosis, secondary diagnosis, procedure, age, sex and discharge status [39]. The coding process has three phases: the first one, when the patient receives the medical care, which involves the registry of the patient characteristics, the diagnosis and medical procedures [41]. The second phase includes the codification of the patient into a specific DRG – this is made by trained physicians in a unit dedicated to this task – *Gabinete de Codificação* [41]. The last phase is the audit process: it is made by other physicians who validate the collected data and conclude the process [41].

Every month, the information from all the Portuguese national health system is gathered to be added to the Portuguese DRG database, which is hosted and managed by ACSS.

2.3.2.1 Diagnosis and Procedures coding

The diagnosis and procedures coding is essential to get an overview of the patient condition and clinical care received, under the DRG in which the patient is included [39]. In Portugal, the international standard from WHO – the International Classification of Diseases (ICD) was adapted to be used. From 1989 until 1st of October 2016 all episodes were coded according to ICD-9-CM ("International Classification of Diseases, 9th Revision, Clinical Modification") which was adapted from the US ICD-9 original codification; after the 1st of October 2016, the diagnosis and procedures started to be coded into ICD-10-PCM ("International Classification of Diseases, 10th Revision, Clinical Modification/Procedures).

Currently, the DRG grouping system used in Portugal is the All-Patient Refined DRG (APR DRG) [42]. This grouping system was enhanced by the creation of two new variables for each DRG episode: severity, which represents the degree of acuteness for the patient; and mortality, which represents the probability for death to occur for that patient. These two variables have four different levels, from 1 to 4, which represent the increasing severity and risk of death, corresponding to minor, moderate, major and extreme. These subclasses are calculated taking into account the secondary diagnosis, their relationship, as well as the principal diagnosis the age and sex of the patient, and all the procedures to which he/she was subject.

3 Literature review

This Dissertation aims to develop an analysis of the benefits, in terms of survival, for patients who have cancer in Portugal, taking into account the implementation of the oncology RCs. As such, in a first stage, an overview of the literature regarding RCs is presented, with a particular focus on oncology RCs. Afterwards, a description of the main survival analysis methods is made, complemented with a comparison between them. The technological tools available to produce these analyses are also presented. Finally, a bibliographic review of survival analysis research in cancer is presented, in particular for the six types of cancer which are studied in the current Dissertation.

3.1 Reference Centres

Reference Centres (RCs) are also referred to in the literature as Centres of Excellence or Highly Specialized Centres [3][9]. The creation of these centres has been incentivized by the demand from patients to have better health care services available, supported by highly-specialized personnel and technology, while taking into account existing financial constraints [9]. This is driving governments and institutions to focus on differentiation and specialization to achieve benefits both for healthcare institutions and for patients [9]. Elrod et al. [9, p. 16] define a Centre of Excellence as "*a program within a healthcare institution which is assembled to supply an exceptionally high concentration of expertise and related resources centred on a particular area of medicine, delivering associated care in a comprehensive, interdisciplinary fashion to afford the best patient outcomes possible*". The areas which have been taken into account for RC creation include cardiology, oncology, ophthalmology or neurology.

The next sub-section presents a review of the literature regarding oncology RCs creation, as well as other aspects related with RCs such as the effect of the volume of surgery in patients' outcomes, or the importance of controlling hospital infections, especially in patients often immunosuppressed (as are usually cancer patients).

3.1.1 Oncology Reference Centres

In its 2020's Report on Cancer [12], the WHO refers that "*centralized cancer centres can provide leadership in care networks, concentrate expertise for training professionals and formation of multidisciplinary teams, promote efficient use of technology and evaluate outcomes more systematically*" [12, p. 139]. The WHO also mentions the creation of referral networks, which connect primary, secondary and tertiary units, to provide diagnosis and treatment of cancer patients [12]. WHO statues that "*it is generally accepted that detection and diagnosis are done in primary and secondary health service facilities, expert treatment in tertiary centres and maintenance therapy and*

care for survivors in outpatient settings" [12, p. 138]. Among the benefits included from a centralized cancer centres approach, the availability of a specialized multidisciplinary team, taking advantage from economies of scale, and the promotion of research and training to improve overall practice [12] are mentioned. In this report, the WHO also mentions some risks of this approach, namely the "*super-specialization*" [12, p. 138], which can lead to an unbalanced distribution of the workforce, or the accessibility reduction to this centres of rural areas residents, as they tend to be located in urban areas [12].

There are several articles mentioning the benefits for cancer patients of RC implementation. For example, J. Blay et al. [24] developed a study over a French clinical network for sarcomas (NETSARC), which includes 26 sarcoma RCs, to evaluate the benefits for sarcoma patients. In this study, a total of 29497 sarcoma patients were included to develop a survival analysis study. The results allowed to conclude that in order to increase survival and decrease relapse, sarcoma patients must be treated by a multidisciplinary team with experience in managing sarcoma, since the diagnostic phase up until at least the first surgery [24]. When sarcoma patients are not followed in centres with these characteristics, best clinical practice guidelines are usually not followed, which is associated with an increase in the risk of relapse and death [24]. Hoekstra et al. [21] studied the treatment of patients with soft tissue sarcoma (STS) in the Netherlands, diagnosed from 2006 to 2011, concluding that the STS treatment can be improved by the centralization of sarcoma patients treatment on dedicated centres. This centralization can result in an improvement in the adherence to treatment guidelines and overall disease outcome [21].

3.1.1.1 Volume of cancer surgery

Hospital surgical volume and its outcomes have been widely investigated in numerous studies in Europe and USA [43]. In these studies, there are signs that surgery centralization in high-volume surgical centres or high-volume surgeons can deliver better outcomes, in particular for cancer patients [43].

Coupland et al. [43] studied the association between hospital volume, the rate of surgical resection and the survival of oesophageal and gastric cancer patients in England. In this study, the authors analysed the mortality for a total of 62811 patients diagnosed with oesophageal or gastric cancer between 2004 and 2008 [43]. Among all the patients in the sample, 13189 underwent surgical resection (21%) [43]. Regression models were developed to assess all-cause mortality related with hospital volume and resection rate, adjusted for covariates such as sex, age or comorbidities [43]. The authors found that lower mortality was associated with resected patients, who had their surgical resection made in higher-volume hospitals [43]. These results support the centralization of oesophageal and gastric cancer healthcare services in highly-specialized centres in England [43].

Gooiker et al. [44] developed a systematic review of the literature regarding the volume-outcome for surgical treatment of breast cancer patients. According to their research, better outcomes are associated with higher-experience and higher volume healthcare providers [44]. The results of this systematic review allowed to examine the effect of surgeon volume and hospital volume on the outcomes for breast cancer surgery. The conclusions indicate that better outcomes are associated with higher volume surgeons and hospitals [44]. Nevertheless, the authors found a higher impact for higher-volume surgeons, when compared to higher-volume hospitals [44]. According to Gooiker et al. [44], the findings support the assumption that centralization of cancer care has the potential for improving quality of care for cancer patients.

3.1.1.2 Hospital-acquired infections

Hospital-acquired infections, also known as nosocomial infections, are a type of infection which is acquired by a patient admitted to a hospital or to another healthcare facility for treatment for a reason other than the infection itself [45][46]. These type of infections represent a major problem in healthcare delivery and are a significant source of adverse healthcare events [47]. The condition of the patients who undergo treatment is among the main reasons for nosocomial infections.[46] These patients are often immunocompromised and are also often subject to invasive examination and treatments. As such, the hospital itself may present a transmission path for microorganisms between different patients, potentially promoting the occurrence of nosocomial infections [46].

Urinary infections are the most common nosocomial infections, with a significant part of those associated with the usage of indwelling bladder catheter [46]. Although urinary infections are usually less associated with morbidity, they can however lead to bacteraemia and death [46]. Surgical site infections are another type of nosocomial infections, which are associated with surgical procedures that the patient undergoes, and can vary based on the type of operation and patient status [46]. Other type of hospital acquired infection is the nosocomial bacteraemia related with the infection with multi-resistant organisms, such as *Staphylococcus* and *Candida spp* [46]. Although these infections are not so frequent, their fatality rates are particularly high (more than 50%, for some microorganisms) [46]. Finally, other types of nosocomial infections can affect skin and soft tissue. There is gastroenteritis, sinusitis and other enteric infections, as well as endometritis and infections which can affect reproductive organs after child birth [46].

Some studies refer that in the 1980s, 9-12% of cancer patients developed nosocomial infections [45]. Most of these patients have underlying diseases or conditions and are immunosuppressed due to the treatments which they undergo, rendering them especially vulnerable to these type of infections [45]. Furthermore, several of these patients are exposed to risks of suffering an infection, which include the utilization of invasive devices (e.g. mechanical ventilation, central venous catheters or urinary

tract catheters) [47]. As such, it is of utmost importance to evaluate and implement best practices to avoid the risk for cancer patients to suffer hospital-acquired infections, to decrease associated morbidity and mortality rates.

3.2 Survival Analysis

Survival analysis is a longitudinal statistical method used to study the occurrence and timing of a specific event [7]. Originally presented in the 18th century as a methodology to study mortality of individuals, survival analysis has been widely used to study general life events occurrence and timing, such as onset of disease, cognitive decline and dementia or nursing home admissions [7][8]. The event is usually referred as failure, because it is often related with a negative occurrence, such as death or disease progression [8]. The time variable is usually referred as the survival time, as it gives the time which the individual has resisted without experiencing the event [8].

Survival analysis takes into account a specific analytical problem which is censoring [8]. Censoring occurs when there is incomplete information about an individual and the survival time is unknown [8]. There are three reasons which lead to censoring: a) an individual does not experience the event before the study ends; b) an individual is lost to follow-up over the study period; c) an individual is withdrawn from the study, due either to having experienced death, in the case death is not the event, or for another reason (such as adverse reaction the medication) [8]. In these cases data is said to be incomplete at the right-side, related with missing information about the exact survival time of a patient – data is cut-off (censored) at the right side of the time interval of the patient [8]. Although there is no information about the exact survival time for censored observations, the information up to the time of censoring can still be used for survival analysis [8].

In figure 3, one can observe a schematic example of a study including six patients, over a 12-week follow-up period. While patients A and F have experienced the event, patients B,C,D and E are right censored: patients B and D have reached the end of the study without experiencing the event; patients C and E have missing data, from withdraw and lost to the study respectively [8].

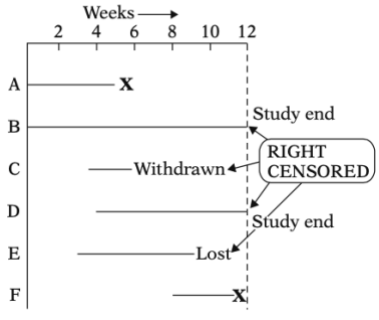

Figure 3 – Schematic example survival analysis for 6 patients – taken from [8]

Every survival analysis can include two important quantitative terms, independent of the chosen methodology: the survival function, $S(t)$, and the hazard function, $h(t)$.

The survival function $S(t)$, shown in Equation 1, is also known as survivor function and gives the probability of an individual to live longer (random variable T) than a specific time t [8].

Equation 1 – Survivor Function

$$S(t) = P(T > t)$$

Figure 4 shows an example of a theoretical survival function. At the beginning, the probability of surviving after time t is equal to one, that is $S(0) = 1$, as no patient has experienced the event yet; after this moment, the survival function $S(t > 0)$ tends to decrease; finally, $S(t \to \infty) = 0$ as eventually all patients will experience the event (e.g. death) [8].

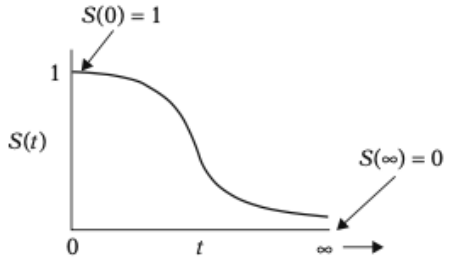

Figure 4 – Example of a theoretical survival function –taken from [8]

The hazard function $h(t)$, presented in Equation 2, gives the instantaneous potential for a patient to experience the event (e.g. death), given the fact that the patient has survived up until t [8]. The higher the value of the hazard, the lower the probability of the patient to survive [8].

Equation 2 – Hazard Function

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

While the survival function is focused on the probability of not experiencing the event, the hazard function is focused on the probability of experiencing the event. Thus these can be seen as the opposite of one another [8]. Given the division per a unit time, the hazard is no longer a probability as the survival function, but rather a rate, ranging from 0 to infinity [8].

In Figure 5, one can observe three examples of hazard functions. Unlike the survival function, which starts always at 1 for $t = 0$, hazard functions can take any value over time. They can assume any positive value, having no upper bound [8].

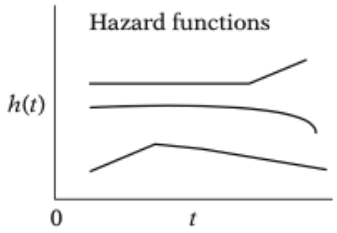

Figure 5 – Example of hazard functions – taken from [8]

3.2.1 Survival Analysis Methods

There are different types of survival analysis methods, which can be selected for application based on the data which will be subject to analysis. Survival analysis can be made through the application of non-parametric, semi-parametric or parametric methods [7].

The oldest non-parametric method for survival analysis is the life-table, presented by Berkson and Gage [48]. This method allows to obtain a table with patients' survival over different sub-intervals [48]. Nowadays, the most frequent non-parametric method for survival analysis is the Kaplan-Meier's (KM), which allows to obtain and compare two or more survival distribution populations [7].

Faced upon the need to study how multiple variables influence the survival of a population, there are semiparametric and parametric methods, to be selected taking into account the information of the data to model.

The application of semi-parametric methods allows data to be free to change over follow-up time, as there is no need to fit data to a specific probability distribution [7]. The most commonly used semi-parametric method for survival analysis is the Cox Proportional Hazards model.

In the case of parametric methods for survival analysis, one needs to know the probability distribution of data to fit the model [7]. An example of a parametric method is the accelerated failure-time model (AFT), which assumes that data can be modelled against a known probability distribution (e.g. Weibull, exponential, gamma) [7].

3.2.1.1 Life Table

The life table method was originally presented by Berkson and Gage [48] and is considered to be the oldest and most straightforward method for cancer survival analysis [49]. This method can be implemented by dividing the range of survival times for all patients into several subintervals. Each of the subintervals is constructed and used to calculate the number and proportion of patients who were alive in that subinterval, the number and proportion of patients who have died in that subinterval, as

well as the number and proportion of patients who were censored in that subinterval [49]. The analysis of these numbers and proportions allows the construction of a survival analysis model [49].

As an example, Schea et al. [50] published a study comparing the survival of limited-stage small-cell lung cancer patients who underwent chemotherapy and radiation therapies with and without a specific treatment protocol. A total of 81 adult patients who underwent chemotherapy and radiation therapy between 1987 and 1992 were selected and reviewed retrospectively [50]. Survival analysis was obtained by using Berkson and Gage Life Table method [50]. The results of this analysis allowed the authors to conclude that patients in the protocol group present an improved survival when compared with the non-protocol group, with a high statistically significant difference between them [50].

3.2.1.2 Kaplan-Meier

The Kaplan-Meier (KM) method was originally presented in 1958, by Edward L. Kaplan and Paul Meier, as a technique to deal with incomplete observations [51]. The KM method allows to obtain estimators of survival data and curves, in particular for cases which have missing data from individuals (censored cases) [51].

The formula for the Kaplan-Meier survival probability at failure time t_j can be written as:

Equation 3 – Kaplan-Meier survival probability function [8]

$$\mathcal{S}(t_j) = \mathcal{S}(t_{j-1}) \times \mathcal{P}(T > t_j \mid T \geq t_j)$$

which can be interpreted as the probability of a patient surviving past t_{j-1} multiplied by the conditional probability of the patient surviving past t_j, knowing that he/she survived to at least t_j [8].

A KM estimator includes survival curves which display, at each time, the information about the patients who are event-free (death-free, when the event of interest is death) [52].

In Figure 6, two examples of KM survival curves are presented, for two different treatment arms of a randomized clinical trial (RCT) – TRT A and TRT B [52]. Each time a patient dies, in each of the treatment groups, the curve drops accordingly, following the change in the KM conditional probability, as described in Equation 3. Additionally, the median survival times are computed using a software solution, which in the case of TRT A is 6 months, while for TRT B it is 11.4 months. On the other hand, the HR presented corresponds to an univariate analysis using the Cox model, which is described in the next sub-chapter but, put in a simplified way, compares the TRT B efficacy in terms of survival to TRT A [52].

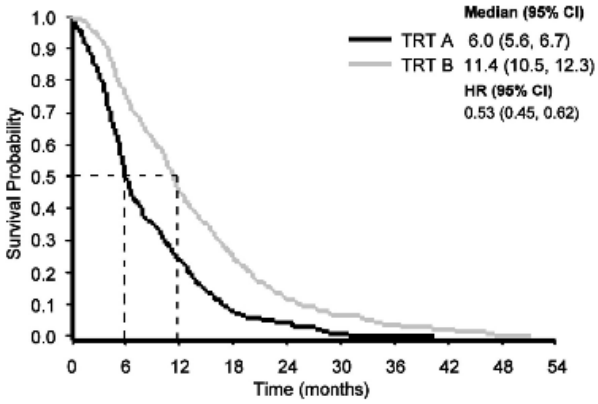

Figure 6 – Example of KM curves for two different treatment groups – TRT A and TRT B – taken from [50]

For testing if two or more KM curves are statistically equivalent, the most popular method is the log-rank test [8]. The aim of this test is to assess whether two survival curves are different or equivalent, considering the null hypothesis of no overall difference between them [8]. For the comparison of two survival curves, the log-rank statistic can be calculated as:

Equation 4 – Log-rank statistic formula [8]

$$Log - rank\ statistic = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}$$

with

$$O_2 - E_2 = summed\ observed\ minus\ expected\ score\ for\ group\ 2$$

When there are more than two survival curves to be compared, the null hypothesis considered is that all the curves are statistically equivalent [8]. The test statistic for this case is more complex, involving the calculation of variances and covariances between the groups [8]. In this case, the test statistic can be obtained through a matrix formula, computed with the technological tools available [8].

3.2.1.3 Cox Proportional Hazards Model

The Cox Proportional Hazards (Cox PH) model is a univariate and multi-variate semiparametric regression model widely adopted in clinical studies to evaluate disease progression, allowing to control for the effect of covariates [49]. The Cox PH model is popular in survival analysis because of its simplicity: it does not assume any statistical curve shape for the survival distribution, but rather the model calculates the hazard rate (instead of the survival time) as a function of the independent variables – the covariates [49]. The Cox PH model is the most used survival analysis methodology in cancer research for comparing survival characteristics of two or more different treatment groups [49].

The function for the Cox PH model can be written as:

Equation 5 – Cox PH formula

$$h(t, X) = h_0(t)\, e^{\sum_{j=1}^{p} \beta_j X_j}$$

with $X = (X_1, X_2, \dots, X_p)$ as the independent covariates to predict an individual's hazard, and β = (β_1, β_2,..., β_p) as the vector of coefficients [8]. The Cox PH model gives an expression for the hazard of an individual at time t, given a set X of covariates [8]. This expression is composed by two quantities: the baseline hazard function at time t, $h_0(t)$, which only depends on time; and the exponential of the linear sum of $\beta_j X_j$, corresponding to the sum of the explanatory variables multiplied by the coefficients.

The baseline hazard, $h_0(t)$, is an unspecified function associated with the fact that Cox PH is a semi-parametric model [8].

The second quantity only depends on the coefficients and explanatory variables values for an individual, being independent of time [8]. As such, these variables are called time-independent variables [8]. One example of a time-independent variable is sex [8]. This property is what constitutes the proportional hazards (PH) assumption: the baseline hazard shall be time-dependent, while the variables shall be time-independent [8].

The coefficients can be obtained through estimates – maximum likelihood (ML) estimates. These coefficients are derived from the maximization of a likelihood function, taking into account the probability of obtaining the data which was actually observed [8].

The hazard ratio (HR) represents the hazard of an individual, divided by the hazard of a different individual. It can be written as [8]:

$$\widehat{HR} = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)}$$

where $X^* = (X_1^*, X_2^*, \dots, X_p^*)$ and $X = (X_1, X_2, \dots, X_p)$ are the p variables for each individual.

The PH assumption requires that the hazard for one individual shall remain proportional and constant to the hazard of other individuals when the hazard has no dependency of time [8][53]. That is, with only time-fixed covariates the relative hazard for any two subjects i and j meets the following relationship described in Equation 7:

Equation 6 – Relative Hazard Ratio for time-independent covariates [53]

$$\frac{h_0(t)\, e^{X_i \beta}}{h_0(t)\, e^{X_j \beta}} = \frac{e^{X_i \beta}}{e^{X_j \beta}}$$

which is also independent of time.

Cox PH model is particularly simple to use for modelling survival based on covariates which are time-independent. There are however cases of models which include variables that can change over the follow-up time. In this case, one should consider the extension of the Cox model which is described in the next section.

There are three general approaches to assess the PH assumption: a graphical approach, the use of a goodness-of-fit test, and the inclusion of time-dependent covariates in an extended Cox PH model [8].

As for the graphical test, the most popular method is the plot of the "log-log" survival curves [8]. This can be done by observing survival curves: if proportional hazards hold, the log survival curves should remain apart over follow-uptime, not intercepting each other [53]. Kaplan-Meier survival curves can be used to check the proportional hazard assumption when there are not many levels: the curves shall be approximately parallel [53].

Another graphical approach is to plot the Schoenfeld partial residuals versus time, which present the difference between the observed and expected values for the β coefficients at any given time [53]. If the proportionality hazard assumption holds, then the points shall be evenly distributed, displaying an approximation over the horizontal axis, which can be seen as an approximate constant impact of one or more covariates on the hazard over time [53] [8]. As described by Therneau, "*the restriction* $\beta(t) = \beta$ *implies proportional hazards; if proportional hazards holds then a plot of* $\beta_j(t)$ *versus time will be a horizontal line."* [53, p. 130]. As pointed out by this author, the plots and tests for coefficients variation over time, such as the Schoenfeld partial residuals versus time, are a "*powerful tool for testing and understanding proportional hazards"* [53, p. 140].

A second approach for evaluating the PH assumption is to use the goodness-of-fit (GOF) statistical tests. These tests compute chi-square statistics for each covariate in the model, deriving p-values [8]. In the case of non-significant p-values for the covariates, the PH assumption is reasonable, whereas significant p-values suggest PH not to be satisfied [8]. Nevertheless, as pointed out by Kleinbaum et al. [8, p. 137], "*a GOF test may be too "global" in that it may not detect specific departures from the PH assumption that may be observed from the other two approaches (graphical and time-dependent covariates)*".

A third approach is the inclusion of time-dependent covariates in the Cox model. However, in this case, as pointed out by Kleinbaum et al. [8, p. 95], "*the Cox model form may still be used, but such a model no longer satisfies the PH assumption, and is called the Extended Cox Model".*

3.2.1.4 Extended Cox Model

The original Cox PH model includes variables which are time-independent (e.g. sex or age at diagnosis). When there is the need to develop a survival analysis for a specific follow-up time, including variables which can change over time, i.e. are not *time-independent*, an extended Cox Model shall be used. The formula for the extended Cox Model can be written as:

Equation 7 – Extended Cox PH Model Formula [8]

$$h(t, X(t)) = h_0(t)\, e^{\sum_{i=1}^{p1} \beta_i X_i + \sum_{j=1}^{p2} \delta_j X_j(t)}$$

where

$$X(t) = (X_1, X_2, \dots, X_{p1}, X_1(t), X_2(t), \dots, X_{p2}(t))$$

includes the time-independent covariates X_{p1} and the time-dependent covariates $X_{p2}(t)$. As with the original Cox PH model, the extended version also includes the baseline hazard function $h_0(t)$ [8].

Regarding time-dependent variables, they can be “internal” or “ancillary” [8]. An internal time-dependent variable changes according to changes of “internal” characteristics or behaviour of the individual [8]. Some examples of internal time-dependent variables are employment status at time t, or smoking status at time t [8]. On the other hand, an “ancillary” time-dependent variable corresponds to a change for an individual which does not directly depend on the individual characteristics or behaviour [8].

One assumption of the Extended Cox Model is that the effect of a time-dependent variable $X_j(t)$ on an individual survival probability for time t only depends on the variable value at time t, not on the value of that variable before or after that time [8].

For the case of time-dependent covariates, the relative hazard for two individuals i and j is given by:

Equation 8 – Relative hazard formula between two individuals

$$\frac{h_0(t)\, e^{X_i(t)\beta}}{h_0(t)\, e^{X_j(t)\beta}} = \frac{e^{X_i(t)\beta}}{e^{X_j(t)\beta}}$$

Nevertheless, the relative impact of a variable, whatever the value, shall still be summarized by a single coefficient β for that variable [53].

In a recent article by Stensrud et al. [54], the authors refer that the assumption that the hazard ratio remains constant from the beginning until the end of the follow-up period does not hold, in practice, for most medical interventions. Instead, the overall hazard ratio shall be seen as a weighted average of the time-varying hazard ratios [54]. As pointed by the Stensrud et al. [54, p. E1], “*the hazards are*

not proportional when the treatment effect changes over time". The authors recommend supplementing hazard ratio reports with absolute effect measures, such as Restricted Mean Survival Time (RMST) difference, to support clinical-decision making and to make it more understandable for patients and doctors [54]. Also, as Therneau [53] refers, faced upon non proportionality for one or more covariates, one should assess if it "matters" and if it is real [53, p. 142]. As Therneau points out, "*a "significant" nonproportionality may make no difference to the interpretation of a data set, particularly for large sample sizes*" [53, p. 142].

3.2.2 Comparison of Survival Analysis Methods

The life table method can be suitable for an homogeneous sample, but does not address a primary goal of cancer research, which is evaluating the impact of a continuous/dichotomous variable on the patients' survival [49]. To account for multiple variables, regression models began to be developed for survival analysis. There were however two limitations for the traditional regression models: survival times are not usually normally distributed and there are missing values for the dependent variable (survival time), related with censoring of patients [49]. These limitations drove research which led to the development of non-parametric methods such as Kaplan-Meier, and semi-parametric methods such as the Cox Models [49].

