UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE,
ECONOMICHE E SOCIALI

Corso di Laurea Magistrale in Data Science and Economics

Classe n. LM-91: Tecniche e Metodi per la Società dell'Informazione

Survival Analysis on the SEER Breast Cancer Dataset: When Does Machine Learning Become Useful?

Relatore:

Chiar.mo Prof. Federico Ambrogi

Correlatrice:

Chiar.ma Prof.ssa Silvia Salini

Tesi di Laurea Magistrale

Lirida Papallazi

Matricola n. 938943

Anno Accademico 2020-2021

List of Tables

List of Figures

Contents

Ringraziamenti

Prima di procedere con la trattazione, vorrei dedicare questo spazio a tutti coloro che mi sono stati vicini in questo percorso di crescita professionale, personale ma soprattutto umano.

Un enorme grazie al mio relatore, il professor Federico Ambrogi per la sua disponibilità, il suo incoraggiamento, il suo supporto, la sua tempestività nel rispondere ad ogni mio dubbio con enorme pazienza e chiarezza. Grazie per la sua costante presenza nonostante la distanza forzata a causa della pandemia.

Grazie a lui e alla mia correlatrice, la professoressa Silvia Salini, perché sono stati in grado di tramsettermi la propria passione per la materia in primis ma soprattutto per il proprio lavoro e grazie per aver rafforzato la stima che nutro nella figura del professore, che aiuta e accompagna i propri studenti nel proprio percorso di studi ma anche di vita. Grazie a tutti i professori incontrati durante questo difficile ma stupendo percorso perché mi hanno insegnato veramente tanto.

Grazie ai miei genitori per gli innumerevoli sacrifici fatti per garantire a me e a mio fratello una vita più agiata e meno sofferta della loro. Grazie per tutto ciò che mi avete insegnato e per i valori che mi avete trasmesso: non sarei nessuno senza di voi.

Grazie al mio fratellino che spesso e volentieri è più un fratellone, per il supporto e per credere in me più di quanto ci creda io.

Grazie al mio amore Simo, sei veramente la cosa più bella che mi sia successa dopo tanti momenti tristi e non smetterò mai di dirtelo e di ringraziarti per questo. Grazie anche alla tua famiglia per tutto l'amore ed il rispetto che mi dimostrano ogni volta.

Grazie alle mie ragazze, Paola, Giulia ed Elena (e alle loro famiglie) per non avermi mai abbandonata e per esserci sempre state per me e sappiate che ci sarò

sempre anche io per voi.

Grazie ai miei ragazzi, Enrico (sei il mio cugino preferito, dai che lo sai), Andrea, Fabio e Simo perché siete dei veri amici ma soprattutto delle bravissime persone.

Grazie Ale e alla mia oramai ex coinquilina Marghe perché vi siete dimostrate delle splendide amiche e delle splendide e coraggiose persone.

Grazie ai miei compagni di università: Ali, Mary, Cri, Simo Quadro, Simo Barba, Andre, Thomas e Nico, Carlo e Marco, Ricardo e Andrea Adami, che hanno condiviso con me questo splendido ma difficile percorso. Non avrei potuto desiderare dei compagni di viaggio migliori.

Grazie alla mia famiglia e agli amici tutti per l'affetto e la stima che provate per me e che io provo per voi.

Grazie a chi ha sempre creduto in me e a chi non l'ha fatto.

Semplicemente Grazie. A tutti.

Abstract

Breast cancer is the second leading cause of mortality, after lung cancer, in women. According to the American Cancer Society, the incidence rate increases by 0.5% every year [1]. An early diagnosis is essential for the course of the disease for this reason periodic checks are essential.

Survival analysis allows us to estimate the effects of tumor prognostic factors on the survival of women affected by this disease.

Survival analysis, or time to event analysis is a branch of statistics that studies the time it takes for an event of interest to occur. One of the most peculiar characteristics of survival analysis is the so-called censoring which can occur, for example, when a subject leaves the trial. The analytical methods of survival analysis are based on the survival function which describes the probability that an individual survives up to time t. Among the more classical methods we find that of Kaplan-Meier, Cox and Accelerated Failure Time models. In this dissertation we aim to understand if more "modern" machine learning algorithms such as random survival forests, gradient boosted models, support vector machines, neural networks and deep neural networks, adapted to the specific characteristics of survival analysis, bring a real added value to the more classic models.

These models do manage to achieve a better performance than the traditional Cox model, but at the same time they require a sufficiently large dataset and a high computational power and memory capacity for training the models.

Chapter 1

Introduction

Starting from the first three decades of the twentieth century, cancer has begun to represent one of the major causes of death all over the world. In 2020, cancer is still the second leading cause of death after cardiovascular diseases (ischaemic heart disease, stroke), respiratory (chronic obstructive pulmonary disease and lower respiratory infections) and neonatal conditions.[2] Men, in particular, have a higher related mortality. The tumors with the highest incidence for the two genders, male and female, are respectively that of the prostate and that of the breasts. The possibility of undergoing screening and early diagnosis, and the apparent lower aggressiveness of these tumors, reduces mortality to 20% for prostate cancer and 25% for breast cancer.

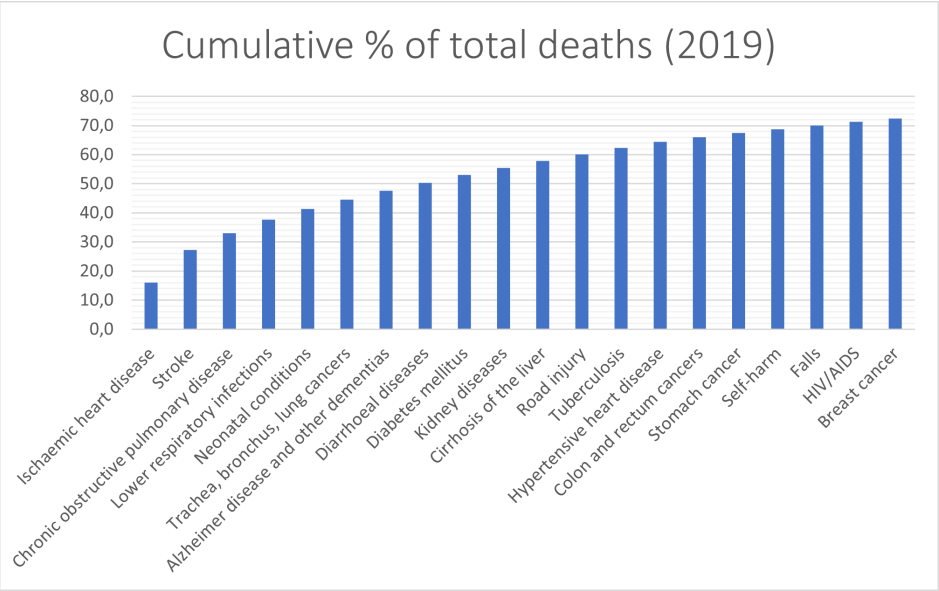


Figure 1.1: Cumulative percentage of total deaths

It is interesting to observe, especially bearing in mind both the purpose and the focus of this paper on breast cancer, that in 2019 this type of neoplasm was the twentieth cause of death, accounting for a percentage of 1,2%. The lack of this cause from the top twenty list for the year 2020 (as it can be observed in Figure 1.2 and from table 1.1) could be considered somewhat reassuring and a confirmation of the negative trend of death rate from cancer (relative to the United States) as reported by the American Cancer Society.[3]

Table 1.1: Global Health Estimates 2020, Deaths by Cause, Age, Gender by Country and by Region, 2000-2020

Rank	Cause	% of total deaths
1	Ischemic heart disease	13,2
2	Stroke	10,7
3	Neonatal conditions	6,2
4	Lower respiratory infections	6,0
5	Chronic obstructive pulmonary disease	5,8
6	Diarrheal disease	5,2
7	Tubercolosis	3,4
8	HIV/AIDS	2,7
9	Trachea, bronchus, lung cancers	2,4
10	Road injury	2,3
11	Cirrhosis of the liver	2,1
12	Diabetes mellitus	1,7
13	Kidney diseases	1,6
14	Self-harm	1,5
15	Stomach cancer	1,5
16	Hypertensive heart disease	1,4
17	Malaria	1,4
18	Congenital anomalies	1,3
19	Colon and rectum cancers	1,2
20	Alzheimer disease and other dementias	1,1

However it seems important to bear in mind the events and the context of this year, namely the outbreak of coronavirus disease caused by the novel Severe Acute Respiratory Syndrom CoronaVirus 2 (SARS-CoV-2). After a rapid spread across continents, Covid-19 was declared by the WHO at first, a Health Emergency of International Concern in January 2020 and afterwards, a global pandemic in March 2020. This emergency has had numerous impacts on the health status of the people besides those infected by the virus. Starting from the health-care systems of the affected countries which were forced to adapt to a situation where a huge number

of patients needed hospitalization and in the worst cases, intensive care.

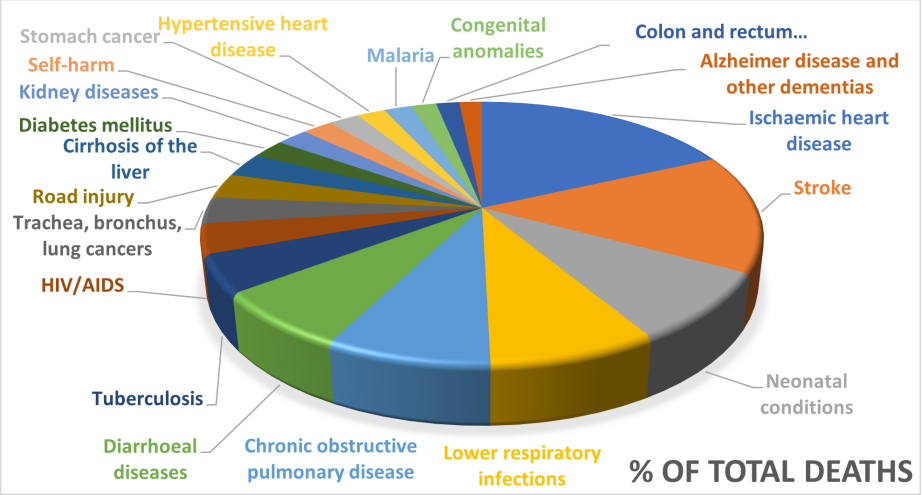


Figure 1.2: World-wide percentages of causes of deaths

The inability to maintain the distance to protect both users and operators, the lack of the former due to the decrease in the availability and capacity of public transportation but also fear, the lack of the latter, relocated and reassigned to support Covid-19, and lastly, the need to reconvert departments and even entire structures..

All these factors have made the screening programs, deferred activities, suspending them at the beginning of the pandemic and subsequently starting again, although slowly, to reactivate them. Huge delays that keep piling up in medical visits, treatments, health services and inevitably, in diagnosis ensued. This situation is still affecting above all patients with chronic and rare disease, and particularly concerning are the possible consequences for cancer patients, whose negative effects of the delays, especially for those categories of tumors that are usually kept under control thanks to prevention programs, will inevitably show over time.

A survey conducted by the WHO concerning 155 countries has found that prevention and treatment services for noncommunicable diseases (NCDs), namely those disease that are not transmissible directly from one person to another such as autoimmune diseases, strokes, cancers, etc., have been severely deranged since the beginning of the Covid-19 pandemic. Circa the 53% of the surveyed countries have partially or completely disrupted services for hypertension treatments while the 42% for cancer treatments. The same survey has identified a correlation between

the levels of disorder and turmoil of services for treating NCDs, and the evolution of the new coronavirus outbreak in the countries hit by the pandemic. To provide further evidence of what has already been stated by the WHO, also the National Screening Observatory reports that, as for what concerns Italy, about 472.389 fewer mammography screening exams were carried out in the first 5 months of Covid-19, in comparison with the same period of 2019, for an estimated 2.0099 fewer cases diagnosed.[4]

In July 2020 the situation pushed the Board of Directors of the European Cancer Organisation - a federation of 31 Member Societies working together with 20 patient groups - to launch the Special Network Impact of Covid-19 on Cancer "as an urgent response to growing evidence and reports of the devastating impact of Covid-19, and associated control measures".[5] This Unit proposes seven urgent points that National Governments, the European Union and WHO Europe are strongly recommended to implement. It will be therefore interesting in the future, to study how Covid-19 might have impacted on the survival rates of those people whose screening and treatments were delayed.

On a side note, it is also interesting to observe the increase in percentage of the "self harm" cause of death, from 1.3 in 2019 to 1.5% in 2020, which seems to be coherent with the impact that the pandemic has had on the mental health of people.