The KM method has established itself as the most popular methodology for estimating survival curves, due to its non-parametric nature and simplicity of use, made easier with the support from the software tools that have been made available over the last years [49]. Nevertheless, KM only allows to obtain a univariate analysis. To produce a multivariable analysis which addresses the impact of multiple variables on the hazard of a function, one needs semi-parametric or parametric methods.

Although the use of a parametric model is preferable when one knows the correct model, this is not always the case, as one is not always certain about the model which is more suitable [8]. This is one of the reasons why Cox models are so popular: good estimates can be obtained for survival analysis, for a variety of data situations, making it a robust model through the approximate results it generates, even though it is a non-parametric model [8]. As Kleinbaum et al. [8] refer "*when in doubt, as is typically the case, the Cox model will give reliable enough results so that it is a "safe" choice of model, and the user does not need to worry about whether the wrong parametric model is chosen"* [8, p. 97].

Another characteristic which makes the Cox models so popular is that, while the logistic model only considers the outcome (alive or death), the Cox models consider more information, such as survival times and censoring [8]. Nevertheless, the Cox PH Model assumes covariates to be time-independent, limiting the usage of covariates which may change value over time. As such, the Extended Cox Model, which allows the usage of time-dependent covariates, is a viable option for multivariate survival studies with this type of covariates.

3.2.3 Technology solutions to support Survival Analysis

There are several tools for supporting the development, application and testing of survival analysis methods. These tools can be based on software solutions, either commercial or open-source. The decision about which tool to choose can take into account several factors, such as simplicity, training availability, licensing cost, existence of literature and software packages to facilitate its usage as well as community support. Each of the existent solutions has its strengths and weaknesses.

The most common technological solutions which allow the implementation of survival analysis are R, SAS, SPSS and STATA.

R is an open source statistics language and environment for statistical studies and graphics [49]. It is widely used for data analysis and statistics. It is supported by a large community of developers, among the scientific community, which contribute to several open-source packages (15.587 in R repository – CRAN, as of May 2020). R has and integrated development environment (IDE), Rstudio, which is available in different platforms (e.g. Windows, MacOS). R also has a specific add-on package called “survival”, developed by Therneau [55], which has a set of ready-to-use functions to support survival analysis studies [49][55]. On the other hand, R has a specific syntax, which can potentially be a barrier to entry for beginners. There are several survival analysis which have been made using R, as well as many books available describing how to use it [53].

SAS is a statistic software which allows the development of statistical analysis, commonly used in clinical research and in the banking sector. It is presented as a commercial solution, having a license cost associated. However, it includes extensive documentation and professional support. As for SAS application to survival analysis, for example Walsh et al. [56] implemented a Cox proportional hazard model and obtained hazard ratios and confidence intervals using this solution.

SPSS is a professional product which is particularly easy to use, being one of the most adopted solutions for statistical analysis. It is easy to learn, having several training materials available online. However, as SAS, it is a commercial solution which has a licensing cost associated. Blay et al. [24] made a statistical analysis of the impact on relapse and overall survival of sarcoma patients in reference centers using SPSS (version 22.0), plotting survival curves with the Kaplan-Meier method, and univariate and multivariate using the Cox proportional hazards model.

STATA is a statistical software which is also available at a multi-platform level (Windows, MacOS), being mostly used by econometricians. As SAS and SPSS, STATA is also a commercial solution with a licensing cost associated. Lanza et al. [57] have produced a survival analysis of hepatocellular carcinoma patients who have been subject to different therapy regimens. Using STATA, the researchers were able to obtain survival curves using KM methods, and to compare them with log-rank tests. Also, a Cox PH model was used for multivariate analysis [57].

For the purpose of this Dissertation, due to my previous experience with R and the availability of specific survival analysis packages which have extensive literature associated (such as the "survival" and "survminer" R packages [55][58]), R was the chosen solution for the data analysis, model development and implementation.

3.3 Survival Analysis in Cancer

There are several studies in the oncology area which aim to analyse how different variables affect the overall survival of cancer patients. Over the next sub-sections, an overview of survival analysis research articles is presented, particularly related with cancer types, with a specific focus on survival analysis studies using Kaplan-Meier and Cox models, as those are the most common in the literature and will be also used in the survival analysis undertaken in this Dissertation.

3.3.1 Survival Analysis in Hepatobiliary Cancer

Kudo et al. [59] published a study which analysed the survival of Hepatocellular Carcinoma (HCC) patients in Japan. In this study, a total of 173378 patients were followed between 1978 and 2005, with data collected from a nationwide registry of all patients in Japan with hepatocellular carcinoma [59]. These patients were grouped by years of diagnosis, for whom the overall survival and 5-year overall survival were calculated [59]. The method used by the authors to calculate overall survival was the KM's, comparing the curves with the log-rank test [59]. The results allowed to understand that, over 28 years, 5-year survival rates have improved for patients with HCC, associated with appropriate treatment such as resection or ablation, as well as the reduction of palliative treatment [59].

Lanza et al. [57] evaluated locoregional therapies for HCC such as transarterial chemoembolization (TACE) and bland embolization (TAE). The authors performed a retrospective study for patients who have received TAEs between the 1st of January 2011 and 28th February of 2018 to analyse the benefits of TAE therapy [57]. A total of 270 patients who received liver embolizations for unresectable HCC were included in the study. Survival curves were obtained using KM method and the curves were compared using log-rank test. A multivariate model was developed using Cox PH. Although the authors refer that the objective of the study is not to provide "*a direct comparison of TAE vs TACE in the same group*" [57, p. 9], the study also aims to question the need for chemotherapeutic drugs for endovascular treatments to HCC. The results showed that TAE has presented overall survival rates aligned with the highest survival rates existing in the literature [57].

3.3.2 Survival Analysis in Pancreatic Cancer

Forssell et al. [20] have studied the definition of a model for survival time prediction in patients with unresectable pancreatic cancer, to optimize patients' care. A total of 132 patients with unresectable pancreatic cancer were recruited from a single centre in Sweden. These patients had been diagnosed with pancreatic cancer between January 2003 and May 2010 [20]. The authors wanted to evaluate if the presence of liver metastases at initial radiographic imaging affected survival, taking into account the decision to start chemotherapy, in patients with unresectable pancreatic cancer [20]. Overall survival estimates were obtained using the KM method, assessing the difference between groups with the log-rank test [20]. Independent factors influence on overall survival, such as tumour diameter, presence of liver metastases or chemotherapy initiation, were evaluated with the Cox PH [20]. The authors have built three risk groups (low, medium and high), taking into account the liver metastases presence, performance status for chemotherapy initiation and tumour size [20]. With the developed model, the authors believe that it is possible to identify and differentiate patients with short and longer expected survival time [20].

Wahutu et al. [18] developed a prospective survival analysis for a total of 6291 pancreatic cancer patients living in Oklahoma, United States, between 1997 and 2012 [18]. The objective was to study risk factors, such as age at diagnosis, sex or stage at diagnosis, as well as the survival time for pancreatic cancer patients [18]. The authors obtained survival curves using the KM method; the difference between survival curves was tested using the log-rank method; also, hazard ratios were obtained using Cox PH [18]. The results show improvement in survival times for pancreatic cancer patients in Oklahoma [18]. Covariates included in the model, such as age, stage of diagnosis, as well as behavioural factors diabetes and smoking status, showed statistical significance [18].

3.3.3 Survival Analysis in Oesophageal Cancer

Tustumi et al. [60] analysed patients diagnosed with oesophageal cancer in Brazil, from 2009 to 2012, in an oncology RC, with the objective of investigating how the demographic, clinical and pathological factors affect overall survival and prognostic. Patients were grouped based on the histological diagnosis – either squamous cell carcinoma or adenocarcinoma. The overall survival percentage of patients after five-years of follow-up is close to 20%, which shows the weak prognosis for oesophageal cancer [60].

Cao et all. [27] aimed to construct a clinical nomogram for analysing and predicting the survival of oesophageal cancer patients after esophagectomy. A total of 4281 patients who went through esophagectomy from 1988 and 2007 were analysed. Data was collected from the National Cancer Institute, as well as from the Surveillance, Epidemiology and End Results (SEER) [27]. A univariate analysis was produced using the Cox PH model, allowing to identify statistically significant covariates.

KM curves were also plotted for the different risk groups. Multivariate analysis was developed using the Cox PH model. This analysis allowed to identify age, race, histology, tumour site, tumour size, grade and depth of invasion, as well as the number of metastases and retrieved nodes, as prognostic factors [27].

Kim et al. [28] aimed to find the conditional survival of oesophageal cancer patients' in the United States, to test if it improves over time, for more advanced stages, as it happens with other gastrointestinal-related cancers. The SEER database was used, collecting information from 63433 patients diagnosed with oesophageal cancer between 1973 and 2011. The KM method was used to obtain survival curves. Cox PH was implemented for each of the covariates, being tested with the log-rank test to identify statistically significant prognostic variables. The variables which were found to be statistically significant in the univariate analysis were considered for inclusion in the Cox PH multivariate model [28]. The results allowed the authors to conclude that conditional and cause-specific survival can support the hypothesis of changing prognosis for patients over time [28].

3.3.4 Survival Analysis in Sarcomas

Blay et al. [24] have studied the impact in terms of relapse and overall survival for sarcoma patients who undergo surgery in a RC. In particular, for the case of soft tissue sarcoma (STS), surgery is the mainstay of curative treatment, being the initial surgery a prognostic variable for recurrence-free survival and overall survival [24]. Across Europe, sarcoma patients can be treated in primary oncology clinics (as in the case of France), while in other countries sarcoma patients must be treated in dedicated RCs [24]. As such, Blay et al. [24] proceeded to study the impact on relapse and overall survival of patients treated in RCs, when compared to other patients. The survival analysis study included a total of 29497 patients with sarcoma, with an initial diagnosis from 1st January of 2010 to 1st of May 2018. The authors obtained survival curves using the KM method. Those curves were compared using the log-rank test. Univariate Cox PH was implemented, considering different classical prognostic factors, such as age, gender, grade, size, or histotypes, and also pre-existing conditions such as previous cancer or previous radiotherapy. The covariates which were found to be statistically significant in univariate analysis ($p < 0.05$) were included for multivariate analysis, using a Cox PH model [24]. The results of this study allowed to conclude that when surgery is made in a RC the survival of patients improves [24]. When patients have surgery in a non-RC, re-surgery is two and a half times more frequent than in patients who had the initial surgery in a RC [24].

Bagaria et al. [61] published a study aiming to analyse the impact of hospital surgical volume and adherence to best practice treatment guidelines on outcomes for soft tissue sarcoma (STS) patients. Taking into account a total of 13864 patients, using the KM method, survival curves were obtained for different stratifications: hospital surgical volume (low, medium and high), as well as compliance or not with best practice treatment guidelines [61]. The survival curves were assessed using the log-

rank test. A multivariate analysis was developed using the Cox PH model to obtain overall survival, adjusted hazard ratios (HRs) and 95% confidence intervals. The results of this analysis allowed the authors to conclude that higher volume hospitals usually result in higher adhesion to guidelines [61]. Nevertheless, for the case of lower-volume, those who follow guidelines seem to achieve comparable survival, which the authors believe to point towards a focus on best-practices guidelines, rather than a focus on volume [61].

3.3.5 Survival Analysis in Onco-ophthalmology

Sant et al. [30] developed a survival analysis for 954 patients diagnosed with retinoblastoma between 1978 and 1989, recruited from 17 European countries. Taking into account the very low incidence of retinoblastoma and the need of a sample with a sufficient number of individuals, patients were recruited among different European countries through the EUROCARE framework, which includes survival and care information for cancer patients in Europe [30]. With this data, Sant et al. [30] obtained survival curves using the KM method. Multivariate analysis using the Cox PH model was developed to estimate the impact of different prognostic factors such as age at diagnosis, sex and period of diagnosis. The results showed that, although there was some inter-country variance among countries in Europe, overall survival exceeded 90% in most of them [30]. The results of the study also showed an increase in survival, which the authors believe to be related with an increased efficacy of therapies during the period under analysis [31].

Rantala et al. [33] developed a systematic review and meta-analysis for the overall survival of patients with metastatic uveal melanoma. The authors extracted individual-level survival data by pooling peer-reviewed articles, from 1st of January 1980 to 29th March 2017 [33]. From a total of 78 articles, individual-level data was obtained for 2494 patients [33]. Median overall survival time after metastatic uveal melanoma was 1.07 years, ranging from 0.84 to 1.34 years [33]. The results of Rantala et al. [33] refer that there is no evidence of longer median overall survival time for patients with metastatic uveal melanoma related with any treatment modality.

3.3.6 Survival Analysis in Testicular Cancer

Verhoeven et al. [37] studied the survival of patients with testicular cancer (TC) in Europe and USA. The data from patients in Europe was gathered from the European Network for Indicators on Cancer (EUNICE) Survival Cooperation database, containing information on incidence and follow-up of TC patients from 12 European cancer registries [37]. The data from the TC patients in the USA was gathered from the SEER database. All the patients diagnosed from 1988 to 2007 with TC, aged 15-84 years, both from EUNICE and SEER database, were collected for analysis [37]. The data included information about age, sex, race (in the case of USA) and calendar period-specific life tables [37]. Then, 5-year period based relative survival estimates were calculated for different periods, using the

saturated Poisson regression model for relative survival [37]. This analysis allowed to conclude that for European and American TC patients the 5-year relative survival was high, mainly for patients aged less than 54 years (higher than 96% survival 5-year relative survival). For patients older than 54 years, the 5-year relative survival seemed to decrease with age [37]. Also, the 5-year relative survival for seminoma was higher than for nonseminoma patients, which the literature suggests to be attributed to the greater propensity of nonseminomas to metastasize [37].

Walsh et al. [56] analysed the possibility for higher risk of TC in men seeking infertility treatment, when compared with the general population. To produce this study, 22562 male partners of couples seeking infertility treatment from 1967 to 1998 in California infertility centres were recruited [56]. The data on these patients was collected from California Cancer Registry. Data from an age-matched sample of men was collected from SEER database [56]. To analyse the risk for TC, a Cox PH model was developed, including men with and without infertility treatment, controlling for age, duration of treatment and site of treatment [56]. The results of the application of this model pointed to an increased risk of testicular germ cell cancer in patients in fertility treatment, when compared to the general California population [56]. In particular, following the Cox PH model results, men with factor infertility had 2.8 times the risk of testicular cancer of men without factor infertility (HR 2.8, CI 1.3-6.0) [56].

Hartmann et al. [38] evaluated the prognostic variables associated with patients with seminomatous and non-seminomatous extragonadal germ-cell tumours (EGCT), to identify factors which affected overall long-term outcome of cisplatin-based chemotherapy. The medical records of 635 patients with EGCT were reviewed, from the United States and Europe, which were followed from 1976 [38]. Different patient characteristics, such as patient age group, location of the primary tumour, or presence of additional metastasis, were studied for potential prognostic factor review. The study of these patient characteristics was made applying an univariate analysis; survival time was estimated using the KM method; log-rank test was used for comparing the curves [38]. Following these methods, a Cox PH model was built in a stepwise forward selection fashion, choosing significant factors ($p < 0.05$), to evaluate the effect of different covariates [38]. The authors refer the difficulties that arise from the small number of patients in this kind of analysis [38], although prognostic variables for the outcome and response to chemotherapy could be identified [38]. Nevertheless, this analysis showed the heterogenous profile of prognostic factors for EGCT patients [38].

4 Materials and Methodology

This chapter describes the materials used in this Dissertation, i.e. the raw DRG datasets with information about patients from a set of cancer types in Portugal, between 2010 and 2019. It also describes the steps followed for cleansing and preparing the datasets to be analysed. In a second section, the methodology used for describing and summarizing the datasets, as well as the methods for developing the survival analysis for each cancer type are also presented.

4.1 Data

This section describes the different data sets for each cancer type used in the analysis. The steps for data cleansing and preparation for analysis, in order to have the data ready for the methodology application, are also described.

4.1.1 Raw Data

An initial meeting was held in the ACSS headquarters to discuss the objectives of the Dissertation, namely to analyse the impact in terms of survival for cancer patients in Portugal, following the RC model implementation. After this meeting, a formal data request was made for existing data in the Portuguese DRG Database, in particular for the following set of cancer types which have seen the creation of RCs in Portugal:

- Adult oncology – Hepatobiliary cancer;
- Adult oncology – Pancreatic cancer;
- Adult oncology – Sarcomas;
- Adult oncology – Oesophagus cancer;
- Onco-ophthalmology;
- Adult oncology – Testicular cancer.

This data request specified the primary diagnosis for these cancer types, using the ICD-9-CM and ICD-10-CM codes, obtained from the "*Circular Normativa nº 8/2018/DPS/ACSS*". This has resulted in a total of 104 distinct diagnosis codes, corresponding to primary tumours/malignant neoplasms for these types of cancers, which were used as filter for sub-setting the data from the Portuguese DRG Database. The requested variables for each episode were the ones which are usually available in the DRG database, such as year of the episode, date, sex, age, total of diagnosis for that episode, or the discharge status. The year interval included in the data request was from 2010 up until the moment of request (November of 2019), to allow to have a relevant number of cases for analysis.

Upon the request, ACSS has made available two different files:

- Excel file including the DRG cancer types (for the requested primary diagnosis codes) from 2010 until 2019 (25th of November 2019, which corresponds to the cut-off date from the moment the datasets were made available). The excel sheets included data for each cancer type described before, where each row corresponded to a hospital episode for a cancer patient. Each excel sheet had 147 columns, corresponding to the different variables for each episode (e.g. patient's age, first diagnosis, discharge status, etc.).
- Excel file including information about the metadata of the shared data, such as the description of the variables (columns), the codes and corresponding names for hospitals and patients' home residence. This file also included the codes and descriptions for diagnosis, procedures, oncology morphology and external causes, in ICD9-CM format. As the codification of diagnosis and procedures changed in 1st of October of 2016 from ICD-9-CM to ICD10CM/PCS, the codes and descriptions for diagnosis and procedures in this format were also included in this file. Finally, the codes and description of the Major Diagnostic Categories (MDC) and Diagnosis-Related Groups (DRG) were also made available in this file.

Each of the cancer types included several rows, with each row corresponding to a hospital episode, which could be of the ambulatory or hospital stay type. As mentioned before, for each episode there is a total of 147 variables, which can be filled or not, depending on the information available. These datasets include information only for adult patients thus the age is always equal or superior to eighteen years old.

The first step after receiving the datasets was to separate each of the excel sheets into different excel files, as the original file had a particularly large size, which was computer intensive to process. Furthermore, the first sheet, which included the hepatobiliary and pancreatic cancer hospital episodes, was split into two different datasets, one for each cancer type (hepatobiliary cancer and pancreatic cancer). This sub setting was made using the ICD-9-CM and ICD-10-CM codes for primary diagnosis as hepatobiliary and pancreatic cancer.

This procedure resulted in six raw datasets, in distinct files, one for each cancer type, which were then loaded into the R statistical programming language to be cleansed and prepared to be statistically analysed. The steps for cleansing and preparing these raw datasets are described in the next section.

4.1.2 Data Cleansing

After having separated the different datasets into different files, each of the raw datasets was loaded separately into a R programming language script, specifically developed by the author for this Dissertation and for each cancer type. The first part of the script was focused on data cleansing and sub-setting of data, according to the following steps:

- Data subset with the condition *N_ficticio_utente > 0*
    - The variable *N_ficticio_utente* is a unique positive numeric identifier for each of the patients included in the dataset. Based on the column analysis, there were several rows with -1 value, which are episodes which had no unique patient identifier assigned. As such, these rows were removed, as it is important to do a follow-up of each patient through time.
- Data subset with the condition *unique(seq_number)*
    - The variable *seq_number* is a unique identifier for each episode registered in the dataset. If the *seq_number* appeared more than once, it means those episodes were repeated. These cases shall be removed, as otherwise there will be inconsistencies in the data analysis produced.
- Data subset to remove duplicate rows with distinct seq_number
    - The variable *seq_number* is the identifier for the event. However, there were distinct rows with all the variables equal, only differing in terms of seq_number. Those cases are duplicate episodes with different seq_number, therefore they were removed from the dataset, keeping only one of them.
- Data subset with the condition *dias_int > 0*
    - The variable *dias_int* is an identifier for the length of stay of the patient in each episode. As such, this shall either be zero (for ambulatory cases) or positive (related with an hospital stay). Therefore, the episodes with *dias_int < 0* shall be removed from the dataset, as those are probably inconsistencies in the dataset.
- Data subset with check for all variables with pre-specified valid values
    - Each cancer dataset has a group of variables which have a set of pre-specified valid values. Every episode in which each of these variables does not assume one of the pre-specified valid values is assumed to be incorrectly recorded and will be discarded. The following variables are checked for the pre-specified valid values:
        - *Dsp* – discharge status (e.g. death, discharged to home, etc.);
        - *Adm_tip* – nature of the patient admission to an hospital (e.g. programmed admission, urgent admission, etc.);
        - *Mot_transf* – reason for hospital transfer (e.g. no transfer, transfer for examinations, etc.);
        - *Severidade_APR31* – severity of the episode (can take an integer numeric value from 1 to 4, corresponding to increasing severity);
        - *Mortalidade_APR31* – mortality risk associated with the episode (can take an integer numeric value from 1 to 4, corresponding to increasing risk of mortality).

- Data subset to remove patients who have more than one episode with *dsp = 20* (patient discharged as dead)
    - The *dsp* variable is related with the discharge status for each episode, with the value "20" corresponding to the dead discharge status. Since one patient can have many episodes, but the episodes shall be distinct registries, the death discharge shall be associated with one episode, for data consistency. Therefore, the cases of patients who have more than one episode with "*dsp = 20*" (discharge status equals to dead) are removed from the dataset.
- Data subset to remove patients who have one episode with *dsp* = 20 (discharge status equal to death), with posterior alive episodes
    - The cases of patients who had one episode with discharge status equals to death (*dsp*=20) and have posterior episodes (with future dates) are considered to be inconsistent. The episodes from these patients are removed, since they are considered as recording mistakes.

The number of patients and episodes, for each cancer dataset, before and after cleansing, can be seen in Table 2.

Table 2 – Cancer datasets – raw and cleansed datasets information

Cancer Datasets	Raw Datasets		Cleansed Datasets	
	# of Patients	# of Episodes	# of Patients (% from raw)	# of Episodes (% from raw)
Hepatobiliary	18888	69611	18865 (99.9%)	66561 (95.6%)
Pancreatic	14948	84545	14932 (99.9%)	80883 (95.7%)
Sarcomas	6353	34658	6332 (99.7%)	31901 (92.0%)
Oesophagus	7595	49419	7572 (99.7%)	47349 (95.8%)
Onco-ophthalmology	279	723	277 (99.3%)	697 (96.4%)
Testicular	2280	15740	2269 (99.5%)	15014 (95.4%)

4.1.3 Data Preparation

After having applied the cleansing steps described above, the different cancer datasets can be prepared, with the expected structure, for the statistical analysis to take place.

The original dataset variables, such as diagnosis, procedures, were included in the dataset according to their respective ICD-9 and ICD10-CM codes. To make data easier to interpret, the original dataset was enriched with additional columns, including the descriptions for each of the variables' codes (e.g. diagnosis, procedures, external causes, etc.). This information was matched with the metadata information which was made available by ACSS.

Another column was added related with the mortality status at the end of the episode, as a binary variable – 0 for alive discharge and 1 for death. This variable was inferred from the episode discharge status variable *dsp*: if *dsp* assumes the “20” value, then the patient is discharged dead at the end of the episode and the new variable will have the value 1; any other discharge code will correspond to a value of 0 for this new variable (non-dead discharge).

Another variable of interest is the age at diagnosis of a patient, as it can be a potential predictor for survival. This variable is not originally included in the dataset. To include it, a new variable was created and added to the dataset, capturing the age of the patient in the first episode (with the earlier date) of the patient in the dataset, making the assumption that this episode is the first and when the diagnosis occurs. This variable assumes the same value for all episodes, for the same patient.

Furthermore, taking into account the objective of analysing the number of surgeries impact on survival, a new numeric variable was added with the number of surgeries for each episode. The number of surgeries was derived from the manual analysis and review of the procedures in the dataset, as well as the codification for each episode (variable “*tipo_GDH_APR31”*, “C” for surgical episode, and “M” for medical episode, without surgery). Taking into account that this is a longitudinal study, with information about several episodes, a new column was added to capture the cumulative number of surgeries for each patient, which is the cumulative sum of surgeries of the patient up to each episode.

Another variable of interest to be included is the number of hospital infections registered for each episode. To populate this variable, the diagnosis names (where hospital infections are coded in the database) were manually reviewed to identify hospital infections and count the diagnosis corresponding to those hospital infections. Another variable was also added to include this information about the cumulative sum of hospital infections the patient has experienced up until each episode.

Each of the episodes registered had also the information about the length of stay (variable “dias_int”, 0 in the case of an ambulatory episode); nevertheless, taking into account the objective of studying the survival of patients, it makes sense to have a view of the cumulative length of stay, which derives from the sum of the length of stay of the different episodes up to the last episode; therefore, this new variable was also created.

Finally, taking into account the objective of studying the impact of the creation of the RCs in terms of survival of cancer patients, new information related with RCs needed to be added to the original dataset, as the raw data does not include information as to whether an episode occurred in a RC or not. As such, based on the information gathered from the existing RCs for each cancer type analysed, two new columns were added to the dataset:

a) All episodes in RC – binary variable, which can take value 0 or 1, taking value 1 if the patient had all episodes in a RC, while it assumes 0 value if the patient had one or more episodes in a Non-RC.
b) Referred to RC – binary value, which can take value 0 or 1, being 0 for a patient until the moment he/she has the first episode in a RC; from that moment, all the posterior episodes of this patient have this variable with value 1.

4.1.3.1 Data Set Variables

All the variables in each of the cleansed and prepared cancer datasets can be categorized in two different types: predictor or explanatory variables and the response variables.

The response or dependent variable is the variable with information about the waiting time (survival time) until the occurrence of the event of interest (i.e. the death). Some observations are censored, since the event of interest has not occurred during the follow-up time for each dataset.