So, given this context it seems quite natural to wonder what survival analysis is and how it can possibly help us. We will only give a short introduction to the concepts of Survival analysis which will be treated more carefully and with a greater attention in a specific chapter.

Survival analysis is a branch of statistics used primarily to study mortality in biological organisms and therefore for cancer studies, but also failures in mechanical systems and sociological events where it is known as *event history analysis*. Survival analysis also corresponds to a set of statistical approaches used to investigate the time it takes for an event of interest to occur. This second definition may help in understanding why survival analysis is also known as *time to event analysis*, where the event of interest could be, with reference to the previous definitions, death or the system's failure, and the analysis involves modeling time with event-data.[6]

Hence survival analysis attempts to answer certain questions, such as what is the proportion of population which will survive past a certain time? Can and how can multiple causes of death be taken into account? How do particular events, circumstances or behaviors affect the probability of survival?

The initial purpose of this analysis is that of trying to conduct a complete survival analysis (as far as possible given the vastness of the topics) on a specific subset of the SEER database, with focus on breast cancer patients, starting by implementing the classical survival analysis methods, that will be explained later on. We will later show how machine learning algorithms can be also used, if their use is actually useful to improve the performance and when these algorithms really outperform the classical methodologies, by taking into consideration variables that in survival analysis are not considered.

Chapter 2

Principles of Oncology

In this chapter we will deal with a very superficial delineation of the disease (as this is not the place to further explore the subject), including epidemiology, pathology and possible therapies. Finally, a brief focus on breast cancer as the subject of our research, will follow.

2.1 Epidemiology

As we mentioned in the introduction section, the constant growth in the incidence of tumors has contributed to increasing the importance assumed by oncology in the medical field.

This growth in the number of patients with cancer (which can be noticed from the graph in Figure 2.2), may also be due to a progressive improvement of screening programs and in general of medical technologies that have made it possible over time to diagnose the disease from the earliest stages of its advancement. However the analysis over time of the incidence seems to prove that this is still higher also in relation to the population.

The disease has become the focus of numerous studies aimed at improving treatments, which have become increasingly complex and specific over time.

Figure 2.2 shows instead the trend over time (from 1990 to 2019) of deaths by type of cancer in the United States.

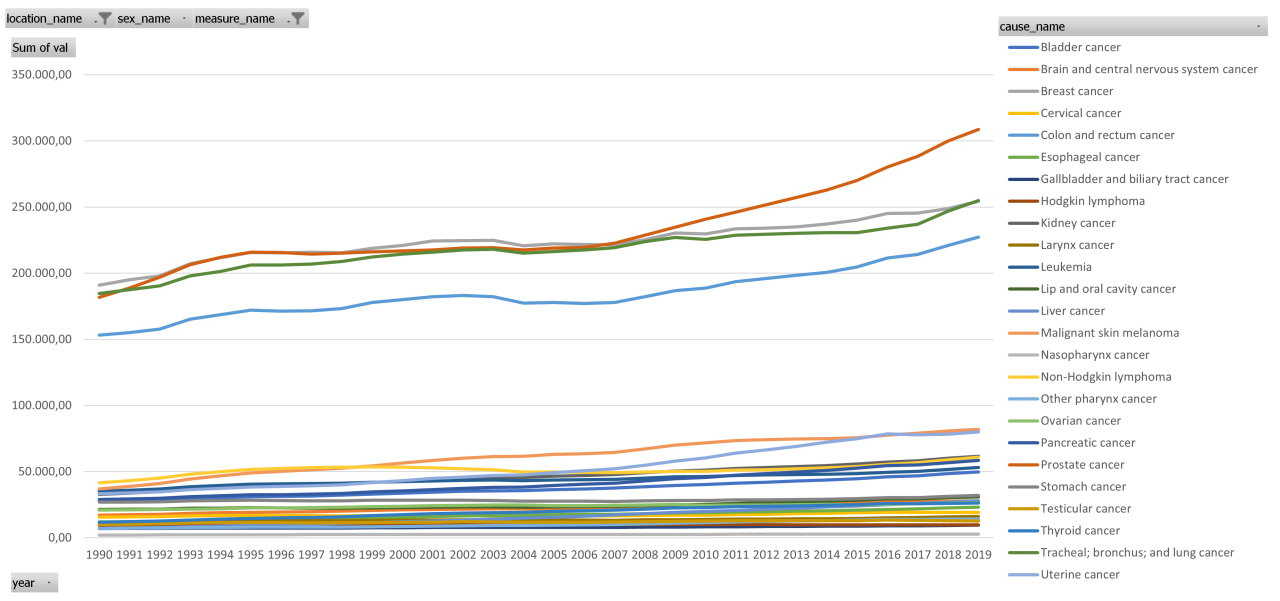


Figure 2.1: Incidence of tumors per type in the USA from 1990 to 2019 for both males and females. [7]

2.2 Pathogenesis

From the pathological point of view, the words tumor and neoplasm indicate a mass of cells that almost always originated from a single somatic cell of the organism, affected by a sequence of genomic alterations, transmissible to the daughter cells.[8]

In order to provide an aid to less experienced readers, we report a definition of somatic cells, which however requires a reference to some concepts of physiology, cytology (i.e. the study of the structure and functioning of cells) and more generally of cell biology.

2.2.1 Physiology

It is generally known that cells are the building blocks of all plants and animals, including humans, are formed and each new cell is born from the division of pre-existing cells. In the human body in particular, two general classes of cells are distinguished: the sexual ones (spermatozoa and oocytes), and the somatic ones which include all the other cells of the human body.

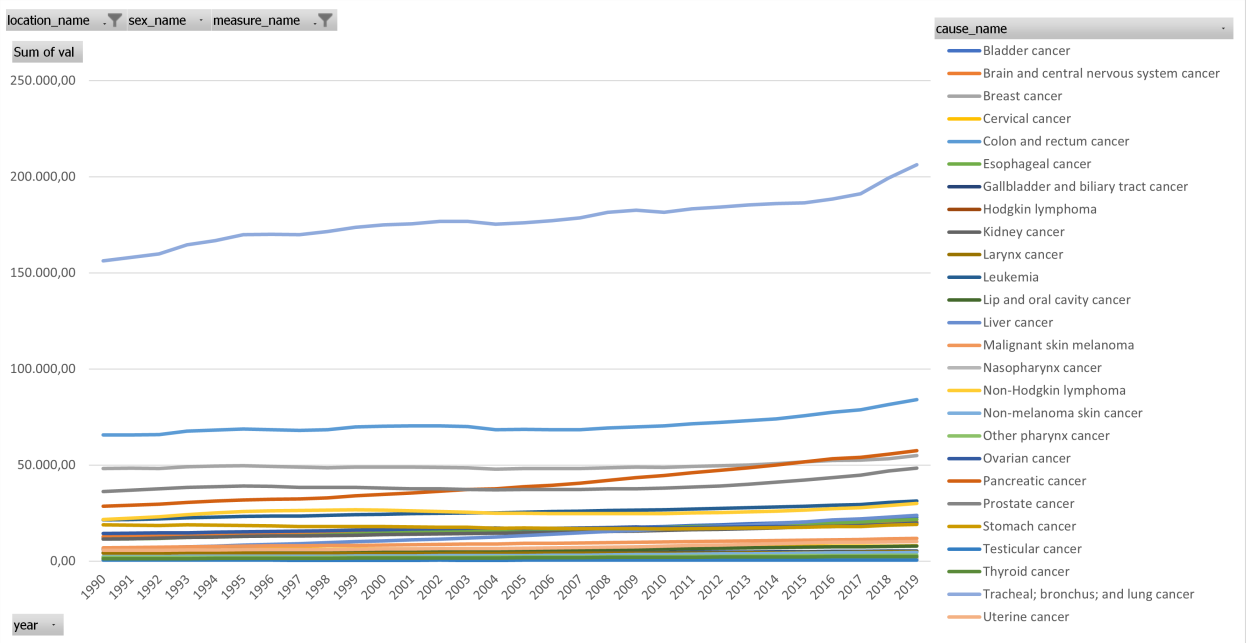


Figure 2.2: Deaths per type of tumor in the USA from 1990 to 2019 for both males and females. [7]

The zygote: the first embryonic stem cell

At fertilization, the fertilized egg cell, i.e. the zygote, is a single cell that contains all the genetic potential and is the cell from which, through a succession of cell divisions and differentiation processes, all the tissues and organs of the new organism will be constituted. This cyto-generative capacity is called totipotency. From this first totipotent cell, pluripotent cells are formed during the development process. These cells have a lower differentiation capacity than the zygote and make up the three germinal sheets: ectoderm, mesoderm and endoderm, which are responsible for the development of the various organisms.

This process of progressive specialization and differentiation involves the inhibition of genes that prevent the cell from synthesizing a particular protein and therefore from carrying out any function deriving from it. For example, nerve cells, bone cells and hepatic cells all have the same genetic composition, one different from the other because in each type a series of genes encoding proteins functional to the activities of the others has been deactivated.

Only one clone of that cell can be obtained from the division of a specialized cell, so the specialization process is generally not reversible. Precisely for these capacities, the zygote is considered to be a real stem cell, that is a primitive non-specialized

cell, with the sole purpose of generating daughter cells.

Adult stem cells

In addition to embryonic stem cells, recent studies show the existence of adult stem cells, which represent a very small percentage of human tissue cells and are randomly dispersed throughout the body. These cells replicate exceptionally, that is, if and only if it is necessary to replace adult stem cells that are missing due to programmed cell death (which we will soon cover but very shortly) or due to pathological processes and in this case therefore, to initiate a repair process. One of the best known examples of adult stem cells are the cells present in the bone marrow, endowed with multipotency - that is, capable of forming different cells but always within their own tissue - as they allow, always through differentiation processes, to give origin to different types of blood (blood) cells.

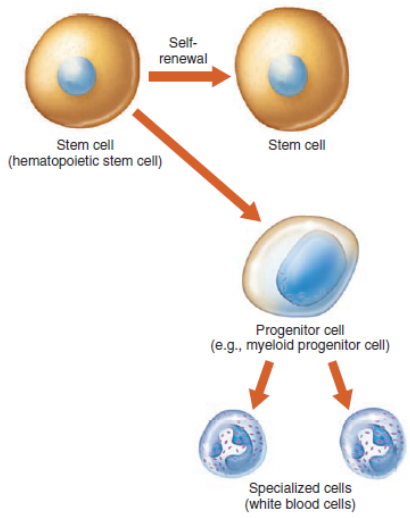


Figure 2.3: Stem cells and progenitor cells. A true stem cell divides mitotically into two stem cell daughters, specifically a stem and a progenitor cell, which may show the beginnings of differentiation.[9]

Cell replication

Cell replication occurs through cell division of stem cells, and during this process a single cell produces a pair of daughter cells, of smaller size, one stem cell again and the second one with less differentiation capacity and which may already have some hint of specialization. Moreover, not all cells that enter the second type can replicate

by cell division. Cell division takes place successfully only if the genetic material within the nucleus is replicated and assigned to the daughter cells without errors: the phase in which DNA duplication takes place is called mitosis. Depending on the specialization of the cell, its life span can vary from a few hours, such as those of the skin, to even a few decades, such as muscular ones or neurons. A very important aspect to highlight is that many of them are programmed to self-destruct after a certain period of time through the activation by their nucleus of suicide genes.

Control of cell division and the role of Telomerase

As for the control of cell division in normal tissue, the index of cell loss or destruction should be counterbalanced by the mitotic index, an index calculated on the number of cells that are in the phase of mitosis (easily identifiable) in a certain moment. There are some internal and external stimuli to the cells that can promote cell division but also the growth of the cells themselves. Among the internal factors we can indicate the presence of proteins such as peptides, but also the size of the cell, in fact the larger the cell, the greater the energy and therefore the nutrients it needs to develop. When the cell begins to occupy too much volume, cell division is stimulated which determines the birth of copy daughter cells but of smaller dimensions. Among the external factors we find instead the availability of space in which the cell can grow.

The number of possible divisions for cells seems to be regulated by a component of the chromosome, the telomeres, terminal segments of DNA to which proteins are associated. These telomeres are synthesized not by DNA, but by an enzyme called telomerase, which becomes more and more inactive over time or during the aging process. Each time the cell divides these segments "consume" becoming shorter and shorter and when they reach a certain threshold, the repressor gene causes the cell to stop reproducing. However, if for some reason, a cell with shortened telomeres does not respond normally to repressor genes, it will continue to divide resulting in damages to DNA chains, mutations and chromosomal abnormalities. It also follows a reactivation of telomerase which then causes the cells to divide with an abnormally rapid frequency.