The predictor or explanatory variables are those variables which are assumed to have an effect on the response variable. Some examples of independent variables included in each cancer dataset are the demographic variables, such as sex, age and home residence of the patient.

In this study, the predictor variables believed to have an influence in the survival of cancer patients and thus included in the analysis are:

a) Sex – dichotomous variable, corresponding to the sex of the patient;
b) Age at diagnosis – continuous variable, corresponding to the patient's age in years on the first episode in the registry, which is assumed to be the episode in which the diagnosis occurs.
c) Cumulative number of surgeries – continuous variable, corresponding to the number of surgeries the patient had up to the episode in analysis;
d) Cumulative number of infections – continuous variable corresponding to the cumulative number of infections the patient had up to the episode in analysis;
e) Cumulative length of stay – continuous variable corresponding to the cumulative length of stay in days, for the total number of episodes, up to the episode in analysis;
f) Severity – categorical variable corresponding the severity attributed to the episode, which can take the integer numeric values of 1 up to 4, with increasing severity.
g) All episodes in RC – binary variable which can take value 0 or 1, being 1 if the patient had all episodes in a RC and 0 if the patient had at least one episode in a Non-RC.
h) Referred to RC – binary variable which can value 0 or 1, being 0 for a patient up until the moment (if it happens) when the patient has the first episode in a RC. If it happens, the patient

is considered to be “referred”, and thus the variable assumes value 1 for all the posterior episodes.

These different variables which were included in each of the cancer datasets served as the basis for the application of the methodology described in the following section.

4.2 Methodology

As mentioned before, this Dissertation intends to study the time to event (in this case death), in order to develop a survival analysis for the patients with each of six different cancer types, which were followed in Portugal and had a discharge date between 1st of January of 2010 and 25th of November 2019. After cleansing and preparing the data, as described in the previous section, the proposed methodology starts with a descriptive analysis for each data set, in order to obtain an overview of the patients and episodes included. In a second phase, survival curves are obtained and compared using the KM method. Finally, the third phase of the methodology is the implementation of two Extended Cox Models for each cancer type, to obtain an adjusted survival analysis for the different cancer patients. These steps can be synthesized in Figure 7.

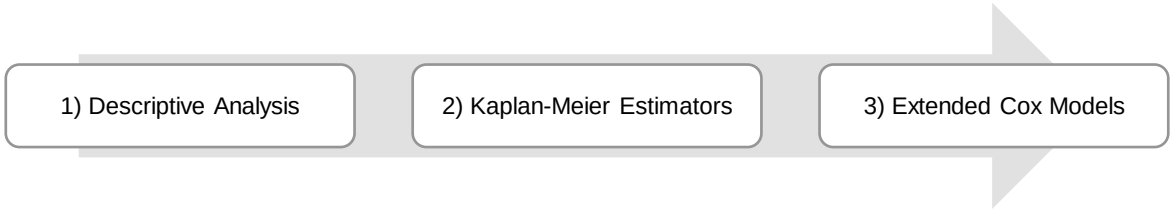

Figure 7 – Methodology overview

All the steps for the operationalization of this methodology were carried out using the R statistical programming language. Several R packages were used to implement this methodology: raw data was loaded from excel files using the R “xlsx” package; datasets were processed in data table, for enhanced performance, using the “dplyr” and “data.table” R packages; descriptive analysis was conducted with the summarization R package “skimr”; plots were obtained using the “ggplot2” R package; survival analysis was conducted with the “survival” and “survminer” packages [55]. More details about the steps of this methodology are described in the following sub-sections.

4.2.1 Descriptive Analysis

The objective of a descriptive analysis is to obtain a summary of the data which will be subject to analysis. In this case, a descriptive analysis was produced for each of the six data sets (one for each cancer type). There were variables which were found to have a completeness rate of zero, meaning those columns were empty, resulting in them being discarded. For continuous numeric variables, mean, minimum and maximum values, as well as the standard deviation were computed. For date

variables, the earliest (minimum) dates, median and latest (maximum) dates were also obtained. For the case of character variables, i.e. variables which can take any character string, the distinct character variable values occurrence was also analysed. Regarding categorical variables, i.e. variables which can take any value from a pre-defined list of possible values, the number of distinct values was also studied.

For each cancer dataset, the predictor variables described before were also statistically analysed, such as the sex or age at diagnosis. The analysis of other variables of interest, such as the most common diagnosis, surgeries and hospital infections, is shown in the appendix for each cancer type. Due to the volume of data and variables, only the more relevant information from the descriptive analysis was included in the appendix, for each cancer dataset.

4.2.2 Kaplan-Meier estimators

Kaplan-Meier (KM) is a method for survival analysis, allowing to study the time until an event occurs [8]. For this Dissertation, the variable of interest is the survival time of the cancer patient, before the event (death of the patient) occurs, also referred as failure [8].

As mentioned before, the cleansed and prepared datasets described in the previous section include patients who have experienced the event (death), but also patients who have not experienced it (censored). Nevertheless, this does not necessarily mean that the patients in this last group have not died in this period – these patients may have died outside the hospital setting, information which the dataset does not include. Furthermore, patients may also have died after the end of the datasets follow-up period (2010 to 2019), for instance in the beginning of 2020 (e.g. related with Covid-19 complications or for lack of cancer treatment, due to fear of being infected with Covid-19 at an hospital setting), so the true survival time can be longer than the observed one. As such, right censoring is applied, meaning that patients who have not experienced the event are considered censored, instead of assuming to be dead or alive outside the study follow-up time.

The general formula for the KM survival probability at failure time t_j is given by:

Equation 9 – Kaplan-Meier survival probability function [8]

$$\mathcal{S}(t_j) = \prod_{i=1}^{j} \mathcal{P}[T > t_i \mid T \geq t_i]$$

$$= \mathcal{S}(t_{j-1}) \times \mathcal{P}(T > t_j \mid T \geq t_j)$$

For data to be ready to be used in KM survival analysis, each patient needs to be characterized by three variables [51]: 1) the serial time, which, in this case is the follow-up duration for that specific patient (i.e. calculated based on the difference between the start date of the first episode and

discharge date for the last episode of that patient in the dataset); 2) the study group that patient belongs to, for a particular KM analysis; 3) the status at the end of the serial time, i.e. dead or censored.

The study group shall be coded into categorical variables, as the KM results in a survival curve for each group. In the case of the numeric variables of interest, such as age and length of stay, these need to be aggregated into different groups, to allow the graphical visualization of the KM survival curves.

Taking into account that each cancer dataset can include several rows (and often does), corresponding to different hospital episodes for each patient, data needed to be prepared in order to load each patient. This means that longitudinal data, corresponding to the different observations for a patient, was combined into summary measures, to allow the study of time-to-event for patients.

The study groups which were analysed for each cancer type were sex, age at diagnosis, total number of infections, total number of surgeries, total length of stay and average severity. The objective is to obtain KM estimators with survival curves for these different groups, in order to analyse if there is a statistically significant difference between them.

For each of these study groups, the values were aggregated to obtain the total cumulative values for the total number of infections, surgeries and length of stay, which were aggregated by summing these variables for the whole episodes, per patient. For the numeric variables, such as age at diagnosis and length of stay, different study groups were tested to identify distinct separate survival curves, to allow an easier interpretation. Furthermore, different groups which were seen in survival analysis literature were also considered (e.g. age groups for a specific cancer dataset – as in the case of pancreatic cancer, whose age groups were the same as those used by Wahutu et al. [18], allowing to compare the results). As severity can take up to four different values for each episode, a mean rounded numeric value was calculated across all episodes.

Finally, after analysing the previous study groups related with those predictor variables (e.g. sex, age at diagnosis, etc.), another KM estimator was obtained for evaluating specifically the oncology RCs model introduction: this KM estimator aims to study the survival analysis of patients who had all hospital episodes in a non-RC (“No Episode in RC”), patients who had at least one episode in an officially recognized RC for their cancer type and one or more in non-RC (“Referred to RC”), and patients who had all episodes in a officially recognized RC for their cancer type (“All Episodes in RC”).

Taking into account the available data, which includes information of all hospital episodes from cancer patients who were discharged after the 1st of January of 2010, as well as the information about the dates of official recognition of RCs (at least after 2015 for onco-ophthalmology, and 2016 for other cancer types), data was sub-setted to only consider patients who had their first episode after the moment of the official recognition of the first RC for their cancer type.

The KM methodology for obtaining the survival curves is based on the number at risk (or risk set) in each KM estimator, which represents the group of patients who have survived at least to that time, based on the fact that those patients have at least another episode after that time. One important aspect to consider is the fact that censoring of patients can occur due to aspects such as loss to follow-up, withdrawal from study, or the patient having not died (at least from the registered data point of view) during the follow-up time. Nevertheless, the analysis shall be cautious, as the number of censored patients can affect the interpretation of the results. As pointed by Rich et al. [51, p. 6] "*survivor function at the far right of a Kaplan-Meier survival curve should be interpreted cautiously, since there are fewer patients remaining in the study group and the survival estimates are not as accurate… the estimations of survival from Kaplan-Meier analysis are most accurate at the time point when most patients are still present*".

The structure of the data which was loaded for obtaining the KM estimators can be observed in Table 3, with some examples of dummy data for different patients (the *Patient_ID* is a unique identifier for each patient in the dataset).

Table 3 – Example of data structure for KM analysis for the sex study group

Patient_ID	Serial time (in years)	Study Group – Sex (1=Male; 2=Female)	Status (1=Dead; 0 = Censored)
258219838	0.558521561	1	0
262119333	0.019164956	2	1
256781745	4.574948665	1	0
…	…	…	…

After preparing the data for each of the study groups, survival curves were obtained and exported in R, using the R survival package [55].

To assess if two or more KM curves are statistically equivalent, the method used was the log-rank test, which is considered to be the most popular method for evaluating KM curves [8]. This test assumes the null hypothesis of no difference between two survival curves [8]. The formula for the log-rank statistic can be observed below:

Equation 10 – Log-rank statistic formula [8]

$$Log-rank\ statistic = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}$$

with

$$O_2 - E_2 = summed\ observed\ minus\ expected\ score\ for\ group\ 2$$

For comparing more than two survival curves, the null hypothesis is that all the curves are the same [8]. The test statistic for this case is more complex, involving the calculation of variances and covariances between each group [8]. In this case, the test statistic can be obtained through a matrix formula, computed with the support from technological tools which are available [8]. This test was obtained using the R "survival" package [55] for each of the different survival curves. The p-value for this test is presented over each KM estimator (*p=value*).

4.2.3 Extended Cox Models

The Cox PH model is considered to be the most commonly used method for multivariate survival analysis in medical research, allowing to model the relationship of covariates with survival of patients [53][62]. The general form for the Cox PH model, given the value for the hazard of the patient for time t, is given as:

Equation 11 – General form of the Cox PH Model [8]

$$h(t) = h_0(t)\, e^{\sum_{i=1}^{p1} \beta_i X_i}$$

where $h_0(t)$ is an unspecified non-negative function of time called the baseline hazard function of time, X_i the predictors or covariates and β_i a set of coefficients. An important aspect of the general form of the Cox PH model is that while the baseline hazard function depends on time, the coefficients and covariates do not. In this form, the covariates X_i are called time-independent covariates [8].

There are however cases in which there is the need to include time-dependent covariates to perform survival analysis, as is the case in this Dissertation. The predictors or covariates are age at diagnosis, number of surgeries, number of hospital infections, length of stay and severity of episodes. Furthermore, besides these covariates, there are others which will be taken into account: a) All episodes in RC and b) Referred to RC, as described in section 4.1.2.

The age at diagnosis and all episodes in a RC are the only time-independent covariates: they remain constant through time for each patient; all the other variables can and most probably change over time for each patient. Therefore, these variables are time-dependent covariates. As such, the traditional Cox model shall be extended to allow the inclusion of both time-dependent and time-independent covariates, through an Extended Cox Model. The formula for this Extended Model can be observed in Equation 12.

Equation 12 – Extended Cox Model Formula [8]

$$h(t, X(t)) = h_0(t)\, e^{\sum_{i=1}^{p1} \beta_i X_i + \sum_{j=1}^{p2} \delta_j X_j(t)}$$

where:

$$X(t) = (X_1, X_2, \dots, X_{p1}, X_1(t), X_2(t), \dots, X_{p2}(t))$$

This formula includes the time-independent covariates X_{p1} and the time-dependent covariates $X_{p2}(t)$. As happens with the original Cox PH model, the extended version also includes the baseline hazard function $h_0(t)$ [8].

Given the objective of studying both the cases when patients have all episodes in a RC and when patients have at least one episode in a RC (i.e. have been "referred"), two Extended Cox Models were selected for evaluating the hazard of the different cancer patients, as follows:

- Multivariate Model 1 – All Episodes in a RC
    - This Extended Cox Model aims to evaluate the hazard of cancer patients who have had all the episodes in a formally recognized RC for their cancer type, compared with patients who had no episode in a RC. The covariate which incorporates this information is the "All_Episodes_RC" covariate, which is a dichotomous variable, assuming value 0 if the patient had all episodes in a Non-RC, being 1 if the patient had all the episodes in a RC for that cancer type.
- Multivariate Model 2 – Referred to a RC
    - This Extended Cox Model aims to evaluate the hazard of cancer patients who have been referred, i.e. had at least one episode in a RC for their cancer type and one or more episodes in a non-RC, compared to patients who had all episodes in a Non-RC. The covariate which provides this information is the "Referred_to_RC" dichotomous covariate, which assumes value 0 if the patient up to that time had no episode in a RC, having value 1 when the patient had at least one episode in a RC for its cancer type.

In order to apply the two Extended Cox Models, each cancer dataset was prepared to be in the correct format, to allow time-dependent covariates to assume the values over the different time intervals. The original cancer datasets included a row for each hospital episode of a patient. To take into account the expected format for developing an Extended Cox model, using the R survival package, data was prepared to code time-dependent covariates using intervals of time, as described in the R vignette by Therneau et al. [55][63]. The built datasets have multiple rows, each row corresponding to a time interval, in the form "]time1; time2]" (open on the left and closed on the right), meaning that the different covariate values apply for that time interval, as described by Therneau et al. [55][63]. Each row has an event column, which takes value 1 if the event (death) happened for that patient during that interval, being 0 otherwise [63].

The time horizon for application of the two Extended Cox models has been defined as beginning when the first RC for the cancer type in study was officially recognized, up to the end of the follow-up period. This means that data has been sub-setted to include only patients who had their first episode

after the recognition of the first RC. The objective is to produce an adjusted analysis and compare the hazard ratio of patients who have been referred to a RC or who had all episodes in a RC.

For both models, the time-dependent covariates are the cumulative number of surgeries (starts at zero and can increase through the follow-up period with the number of surgeries the patient has), the cumulative number of hospital infections and cumulative length of stay (covariates with the same logic as the number of surgeries). The severity variable can also change and assumes the value of the severity attributed to the last hospital episode. The age at diagnosis assumes the same value throughout the follow-up period.

For the selection and validation of the covariates for the two Extended Cox Models described above, a stepwise forward selection procedure was implemented, as applied by Hartmann et al. [38]. A univariate analysis was carried out for each of the predictors, to find the ones which were statistically significant – a p-value < 0.05 was employed. All the predictors found to be statistically significant were included in the multivariate models, while the statistically non-significant predictors were discarded.

In the case of the two predictors related with RCs, which serve as the basis for each of the two Extended Cox Models, if any of those were found to be statistically non-significant in univariate analysis, the respective Extended Cox Model was not implemented, as there is not enough information for the multivariate adjusted model to statistically compare the results, from a RC episodes point of view. In this case, the results shall be interpreted through the univariate analysis produced using the KM estimators. For the cases in which these covariates are found to be not statistically significant, additional data may be needed to allow the implementation (e.g. longer follow-up time, since the creation of the first RC).

The results for each Extended Cox Model are presented in Chapter 5, with relative risk ratios (hazard ratios, HRs) and 95% Confidence Intervals (95% CI) for each of the covariates.

Regarding the proportional hazards assumption, since there are covariates which depend on time (e.g. number of surgeries, number of infections and length of stay), the relative hazard is also time-dependent, resulting in this model no longer being a proportional hazards model. Therefore, an Extended Cox model with time-dependent variables no longer satisfies the proportional hazards assumption [8]. As described by Stensrud et al. [54], the assumption that the hazard ratio remains constant from the beginning of the study until the end of follow-up does not happen in practice for most medical interventions, which is the same for the cancer patients in this study (the hazard can be influenced by the number of surgeries, number of infections, among others). Thus, the obtained overall hazard ratios shall be seen as a weighted average of the time-varying hazard ratios, for each cancer type [54].

Following the suggestions by Stensrud et al. [54], for the cancer types for which the Extended Cox models are implemented, the Restricted Mean Survival Time (RMST) difference will be included, to

support clinical decision making and to make it more understandable to interpret. The Restricted Mean Survival Time (RMST) is a measure which indicates the average survival time up to a pre-specified, clinically important time, which in the case of the comparisons to be made, will be the last follow-up time which is common to the two different study groups (e.g. patients with all episodes in a RC vs. patients with no episode in a RC; patients with at least one episode in a RC vs. patients with no episode in a RC) [64]. The RMST difference represents the gain (if positive) or loss (if negative) in terms of survival time for a group of interest (e.g. patients with all episodes or at least one episode in a RC) [64].

5 Results

This chapter presents the results obtained from the application of the methodology described in the previous chapter, with the aim to study the survival of cancer patients in Portugal which had at least one or all hospital episodes in a oncology RC, specialized on their cancer types, compared to patients who had no episodes in RCs. The application of this methodology is made recursively for each of the six analysed cancer types: a first section describes the statistical analysis produced; a second section presents the survival curves obtained using the KM method; finally, the results of the Extended Cox models are presented and interpreted.

5.1 Hepatobiliary Cancer

As explained in section 4.1.1, the hepatobiliary cancer dataset was derived as a subset from the hepatobiliary and pancreatic data made available by ACSS, according to the primary diagnosis described in "*Circular Normativa ACSS 22/2014*" [42]. As the other datasets, this one was subject to cleansing and preparation and then the chosen methodology was applied. The results are presented in the next sub-sections.

5.1.1 Descriptive Analysis

Analysing the cleansed hepatobiliary dataset, there are a total of 18865 hepatobiliary cancer patients, which represent 66561 episodes, with discharge date in the follow-up period from 1st of January of 2010 to 25th November 2019. There are a total of 12882 (68%) males and 5983 (32%) females, in line with previous studies which report hepatobiliary disease (in particular hepatocellular carcinoma – HCC) to have a higher incidence on males, due to a higher incidence of liver cirrhosis [16]. From the total number of patients, 8662 (46%) died during the follow-up period, while 10203 (54%) were censored. The median follow-up time for censored patients is 30 days (ranging from 0 to 3411 days), while the median time of death is 48 days (ranging from 0 to 3094 days). The patients with follow-up time or time of death of 0 are patients who have only one single episode registered in the dataset, corresponding to death or censoring, respectively. On the other hand, the maximum values for follow-up time and time of death correspond to outliers, i.e. patients who had the first and last registered episodes separated by several years (more than 8 years apart, which is close to the difference between the beginning and end of the total follow up period), hence the relevance of observing the median (instead of the average). Ever since the official recognition of the first hepatobiliary oncology RC in 2016, there have been 3451 (49%) patients who had no episode in a RC, 704 (10%) patients who have been referred to a RC and 2933 (41%) patients who had all episodes in a RC. Further details regarding descriptive analysis for hepatobiliary cancer patients are included in Appendix A.

5.1.2 Kaplan-Meier estimators

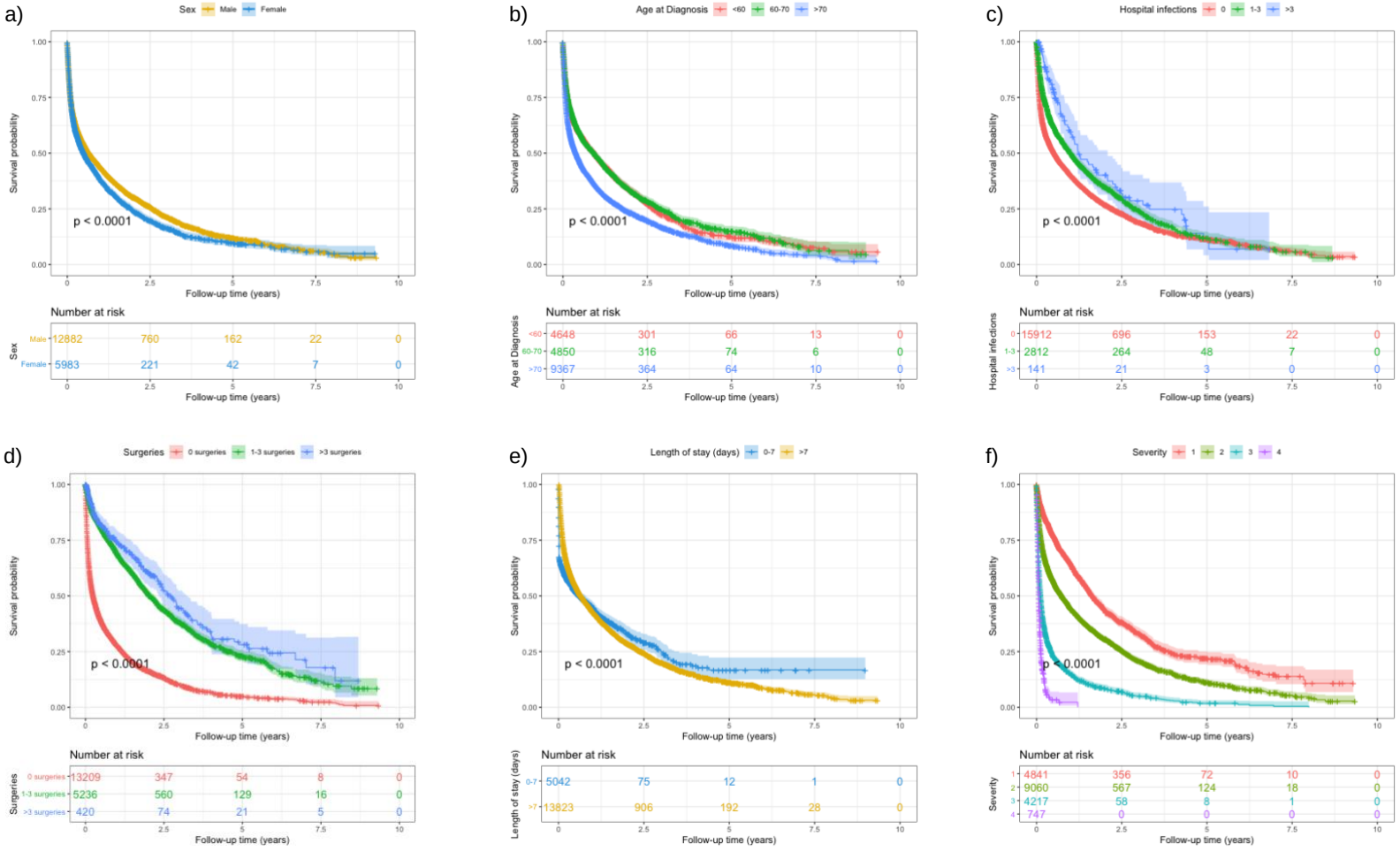

Figure 8 – Comparison of Hepatobiliary Cancer KM estimators for the different predictors

Figure 8 presents the KM estimators, including survival curves for the different predictors: sex, age at diagnosis, number of infections, number of surgeries, length of stay and severity. All the survival curves obtained are statistically different ($p < 0.05$), based on the application of the log-rank test.

The analysis of the KM survival curves allows, at each moment, to compare the survival probability functions for individuals in two or more different study groups; the upper curves mean the individuals in that study groups have higher survival probability than those in the lower survival curves. A table with the number of patients at risk is presented below each survival curve: these figures represent the patients, from the initial sample, who have not yet died or been censored. For each survival curve displayed as a solid line, the 95% confidence intervals are presented, with lower and upper bounds around the survival curve in the same, but brighter, colour. When there is a higher number of patients at risk, the 95% confidence intervals are narrower, which can make them less visible around the survival curves. This interpretation guide is applicable to all the cancer types analysed in this Dissertation.

The KM survival curves for the sex variable (Fig. 8.a) show a survival probability function for hepatobiliary cancer patients similar for males and females, with a slight difference between the second and third years of follow-up, moment upon which the survival probability appears to be more favourable for males than females. Regarding age of diagnosis (Fig. 8.b), taking into account its mean value of 69.3 years, three groups were created: “< 60 years”; “60-70 years”; “> 70 years”. As expected, survival appears to be less favourable to older patients, which may be explained by existing associated co-morbidities, which are not considered in this univariate analysis. This can be further analysed in the multivariate analysis, performed through Extended Cox Models adjusted for severity, as will be done in the next section.

As for hospital infections (Fig 8.c), the patients who have experienced one to three hospital infections or more than three hospital infections appear to have a more favourable survival probability in the first years (until close to four years), decreasing after that period, which can possibly be related with the hospital infections having an impact at a later stage in the follow-up period, when the health condition of the patient may be worse. There is however a wide confidence interval of 95% (denoted by the lower and upper bounds around the solid line), which can be related with censoring of patients; the interpretation of this curves needs to take this fact into account.

Analysing the survival curves for the number of surgeries (Fig 8.d), it appears that the higher the number of surgeries, the higher the survival probability of the patients. This can be related with the potential curative aspect of surgeries, such as resection, local ablation and liver transplantation, as described in the literature [59].

Analysing the KM survival curves for the length of stay (Fig. 8.e), at an initial follow-up period it appears that patients with less days of hospital stay have a less favourable survival function.

Nevertheless, considering that these curves take into account the length of stay for the total follow-up period, the patients with higher number of length of stay may be having those episodes with longer hospital stay later in the follow-up period, justifying the decrease in the survival curve at a later stage.

Regarding the KM survival curves for the mean severity for each patient across all episodes (Fig 8.f), as expected the higher the mean severity the worse is the survival probability function for that patient.