Cell death

Cell death, despite the name, plays an extremely important role as it allows the elimination of cells that threaten the harmonious relationship between size and function between the various tissues and organs of the body, but also for the removal of those cells born or become dangerous for the organism. According to the literature, there are three ways in which cells can die:

- Apoptosis or programmed cell death, which represents a form of self-destruction or programmed suicide carried out "voluntarily" by the cell.
- by Autophagy, i.e. performed by the cells themselves to eliminate damaged molecules or organelles. If performed incorrectly it can lead to cell death.
- Necrosis, which the cell undergoes without being able to oppose, and is typically a consequence of an accidental, traumatic or pathological event.

Apoptosis in particular is a genetically programmed phenomenon, for which cells predispose a certain energy consumption, and can involve both damaged cells and healthy cells, aimed at achieving correct cell differentiation and in some cases also guaranteeing the survival of the organism.

2.2.2 Pathophysiology

Cancer is a disease characterized by mutations that radically modify the normal control mechanisms, producing malignant cells. In particular, normal cells become malignant when a mutation occurs in a gene involved in the phases of growth, differentiation or cell division, which we have previously described. The mutated genes are called oncogenes and cause cells to assume abnormal shapes and sizes. Since these mutations can affect both the genes that code for proteins responsible for DNA replication, but also those that code for enzymes responsible for repairing the damage suffered, genomic damage continues to accumulate.

The first case refers to damages to tumor suppressor genes (oncosuppressors), i.e. the genes predisposed to control the correct functioning of mitosis processes; when these are damaged or deactivated, the cell cycle becomes uncontrolled, leading to an uncontrolled division of cells and consequently to the genesis of tumors.

Based on numerous experiments, the literature identifies three evolutionary stages in the process of tumor formation:

- Initiation: one or more mutations transform a normal somatic cell into a neoplastic tumor cell, still lacking the multiplicative capacity.

- Promotion: following further genomic damage the cell begins to reproduce, giving life to a population of cells which, as copies of the same, will be diseased.

- Progression: further accumulation of mutations in the genome of some cells of the tumor population, which in turn can take over the reproductive phenomena giving life to a population of tumor cells that present the specific mutation of that cell. In fact, it is possible that the tumor mass is composed of different types of cancer cells which, as we will see later, implies, for example, that a chemotherapy can be particularly effective in defeating a particular type of tumor cell, but absolutely ineffective with another population. tumor belonging to the same mass

In an early stage of development, the growth of the tumor alters the tissue but the organization of the same still remains unharmed. Subsequently, however, the number of cancer cells explodes from the primary tumor and invades the surrounding tissues. Cancer cells can then enter the lymphatic system through which they are transported to the lymph nodes (which is why in the project we also take into consideration the two variables "examined lymph nodes" and "positive lymph nodes") or into the blood vessels, thus being able to circulate throughout the body, creating secondary tumors in other locations and initiating the process of metastasis.

Moreover, since tumors need enormous amounts of energy to grow and multiply, they are able to stimulate the creation of new blood vessels in the areas where they reside, which can "feed" their growth and metastasis process, ultimately compromising the organ functionality.

2.3 Pathology

Below we take care to list the main characteristics of the disease.

2.3.1 Benign and malignant tumors

On the basis of the morphological and biological characteristics, neoplasms are typically divided into two classes:

- Benign, since lacking the ability to spread to surrounding tissues can be removed without usually fearing recurrence, i.e. reappearance in the same site as a tumor with the same characteristics as the one previously removed.

- Malignant, which instead have the ability to affect the surrounding tissues and once removed, if possible, there is still the possibility that these can return.

However, this does not mean that even benign tumors do not cause damage to health: the size and growth of some of these can in fact generate the so-called "mass effect", that is compressing the surrounding tissues with their own mass causing damage to the nerves, reduction of blood flow which can also lead to ischemia or tissue necrosis. Benign brain tumors, for example, still remain among the most life threatening. In addition, the possibility remains that benign tumors become more aggressive and eventually turn into malignant tumors (this typically occurs through a process known as tumor progression).[10]

2.3.2 Etiology of tumors

It is interesting to note how the frequency and type of most common cancers changes from country to country, as numerous studies have confirmed that they are linked to environmental conditions, habits, traditions and lifestyles. The site "Our World in data", provides numerous interactive graphs on the incidence and mortality over time of cancer by type, country and gender that allow you to have an excellent overview and visualization of the aspects mentioned above.[11]

Although we often hear about factors or events that cause cancer, it would be more correct to express it in terms of risk factors that increase the possibility of developing cancer, inducing alterations in the DNA and these factors can be:

- External or Environmental causes
  - Chemical agents

  - Physical agents (e.g. radiation), as a matter of fact undergoing a first cycle of radiotherapy increases the chances of developing another tumor.
  - Biological agents (e.g. oncogenic viruses)

- Endogenous cases
  - Mutations transmitted from parents to children or that occur due to errors in DNA duplication.
  - Hormonal imbalances
  - Agents that develop in the body and that are not neutralized.

WHO also reports that approximately one third of cancer deaths are a consequence of 5 main types of behaviours considered dangerous:

- High body mass index,
- Unhealthy diet with low fruit and vegetable intake,
- Lack of physical activity,
- Smoke, active or passive, which is also the most important risk factor accounting for about 22% of cancer deaths and 85-90% circa of all lung cancers,
- Alchohol usage.

Besides these, it is important to highlight also the following:

- Age, as it is shown that different types of cancer present with a different incidence in different age groups and in general the incidence of tumors increases with aging.
- Progressive aging of the population,
- Longer duration of the fertile period and therefore prolonged exposure to the proliferating stimuli of estrogen (the main sex steroid hormones in women) and progesterone. The former regulates the menstrual cycle and ovulation, also helping in preparing the reproductive system for a potential pregnancy. The latter, on the other hand, is a hormone that works together with estrogen to prepare the uterus for pregnancy, promoting the implantation of the fertilized egg and is a crucial element for maintaining pregnancy.

- Genetic predisposition and in particular first-degree family members of patients with cancer generally have an increased risk of developing the disease themselves. However, it is important to emphasize that individuals do not inherit the tumor from their parents but an alteration of the genome which thus becomes a predisposing factor and it is interesting that in most cases the transmitted mutation concerns the tumor suppressor genes.

- Previous cycle of radiotherapy and therefore previous diagnosis of tumor, even benign.

- Hormone replacement therapy, such as that used to counteract the effects of menopause.

In such a context the importance that prevention becomes clear, even before the complicated pharmacological and surgical treatments aimed at curing the disease that has already developed; the same WHO estimates that between 30 and 35% of today's cancers can be prevented by avoiding risk factors and implementing existing and tested prevention strategies.

2.3.3 Symptoms and Diagnosis

Let us now list only the main and most common symptoms typically found since the symptomatology is varied both in the form and in the timing of onset, being able to appear even only in the advanced stages of the disease; moreover it is also strongly conditioned by the type of tumor and its location.

One of the most common symptoms is probably the feeling of fatigue, general tiredness, depression and difficulty falling asleep and sleeping. Other symptoms may be uncontrolled weight loss, skin lesions, difficulty in swallowing solid and/or liquid foods (dysphagia), difficulty in breathing (dyspnoea) and nausea.

Dry cough is the most common symptom associated with lung cancer which can degenerate into hemoptysis, i.e. the emission of blood from the mouth often during the most "serious" cough episodes. Persistent fever can be another symptom, as well as anemia.

Finally, in more advanced states it is also possible the appearance of pain in specific points such as muscles or skeleton and bones, and may be a consequence of

the invasion by the tumor of nearby organs and tissues or of the therapies to which the patient is subjected.

Diagnosis

As for the diagnostic phase, the role of general practitioners plays a fundamental role since on the basis of the patient's reported symptoms they will have to prescribe further investigations such as diagnostic imaging, i.e. radiography, PET, magnetic resonance, mammography and ultrasound or a histological examination (i.e. the so-called biopsy) if the tumor was easily reachable.

2.3.4 Treatments

Once a tumor has been diagnosed, the specialist will have to define the best cure, and we list the main ones below, with a brief explanation:

- Surgical Therapy
- Radiotherapy
- Medical Therapy

Surgical Therapy

For a long time considered the only possible solution to the treatment of tumors, this solution has progressively evolved especially following the confirmation by numerous studies that the only removal of the tumor in about 70% of cases is unsuccessful due to the onset of metastases that develop as a consequence of the presence of neoplastic cells in the tissues surrounding the site of the disease. The question that arises spontaneously is therefore how much surrounding tissue must be removed during the operation to reduce the likelihood of recurrence, also compatibly with the new concepts of preserving the patient's image and identity.

It is for this purpose that integrated therapies and combinations of therapeutic treatments are involved. A fundamental part of this form of therapy is the removal, in the presence of the disease, of the regional lymph nodes, which often become the site of metastases of solid tumors. This procedure has also evolved over time, replacing the "classic" removal of all lymph nodes, with initially the analysis of the

first lymph node closest to the site of the neoplasm, the so-called "sentinel lymph node", and consequently the first to receive the lymph coming from the tumor itself: if this is negative, then most likely the following ones will also be negative and therefore it is not necessary to remove it.

Radiotherapy

This type of therapy uses ionizing radiation and in particular, a beam of penetrating photons to damage the genetic heritage of diseased cells, preventing their proliferation.[12]

Also in this type of therapy it is important to find a balance between the dose of radiation administered and the impact that these have on the surrounding tissues, which in fact leads back to an optimization problem whose solution can be implemented on the radiotherapy machines as also explained Holder and Ehrgott in their interesting study entitled "Operations Research Methods for Optimization in Radiation Oncology".[13]

Medical Therapy

It is possible to distinguish among:

- Medical oncological therapy, which involves the use of chemotherapy, drugs, biological and hormonal agents also in combination, based on the type of tumor. The aim is that of blocking or delaying the multiplication process of neoplastic cells. The downside of these drugs is that they are unable to distinguish between diseased and healthy cells and damage the latter by speeding up the cell cycle. For example, as regards tumors influenced by estrogen, the most logical solution is to deprive them of these hormones, this can be achieved:
  - By preventing the cancer cell from using the estrogen produced or
  - By inhibiting the production of estrogen.
- Hormonal therapy, originates from the discovery that the growth of some neoplasms (including breast cancer but also prostate cancer) depends on the level of hormones.

- Immunological therapy: in recent years, a number of molecularly targeted biological drugs have been introduced that affect cell growth, and the functions that control cell growth and reproduction. Probably the best known are monoclonal antibodies (whose fame recently increased thanks to Sars-Cov-2) capable of recognizing and destroying the tumor cell, leaving the adjacent healthy tissues intact.

- Stem cell transplant, which involves the transplantation of stem cells from umbilical cord or bone marrow in patients typically suffering from solid tumors or haematological diseases (for example, leukemia). For the transplant to be effective it is first of all necessary to remove the patient's bone marrow; once the donor cells are transplanted they will begin to produce new blood cells.[14]

Among other things, the combination of different therapies, obtaining the so-called integrated treatments, is increasingly common. The purpose of an integrated treatment is to increase therapeutic efficacy and as a consequence, increase of the patient's probability of survival while trying to administer a treatment that is not too invasive, debilitating or harmful to other vital functions. The WHO also contributes to underline the importance of this aspect, which since 1986 has defined guidelines to be followed in the prescription of painkillers, which play a primary role especially during the so-called pain therapy.

2.3.5 Gradation and Staging of tumors

The procedure that determines and evaluates the degree of severity, interpreted as the size and severity of malignant tumors, is called gradation and plays a fundamental role in the design of the treatment and in the evaluation of the results of the therapy.

To this end, the gradation process identifies with four degrees indicated in Roman letters, where the first degree I indicates well-differentiated tumors while the IV indicates non-differentiated tumors (where by differentiation we mean exactly what was previously described when talking about cell differentiation, i.e. acquisition by cells of specific functions, characteristics and consequently specialization).

In addition to the assessment of the degree of the disease, the extent of the tumor

is also considered, which is assessed using the TNM system. This system identifies 3 fundamental categories:

- T = extension of the primary tumor.
- N = absence or presence and extension of metastases to regional lymph nodes.
- M = absence or presence of distant metastases.

The numbers from 1 to 4 that follow these letters indicate the extent of the tumor and the combination of the three categories makes it possible to classify the disease into different stages. In addition to the four numbers the following writings are also used:

- T0 = in the presence of metastases, the primary tumor was not identified.
- Tis = carcinoma in situ.
- N0 = absence of lymphodal involvement.
- M0 = absence of metastases.

2.4 Breast cancer

Since this dissertation focuses on breast cancer, it seems appropriate to dwell briefly on the main characteristics of this organ and how the disease affects it.