Analysing the KM survival curves related with RCs in Figure 9, one observes that patients who had all episodes in a RC and patients referred to RC have higher survival probability than those patients who had no episode in a RC. This is particularly evident at an earlier stage of the follow-up period for patients. At a later stage, the curves seem to converge, but at that moment there are fewer patients in the risk group (there are more censored patients), with wider 95% confidence intervals, which can be misleading.

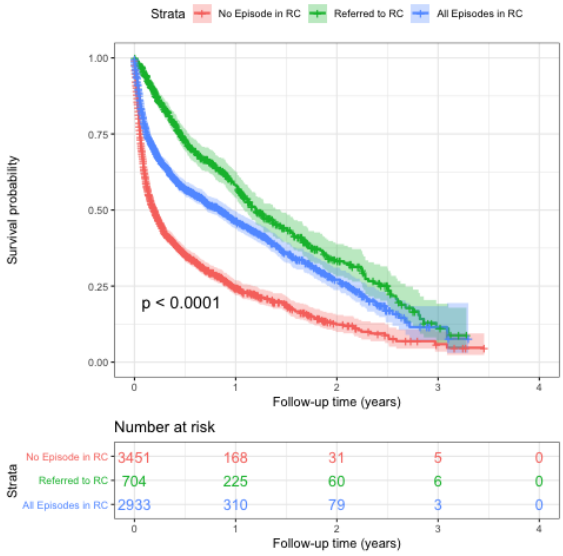

Figure 9 – KM estimators for Hepatobiliary Cancer patients with no episode in RC, referred to RC and with all episodes in RC since the recognition of the first hepatobiliary oncology RC

5.1.3 Extended Cox Models

Following the defined methodology, a stepwise forward selection procedure was implemented, starting with a univariate analysis for each of the covariates. All the covariates except sex were found to be statistically significant ($p < 0.05$) and considered for the two Extended Cox models.

The Extended Cox Models 1 and 2 include the same covariates, which were found to be statistically significant in the univariate analysis, but while Model 1 includes the covariate with information about patients who had all episodes in a RC ("All_Episodes_RC"=1) or no episode in a RC ("All_Episodes_RC"=0), Model 2 includes the information of patients who had one or more episodes in a RC and one or more in a Non-RC (i.e. patients referred to RC – "Referred_to_RC"=1) compared

to patients who had no episode in a RC (“Referred_to_RC”=0). For each Model, a forest plot is presented, with information about the Hazard Ratios (HRs), the 95% Confidence Intervals (95% CI), as well as the p-values. In the case of “All_Episodes_RC”, “Referred_to_RC” and “Severity”, which are categorical covariates, the HRs are compared to the reference group (e.g. the HR for patients with all episodes in a RC – “All_Episodes_RC”=1 is compared to the patients who had no episodes in a RC – “All_Episodes_RC”=0); the same for severity (the HR for “Severity”=4 is compared to the reference “Severity”=1). In case the HR is less than 1, there is a reduction of the hazard when compared with the reference – the covariate is assumed to be a good prognostic factor. Conversely, if the HR is superior to 1, there is an increase in the hazard, therefore the covariate can be seen as a bad prognostic factor.

In the case of the numeric covariates included in both models (i.e. age at diagnosis, number of surgeries, number of infections and length of stay), if the HR is higher than 1 the higher value of that variable corresponds to an higher hazard, thus can ben as a bad prognostic factor; conversely, for an HR less than 1, a higher value of the variable corresponds to a smaller hazard, thus can be seen as a good prognostic factor. This interpretation is applicable to the two Extended Cox Models included for this cancer type and to all subsequent Extended Cox models.

Taking into account this interpretation, the results for Models 1 and 2 can be seen in Figure 10.

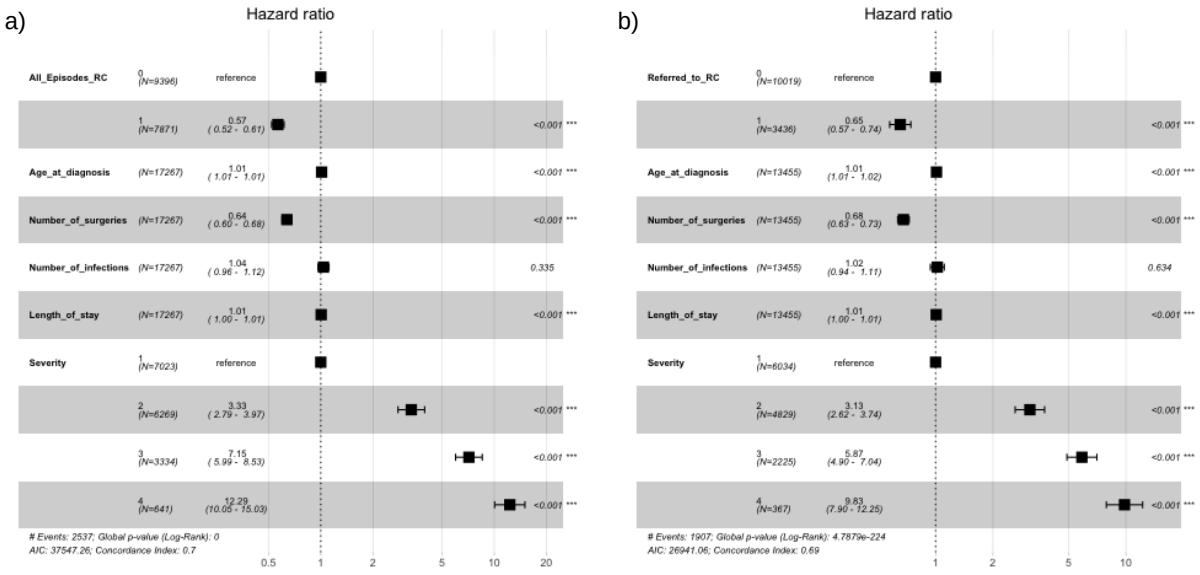

Figure 10 – Hepatobiliary Cancer – Multivariate Model 1 (a) and Model 2 (b) results

In Multivariate Model 1 (Fig. 13.a) all covariates are statistically significant, except for the number of infections (p = 0.335). Patients who had all episodes in an officially recognised RC (All_Episodes_RC=1) have a Hazard Ratio (HR) of 0.57 (95% CI 0.52-0.61), indicating a strong

relationship between having all episodes in a RC and decreased risk of death, when compared to patients who had no episodes in a RC. The increase in the number of surgeries has a strong relationship with better survival outlook, having an associated HR = 0.64 (95% CI 0.60-0.68). The length of stay and age at diagnosis covariates have an HR of approximately 1, therefore do not seem to be associated with better or poorer survival in the context of this model. Finally, the higher the severity, the higher the HR, as expected (the higher the severity, the higher the risk of death). As a complement to the information displayed in the figures above, the RMST difference was computed between the patients who had all episodes in a RC ("All_Episodes_RC"=1 – Group A) and patients who had no episode in a RC ("All_Episodes_RC"=0 – Group B). For a minimum of the largest observed time in each of the two groups of 3.302 years (i.e. the minimum of the largest time when death occurred in both groups of patients), the computed RMST obtained is 0.483 years (95% CI 0.385-0.581, with $p < 0.05$), meaning that patients in Group A, on average, live approximately an additional 176 days (0.483 years) than patients in Group B.

Analysing Multivariate Model 2 (Fig. 13.b), one can see that, as in Model 1, all covariates are statistically significant, except for the number of infections ($p = 0.63$). The patients who have been referred to a RC ("Referred_to_RC"=1) have a HR of 0.65 (95% CI 0.57-0.74), also indicating a strong relationship between being referred to a RC and a decreasing risk of death, when compared to patients who had no episodes in a RC ("Referred_to_RC"=0). Similar to Model 1, the increase in the number of surgeries is associated with better survival – HR = 0.68 (95% CI 0.63-0.73). Regarding the length of stay, age at diagnosis and the severity covariates, the conclusions for Model 1 apply to Model 2, due to the similarity of results. Additionally, the computed RMST difference between patients who have been referred to a RC ("Referred_to_RC"=1 – Group C) and patients who had no episode in a RC ("Referred_to_RC" = 0 – Group B), for a minimum largest observed time in each of the two groups of 3.283 years, is 0.763 years (95% CI 0.642-0.883, with $p < 0.05$), meaning the patients in Group C, on average, live approximately an additional 279 days (0.763 years) than patients in Group B.

5.2 Pancreatic Cancer

Similarly to the hepatobiliary cancer, the pancreatic cancer dataset was derived from the original hepatobiliary and pancreatic dataset made available by ACSS, through the primary diagnosis codes. The application of the chosen methodology to the pancreatic cancer dataset is described in the following sub-sections.

5.2.1 Descriptive Analysis

Analysing the cleansed pancreatic dataset, there are a total of 14932 pancreatic cancer patients, which represent 80883 episodes, with discharge date between the follow-up period from 1st of

January of 2010 to 22nd November 2019 (corresponding to the day of the last episode registered in this dataset). There is a total of 8077 (54%) males and 6855 (46%) females. From the total number of patients, 7274 (49%) died during the follow-up period, while 7658 (51%) were censored. The median follow-up time for censored patients is 31 days (ranging from 0 to 3502 days), while the median time of death is 40 days (ranging from 0 to 2896 days). As in the case of the other cancer types, the patients with follow-up time or time of death of 0 are patients who have only one single episode registered in the dataset, corresponding to death or censoring for each case, respectively. Ever since the official recognition of the first pancreatic oncology RC, there have been 3226 (55%) patients who had no episode in a RC, 478 (8%) patients who had been referred to a RC and 2174 (37%) patients who had all episodes in an RC. Further details regarding descriptive analysis for pancreatic cancer patients are included in Appendix B.

5.2.2 Kaplan-Meier estimators

Analysing Figure 11, all the KM estimators obtained are statistically significantly different ($p < 0.05$), based on the application of the log-rank test, with exception for the sex estimator. This estimator (Fig. 11.a) has a p-value of 0.92, indicating that the sex groups do not differ significantly in terms of survival. Regarding the age at diagnosis (Fig. 11.b), the age groups created were the same as in the study by Wahutu et al. [18]: ("< 55 years"; "55-70 years"; "> 70 years"). Survival appears to be less favourable for older patients, in the study group of patients older than 70 years, which is similar to the results obtained by Wahutu et al. [18].

Analysing the KM survival curves for the hospital infections study groups (Fig 11.c), the patients who experienced one to three or more than three hospital infections appear to have a more favourable survival probability in the first years, decreasing after that period. There is however a wide confidence interval of 95% (denoted by the upper and lower bounds around the solid lines), influenced by the lower number of patients at risk, thus some caution interpreting these curves is needed.

Analysing the KM survival curves for the number of surgeries (Fig 11.d), the patients in the study group who had no surgery throughout the follow-up period appear to have a worse survival function than those patients who had one to three surgeries or more than three surgeries. These results are aligned with those reported by Huang et al. [17], who mentions the improvements in survival for patients who were subject to surgery, in particular for resected patients.

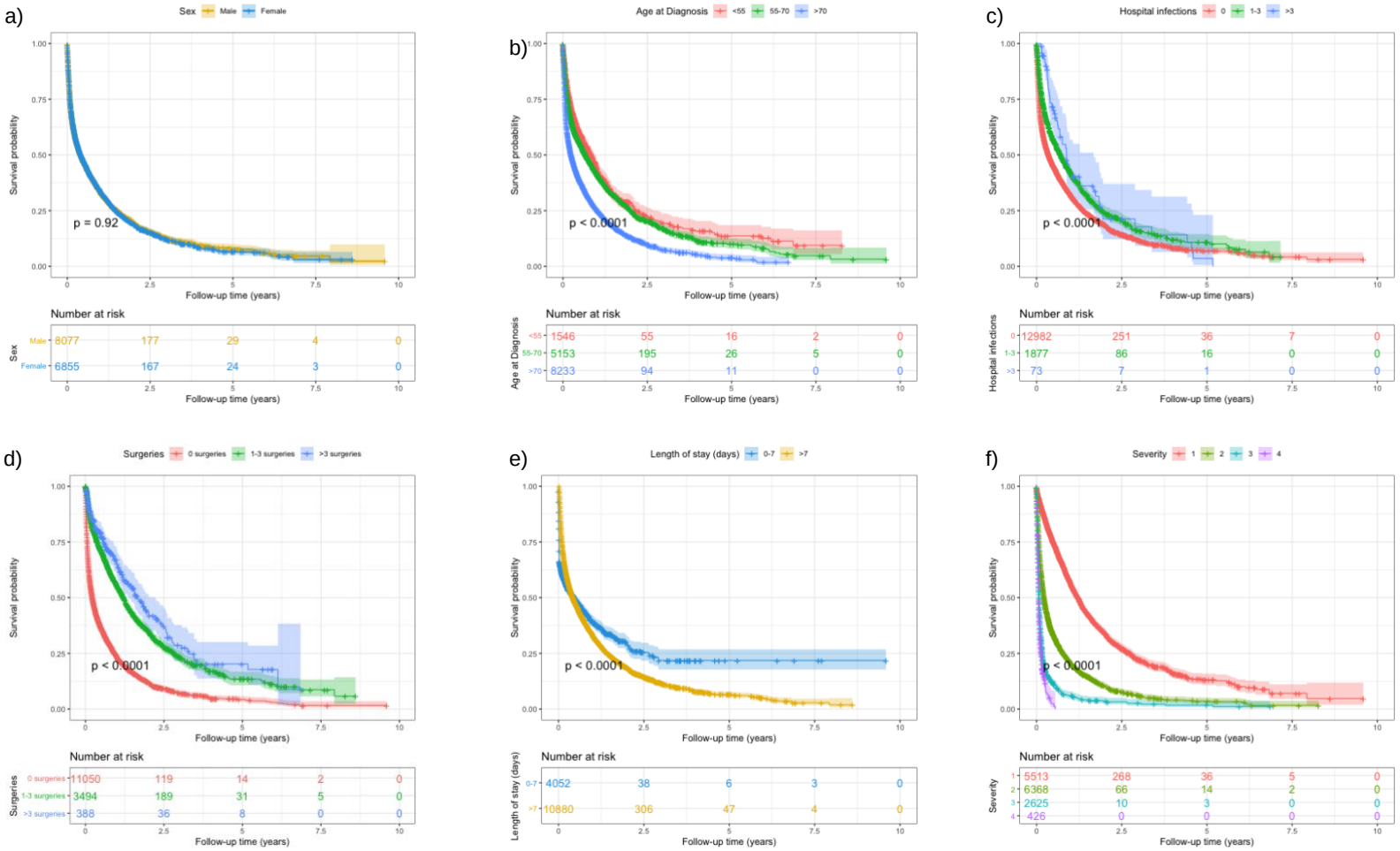

Figure 11 – Comparison of Pancreatic Cancer KM estimators for the different predictors

Analysing the KM survival curves for the length of stay (Fig. 11.e), it appears that at an initial follow-up period, patients with less than seven days of hospital stay have a less favourable survival function than those with a larger number of days in terms of hospital stay. Nevertheless, as in the case of hepatobiliary cancer, patients with higher number of days may be having those episodes later in the follow-up period, justifying the decrease in the survival curve at a later stage.

The KM survival curves for the mean severity estimator for each patient across all episodes (Fig 11.f) show that, as expected, the higher the mean severity for all episodes, the worse the survival probability function for the patients.

The KM survival curves in Fig. 12 are obtained for a sub-set of the patients from the total follow-up period (2010-2019), to include only those who had their first hospital episode after the official recognition of the first pancreatic oncology RC. The patients who had all episodes in a RC (blue curve) or who were referred to a RC (green curve) have a more favourable survival function than those who had no episodes in a RC (red curve). For later stages in the follow-up period, the 95% confidence intervals are wider (denoted by the lower and upper bounds around the solid line), due to the higher number of censored patients and lower number of patients at risk. Similarly, the patients referred to a RC appear to have a more favourable survival function than the patients who had all episodes in a RC, which can be related with differences in other covariates, such as the severity of the episodes.

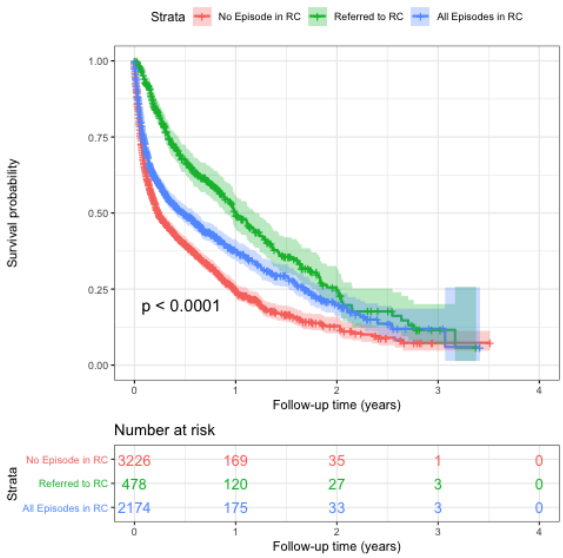

Figure 12 – KM estimators for Pancreatic Cancer patients with no episode in RC, referred to RC and with all episodes in RC since the recognition of the first pancreatic oncology RC

5.2.3 Extended Cox Models

Following the defined methodology, a stepwise forward selection procedure was implemented for the pancreatic cancer dataset. All predictors were found to be statistically significant ($p < 0.05$) in the univariate analysis, except for sex ($p = 0.92$) which was not considered for the two Extended Cox models. The results from the univariate analysis for each of the predictors can be viewed in more detail in Appendix B. The results from the two Extended Cox Models can be seen in Figure 13.

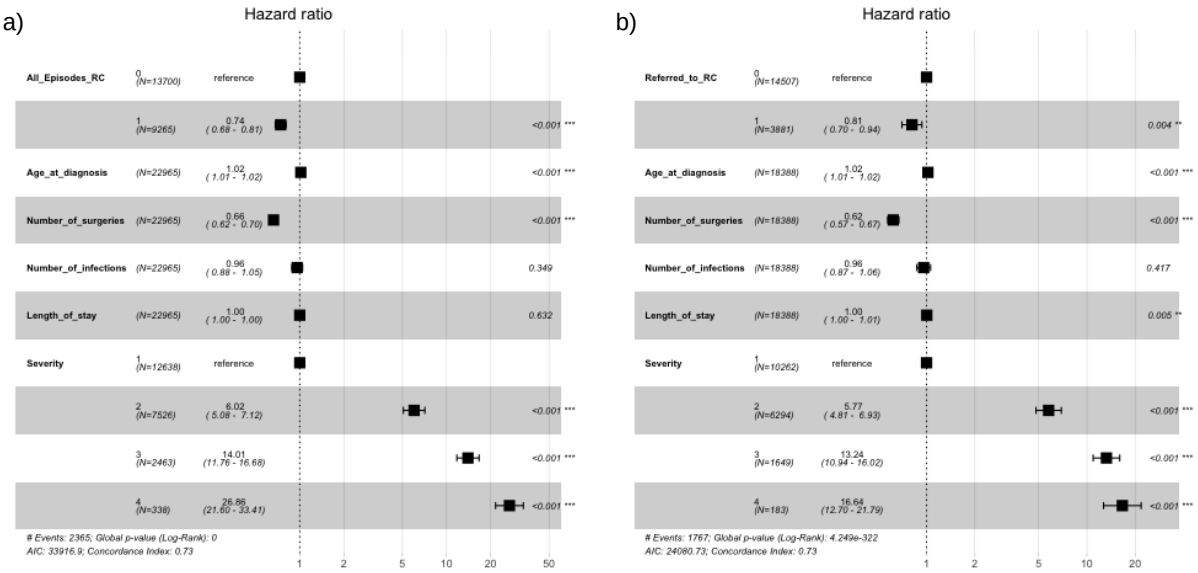

Figure 13 – Pancreatic Cancer – Multivariate Model 1 (a) and Model 2 (b) results

Analysing the results from Model 1 – "All_Episodes_RC" – Fig. 13.a), we observe that all covariates are statistically significant ($p < 0.05$), except for the number of infections and length of stay. The patients who had all episodes in a RC ("All_Episodes_RC"=1) present a HR of 0.74 (95% CI 0.68-0.81), being thus associated with reduced risk of death when compared to patients who had no episode in a RC ("All_Episodes_RC"=0). The age at diagnosis has an HR of 1.01 (95% CI 1.01-1.02), which does not appear to be associated with better or poorer survival in this model. The number of surgeries has an HR of 0.66 (95% CI 0.62-0.70), indicating a strong relationship between an increase in the number of surgeries and a reduced risk of death. The number of hospital infections has an HR of 0.95 (95% CI 0.91-0.99), which appears to have little to no impact on the survival outlook of pancreatic patients in this model. The higher the severity, the higher is the HR, as expected (the higher the severity, the higher the risk of death associated). To complement the information displayed in the figures above, the RMST difference between patients who had all episodes in a RC ("All_Episodes_RC"=1 – Group A) and patients who had no episode in a RC ("All_Episodes_RC"=0 – Group B) was computed. For a minimum largest observed time in each of the two groups of 3.417 years (i.e. the minimum of the largest time when death occurred in both groups of patients), the

computed RMST is 0.257 years (95% CI 0.145-0.369, with $p < 0.05$), meaning that the patients in Group A, on average, live approximately 94 additional days (0.483 years) than patients in Group B.

Analysing the results from Model 2 – "Referred_to_RC", all covariates are also found to be statically significant ($p > 0.05$), except for the number of infections covariate. The HR for pancreatic cancer patients who are referred to RC ("Referred_to_RC" = 1) is 0.81 (95% CI 0.70-0.94), which indicates a better survival prognosis for these patients when compared to patients who had no episode in a RC ("Referred_to_RC" = 0). All the other covariates have similar HRs to those in Model 1, thus the conclusions from Model 1, in terms of survival prognosis, are similar. Additionally, to complement the information displayed in the Figure above, the RMST difference between patients who have been referred to a RC ("Referred_to_RC"=1 – Group C) and patients who had no episode in a RC ("Referred_to_RC" = 0 – Group B) was also obtained. For a minimum largest observed time in each of the two groups of 3.379 years, the computed RMST difference is 0.542 years (95% CI 0.400-0.685, with $p < 0.05$), meaning that the patients in Group C live on average approximately 198 additional days (0.542 years) than patients in Group B.

5.3 Sarcomas

The sarcomas dataset was subject to cleansing and preparation, according to the procedures described in the previous chapter. After this preparation step, the sarcomas dataset was subject to application of the defined methodology, yielding the results which are presented in the next sub-sections.

5.3.1 Descriptive Analysis

Analysing the cleansed sarcomas dataset, there are a total of 6332 sarcomas patients, which represent 31901 episodes, with discharge date between the follow-up period from 1st of January of 2010 to 5th November 2019 (corresponding to the day of the last episode registered in this dataset). There is a total of 3402 (54%) male and 2930 (46%) female patients. From the total number of patients, 1154 (18%) died during the follow-up period, while 5178 (82%) were censored. The median follow-up time for censored patients is 13 days (ranging from 0 to 3422 days), while the median time of death is 51 days (ranging from 0 to 2817 days). As in the previous cancer types, patients with follow-up time or time of death of zero are patients who have only one single episode registered in the dataset, corresponding to death or censoring, respectively. Ever since the official recognition of the first sarcomas RC, there have been 1036 (53%) patients who had no episode in a RC, 77 (4%) patients who had been referred to a RC and 849 (43%) patients who had all episodes in an RC. Further details regarding descriptive analysis for sarcomas patients is included in Appendix C.

5.3.2 Kaplan-Meier estimators

Figure 14 presents six KM estimators associated with sarcoma patients, for the following predictors: sex, age at diagnosis, number of infections, number of surgeries, length of stay and severity. The log-rank test for difference in survival for the groups in the predictors age at diagnosis, hospital infections, surgeries and severity yield statistically significant p-values ($p < 0.05$), indicating that those survival curves differ significantly. On the other hand, the log-rank test for the sex and length of stay predictors give a non-statistically significant p-value ($p > 0.05$), indicating that these survival curves do not differ significantly.

The study groups defined for the age at diagnosis survival curves were the same as in Blay et al. [24]. Analysing the KM survival curves in Figure 14.b throughout the follow-up period, one observes that the confidence intervals (denoted by the lower and upper bounds around the solid lines) become wider for later stages in the follow-up period, which is related with censoring (patients lost to follow-up) and with a decrease of the numbers of patients at risk. Nevertheless, at an earlier stage of follow-up, the survival probability function is more favourable for younger patients.

Analysing the KM estimator for hospital infections (Fig. 14.c) and the numbers at risk for each survival curve (presented in the table below the survival curves), one observes that patients who do not experience any hospital infections represent the vast majority of patients. According to this KM estimator, patients who experience one to three hospital infections have worse survival than patients who had no hospital infections. The survival curve for patients who experienced more than three hospital infections appears to be more favourable at an early stage but, nevertheless, the confidence interval is wide (denoted by the lower and upper bounds around the solid line) and there is a steep decrease in the number of patients at risk. This fact indicates a relevant number of censored patients; therefore, any interpretation of this survival curve needs to be carefully made.

Analysing the KM estimator for the number of surgeries (Fig. 14.d), the survival curve for patients who had no surgery presents a less favourable prognosis than the survival curves for patients who had one or more surgeries. This is more evident at an early stage of follow-up than at later follow-up stages, which can be related with the decrease of the number of patients at risk, due to censoring. As described by Blay et al. [24], the surgical removal of sarcoma with resection is considered to be the mainstay curative treatment for sarcoma, which can be associated with improved survival and is aligned with the results obtained.

Finally, the survival curves for the severity KM estimator (Fig. 14.f) are aligned with the expected results – the higher the severity of the patient, the worse the survival probability for that patient.

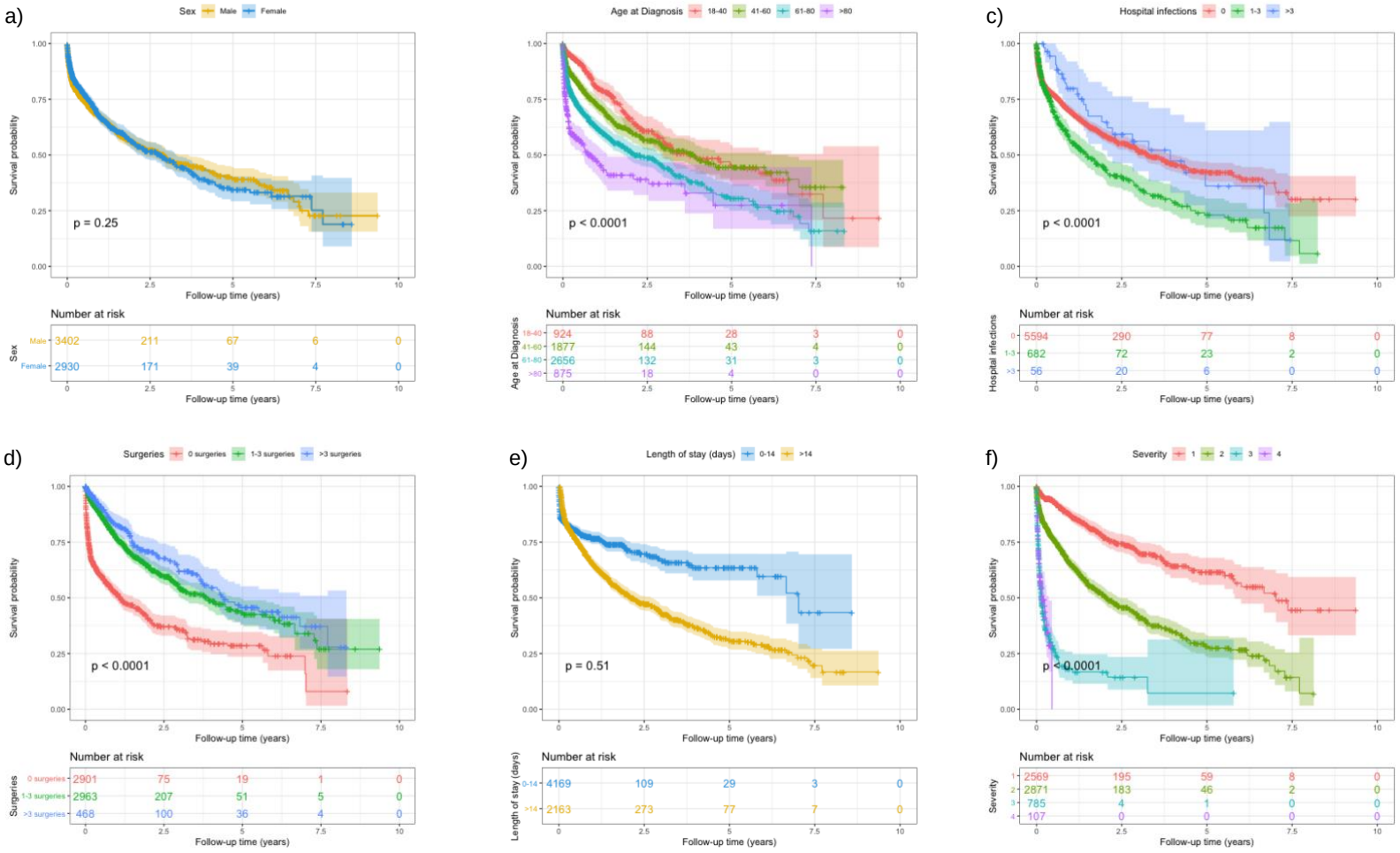

Figure 14 – Comparison of Sarcomas KM estimators for the different predictors

Figure 15 presents the KM survival curves for sarcoma patients who had no episode in a RC (red curve), who were referred to a RC (green curve) and who had all episodes in a RC (blue curve), since the official recognition of the first sarcoma RC. The survival curves for sarcoma patients who had all episodes in a RC and who were referred to a RC appear to have a more favourable survival probability than those patients who had no episode in a RC. This is particularly evident at an early stage of the follow-up period, when there are more patients in the groups at risk (as shown in the table in Figure 15, below the survival curves), than at later stages in the follow-up period, when there are less patients (increase of censored patients).

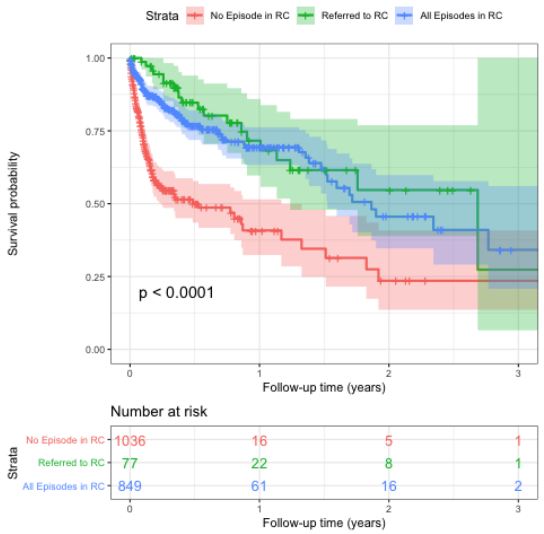

Figure 15 – KM estimators for Sarcoma patients with no episode in RC, referred to RC and with all episodes in RC since the recognition of the first sarcoma oncology RC

5.3.3 Extended Cox Models

Following the defined methodology, a stepwise forward selection procedure was applied to select the covariates to be included in the sarcoma Extended Cox models. The univariate analysis for these covariates shows statistical significance for the covariate “All_Episodes_RC” ($p < 0.05$) and statistical non-significance for the covariate “Referred_to_RC” ($p = 0.64$). Therefore, according to the methodology, since multivariate models 1 and 2 depend on these covariates (respectively) being statistically significant, Model 1 was implemented, while Model 2 was not, due to the covariate of interest not being statistically significant ($p > 0.05$). The results of the univariate analysis can be viewed in more detail in Appendix C.

Analysing the results for Multivariate Model 1 – “All_Episodes_RC”, presented in Figure 16, all covariates are found to be statistically significant ($p < 0.05$), except for the number of infections and length of stay. The patients who had all episodes in a RC (“All_Episodes_RC”=1) have a HR of 0.60 (95% CI 0.46-0.79), being associated with reduced risk of death, when compared to patients who had

no episodes in a RC (“All_Episodes_RC”=0). The age at diagnosis, which has an HR of 1.01 (95% CI 1.01-1.02), does not appear to be associated with better or poorer survival for sarcoma patients, according to these results. The number of surgeries has an HR of 0.81 (95% CI 0.73-0.89), indicating a strong relationship between an increase in the number of surgeries and a reduced risk of death. The higher the severity level, the higher the HR, indicating a higher risk of death, which is aligned with the expectations.

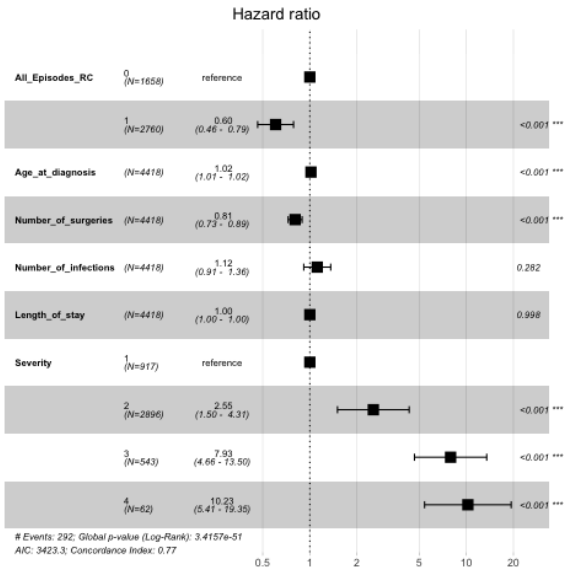

Figure 16 – Sarcomas – Multivariate Model 1

To complement the information displayed in Figure 16, the RMST difference between patients who had all episodes in a RC (“All_Episodes_RC”=1 – Group A) and patients who had no episode in a RC (“All_Episodes_RC”=0 – Group B) was obtained. For a minimum largest observed time in each of the two groups of 3.261 years (i.e. the minimum of the largest time when death occurred in both groups of patients), the computed RMST is 0.711 years (95% CI 0.344-1.077, with $p < 0.05$), meaning the patients in Group A live on average approximately 260 additional days (0.711 years) than patients in Group B.

5.4 Oesophageal Cancer

Similar to the previous cancer types, the methodology described in chapter four was applied to the cleansed oesophageal cancer dataset. The results from the methodology application to this data set are presented in the following sub-sections.

5.4.1 Descriptive Analysis

There is a total of 7572 oesophageal cancer patients, which represent 47349 episodes, with discharge date between the follow-up period from 1st of January of 2010 to 25th November 2019 (corresponding to the day of the last episode registered in this dataset). There is a total of 6343 (84%) males and 1229 (16%) females. This type of cancer affected mainly male patients, which is aligned with the literature regarding sex prevalence [65]. From the total number of patients, 3116 (41%) died during the follow-up period, while 4456 (59%) were censored. The median follow-up time for censored patients is 44 days (ranging from 0 to 3212 days), while the median time of death is 101 days (ranging from 0 to 2935 days). As in the case of the other cancer types, the patients with follow-up time or time of death of 0 are patients who have only one single episode registered in the dataset, corresponding to death or censoring for each case, respectively. Ever since the official recognition of the first oesophageal RC, there have been 1410 (55%) patients who had no episode in a RC, 253 (10%) patients who had been referred to a RC and 884 (35%) patients who had all episodes in a RC. Further details regarding descriptive analysis for oesophageal cancer patients are included in Appendix D.

5.4.2 Kaplan-Meier estimators

Figure 17 presents the results for the six oesophageal cancer KM estimators. The log-rank test for all KM estimators yields significance with $p < 0.05$, except for the sex estimator (Fig 17.a), which indicates that sex survival curves differ significantly in survival ($p = 0.11$).

Regarding the age at diagnosis (Fig. 17.b), the age groups chosen were less and more than 50 years, which are the same age groups used by Tustumi et al. [60]. According to the KM survival curves obtained for the different age groups, the survival probability appears to be more favourable for younger patients (less than 50 years group) than for older patients (more than 50 years). For later stages in the follow-up time, the survival probability of the two age groups appears to converge, but the confidence interval for younger patients also increases (denoted by the upper and lower bounds around the solid line), due to the fewer number of patients at risk. Thus, the interpretation of the survival curves at this stage of follow-up shall take this fact into account.

Analysing the KM estimator for hospital infections (Fig. 17.c), the curves for patients who had no hospital infection and patients who had between one and three infections appear to be similar. On the other hand, the survival curve for patients who had more than three hospital infections is more difficult to interpret due to the fewer number of patients at risk, presenting wider confidence intervals (denoted by the lower and upper bounds around the solid line). Therefore, this fact shall be considered when interpreting this KM estimator.

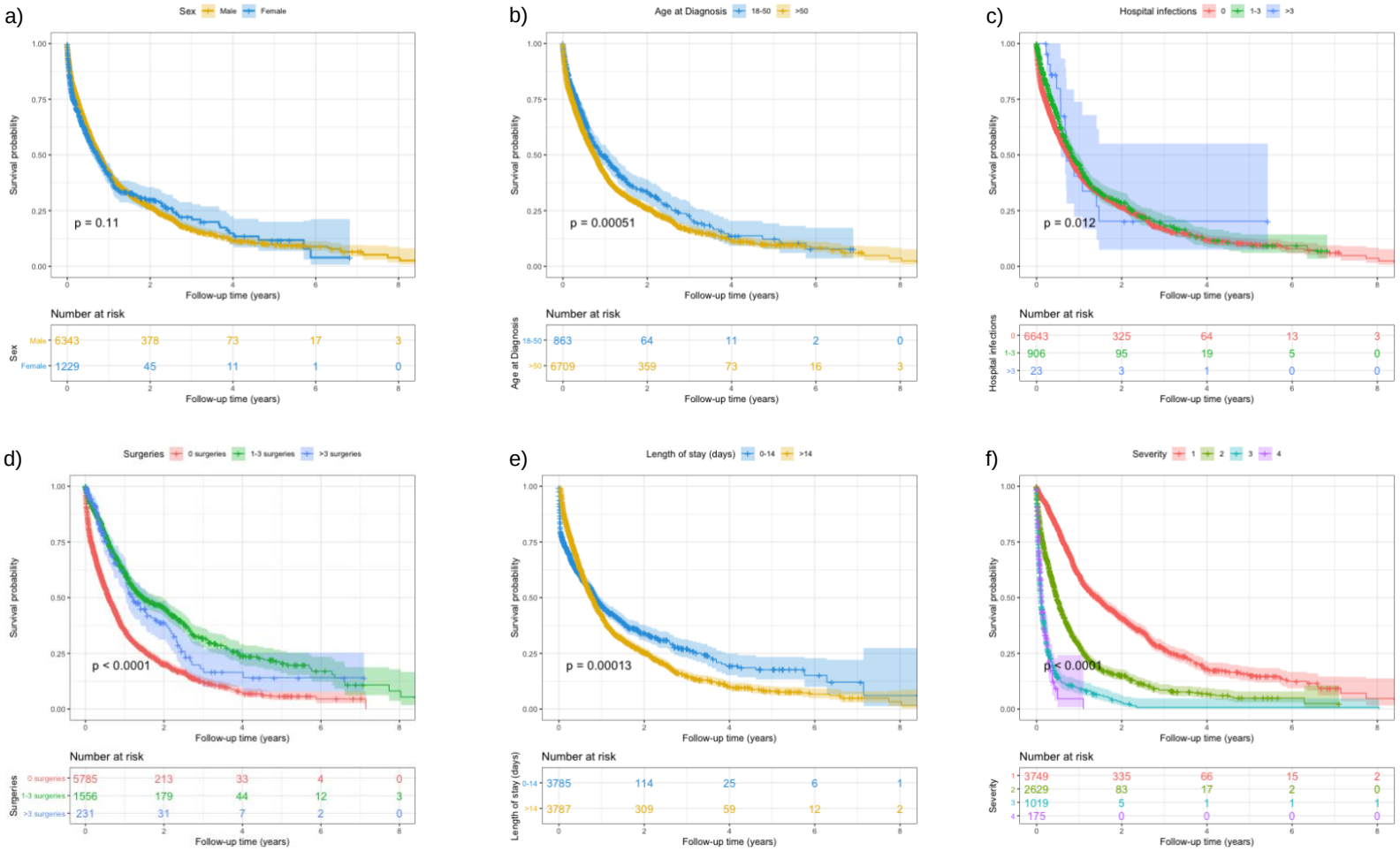

Figure 17 – Comparison of Oesophageal Cancer KM estimators for the different predictors

Analysing the KM estimator for the number of surgeries (Fig. 17.d), the survival probability appears to be more favourable for patients who have a higher number of surgeries – the survival curve for patients who had no surgery is associated with lower survival probability when compared with the survival curves for patients who had one to three surgeries and patients who had more than three surgeries. This can possibly be related with the obtained benefits, in terms of survival, from curative surgery [66].

The KM estimator for length of stay (Fig.17.e) has been plotted with two survival curves – one for patients who had less than fourteen days of total length of stay, and another for patients who had more than fourteen days of length of stay. The survival probability for patients who had a total length of stay higher than fourteen days appears to be more favourable at an earlier stage than for patients with fewer total days of hospital stay. This changes at later stages of the follow-up time, which can be related with patients who have higher length of hospital stay at later stages in the follow-up time.

Regarding the severity KM estimator (Fig. 16.f), the survival curves show poorer survival probabilities for patients with higher mean severity, which is in accordance with the empirical expectations.

Finally, Figure 18 presents the survival curves for oesophageal cancer patients who have been referred to RC (green curve), had all episodes in a RC (blue curve) and who had no episode in a RC (red curve), since the recognition of the first oesophageal RC. As with previous cancer types, the survival curves, at an earlier stage of the follow-up period, show a more favourable survival probability for patients who were referred (green curve) and had all episodes in a RC (blue curve), when compared with patients who had no episodes at a RC (red curve).

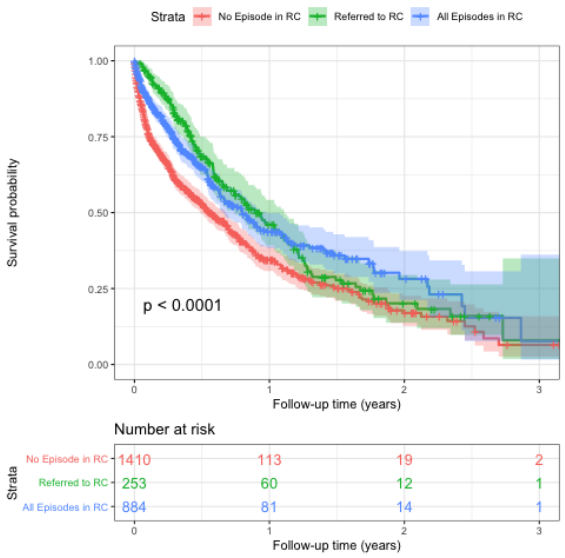

Figure 18 – KM estimators for Oesophageal Cancer patients with no episode in RC, referred to RC and with all episodes in RC since the recognition of the first oesophageal oncology RC

5.4.3 Extended Cox Models

As defined in the methodology, a stepwise forward selection procedure was applied to select the covariates to be included in the Extended Cox models. The univariate analysis for these covariates has given a statistically significant p-value for the covariate “All_Episodes_RC” ($p < 0.05$) and a statistically non-significant p-value for the covariate “Referred_to_RC” ($p = 0.63$). Therefore, since each of the Models 1 and 2 depend on these covariates, Model 1 was implemented, while Model 2 was not, due to the covariate of interest for the latter not being statistically significant. The results of the univariate analysis for the oesophageal cancer dataset can be viewed in more detail in Appendix D.

The results obtained for Multivariate Model 1 for oesophageal cancer presented in Figure 19 show all covariates to be statistically significant ($p < 0.05$), except for sex ($p = 0.177$) and number of hospital infections ($p = 0.477$). Regarding the main covariate of interest in this model, “All_Episodes_RC”, the HR is 0.64 (95%CI 0.56-0.75) indicating a strong relationship between having all episodes in a RC (“All_Episodes_RC” = 1) and a decreasing risk of death, when compared to patients who have no episode in a RC (“All_Episodes_RC” = 0). The number of surgeries has a HR of 0.87 (95% CI 0.80-0.90), showing a decreasing risk of death when the number of surgeries increases, which can be associated with the potential curative aspect of surgery for oesophageal cancer patients. The length of stay covariate has an HR close to 1 (95% CI 1-1.01), indicating that, according to this model, it does not have a significant impact on the survival of oesophageal cancer patients. The higher the severity covariate, the higher the HR, which indicates a higher risk of death, aligned with the empirical expectations.

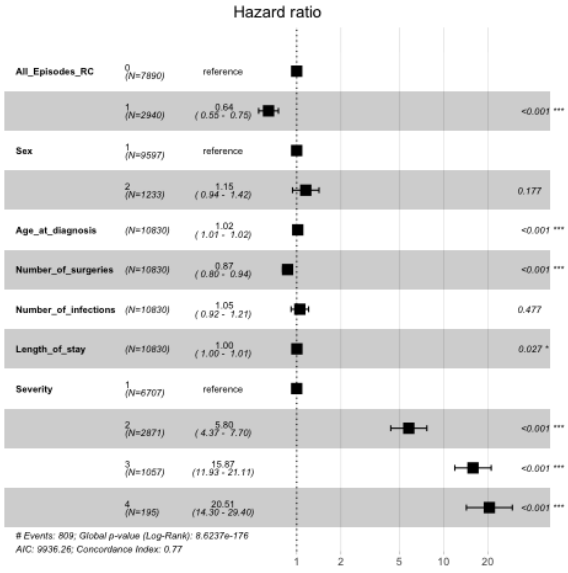

Figure 19 – Oesophageal Cancer – Multivariate Model 1

To complement the information displayed in Figure 19, the RMST difference between patients who had all episodes in a RC ("All_Episodes_RC"=1 – Group A) and patients who had no episode in a RC ("All_Episodes_RC"=0 – Group B) was obtained. For a minimum largest observed time on each of the two groups of 3.149 years (i.e. the minimum of the largest time when death occurred in both groups of patients), the computed RMST is 0.237 years (95% CI 0.052-0.442, with $p < 0.05$), meaning that the patients in Group A live on average approximately an additional 87 days (0.237 years) than patients in Group B.

5.5 Onco-ophthalmology

This section presents the results from the methodology application to the cleansed onco-ophthalmology dataset. This dataset is relatively small (277 patients for follow-up period 2010-2019), when compared with the other analysed cancer datasets. This presents challenges to obtain the desirable conclusions regarding the objective of producing a survival analysis of patients who had no episode, have been referred, or had all the episodes in a onco-ophthalmology RC. The next sub-sections present the results of the analysis.

5.5.1 Descriptive Analysis

Analysing the cleansed dataset, there are a total of 277 onco-ophthalmology patients, who had a total of 697 episodes during the follow-up period, with discharge date between the follow-up period from 25th of January of 2010 to 6th November 2019 (corresponding to the day of the last episode registered in this dataset). There is a total of 141 (51%) male patients and 136 (49%) female patients. From the total number of patients, 31 (11%) died during the follow-up period, while 246 (89%) were censored. The median follow-up time for censored patients is 12 days (ranging from 0 to 711 days), while the median time of death is 4 days (ranging from 0 to 2248 days).

For the whole follow-up period, there are a total of 266 (~ 95%) patients who had no episode in a RC, a total of 1 (~ 0.3%) patient who was referred to a RC and 10 (~ 4%) patients who had all episodes in a RC. From the total number of patients who were referred or had all episodes in the onco-ophthalmology RC (11 patients), all have been censored (none has registered death in the data during the follow-up period). Further details regarding the descriptive analysis for onco-ophthalmology patients are included in Appendix E.

5.5.2 Kaplan-Meier estimators

Analysing Figure 20, which includes the results for the KM estimator for onco-ophthalmology, there are three KM estimator for which the log-rank test indicate that the survival curves differ significantly ($p < 0.05$). Those estimators are the number of hospital infections, number of surgeries and severity.

On the other hand, there are other three KM estimators for which the log-rank test yields a p-value > 0.05, which means that the survival curves do not differ significantly – those estimators are sex, age at diagnosis and length of stay. Taking into account that the survival curves from these last three KM estimators do not differ significantly, it is not possible to take statistically significant conclusions for them.

All the KM estimators have wide confidence intervals, denoted by the lower and upper bounds around the solid lines, which can be related with a high number of censored patients, but also with the low number of patients who have died during the study (11% of the total patients).

Concerning the hospital infections KM estimator (Fig. 20.c), the patients who have experienced one to three hospital infections have worse survival probability than the patients who have not experienced any hospital infection. There are however wide confidence intervals for the two curves, denoted by the upper and lower bounds around the solid lines, which can be related to the censoring of patients.

Analysing the KM estimator for the number of surgeries (Fig. 20.d), at an initial stage of the follow-up period the patients who had at least one to three surgeries or more than three surgeries have a more favourable survival probability than those patients who had no surgery. The survival curves for patients who had no surgery and had one to three surgeries end earlier than the curve for patients who had more than three surgeries (blue curve), which is related with patients who do not have any further episode registered after that follow-up time.

Analysing the KM estimator for severity (Fig. 20.e), as in the previous cancer types the survival curves are according to the expectations – the higher the mean severity of a patient, the worse is the survival probability for that patient.

Regarding the KM estimator which compare the survival curves of patients who had no episode in a RC, were referred to a RC or had all the episodes in a RC, since there are no deaths recorded for patients who had at least one episode in a RC the survival probability for those patients is one for all the follow-up period of time. This makes it difficult to effectively compare the survival curves for these study groups. Therefore, this KM estimator was not possible to obtain and has not been included in this Dissertation.

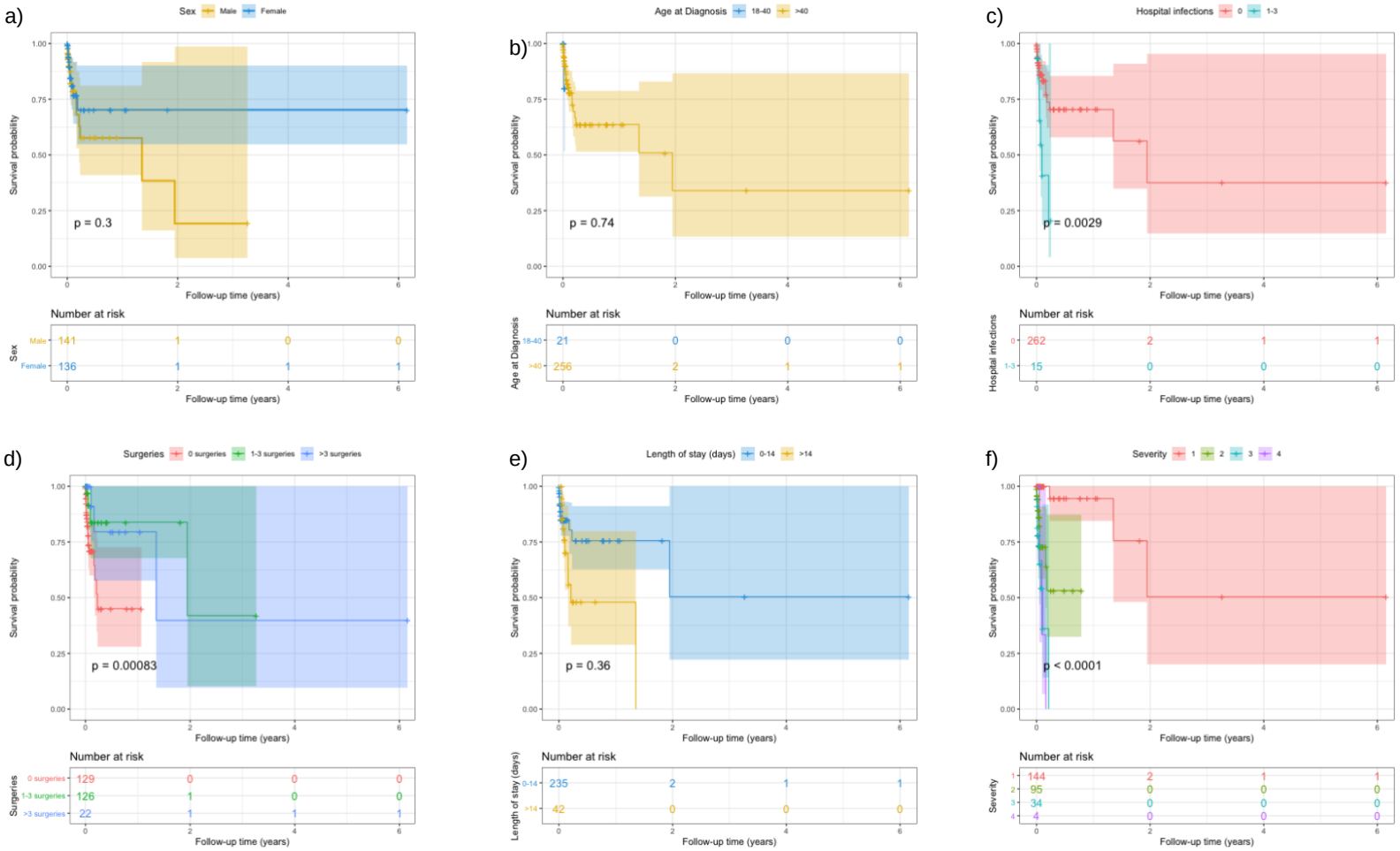

Figure 20 – Comparison of Onco-Ophthalmology KM estimators for the different predictors

5.5.3 Extended Cox Models

The implementation of the Extended Cox models, as described in the methodology, depends on the existence of study groups with relevant information for the survival analysis. In the case of the predictors “Referred_to_RC” and “All_Episodes_RC” for the onco-ophthalmology dataset, the first predictor has only one patient who was referred to a RC, and the second predictor has only ten patients who had all episodes in a RC – none of these eleven patients have died during the follow-up period. Therefore, it is not feasible to study and compare the survival and hazard of the patients including the information from these covariates. Nevertheless, the results from the univariate analysis can be viewed in Appendix E.