As also Wikipedia reports, the breast is a glandular organ that secretes milk in female mammals.[15]

The development and activity of the breast are controlled by hormones produced by the ovary, pituitary and adrenal glands and in addition, during puberty in the female gender, the production of estrogen determines the increase in breast volume. The posterior and lateral lymphatic vessels lead to the axillary lymph nodes, while the medial ones drain the lymph into the internal mammaries.

2.4.1 Diagnosis

The diagnoses of breast diseases go through three diagnostic tests: clinical examination, cytology and mammography and the positivity of even one of these tests motivates the surgical biopsy.

Breast carcinomas as they grow, are typical to form nodules which, if sufficiently developed, are able to be felt through the breast palpation procedure. However, since non-palpable breast lesions are also possible, mammography screening and therefore the possibility of an early diagnosis, remains today the main "weapon" against breast cancer.

2.4.2 Risk Factors

Even for breast cancer, the main risk factors remain those previously listed with particular attention to the side of familiarity and therefore to the "genetic predisposition". In the 1990s, two genes known as BRCA1 and BRCA2 associated with the development of breast and ovarian cancers were identified and it was estimated that those carrying mutations to these two genes risk developing breast cancer in 40-65% of cases. Other studies find that more than 50% of women with BRCA gene mutations develop breast cancer before age 50. Today it is possible to offer people deemed most exposed to risk a genetic analysis aimed at identifying mutations in the BRCA genes.

2.4.3 Pathology

There are two main groups of breast cancer, those arising from the ducts of the mammary gland (ductile) and those arising from the lobules (lobular); both can also be infiltrating or not (i.e. in situ).[16]

In the section which focused on therapies and in particular on medical methodologies we mentioned how some tumors are susceptible to the level of particular hormones such as estrogen and progesterone. This dependence is usually determined by the presence on the surface or inside the cells of receptors, that is specialized proteins able to recognize and bind other substances, precisely estrogen and progesterone.

When an invasive breast cancer is diagnosed or in the presence of relapse, the specialist could prescribe a test that investigates the state of the tumor receptors; this test typically takes place by analyzing a sample of cells extracted by biopsy. The result of this test will indicate whether certain treatments, especially those of a hormonal and/or chemotherapy type, are appropriate for the type of tumor.

Hormone-sensitive breast cancers and appropriate tests

Hormone-sensitive breast cancers can therefore be:

- Positive for the estrogen receptor and positive for the progesterone receptor (respectively, ER + and PR +)
- Positive only for the estrogen receptor (ER + and PR-)
- Positive to the progesterone receptor only (ER- and PR +)
- Both estrogen and progesterone receptors are negative (ER- and PR-)

In general, the fewer the number of receptors, the less effective the hormone therapy will be and the same happens also in case of negativity of the test due to the presence of one of the two receptors.

Another test that can help define the correct therapy to follow is that on the positivity or negativity of the tumor for a mutation of the HER2 receptor, epithelial growth factor (Human Epidermal Growth Factor Receptor 2) for which a biopsy is always required. Under normal conditions this gene encodes a protein responsible for promoting cell growth; a positive test result implies that the gene is over-expressed and it follows that the tumor is more aggressive and less predictable in its response to therapies.[17]

Lastly, mastectomy indicates the invasive surgical operation which involves the removal of all the breast often including the nipple if there is the risk that this has been also affected by cancer cells Mastectomy ensures greater certainty that the tumor has been completely removed.

Those listed are just some of the numerous tests that patients can undergo, typically following a diagnosis of the disease in order to define the prognosis and outline the best possible therapy. We care to remind the reader once again that this chapter only wants to give a superficial view of the problem that is actually much more complex and worthy of further description and we invite the reader, if interested, to consult the existing literature on the subject.

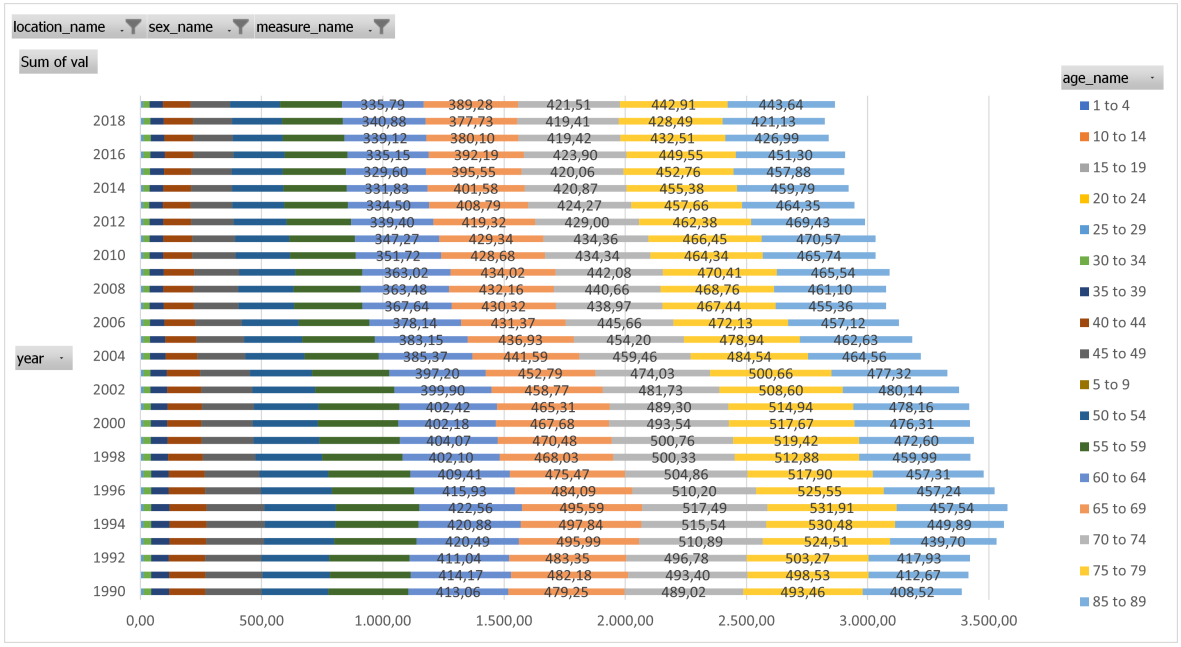


Figure 2.4: Breast cancer incidence rates over time (per 100.000 population) [7]