5.6 Testicular Cancer

This section includes the application of the methodology to the testicular cancer dataset which was made available by ACSS. The results obtained from the application of the methodology to this dataset are presented in the next sub-sections.

5.6.1 Descriptive Analysis

The testicular cancer dataset was subject to cleansing, after which there are a total of 2260 patients, representing a total of 15014 hospital episodes. These episodes had a discharge date between the follow-up period of 1st of January of 2010 to 25th November 2019 (corresponding to the day of the last episode registered in this dataset). Of the total patients, 2162 (95%) are censored, while 107 (5%) have died during the follow-up period. The median follow-up time for censored patients is 28 days (ranging from 0 to 2676 days), while the median time of death is 88 days (ranging from 0 to 3246 days). The patients with follow-up time or time of death of 0 are patients who have only one single episode registered in the dataset, corresponding to death or censoring for each case, respectively. Ever since the official recognition of the first testicular oncology RC, there have been 543 (66%) patients who had no episode in a RC, 53 (6%) patients who had been referred to a RC and 230 (28%) patients who had all episodes in an RC. Further details regarding descriptive analysis for testicular cancer patients is included in Appendix F.

5.6.2 Kaplan-Meier estimators

Figure 21 presents the six KM estimators for the different predictors associated with testicular cancer: sex, age at diagnosis, number of infections, number of surgeries, length of stay and severity. Since there are only male patients, there are no additional survival curves and the KM estimator for sex presents the survival curve only for male patients. For the case of all the other five KM estimators,

the log-rank test yields statistical significance ($p < 0.05$), which means that those survival curves differ significantly.

Analysing the age at diagnosis KM estimator (Fig. 21.b), the survival curve for younger patients (blue curve), which includes patients with an age at diagnosis between 18 and 40 years, represents the vast majority of testicular cancer patients, as can be observed in the table with the number of patients at risk. These numbers are aligned with the literature, which describe testicular cancer to mainly affect younger patients [34]. Comparing the age at diagnosis KM descriptor, the survival curve for patients with an age at diagnosis of less than 40 years old shows a more favourable survival probability than the survival probability for patients diagnosed after 40 years old.

Concerning the KM estimator for hospital infections (Fig. 21.c), the survival curve for patients who had no hospital infections shows a more favourable survival probability than the survival curve for patients who had one to three hospital infections. The survival curve for patients who had more than three hospital infections shall be interpreted with caution, as the number of patients included is very reduced (only 5 patients), originating a wide confidence interval (denoted by the upper and lower bounds around the solid line).

Analysing the KM estimator for the number of surgeries (Fig. 21.d), the survival curves for patients who had one to three surgeries and more than three surgeries show more favourable survival probabilities than the survival curve for patients who had no surgery. The benefits from surgery for patients, in terms of survival, can possibly be associated with the curative effect of surgery (e.g. orchiectomy, with testicular removal).

Regarding the KM estimator for the length of stay (Fig. 21.e), the survival curve for patients who have 0 to 14 days of length of stay shows a more favourable survival probability at each time than the survival curve for patients with more than 14 days of length of stay. This fact can be associated with patients with a worse prognosis, who possibly have a more prolonged length of stay. This extended length of stay can also be associated with a worse survival probability.

Finally, the KM estimator for severity (Fig 21.f) is in line with the analysis produced in the previous cancer types: the higher the mean severity of the patient (according with the respective survival curve), the worse the survival probability for the testicular cancer patient.

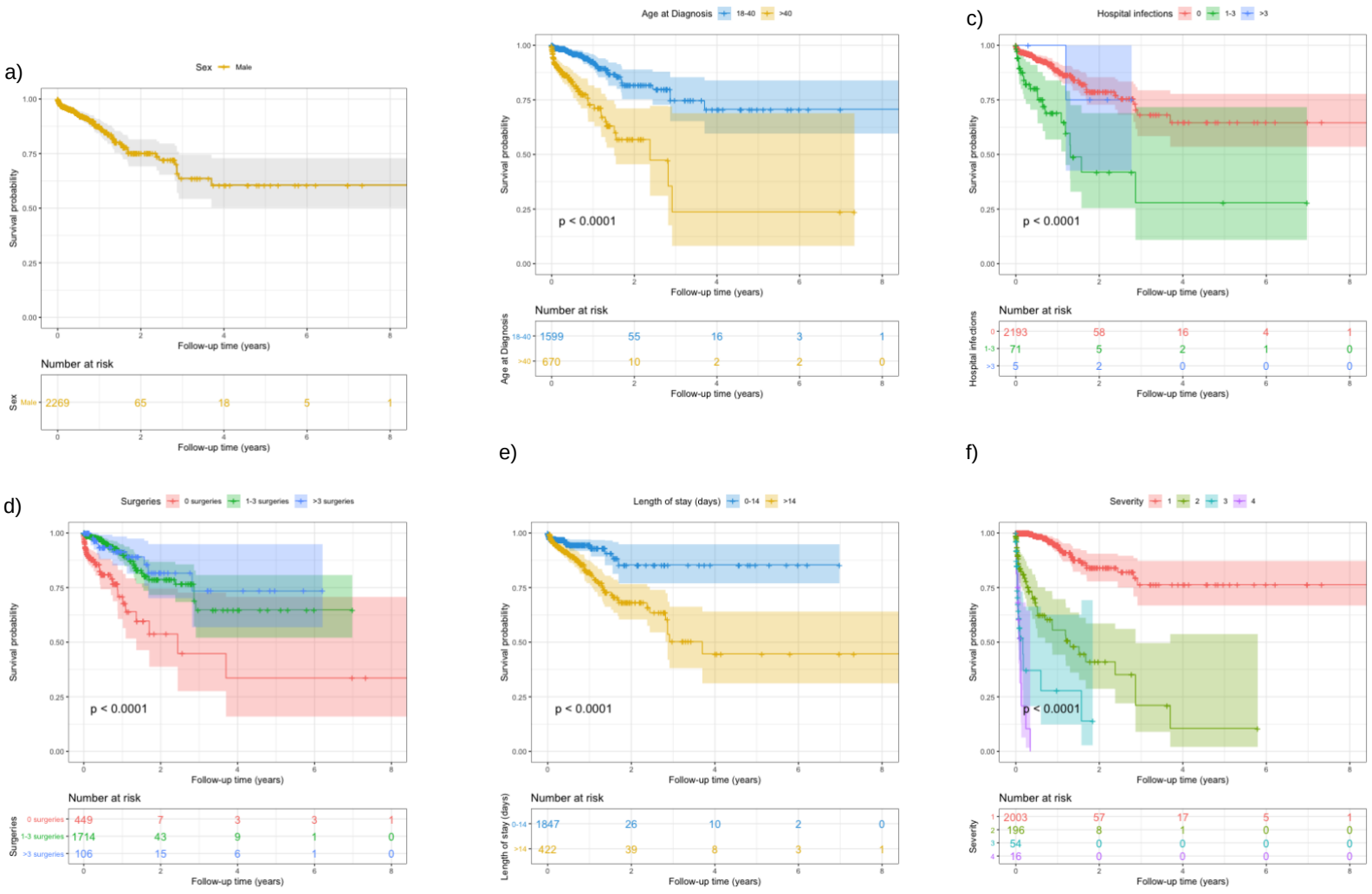

Figure 21 – Comparison of Testicular Cancer Kaplan-Meier estimators for the different predictors

Figure 22 presents the survival curves for testicular cancer patients who had no episode in a RC (red curve), have been referred to a RC (green curve), or had all the episodes in a RC since the creation of the first testicular oncology RC (blue curve). The results from the application of the log-rank test indicate that the presented survival curves do not differ significantly among them ($p > 0.05$). Therefore, from the analysis of this KM estimator, no significant conclusion can be extracted regarding the impact of the creation of the RCs on the survival of testicular cancer patients.

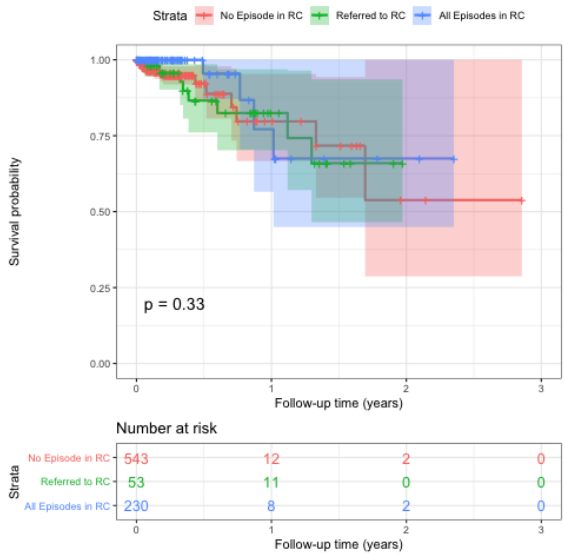

Figure 22 – KM estimators for Testicular Cancer patients with no episode in RC, referred to RC and with all episodes in RC since the recognition of the first testicular oncology RC

5.6.3 Extended Cox Models

According to the described methodology, a stepwise forward selection procedure was implemented to evaluate the variables to be included in the two Extended Cox models. The implementation of an univariate model for each of the variables of interest has given p-values, using the log-rank test, of 0.15 for “All_Episodes_RC” – non statistically significant ($p > 0.05$), and 0.021 for “Referred_to_RC” – statistically significant ($p < 0.05$). Besides these two covariates, all the covariates were found to be statistically significant ($p < 0.05$), except for the number of surgeries and severity covariates. The results from the univariate analysis can be viewed in the table included in Appendix F.

Taking into account the results obtained in the univariate analysis, following the defined methodology, only Model 2 was implemented, with all covariates except the number of surgeries and severity. The results for Model 2 implementation can be seen in Figure 23, with the information about the HRs for the different covariates of the model.

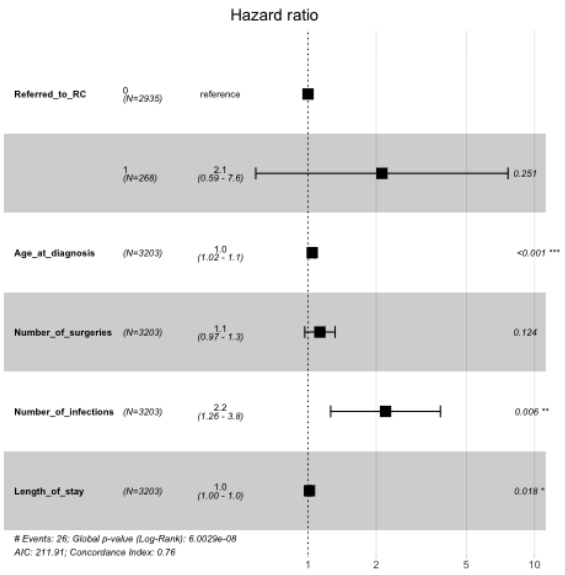

Figure 23 – Testicular Cancer – Multivariate Model 2

Although the covariate “Referred_to_RC” was statistically significant in the univariate analysis, the implementation of the Multivariate Model 2 has shown it not to be statistically significant in the multivariate analysis (p = 0.263). Furthermore, the 95% CI is relatively wide (0.59-7.6). Therefore, no statistically significant conclusion can be extracted from Model 2, regarding the impact on survival for patients referred to testicular oncology RCs, adjusted to other covariates. This can be related with the low mortality associated with testicular cancer patients, when compared to other cancer types described before.

6 Discussion of the results and implications for Public Health Policy

The results from the univariate analysis, using the Kaplan-Meier method, show survival curves with overall better survival probability, at each moment, for patients who have either been referred (i.e. had at least one episode in a RC, with one or more in a non-RC) or patients who had all the hospital episodes in a oncology RC for their cancer type, when compared to the survival probability for patients who had no episode in a RC. These conclusions apply for hepatobiliary, pancreatic, oesophageal and sarcomas cancer types, but are more noticeable in the cases of hepatobiliary and pancreatic cancer. The benefit for these patients, in terms of survival and according to the KM methodology applied, is more evident at earlier stages of the follow-up time, when there is a large number of patients in the groups at risk (i.e. censoring or death of patients has not yet occurred). In the case of testicular cancer and onco-ophthalmology, similar conclusions were not possible to obtain due to the relatively small number of patients included in the dataset and high number of censored patients (i.e. patients who have not died during the follow-up period).

Comparing the survival curves obtained through the KM method, cancer patients who were referred to a RC appear to have better survival probability at each time, when compared with patients who had all episodes in a RC. Since KM methodology produces a univariate analysis, these results may be ignoring the effect of other variables which can affect survival of patients and be different between the two study groups (e.g. severity, number of surgeries, etc.).

Hence, a multivariate analysis using two Extended Cox Models was performed. The results from the multivariate analysis for hepatobiliary and pancreatic cancer have shown a better survival outlook for patients who either had all episodes in a RC (Model 1) and who had at least one episode in a RC (Model 2), when compared to patients who had no episode in a RC. Comparing the results obtained from Models 1 and 2 for hepatobiliary and pancreatic cancer, patients who had all episodes in a RC seem to have better survival prognosis than patients who were referred to a RC, as there is a lower HR for the first group of patients, when compared with the reference group of patients who had no episode in a RC (patients who had no episode in a RC in Model 1 and patients who had no episode in a RC up to that moment – i.e. had not been referred, in the case of Model 2). These results allow to conclude that other covariates (e.g. severity, number of surgeries) may affect overall hepatobiliary and pancreatic cancer patients' survival prognosis, which can explain the difference obtained using the KM method.

For the case of sarcomas and oesophagus cancer patients, only Model 1 was implemented, which compares patients who had all episodes in a RC to patients with no episode in a RC (the application of Model 2 was not possible because the covariate of interest for that Model – “Referred_to_RC” was

found to be statistically non-significant in the univariate analysis for both cases). The results of Model 1 have shown, for both sarcomas and oesophagus cancer types, that patients who had all episodes in a RC have better survival prognostic than patients who had no episode in a RC.

For the cases of onco-ophthalmology and testicular cancer, Models 1 and 2 were not implemented as no statistical significance was obtained beforehand in the univariate analysis for the variables of interest ("All_Episodes_RC" and "Referred_to_RC"). This can be associated with the relatively small number of patients, as well as with the low mortality when compared to other cancer types, making survival analysis more difficult. A similar challenge was found in the survival analysis study of retinoblastoma (related with onco-ophthalmology) developed by Sant et al. [30] – due to low incidence of retinoblastoma and the need for a high number of cases to allow to obtain a reliable survival analysis, data was collected from 28 European population-based cancer registries. In future survival analysis of onco-ophthalmology and testicular cancer, due to their low incidence and relative low mortality, one possible approach may be similar to the one adopted by Sant et al. [30].

Surgeries and hospital infections were manually identified for each cancer type. This was made through analysis of the most frequent diagnosis, the nomenclature for surgical procedures, as well as diagnosis which were potentially associated with a hospital infection. There may be other hospital infections or surgeries which were not considered due to the manual approach and potential associated error. One suggestion arising from this Dissertation, relevant for more rigorous conclusions and corresponding policy implementation, is thus to have these two variables natively coded in the Portuguese DRG database, that is, the number of hospital infections and the number of surgeries per hospital episode.

The official recognition of the first oncology RC happened in 2015, having ever since been recognized other oncology RCs, up to the fifty RCs which exist today in Portugal, for all cancer types. In Table 4, the cancer types analysed in this Dissertation are presented, with information about the number of RC's, the number of cancer patients with hospital episodes after the official recognition of the first RC for their cancer type, as well as the percentage of those patients who were not referred to a RC (i.e. percentage of patients who had no episode in a RC). This table does not include information for other existing oncology RCs in Portugal, such as rectum oncology and paediatric oncology RCs, which were not subject to analysis in this Dissertation.

Table 4 – Percentage of cancer patients not referred to an oncology RC

Cancer Type	# of RC	# of Patients since 1st RC	% of Patients NOT referred since 1st RC
Hepatobiliary	10	7088	49%
Pancreatic	10	5878	55%
Sarcomas	5	1962	53%
Oesophagus	6	2547	55%
Onco-ophthalmology	1	63	62%
Testicular	4	826	66%

Analysing Table 4, there are still close to half or more cancer patients who had no hospital episodes in an oncology RC. Since survival probability can improve for patients referred to a RC, one should look for the reasons for these patients not being referred (e.g. lack of accessibility for patients who live far away from a RC) and propose solutions, from an health management perspective, to increase their accessibility to RCs.

According to the Portuguese National Programme for Oncologic Disease [67], one of the goals for 2020 is to ensure that 75% of the rectum, pancreatic and testicular cancer patients are treated in a RC. Taking into account the results obtained in this Dissertation, the extension of this goal to other cancer types is expected to promote better survival probability for cancer patients, in particular for hepatobiliary, pancreatic, but also for sarcomas and oesophagus cancer patients. Nevertheless, there is still a relevant percentage of these patients who need to be referred to a RC (as can be seen in Table 4), therefore an important effort to achieve this goal is required.

Besides confirming the presumed relationship of survival probability with variables such as age, severity, number of surgeries or number of infections, the current work has also clearly shown the importance of being referred to a RC for cancer patients, especially in the case of hepatobiliary and pancreatic cancer, which have been responsible for approximately 1600 and 1400 new cancer cases in 2018, respectively [6]. An immediate consequence of these results in terms of public health policy is thus the need to improve the network of RCs, as well as patients' access to the services they provide. This can be achieved through investment towards forming and recognizing new oncology RCs, but also by enhancing the referring of cancer patients to existing RCs. A possible strategy to promote the referral of cancer patients can be the creation and recognition of Affiliate Centres in oncology ("*centre which does not comply with the conditions and criteria in order to be officially recognized as national RC but has knowledge on a certain specific area of competencies."* [3, p. 101]*)*, and establish relationship with RCs, to promote and enhance referral pathways.

The annual costs of cancer treatment in Portugal, associated with direct medical costs, account to 867 million euros, representing 5.5% of Portuguese health expenditure. In addition to contributing to

a better survival prognosis for patients, the implementation and promotion of the RC Model can support the achievement of economies of scale and allow to take advantage of the scope and experience economies provided by RCs, which is one of the main objectives for the RC model implementation, as described by Penedo et al. [3]. In addition to contributing to a better quality of life for these patients, such a measure may thus also save money by increasing the efficiency of the treatments and reducing average costs.

7 Conclusions and future work

The ageing of population in Portugal and in Europe will be an important driver for the increase of cancer cases in the next decades. In Portugal, cancer already represents a significant burden due to the associated mortality and morbidity. There has been a relevant increase of healthcare costs associated with cancer, due to new therapies and technologies. Lopes et al. [5, p. 8] mention an increase of 300 million euros in the annual direct medical costs associated to cancer treatment, from 2006 to 2017, “which may be explained by an increase in incidence and the rising cost of drugs “.

With the aim of promoting the concentration of resources and providing differentiated medical care for specific medical conditions, in particular for cancer, following the European Directive 2011/24/EU, Portugal and other European Countries started the implementation of a Reference Centre (RC) Model, with highly-specialized units for treating cancer patients. The first oncology RC was officially recognized in Portugal in 2015. Ever since, several other RCs for different cancer types have been officially recognized, reaching a total of fifty centres, as of today.

Being cancer a disease with overall high mortality, this Dissertation had the objective of studying the RC Model implementation impact on the survival of cancer patients in Portugal, for specific cancer types, by studying patients who had at least one hospital episode in a RC, compared to patients who had no episode in a RC.

This Dissertation analysed survival for six cancer types: hepatobiliary, pancreatic, sarcomas, oesophageal, onco-ophthalmology and testicular cancer. These cancer types have recognized RCs in Portugal, and there were datasets made available by ACSS to perform this research.

The main conclusion from this Dissertation is that, overall, cancer patients who were treated in a RC (at least one or all hospital episodes) present better survival probability at each moment than patients who had no hospital episode in a RC. This is clearer for hepatobiliary and pancreatic cancer patients, but also visible in the case of sarcomas and oesophageal cancer patients (although with wider confidence intervals). In the case of hepatobiliary and pancreatic cancer, the two implemented multivariate models, adjusted for other covariates of interest (e.g. mean severity of the patient, number of surgeries, number of infections, etc.), show a better survival prognosis for patients who had all episodes in a RC than for patients who were referred to a RC. For onco-ophthalmology and testicular cancer, due to the relative small number of patients annually treated in Portugal, it was more difficult to obtain accurate conclusions in terms of survival, based on the applied methodology – there is a relatively low number of patients recorded as dead during the follow-up period, and there is a high number of censored patients.

The conclusions reached support the RC Model implementation for oncology, as the collected data indicates a positive impact in terms of survival for cancer patients who have been referred or had all

episodes in a RC, when compared to patients who had no episodes in a RC. This conclusion is also aligned with the literature for the implementation of the RC Model in different countries [24][68][69][70]. One example mentioned in the literature indicates that, being surgery potentially curative for an important number of cancer types, the increase in hospital volume from centralization of cancer services (which can be potentiated through recognition of RCs) is associated with lower mortality for surgical operated patients [43].

A possible future area of research is the study of the impact on survival of patients treated in RCs with more years of experience – the expectation is that the treatment improves over time, with the increase in volume, resources and experience. This research can also be extended to other cancer types in Portugal, but also to other health intervention areas which have seen the creation of RCs, such as transplantation of solid organs or interventional cardiology and hemodynamics.

The research performed in this Master Dissertation has allowed to analyze and apply a methodology, according to the best practices in the literature, to a real case, which is the implementation of the RC model in Portugal. The application of this methodology yields important information and conclusions to be considered for Public Health Policy, specifically in support of the improvement of cancer patients' survival. This research can be considered a typical biomedical engineering work, as it is focused on using tools and engineering principles to support and promote progress in Medicine and benefits for patients.

8 Bibliography

[1] J. de Almeida Simoes, G. F. Augusto, I. Fronteira, and C. Hernandez-Quevedo, “Portugal: Health System Review.,” *Health Syst. Transit.*, vol. 19, no. 2, pp. 1–184, Mar. 2017.

[2] OECD - European Observatory on Health Systems and Policies, “Portugal: Country Health Profile 2019,” Brussels, 2019.

[3] J. M. Penedo *et al.*, “Reference Centres - Final Report,” 2014.

[4] Direção-Geral da Saúde, “Programa Nacional para as Doenças Oncológicas,” Lisbon, 2017.

[5] J. M. Lopes, F. R. Gonçalves, M. Borges, P. Redondo, and J. Laranja-Pontes, “The cost of cancer treatment in Portugal,” *ecancer*, vol. 11, no. 765, 2017.

[6] Global Cancer Observatory - International Agency for Research on Cancer, “Portugal,” 2019.

[7] N. A. Turiano, “Survival Analysis,” in *The Encyclopedia of Adulthood and Aging*, 2015, pp. 1–5.

[8] D. G. Kleinbaum and M. Klein, *Introduction to Survival Analysis - A Self-Learning Text, Third Edition*. Springer-Verlag New York, 2012.

[9] J. K. Elrod and J. L. Fortenberry, “Centers of excellence in healthcare institutions: What they are and how to assemble them,” *BMC Health Services Research*, vol. 17, no. Suppl 1. BioMed Central Ltd., 11-Jul-2017.

[10] European Commission, “European Reference Networks - Working for patients with rare, low-prevalence and complex diseases,” Luxembourg, 2017.

[11] S. Sandrucci, P. Naredi, and S. Bonvalot, “Centers of excellence or excellence networks: The surgical challenge and quality issues in rare cancers,” *Eur. J. Surg. Oncol.*, vol. 45, no. 1, pp. 19–21, Jan. 2019.

[12] World Health Organization, “WHO report on cancer: setting priorities, investing wisely and providing care for all,” Geneva, 2020.

[13] J. Ferlay *et al.*, “Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012,” *Int. J. Cancer*, vol. 136, no. 5, pp. E359–E386, Mar. 2015.

[14] A. G. Duffy and T. F. Greten, “Treating Hepatobiliary Cancer: The Immunologic Approach,” *Dig. Dis.*, vol. 35, no. 4, pp. 390–396, May 2017.

[15] A. Simão, “The Burden of Hepatobiliary Diseases in Portugal: What Can We Learn from Mortality Data,” *GE - Port. J. Gastroenterol.*, vol. 25, no. 3, pp. 110–111, Apr. 2018.

[16] M. C. Da Rocha, R. T. Marinho, and T. Rodrigues, “Mortality Associated with Hepatobiliary Disease in Portugal between 2006 and 2012,” *GE Port. J. Gastroenterol.*, vol. 25, no. 3, pp. 123–131, Apr. 2018.

[17] L. Huang *et al.*, “Stratified survival of resected and overall pancreatic cancer patients in Europe and the USA in the early twenty-first century: A large, international population-based study,” *BMC Med.*, vol. 16, no. 1, p. 125, Aug. 2018.

[18] M. Wahutu, S. K. Vesely, J. Campbell, A. Pate, A. L. Salvatore, and A. E. Janitz, “Pancreatic Cancer: A Survival Analysis Study in Oklahoma,” *J. Okla. State Med. Assoc.*, vol. 109, no. 7–8, pp. 391–398, Jul. 2016.

[19] P. Marques da Costa, R. Tato Marinho, H. Cortez-Pinto, L. Costa, and J. Velosa, “Twenty-Five Years of Increasing Mortality From Pancreatic Cancer in Portugal,” *Pancreas*, vol. 49, no. 1, pp. e2–e3, Jan. 2020.

[20] H. Forssell, M. Wester, K. Akesson, and S. Johansson, “A proposed model for prediction of survival based on a follow-up study in unresectable pancreatic cancer,” *BMJ Open*, vol. 3, no. 12, p. e004064, Dec. 2013.

[21] H. J. Hoekstra *et al.*, “Adherence to Guidelines for Adult (Non-GIST) Soft Tissue Sarcoma in the Netherlands: A Plea for Dedicated Sarcoma Centers,” *Ann. Surg. Oncol.*, vol. 24, no. 11, pp. 3279–3288, Oct. 2017.

[22] C. A. Stiller *et al.*, “Descriptive epidemiology of sarcomas in Europe: Report from the RARECARE project,” *Eur. J. Cancer*, vol. 49, no. 3, pp. 684–695, Feb. 2013.

[23] G. Mastrangelo *et al.*, “Incidence of soft tissue sarcoma and beyond: A population-based prospective study in 3 European regions,” *Cancer*, vol. 118, no. 21. pp. 5339–5348, 01-Nov-2012.

[24] J. Blay *et al.*, “Surgery in reference centers improves survival of sarcoma patients: a nationwide study.,” *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, vol. 30, no. 7, pp. 1143–1153, 2019.

[25] M. G. Patti, W. Gantert, and L. W. Way, “Surgery of the esophagus: Anatomy and physiology,” *Surg. Clin. North Am.*, vol. 77, no. 5, pp. 959–970, Oct. 1997.

[26] American Cancer Society, “Cancer Facts & Figures 2020,” Atlanta, 2020.

[27] J. Cao *et al.*, “Clinical Nomogram for Predicting Survival of Esophageal Cancer Patients after Esophagectomy,” *Sci. Rep.*, vol. 6, May 2016.

[28] E. Kim, S. Koroukian, and C. R. Thomas, “Conditional survival of esophageal cancer: An analysis from the SEER registry (1988-2011),” *J. Thorac. Oncol.*, vol. 10, no. 10, pp. 1490–1497, Oct. 2015.

[29] J. H. Kauppila, F. Mattsson, N. Brusselaers, and J. Lagergren, “Prognosis of oesophageal adenocarcinoma and squamous cell carcinoma following surgery and no surgery in a nationwide Swedish cohort study,” *BMJ Open*, vol. 8, no. 5, p. e021495, May 2018.

[30] M. Sant *et al.*, “Survival for retinoblastoma in Europe,” *Eur. J. Cancer*, vol. 37, no. 6, pp. 730–735, Apr. 2001.

[31] Union for International Cancer Control, “Retinoblastoma - Review of Cancer Medicines on the WHO List of Essential Medicines,” 2014.

[32] S. Kaliki and C. L. Shields, “Uveal melanoma: Relatively rare but deadly cancer,” *Eye*, vol. 31, no. 2. Nature Publishing Group, pp. 241–257, 01-Feb-2017.

[33] E. S. Rantala, M. Hernberg, and T. T. Kivelä, “Overall survival after treatment for metastatic uveal melanoma,” *Melanoma Res.*, vol. 29, no. 6, pp. 561–568, Dec. 2019.

[34] M. Peixoto *et al.*, “Prognosis and survival in testicular cancer – 10 years in review, a population-based analysis,” *Eur. Urol. Suppl.*, vol. 18, no. 11, p. e3614, Nov. 2019.

[35] A. Rolevich *et al.*, “Trends in incidence, mortality and survival of testicular cancer patients in Belarus,” *Cent. Eur. J. Urol.*, vol. 72, no. 4, pp. 357–368, 2019.

[36] K. B. Zuniga, S. L. Washington, S. P. Porten, and M. V. Meng, “A comparison of stage-specific all-cause mortality between testicular sex cord stromal tumors and germ cell tumors: Results from the National Cancer Database,” *BMC Urol.*, vol. 20, no. 1, p. 40, Apr. 2020.

[37] R. H. A. Verhoeven *et al.*, “Testicular cancer in Europe and the USA: survival still rising among older patients,” *Ann. Oncol.*, vol. 24, no. 2, pp. 508–513, Feb. 2013.

[38] E. L. *et al.*, “Prognostic variables for response and outcome in patients with extragonadal germ-cell tumors,” *Ann. Oncol.*, vol. 13, no. 7, pp. 1017–1028, 2002.

[39] R. Busse, A. Geissler, W. Quentin, and M. Wiley, *Diagnosis-related groups in Europe: moving towards transparency, efficiency and quality in hospitals*. Open University Press, 2011.

[40] R. Busse *et al.*, “Diagnosis related groups in Europe: Moving towards transparency, efficiency, and quality in hospitals?,” *BMJ (Online)*, vol. 347, no. 7916. BMJ Publishing Group, 13-Jul-2013.

[41] P. Barros and G. Braun, “Upcoding in a National Health Service: the evidence from Portugal,”

Health Econ., vol. 26, no. 5, pp. 600–618, May 2017.

[42] ACSS, *Circular Normativa ACSS nº 22/2014/DPS/ACSS*. 2014.

[43] V. H. Coupland *et al.*, “Hospital volume, proportion resected and mortality from oesophageal and gastric cancer: A population-based study in England, 2004-2008,” *Gut*, vol. 62, no. 7, pp. 961–966, Jul. 2013.

[44] G. Gooiker, W. Van Gijn, P. Post, C. J. H. van de Velde, R. Tollenaar, and M. Wouters, “A systematic review and meta-analysis of the volume-outcome relationship in the surgical treatment of breast cancer. Are breast cancer patients better of with a high volume provider?,” *European Journal of Surgical Oncology*, vol. 36, no. SUPPL. 1. Sep-2010.

[45] M. Kamboj and K. A. Sepkowitz, “Nosocomial infections in patients with cancer,” *The Lancet Oncology*, vol. 10, no. 6. pp. 589–597, Jun-2009.

[46] G. Ducel, J. Fabry, and L. Nicolle, “Prevention of hospital acquired infections: a practical guide.,” *Prev. Hosp. Acquir. Infect. a Pract. Guid.*, no. Ed.2, 2002.

[47] P. Cornejo-Juárez, D. Vilar-Compte, C. Pérez-Jiménez, S. A. Ñamendys-Silva, S. Sandoval-Hernández, and P. Volkow-Fernández, “The impact of hospital-acquired infections with multidrug-resistant bacteria in an oncology intensive care unit,” *Int. J. Infect. Dis.*, vol. 31, pp. e31–e34, Feb. 2015.

[48] J. Berkson and R. P. Gage, “Calculation of survival rates for cancer.,” *Proc. Staff Meet. Mayo Clin.*, vol. 25, no. 11, pp. 270–86, May 1950.

[49] F. E. Ahmed, P. W. Vos, and D. Holbert, “Modeling survival in colon cancer: A methodological review,” *Molecular Cancer*, vol. 6, no. 1. BioMed Central, p. 15, 12-Feb-2007.

[50] R. A. Schea, P. Perkins, P. K. Allen, R. Komaki, and J. D. Cox, “Limited-stage small-cell lung cancer: Patient survival after combined chemotherapy and radiation therapy with and without treatment protocols,” *Radiology*, vol. 197, no. 3, pp. 859–862, Dec. 1995.

[51] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. J. Voelker, B. Nussenbaum, and E. W. Wang, “A practical guide to understanding Kaplan-Meier curves,” *Otolaryngol. - Head Neck Surg.*, vol. 143, no. 3, pp. 331–336, 2010.

[52] H. Barraclough, L. Simms, and R. Govindan, “Biostatistics primer: What a clinician ought to know: Hazard ratios,” *J. Thorac. Oncol.*, vol. 6, no. 6, pp. 978–982, Jun. 2011.

[53] T. M. Therneau and P. M. Grambsch, *Modeling survival data: extending the Cox model*. Springer-Verlag New York, 2000.

[54] M. J. Stensrud and M. A. Hernán, “Why Test for Proportional Hazards?,” *JAMA - Journal of the American Medical Association*, vol. 323, no. 14. American Medical Association, pp. 1401–1402, 14-Apr-2020.

[55] T. M. Therneau, “A Package for Survival Analysis in R.” 2020.

[56] T. J. Walsh, M. S. Croughan, M. Schembri, J. M. Chan, and P. J. Turek, “Increased risk of testicular germ cell cancer among infertile men,” *Arch. Intern. Med.*, vol. 169, no. 4, pp. 351–356, Feb. 2009.

[57] E. Lanza *et al.*, “Survival analysis of 230 patients with unresectable hepatocellular carcinoma treated with bland transarterial embolization,” *PLoS One*, vol. 15, no. 1, p. e0227711, Jan. 2020.

[58] A. Kassambara, M. Kosinski, P. Biecek, and S. Fabian, “Survminer: Drawing Survival Curves using ‘ggplot2.’” 2020.

[59] M. Kudo *et al.*, “Survival Analysis over 28 Years of 173,378 Patients with Hepatocellular Carcinoma in Japan,” *Liver Cancer*, vol. 5, no. 3, pp. 190–197, Jul. 2016.

[60] F. Tustumi *et al.*, “Prognostic factors and survival analysis in esophageal carcinoma,” *Arq. Bras. Cir. Dig.*, vol. 29, no. 3, pp. 138–141, Jul. 2016.

[61] S. P. Bagaria, Y.-H. Chang, R. J. Gray, J. B. Ashman, S. Attia, and N. Wasif, “Improving Long-Term Outcomes for Patients with Extra-Abdominal Soft Tissue Sarcoma Regionalization to High-Volume Centers, Improved Compliance with Guidelines or Both?,” *Sarcoma*, vol. 2018, 2018.

[62] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, “Survival analysis part IV: Further concepts and methods in survival analysis,” *British Journal of Cancer*, vol. 89, no. 5. pp. 781–786, 01-Sep-2003.

[63] T. Therneau, C. Crowson, and E. Atkinson, “Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model,” 2020.

[64] D. H. Kim, H. Uno, and L. J. Wei, “Restricted mean survival time as a measure to interpret clinical trial results,” *JAMA Cardiology*, vol. 2, no. 11. American Medical Association, pp. 1179–1180, 01-Nov-2017.

[65] G. Abbas and M. Krasna, “Overview of esophageal cancer,” *Ann. Cardiothorac. Surg.*, vol. 6, no. 2, pp. 131–136, Mar. 2017.

[66] M. F. Chen, P. T. Chen, M. S. Lu, C. P. Lee, and W. C. Chen, “Survival benefit of surgery to patients with esophageal squamous cell carcinoma,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–9, Apr. 2017.

[67] Direção-Geral da Saúde, “Programas de Saúde Prioritários - Metas de Saúde 2020,” Lisbon, 2020.

[68] J. Martin-Broto *et al.*, “Relevance of Reference Centers in Sarcoma Care and Quality Item Evaluation: Results from the Prospective Registry of the Spanish Group for Research in Sarcoma (GEIS),” *Oncologist*, vol. 24, no. 6, pp. e338–e346, Jun. 2019.

[69] I. Ray-Coquard *et al.*, “Improving treatment results with reference centres for rare cancers: Where do we stand?,” *European Journal of Cancer*, vol. 77. Elsevier Ltd, pp. 90–98, 01-May-2017.

[70] R. Leroy *et al.*, “Improved survival in patients with head and neck cancer treated in higher volume centres: A population-based study in Belgium,” *Eur. J. Cancer*, vol. 130, pp. 81–91, May 2020.

Appendix A – Hepatobiliary Cancer Data Set information

For a total of 18865 hepatobiliary cancer patients, which had at least a discharge date between the 1st of January of 2010 and 21st of November of 2019, corresponding to a total of 66561 episodes, a descriptive analysis of the patients and hospital episodes is shown in the following table:

Table 5 - Descriptive analysis for hepatobiliary cancer patients during total follow-up period

Hepatobiliary Cancer Patients and Hospital Episodes Descriptive Analysis						
	Total (%)	**Mean**	**Median**	**Minimum**	**Maximum**	**Std Deviation**
Patient's Sex	-	-	-	-	-	-
Male	12882 (~ 68%)	-	-	-	-	-
Female	5983 (~ 32%)	-	-	-	-	-
Patients' age at diagnosis (years)	-	69.29	70.00	18.00	103.00	12.20
Number of Surgeries per patient	-	0.54	0.00	0.00	9.00	1.01
Number of Hosp. Infections per patient	-	0.25	0.00	0.00	10.00	0.25
Patient's Length of Stay (days)	-	22.57	16.00	0.00	228.00	22.80
Severity per episode	-	-	-	-	-	-
1 – Minor	28100 (~ 42%)	-	-	-	-	-
2 – Moderate	23107 (~ 35%)	-	-	-	-	-
3 – Major	13346 (~ 20%)	-	-	-	-	-
4 – Extreme	2008 (~ 3%)	-	-	-	-	-

At the end of the follow-up period, there are a total of 10203 (54%) patients who have been censored, while 8662 (46%) patients have died.

A descriptive analysis of the most common values for variables of interest associated with hepatobiliary cancer hospital episodes is presented below:

Table 6 - Top discharge status, admission types and episode types for hepatobiliary cancer episodes

	Top 3 Discharge status (# of Episodes ~ %)	Top 3 Admission Types (# of Episodes ~ %)	Top 2 Episode Types (# of Episodes ~%)
1	Discharged to home (55955 ~ 84.07%)	Programmed admission (42488 ~ 63.58%)	Hospital stay (44419 ~ 66.73%)
2	Dead Discharge (8662 ~ 13.01%)	Urgent admission (24032 ~ 35.96%)	Ambulatory (22142 ~ 33.27%)
3	*Other discharge status* (1944 ~ 2.92%)	*Other admission types* (41 ~ 0.06%)	

Table 7 - Top diagnosis per episode and top diagnosed tumours for hepatobiliary cancer episodes (*)

	Top 5 Diagnosis (# of Episodes ~ %)	Top 5 Diagnosed Tumours (# of Episodes ~ %)
1	Admission for anti-neoplastic chemotherapy (22316 ~ 33.53%)	Hepatocellular Carcinoma (HCC) (6862 ~ 10.30%)
2	Primary malignant liver neoplasm (16491 ~ 24.78%)	Adenocarcinoma (4364 ~ 6.56%)
3	Malignant extra-hepatic biliary ducts neoplasm (10139 ~ 15.23%)	Cholangiocarcinoma (3032 ~ 4.56%)
4	Hypertension not specified as malignant or benign (9092 ~ 13.66%)	Metastatic Adenocarcinoma (740 ~ 1.11%)
5	Gallbladder malignant neoplasm (6931 ~ 10.41%)	Unspecified Carcinoma (436 ~ 0.66%)

Table 8 - Top surgeries and hospital infections for hepatobiliary cancer episodes (*)

	Top 5 Surgeries (# of Episodes ~ %)	Top 5 Hospital Infections (# of Episodes ~%)
1	Cholecystectomy (898 ~ 1.35%)	Urinary tract infection, site not specified (1484 ~ 2.22%)
2	Partial hepatectomy (779 ~ 1.17%)	Other postoperative infection (480 ~ 0.72%)
3	Anastomosis of the hepatic channel to the gastrointestinal tract (444 ~ 0.67%)	Infection by Escherichia Coli [E. Coli] NEC and site not specified (468 ~ 0.70%)
4	Proximal Pancreatectomy (397 ~ 0.60%)	Infection by D Group Streptococci (300 ~ 0.45%)
5	Liver Transplantation (366 ~ 0.55%)	Infection by Gram Negative Bacteria (169 ~ 0.25%)

(*) One single episode can have up to 20 different diagnosis and procedures (e.g. surgeries), therefore the percentages in Tables 7 and 8 shall be interpreted as prevalence per episode.

The following table describes the obtained results from applying the forward selection procedure to the hepatobiliary cancer cases. First, a univariate analysis was carried out for each of the predictors, to find those which were significant ($p < 0.05$). All the significant predictors were included in the multivariate analysis models, the Extended Cox Models, while the non-significant predictors were discarded.

Table 9 - Univariate and multivariate Extended Cox Models results for hepatobiliary cancer

	Univariate Analysis			Multivariate Analysis - Model 1			Multivariate Analysis - Model 2		
	beta	**HR (95% CI)**	**p-value**	**Beta**	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**
Age_at_diagnosis	0.02	1.02 (1.01-1.02)	< 0.001	0.01	1.01 (1.01-1.01)	< 0.001	0.01	1.01 (1.01-1.02)	< 0.001
All_Episodes_RC	-0.35	0.71 (0.65-0.76)	< 0.001	-0.57	0.57 (0.52-0.61)	< 0.001	-	-	-
Length_of_stay	0.01	1.01 (1.01-1.02)	< 0.001	0.01	1.01 (1.00-1.01)	< 0.001	0.01	1.01 (1.00-1.01)	< 0.001
Number_of_infections	0.24	1.27 (1.20-1.35)	< 0.001	0.04	1.04 (0.96-1.12)	0.34	0.02	1.02 (0.94-1.11)	0.63
Number_of_surgeries	-0.38	0.69 (0.65-0.73)	< 0.001	-0.45	0.64 (0.60-0.68)	< 0.001	-0.39	0.68 (0.63-0.73)	< 0.001
Referred_to_RC	-0.33	0.72 (0.64-0.81)	< 0.001	-	-	-	-0.43	0.65 (0.57-0.74)	< 0.001
Severity	-	-	-	-	-	-	-	-	-
Severity = 1 (Reference)	1.00	1.00	-	1.00	1.00	-	1.00	1.00	-
Severity = 2	1.19	3.29 (2.80-3.87)	< 0.001	1.20	3.33 (2.79-3.97)	< 0.001	1.14	3.13 (2.62-3.74)	< 0.001
Severity = 3	2.06	7.86 (6.71-9.22)	< 0.001	1.97	7.15 (5.99-8.53)	< 0.001	1.77	5.87 (4.90-7.04)	< 0.001
Severity = 4	2.57	13.02 (10.88-15.57)	< 0.001	2.51	12.29 (10.05-15.03)	< 0.001	2.29	9.83 (7.9-12.25)	< 0.001
Sex	-	-	-	-	-	-	-	-	-
Sex = 1 – Male (Reference)	1.00	1.00	-	-	-	-	-	-	-
Sex = 2 – Female	0.08	1.08 (1.00-1.17)	0.06	-	-	-	-	-	-

Appendix B – Pancreatic Cancer Data Set information

For a total of 14932 pancreatic cancer patients who had at least a discharge date between the 1st of January of 2010 and 22nd of November of 2019, corresponding to a total of 80883 episodes, a descriptive analysis of the patients and hospital episodes is shown in the following table:

Table 10 - Descriptive analysis for pancreatic cancer patients during total follow-up period

Pancreatic Cancer Patients and Hospital Episodes Descriptive Analysis						
	Total (%)	**Mean**	**Median**	**Minimum**	**Maximum**	**Std Deviation**
Patient's Sex	-	-	-	-	-	-
Male	8077 (~ 54%)	-	-	-	-	-
Female	6855 (~ 46%)	-	-	-	-	-
Patients' age at diagnosis (years)	-	70.81	72.00	18.00	107.00	12.11
Number of Surgeries per patient	-	0.51	0.00	0.00	13.00	1.05
Number of Hosp. Infections per patient	-	0.20	0.00	0.00	7.00	0.60
Patient's Length of Stay (days)	-	20.87	16.00	0.00	308.00	20.75
Severity per episode	-	-	-	-	-	-
1 – Minor	50530 (~ 63%)	-	-	-	-	-
2 – Moderate	22131 (~ 27%)	-	-	-	-	-
3 – Major	7344 (~ 9%)	-	-	-	-	-
4 – Extreme	878 (~ 1%)	-	-	-	-	-

At the end of the follow-up period, there are a total of 7658 (51%) patients who have been censored, while 7274 (49%) patients have died.

A descriptive analysis of the most common values for variables of interest associated with pancreatic cancer hospital episodes is presented below:

Table 11 - Top discharge status, admission types and episode types for pancreatic cancer episodes

	Top 3 Discharge status (# of Episodes ~ %)	Top 3 Admission Types (# of Episodes ~ %)	Top 2 Episode Types (# of Episodes ~%)
1	Discharged to home (72220 ~ 89.30%)	Programmed admission (62749 ~ 77.58%)	Ambulatory (48722 ~ 60.24%)
2	Dead Discharge (7274 ~ 8.99%)	Urgent admission (18115 ~ 22.40	Hospital stay (32161 ~ 39.76%)
3	*Other discharge status* (1389 ~ 1.71%)	*Other admission types* (19 ~ 0.02%)	

Table 12 - Top diagnosis per episode and top diagnosed tumours for pancreatic cancer episodes (*)

	Top 5 Diagnosis (# of Episodes ~ %)	Top 5 Diagnosed Tumours (# of Episodes ~ %)
1	Admission for anti-neoplastic chemotherapy (48.485 ~ 60.00%)	Adenocarcinoma, unspecified (6502 ~ 8.04%)
2	Malignant neoplasm of head of pancreas (27053 ~ 33.45%)	Metastatic Adenocarcinoma, unspecified (1794 ~ 2.22%)
3	Malignant neoplasm of pancreas, part unspecified (24080 ~ 29.77%)	Malignant neoplasm (838 ~ 1.04%)
4	Malignant neoplasm of pancreas, unspecified (12847 ~ 15.88%)	Metastatic neoplasm (410 ~ 0.51%)
5	Malignant neoplasm of liver, secondary (8269 ~ 10.22%)	Infiltrating duct carcinoma (408 ~ 0.50%)

Table 13 - Top surgeries and hospital infections for pancreatic cancer episodes during follow-up period (*)

	Top 5 Surgeries (# of Episodes ~ %)	Top 5 Hospital Infections (# of Episodes ~%)
1	Cholecystectomy (615 ~ 0.76%)	Urinary tract infection, site not specified (940 ~ 1.16%)
2	Gastroenterostomy NEC (567 ~ 0.70%)	Infection by Escherichia Coli (E. Coli) NEC and of unspecified site (401 ~ 0.50%)
3	Proximal Pancreatectomy (563 ~ 0.70%)	Other post-operation infection (330 ~ 0.41%)
4	Anastomosis of the Hepatic channel to the gastrointestinal tract (469 ~ 0.58%)	Streptococcus infection NEC and site not specified (144 ~ 0.18%)
5	Radical Pancreaticoduodenectomy (367 ~ 0.45%)	Infection following a procedure, initial encounter (88 ~ 0.11%)

(*) One single episode can have up to 20 different diagnosis and procedures (e.g. surgeries), therefore, the percentages in Tables 12 and 13 shall be interpreted as prevalence per episode.

The following table describes the obtained results from applying the forward selection procedure to the pancreatic cancer cases. First, a univariate analysis was carried out for each of the predictors, to find those which were significant ($p < 0.05$). All the significant predictors were included in the multivariate analysis models, the Extended Cox Models, while the non-significant predictors were discarded.

Table 14 - Univariate and multivariate Extended Cox Model results for pancreatic cancer

	Univariate Analysis			Multivariate Analysis - Model 1			Multivariate Analysis - Model 2		
	beta	**HR (95% CI)**	**p-value**	**Beta**	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**
Age_at_diagnosis	0.03	1.03 (1.03-1.04)	< 0.001	0.02	1.02 (1.01-1.02)	< 0.001	0.02	1.02 (1.01-1.02)	< 0.001
All_Episodes_RC	-0.15	0.86 (0.79-0.93)	< 0.001	-0.30	0.74 (0.68-0.81)	< 0.001	-	-	-
Length_of_stay	0.01	1.01 (1.01-1.01)	< 0.001	< 0.001	1.00 (1.00-1.00)	0.63	< 0.001	1.00 (1.00-1.00)	< 0.001
Number_of_infections	0.14	1.15 (1.07-1.24)	< 0.001	-0.04	0.96 (0.88-1.05)	0.35	-0.04	0.96 (0.87-1.06)	0.42
Number_of_surgeries	-0.34	0.71 (0.67-0.75)	< 0.001	-0.41	0.66 (0.62-0.7)	< 0.001	-0.48	0.62 (0.57-0.67)	< 0.001
Referred_to_RC	-0.25	0.78 (0.68-0.89)	< 0.001	-	-	-	-0.21	0.81 (0.70-0.94)	< 0.001
Severity	-	-	-	-	-	-	-	-	-
Severity = 1 (Reference)	1.00	1.00	-	1.00	1.00	-	1.00	1.00	-
Severity = 2	1.83	6.26 (5.35-7.34)	< 0.001	1.79	6.02 (5.08-7.12)	< 0.001	1.75	5.77 (4.81-6.93)	< 0.001
Severity = 3	2.64	14.01 (11.94-16.44)	< 0.001	2.64	14.01 (11.76-16.68)	< 0.001	2.58	13.24 (10.94-16.02)	< 0.001
Severity = 4	3.01	20.33 (16.70-24.76)	< 0.001	3.29	26.86 (21.60-33.41)	< 0.001	2.81	16.64 (12.70-21.79)	< 0.001
Sex	-	-	-	-	-	-	-	-	-
Sex = 1 – Male (Reference)	1.00	1.00	-	-	-	-	-	-	-
Sex = 2 – Female	-0.02	0.98 (0.91-1.06)	0.62	-	-	-	-	-	-

Appendix C – Sarcomas Data Set information

For a total of 6332 sarcoma patients who had at least a discharge date between the 1st of January of 2010 and 5th of November of 2019, corresponding to a total of 31901 episodes, a descriptive analysis of the patients and hospital episodes is shown in the following table:

Table 15 - Descriptive analysis for sarcoma patients during total follow-up period

Sarcoma Patients and Hospital Episodes Descriptive Analysis						
	Total (%)	**Mean**	**Median**	**Minimum**	**Maximum**	**Std Deviation**
Patient's Sex	-	-	-	-	-	-
Male	3402 (~ 54%)	-	-	-	-	-
Female	2930 (~ 46%)	-	-	-	-	-
Patients' age at diagnosis (years)	-	61.06	63.00	18.00	101.00	17.91
Number of Surgeries per patient	-	1.12	1.00	0.00	20.00	1.56
Number of Hosp. Infections per patient	-	0.21	0.00	0.00	10.00	0.72
Patient's Length of Stay (days)	-	20.20	8.00	0.00	711.00	36.55
Severity per episode	-	-	-	-	-	-
1 – Minor	14504 (~ 45%)	-	-	-	-	-
2 – Moderate	14955 (~ 47%)	-	-	-	-	-
3 – Major	2217 (~ 7%)	-	-	-	-	-
4 – Extreme	225 (~ 1%)	-	-	-	-	-

At the end of the follow-up period, there are a total of 5178 (82%) patients who have been censored, while 1154 (18%) patients have died.

A descriptive analysis of the most common values for variables of interest associated with sarcomas hospital episodes is presented below:

Table 16 - Top discharge status, admission types and episode types for sarcomas episodes

	Top 3 Discharge status (# of Episodes ~ %)	Top 3 Admission Types (# of Episodes ~ %)	Top 2 Episode Types (# of Episodes ~%)
1	Discharged to home (30207 ~ 94.69%)	Programmed admission (28447 ~ 89.17%)	Hospital stay (25089 ~ 78.65%)
2	Dead Discharge (1154 ~ 3.62%)	Urgent admission (3357 ~ 10.53%)	Ambulatory (6812 ~ 21.35%)
3	*Other discharge status* (540 ~ 1.69%)	*Other admission types* (97 ~ 0.31%)	

Table 17 - Top diagnosis per episode and top diagnosed tumours for sarcomas episodes (*)

	Top 5 Diagnosis (# of Episodes ~ %)	Top 5 Diagnosed Tumours (# of Episodes ~ %)
1	Admission for radio-therapy (11496 ~ 36.04%)	Sarcoma, unspecified (1996 ~ 6.26%)
2	Admission for anti-neoplastic chemotherapy (10161 ~ 31.85%)	Leiomyosarcoma, unspecified (1074 ~ 3.37%)
3	Malignant neoplasm of connective and other soft tissue, site unspecified (5838 ~ 18.30%)	Lipomyxosarcoma (1042 ~ 3.27%)
4	Malignant neoplasm of connective and other soft tissue of lower limb, including hip (5060 ~ 15.86%)	Osteosarcoma, unspecified(928 ~ 2.91%)
5	Malignant neoplasm of bone and articular cartilage, site unspecified (2649 ~ 8.30%)	Ewig Sarcoma (920 ~ 2.88%)

Table 18 - Top surgeries and hospital infections for sarcomas episodes (*)

	Top 5 Surgeries (# of Episodes ~ %)	Top 5 Hospital Infections (# of Episodes ~%)
1	Excision of lesion of other soft tissue (899 ~ 2.82%)	Urinary tract infection, site not specified (357 ~ 1.11%)
2	Radical excision of skin lesion (542 ~ 1.70%)	Other post-operation infection (148 ~ 0.46%)
3	Other excision of soft tissue (182 ~ 0.57%)	Infection by Escherichia Coli (E. Coli) NEC and of unspecified site (96 ~ 0.30%)
4	Attachment of pedicle or flap graft to other sites (167 ~ 0.52%)	Infection by Pseudomonas of unspecified site (81 ~ 0.25%)
5	Full-thickness skin graft to other sites (125 ~ 0.39%)	Infection and inflammatory react. due to vascular device, implant, and graft (43 ~ 0.13%)

(*) One single episode can have up to 20 different diagnosis and procedures (e.g. surgeries), therefore, the percentages in Tables 17 and 18 shall be interpreted as prevalence per episode.

The following table describes the obtained results from applying the forward selection procedure to the sarcoma cases. First, a univariate analysis was carried out for each of the predictors, to find those which were significant ($p < 0.05$). The predictor "Referred_to_RC" was found to be non-statically significant in the univariate analysis, therefore Model 2 was not implemented.

Table 19 - Univariate and multivariate Extended Cox Model results for sarcomas

	Univariate Analysis			Multivariate Analysis - Model 1 (*)			Multivariate Analysis - Model 2 (*)		
	beta	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**
Age_at_diagnosis	0.03	1.03 (1.02-1.03)	< 0.001	0.02	1.02 (1.01-1.02)	< 0.001	-	-	-
All_Episodes_RC	-0.76	0.47 (0.37-0.59)	< 0.001	-0.50	0.60 (0.46-0.79)	< 0.001	-	-	-
Length_of_stay	0.01	1.01 (1.00-1.01)	< 0.001	< 0.001	1.00 (1.00-1.00)	< 0.001	-	-	-
Number_of_infections	0.40	1.49 (1.29-1.73)	< 0.001	0.11	1.12 (0.91-1.36)	0.28	-	-	-
Number_of_surgeries	-0.23	0.80 (0.72-0.88)	< 0.001	-0.22	0.81 (0.73-0.89)	< 0.001	-	-	-
Referred_to_RC	-0.11	0.90 (0.56-1.43)	0.64	-	-	-	-	-	-
Severity	-	-	-	-	-	-	-	-	-
Severity = 1 (Reference)	1.0	1.0	-	1.0	1.0	-	-	-	-
Severity = 2	1.08	2.94 (1.74-4.97)	< 0.001	0.93	2.55 (1.50-4.31)	< 0.001	-	-	-
Severity = 3	2.39	10.90 (6.51-18.26)	< 0.001	2.07	7.93 (4.66-13.50)	< 0.001	-	-	-
Severity = 4	2.66	14.36 (7.78-26.49)	< 0.001	2.33	10.23 (5.41-19.35)	< 0.001	-	-	-
Sex	-	-	-	-	-	-	-	-	-
Sex = 1 – Male (Reference)	1.0	1.0	-	-	-	-	-	-	-
Sex = 2 – Female	-0.12	0.89 (0.71-1.1)	0.29	-	-	-	-	-	-

(*) Since the predictor "Referred_to_RC" was not statistically significant ($p > 0.05$) in Univariate Analysis, the Multivariate Model 2 was not implemented.

Appendix D – Oesophagus Cancer Data Set information

For a total of 7572 oesophagus cancer patients who had at least a discharge date between the 1st of January of 2010 and 25th of November of 2019, corresponding to a total of 47349 episodes, a descriptive analysis of the patients and hospital episodes is shown in the following table:

Table 20 - Descriptive analysis for oesophagus cancer patients during total follow-up period

Oesophagus Cancer Episodes and Hospital Episodes Descriptive Analysis						
	Total (%)	**Mean**	**Median**	**Minimum**	**Maximum**	**Std Deviation**
Patient's Sex	-	-	-	-	-	-
Male	6343 (~ 84%)	-	-	-	-	-
Female	1229 (~ 16%)	-	-	-	-	-
Patients' age at diagnosis (years)	-	65.16	64.00	18.00	99.00	12.33
Number of Surgeries per patient	-	0.47	0.00	0.00	13.00	1.09
Number of Hosp. Infections per patient	-	0.18	0.00	0.00	7.00	0.55
Patient's Length of Stay (days)	-	21.80	15.00	0.00	233.00	24.89
Severity per episode	-	-	-	-	-	-
1 – Minor	31859 (~ 67%)	-	-	-	-	-
2 – Moderate	10680 (~ 23%)	-	-	-	-	-
3 – Major	4145 (~ 9%)	-	-	-	-	-
4 – Extreme	665 (~ 1%)	-	-	-	-	-

At the end of the follow-up period, there are a total of 23630 (62%) patients who have been censored, while 14474 (38%) patients have died.

A descriptive analysis of the most common values for variables of interest associated with oesophagus cancer hospital episodes is presented below:

Table 21 - Top discharge status, admission types and episode types for oesophagus cancer episodes

	Top 3 Discharge status (# of Episodes ~ %)	Top 3 Admission Types (# of Episodes ~ %)	Top 2 Episode Types (# of Episodes ~%)
1	Discharged to home (43380 ~ 91.62%)	Programmed admission (38791 ~ 81.93%)	Hospital stay (30214 ~ 63.81%)
2	Dead Discharge (3116 ~ 6.58%)	Urgent admission (8530 ~ 18.01%)	Ambulatory (17135 ~ 36.19%)
3	*Other discharge status* (853 ~ 1.80%)	*Other admission types* (28 ~ 0.06%)	

Table 22 – Top diagnosis per episode and top diagnosed tumours for oesophagus cancer episodes (*)

	Top 5 Diagnosis (# of Episodes ~ %)	Top 5 Diagnosed Tumours (# of Episodes ~ %)
1	Admission for anti-neoplastic chemotherapy (18565 ~ 39.2%)	Squamous cell carcinoma, site not specified (4540 ~ 9.59%)
2	Malignant neoplasm of oesophagus, site not specified (14213 ~ 30.02%)	Adenocarcinoma, site not specified (1410 ~2.98%)
3	Admission for radio-therapy (13891 ~ 29.34%)	Malignant neoplasm (338 ~ 0.71%)
4	Malignant neoplasm of lower third of oesophagus (8366 ~ 17.67%)	Carcinoma, not specified (318 ~0.67%)
5	Malignant neoplasm of middle third of oesophagus (7065 ~ 14.92%)	Metastasized squamous cell carcinoma, site not specified (278 ~0.59%)

Table 23 - Top surgeries and hospital infections for oesophagus cancer episodes (*)

	Top 5 Surgeries (# of Episodes ~ %)	Top 5 Hospital Infections (# of Episodes ~%)
1	Partial esophagectomy (333 ~ 0.70%)	Urinary tract infection, site not specified (259 ~ 0.54%)
2	Total esophagectomy (306 ~ 0.65%)	Other post-operation infection (66 ~ 0.14%)
3	Intrathoracic Esophagogastrostomy (155 ~ 0.33%)	Acute lower respiratory tract infection (65 ~0.14%)
4	Radical excision of other lymph nodes (97 ~ 0.20%)	Infection by Escherichia Coli (E. Coli) NEC and of unspecified site (63 ~ 0.13%)
5	Cholecystectomy (81 ~ 0.17%)	Infection by Pseudomonas of unspecified site (59 ~ 0.12%)

(*) One single episode can have up to 20 different diagnosis and procedures (e.g. surgeries), therefore, the percentages in Tables 22 and 33 shall be interpreted as prevalence per episode.

The following table describes the obtained results from applying the forward selection procedure to the oesophagus cancer cases. First, a univariate analysis was carried out for each of the predictors, to find those which were significant ($p < 0.05$). The predictor "Referred_to_RC" was found to be non-statically significant in the univariate analysis, therefore Model 2 was not implemented.

Table 24 - Univariate and multivariate Extended Cox Model results for oesophagus cancer

	Univariate Analysis			Multivariate Analysis - Model 1 (*)			Multivariate Analysis - Model 2 (*)		
	beta	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**
Age_at_diagnosis	0.023	1.02 (1.02-1.03)	< 0.001	0.02	1.02 (1.01-1.02)	< 0.001	-	-	-
All_Episodes_RC	-0.3	0.74 (0.64-0.86)	< 0.001	-0.44	0.64 (0.55-0.75)	< 0.001	-	-	-
Length_of_stay	0.02	1.02 (1.01-1.02)	< 0.001	< 0.001	1.00 (1.00-1.01)	0.03	-	-	-
Number_of_infections	0.55	1.74 (1.56-1.94)	< 0.001	0.05	1.05 (0.92-1.21)	0.48	-	-	-
Number_of_surgeries	0.07	1.07 (1.00-1.14)	0.035	-0.14	0.87 (0.80-0.94)	< 0.001	-	-	-
Referred_to_RC	0.05	1.05 (0.87-1.26)	0.63	-	-	-	-	-	-
Severity	-	-	-	-	-	-	-	-	-
Severity = 1 (Reference)	1.00	1.00	-	1.00	1.00	-	-	-	-
Severity = 2	1.82	6.17 (4.75-8.03)	< 0.001	1.76	5.80 (4.37-7.70)	< 0.001	-	-	-
Severity = 3	2.78	16.20 (12.52-20.96)	< 0.001	2.76	15.87 (11.93-21.11)	< 0.001	-	-	-
Severity = 4	2.94	18.92 (13.96-25.64)	< 0.001	3.02	20.51 (14.30-29.40)	< 0.001	-	-	-
Sex	-	-	-	-	-	-	-	-	-
Sex = 1 – Male (Reference)	1.00	1.00	-	1.00	1.00	-	-	-	-
Sex = 2 – Female	0.28	1.33 (1.10-1.60)	0.003	0.14	1.15 (0.94-1.42)	0.18	-	-	-

(*) Since the predictor "Referred_to_RC" was not statistically significant ($p > 0.05$) in Univariate Analysis, the Multivariate Model 2 was not implemented.

Appendix E – Onco-ophthalmology Data Set information

For a total of 277 onco-ophthalmology patients who had at least a discharge date between the 25th of January of 2010 and 6th of November of 2019, corresponding to a total of 697 episodes, a descriptive analysis of the patients and hospital episodes is shown in the following table:

Table 25 - Descriptive analysis for onco-ophthalmology patients during total follow-up period

Onco-ophthalmology Patients and Hospital Episodes Descriptive Analysis						
	Total (%)	**Mean**	**Median**	**Minimum**	**Maximum**	**Std Deviation**
Patient's Sex	-	-	-	-	-	-
Male	141 (~51%)	-	-	-	-	-
Female	136 (~49%)	-	-	-	-	-
Patients' age at diagnosis (years)	-	67.09	69.00	18.00	98.00	16.46
Number of Surgeries per patient	-	1.13	1.00	0.00	10.00	1.48
Number of Hosp. Infections per patient	-	0.08	0.00	0.00	3.00	0.35
Patient's Length of Stay (days)	-	7.48	3.00	0.00	93.00	11.82
Severity per episode	-	-	-	-	-	-
1 – Minor	519 (~ 74%)	-	-	-	-	-
2 – Moderate	131 (~ 19%)	-	-	-	-	-
3 – Major	42 (~ 6%)	-	-	-	-	-
4 – Extreme	5 (~ 1%)	-	-	-	-	-

At the end of the follow-up period, there are a total of 246 (89%) patients who have been censored, while 31 (11%) patients have died.

A descriptive analysis of the most common values for variables of interest associated with onco-ophthalmology hospital episodes is presented below:

Table 26 - Top discharge status, admission types and episode types for onco-ophthalmology episodes

	Top 3 Discharge status (# of Episodes ~ %)	Top 3 Admission Types (# of Episodes ~ %)	Top 2 Episode Types (# of Episodes ~%)
1	Discharged to home (659 ~ 94.55%)	Programmed admission (594 ~ 85.22%)	Hospital stay (597 ~ 85.65%)
2	Dead Discharge (31 ~ 4.45%)	Urgent admission (102 ~ 14.63%)	Ambulatory (100 ~ 14.35%)
3	*Other discharge status* (7 ~ 1.00 %)	*Other admission types* (1~ 0.14%)	

Table 27 - Top diagnosis per episode and top Diagnosed tumours for onco-ophthalmology episodes (*)

	Top 5 Diagnosis (# of Episodes ~ %)	Top 5 Diagnosed Tumours (# of Episodes ~ %)
1	Malignant orbit neoplasm (374 ~ 53.66%)	Malignant neoplasm (55 ~ 7.89%)
2	Admission for radio-therapy (350 ~ 50.22%)	Malignant lymphoma (44 ~ 6.31%)
3	Malignant eye neoplasm, site not specified (229 ~ 32.86%)	Squamous cell carcinoma, site not specified (33 ~ 4.73%)
4	Admission for anti-neoplastic chemotherapy (80 ~ 11.48%)	Malignant melanoma (30 ~ 4.30%)
5	Malignant retinal neoplasm (41 ~ 5.88%)	Adenocarcinoma, site not specified (14 ~ 2.01%)

Table 28 - Top surgeries and hospital infections for onco-ophthalmology episodes (*)

	Top 5 Surgeries (# of Episodes ~ %)	Top 5 Hospital Infections (# of Episodes ~%)
1	Excision of lesion of orbit (42 ~ 6.03%)	Urinary tract infection, site not specified (11 ~ 3.97%)
2	Orbitotomy, NEC (41 ~ 5.88%)	Infection by Escherichia Coli (E. Coli) NEC and of unspecified site (3 ~ 0.43%)
3	Biopsy of eyeball and orbit (29 ~ 4.16%)	Other post-operation infection (2 ~ 0.29%)
4	Exenteration of orbital contents, NEC (17 ~ 2.44%)	Infection with drug-resistant microorganisms with multiple drug resistance (1 ~ 0.14%)
5	Exenteration of orbit with removal of adjacent structures (11 ~ 1.58%)	Infection by Pseudomonas of unspecified site (1 ~ 0.14%)

(*) One single episode can have up to 20 different diagnosis and procedures (e.g. surgeries), therefore, the percentages in Tables 27 and 28 shall be interpreted as prevalence per episode.

The following table describes the obtained results from applying the forward selection procedure to the onco-ophthalmology cancer cases. The results of the univariate analysis are show below – but, one were unable to obtain results for the covariate “All_Episodes_RC”, “Referred_to_RC” and “Severity”, due to the low number of cases in some of the covariate values (e.g. “All_Episodes_RC = 1”, “Referred_to_RC=1”, “Severity = 4”). The results of the univariate analysis were not possible for the two covariates “Referred_to_RC” and “All_Episodes_RC”, therefore the multivariate models were not implemented for this cancer type.

Table 29 - Univariate and multivariate Extended Cox Models results for onco-ophthalmology

	Univariate Analysis			Multivariate Analysis - Model 1 (*)			Multivariate Analysis - Model 2 (*)		
	beta	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**
Age_at_diagnosis	0.04	1.04 (1.01-1.08)	0.012	-	-	-	-	-	-
All_Episodes_RC	*N/A*	*N/A*	*N/A*	-	-	-	-	-	-
Length_of_stay	0.01	1.01 (1.00-1.03)	0.06	-	-	-	-	-	-
Number_of_infections	0.60	1.83 (1.10-3.04)	0.02	-	-	-	-	-	-
Number_of_surgeries	-0.16	0.85 (0.66-1.11)	0.24	-	-	-	-	-	-
Referred_to_RC	*N/A*	*N/A*	*N/A*	-	-	-	-	-	-
Severity	-	-	-	-	-	-	-	-	-
Severity = 1 (Reference)	1.00	1.00	-	-	-	-	-	-	-
Severity = 2	*N/A*	*N/A*	*N/A*	-	-	-	-	-	-
Severity = 3	*N/A*	*N/A*	*N/A*	-	-	-	-	-	-
Severity = 4	*N/A*	*N/A*	*N/A*	-	-	-	-	-	-
Sex	-	-	-	-	-	-	-	-	-
Sex = 1 – Male (Reference)	1.00	1.00	-	-	-	-	-	-	-
Sex = 2 – Female	-0.38	0.68 (0.33-141)	0.30	-	-	-	-	-	-

(*) Since the predictors “All_Episodes_RC” and “Referred_to_RC” had not enough cases for univariate analysis to be conducted, the Multivariate Model 1 and 2 were not implemented.

Appendix F – Testicular Cancer Data Set information

For a total of 2269 testicular cancer patients who had at least a discharge date between the 1st of January of 2010 and 25th of November of 2019, corresponding to a total of 15014 episodes, a descriptive analysis of the patients and hospital episodes is shown in the following table:

Table 30 - Descriptive analysis for testicular cancer patients during total follow-up period

Testicular Cancer Patients and Hospital Episodes Descriptive Analysis						
	Total (%)	**Mean**	**Median**	**Minimum**	**Maximum**	**Std Deviation**
Patient's Sex	-	-	-	-	-	-
Male	2269 (100%)	-	-	-	-	-
Female	0 (0%)	-	-	-	-	-
Patients' age at diagnosis (years)	-	38.34	34.00	18.00	95.00	15.55
Number of Surgeries per patient	-	1.52	1.00	0.00	13.00	1.25
Number of Hosp. Infections per patient	-	0.06	0.08	0.00	5.00	0.35
Patient's Length of Stay (days)	-	9.02	3.00	0.00	306.00	17.79
Severity per episode	-	-	-	-	-	-
1 – Minor	12903 (~ 85%)	-	-	-	-	-
2 – Moderate	1799 (~ 12%)	-	-	-	-	-
3 – Major	262 (~ 2%)	-	-	-	-	-
4 – Extreme	50 (~ 1%)	-	-	-	-	-

At the end of the follow-up period, there are a total of 2162 (95%) patients who have been censored, while 107 (5%) patients have died.

A descriptive analysis of the most common values for variables of interest associated with testicular cancer hospital episodes is presented below:

Table 31 - Top discharge status, admission types and episode types for testicular cancer episodes

	Top 3 Discharge status (# of Episodes ~ %)	Top 3 Admission Types (# of Episodes ~ %)	Top 2 Episode Types (# of Episodes ~%)
1	Discharged to home (14789 ~ 98.50%)	Programmed admission (14145 ~ 94.21%)	Ambulatory (9797 ~ 65.25%)
2	Dead Discharge (107 ~ 0.71%)	Urgent admission (780 ~ 5.20%)	Hospital stay (5217 ~ 34.75%)
3	*Other discharge status* (118 ~ 0.79%)	*Other admission types* (89~ 0.59%)	

Table 32 - Top diagnosis per episode and top diagnosed tumours for testicular cancer episodes (*)

	Top 5 Diagnosis (# of Episodes ~ %)	Top 5 Diagnosed Tumours (# of Episodes ~ %)
1	Admission for anti-neoplastic chemotherapy (10865 ~ 72.37%)	Seminoma (1187 ~ 7.91%)
2	Malignant neoplasm of other and unspecified testis (10202 ~ 67.95%)	Carcinoma, embryonal (370 ~ 2.46%)
3	Malignant neoplasm of unspecified undescended testis (1391 ~9.26%)	Transitional cell carcinoma (285 ~ 1.90%)
4	Admission for radio-therapy (1224 ~ 8.15%)	Endodermal sinus tumour (264 ~ 1.76%)
5	Malignant neoplasm of unsp testis, unsp descended or undescended (1216 ~ 8.10%)	Malignant neoplasm (174 ~ 1.16%)

Table 33 - Top surgeries and hospital infections for testicular cancer episodes (*)

	Top 5 Surgeries (# of Episodes ~ %)	Top 5 Hospital Infections (# of Episodes ~%)
1	Unilateral orchiectomy (1143 ~ 7.61%)	Urinary tract infection, site not specified (32 ~ 0.20%)
2	Insertion of testicular prosthesis (694 ~ 4.62%)	Infection by Pseudomonas of unspecified site (10 ~ 0.07%)
3	Resection of right testis, open approach (198 ~ 1.32%)	Infection and inflammatory reaction due to vascular device, implant, and graft (7 ~0.05%)
4	Resection of left testis, open approach (167 ~ 1.11%)	Streptococcus infection in conditions classified elsewhere and of unspec. site (5 ~0.03%)
5	Replacement of Right Testis with Synthetic Substitute (155 ~ 1.03%)	Other post-operation infection (5 ~ 0.03%)

(*) One single episode can have up to 20 different diagnosis and procedures (e.g. surgeries), therefore, the percentages in Tables 32 and 33 shall be interpreted as prevalence per episode.

The following table describes the obtained results from applying the forward selection procedure to the testicular cancer cases. First, a univariate analysis was carried out for each of the predictors, to find those which were significant ($p < 0.05$). Severity and All_Episodes_RC were found to be statistically non-significant in univariate analysis. Thus, Multivariate Model 2 was the only multivariate model implemented.

Table 34 - Univariate and Multivariate Extended Cox Model results for testicular cancer

	Univariate Analysis			Multivariate Analysis - Model 1 (*)			Multivariate Analysis - Model 2		
	beta	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**	**beta**	**HR (95% CI)**	**p-value**
Age_at_diagnosis	0.03	1.03 (1.02-1.05)	< 0.001	-	-	-	0.04	1.04 (1.02-1.07)	< 0.001
All_Episodes_RC	-0.78	0.46 (0.16-1.32)	0.15	-	-	-	-	-	-
Length_of_stay	0.02	1.02 (1.02-1.03)	< 0.001	-	-	-	0.01	1.01 (1.00-1.03)	0.02
Number_of_infections	1.16	3.18 (2.01-5.02)	< 0.001	-	-	-	0.79	2.20 (1.26-3.84)	0.01
Number_of_surgeries	0.30	1.36 (1.17-1.57)	< 0.001	-	-	-	-0.01	1.13 (0.97-1.32)	0.12
Referred_to_RC	1.03	2.81 (1.17-6.72)	0.02	-	-	-	0.75	2.12 (0.59-7.63)	0.25
Severity	-	-	-	-	-	-	-	-	-
Severity = 1 (Reference)	1.00	1.00	-	-	-	-	-	-	-
Severity = 2	17.78	+ ∞	0.99	-	-	-	-	-	-
Severity = 3	19.02	+ ∞	0.99	-	-	-	-	-	-
Severity = 4	19.91	+ ∞	0.99	-	-	-	-	-	-
Sex	-	-	-	-	-	-	-	-	-
Sex = 1 – Male (Reference)	-	-	-	-	-	-	-	-	-

(*) Since the predictor “All_Episodes_RC” was not statistically significant in univariate analysis, the Multivariate Model 1 was not implemented.