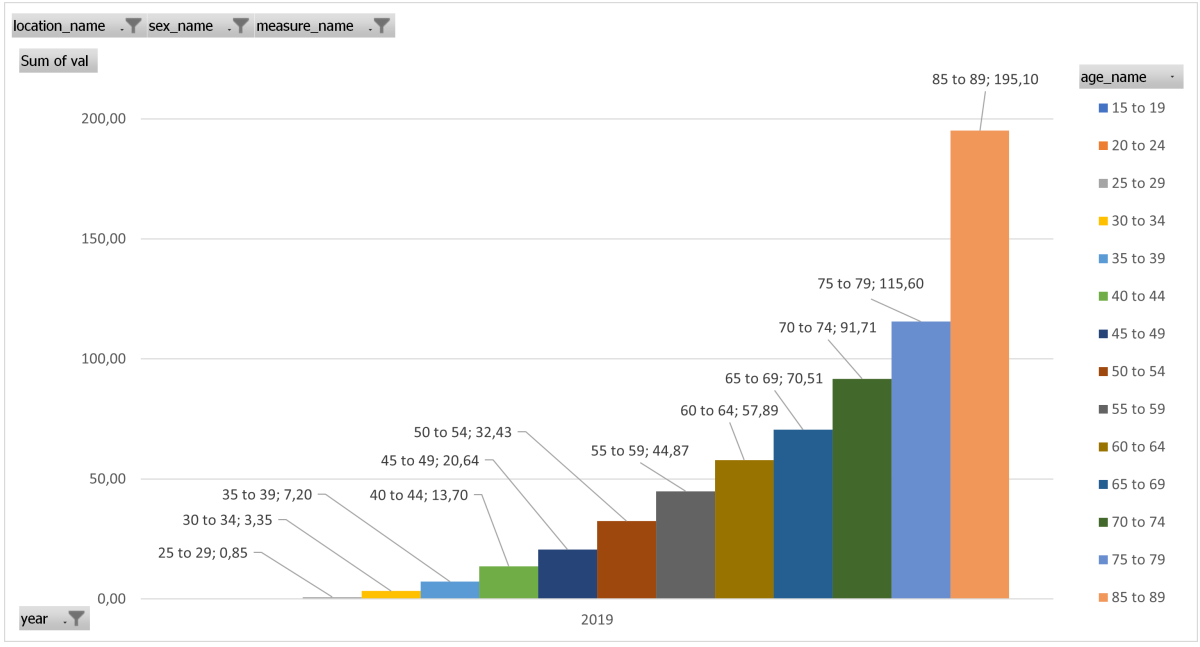


Figure 2.5: Breast cancer death rate in 2019 (per 100.000 population) [7]

Chapter 3

Data Sources, Instruments and Literature

For our analysis we will mainly use the data provided by the Surveillance, Epidemiology, and End Results (SEER) program supported by the Surveillance Research Program (SRP) in the Division of Cancer Control and Population Sciences of the National Cancer Institute (NCI). It is considered to be an authoritative source of epidemiologic information on cancer statistics, incidence and survival "in an effort to reduce the cancer burden among the U.S. population", as the official web site states.

A description of the main characteristics will therefore follow, focusing in particular on the methods of practical use of this database. Furthermore, in this chapter we will also try to collect and go through the main literary sources which, given their vastness, can put even the most expert readers in difficulty and the tools utilised for the analysis. This premise becomes necessary to define a common frame of reference that we will use in this report.

For more and more accurate information, however, we invite the readers to consult the official websites.

3.1 The SEER Program

As reported in the SEER site, the program covers approximately 34,6% of the U.S. population through the contribution of the so called authorized SEER registries.

[18]

It gathers data on:

- Patient demographics,
- Typology of tumor,
- Tumor morphology,
- Year at diagnosis,
- First course of treatment,
- Vital status of the patient at the end of the follow up.

When we subsequently introduce the variables selected for our study, we will see more in detail the information and classifications used by the SEER program to describe both the variables of our interest and of the database in general.

3.1.1 Database updates

Every spring a new standard set of research data is released, in particular two data products, based on the previous November's submission of data from the registries. These products are the Research, which includes the fields and variables that SEER makes publicly available by signing the SEER Data-Use Agreement form, and the Research Plus, typically available later in the release year and which includes additional fields not available in the Research data. It is important to highlight that since 2001, all cases reported to SEER are required to have an ICD-O-3 histology and behaviour code.

All the data are then made available to the registered user for analytical purposes through a software that has to be downloaded, called SEER*Stat, which allows to extract the data of interest with the aid of predefined sessions designed to perform specific calculations. In particular, the site provides the following options:

- *Frequency* session (represented by the symbol #), created to generate the number of records stratified by any variable in the dataset and in particular it allows to obtain:

  - Counts of cases,
  - Percentages (rows, columns and totals),
  - Trends over time based on frequencies (percent change, annual percent change).

- *Rate* session (symbol $\sum$), which allows to compute disease incidence and mortality rates, and in particular:
  - Crude rates,
  - Age-adjusted rates,
  - Trends over time based on rates (percent change, annual percent change),
  - Other statistics related to the calculation of rates.

- *Survival* session (symbol %) which provides data on:
  - Observed survival,
  - Net survival (relative or cause-specific),
  - Crude probability of death (presence of other causes of death),
  - Conditional survival.

- *Limited-Duration Prevalence* session (symbol ***P***), which allows to extract data on:
  - Prevalence counts,
  - Crude (non-adjusted) percentages,
  - Age-adjusted percent.

- *Multiple-Primary Standardized Incidence Ratio (MP-SIR)* session (symbol ÷) which instead provides comparison of cancer incidence in a defined cohort of people diagnosed with cancer, with cancer to cancer incidence in the general population.

- *Left-Truncated Life Tables* session (symbol ***LT***):
  - Cumulative summary page,

  - Detailed life page.

- *Case-Listing* session (represented by a mini-table), probably the most intuitive one since it allows to view the values of the variables for individual cases. It is the session that was used to extract the data on which we performed the analysis.

The official Website of the SEER program also offers some step by step tutorials with practical and reproducible examples.

3.2 ICD-10 and ICD-O-3

Since 2001, all cases reported to SEER are required to have an ICD-O-3 histology and behaviour code.

ICD-O-3 is the acronym used to indicate the third (and current) version of the *International Classification of Diseases for Oncology*, an extension of the *International statistical Classification of Diseases and Related Health Problems* (ICD-10, namely the 10th version), that focuses on tumors and which is a classification widely used by cancer registries, promoted by the World Health Organization (WHO). In particular it is used for coding the site (topography), standardized with the C section of the ICD-10, and the histology (morphology) of the neoplasm, usually obtained from a pathology report.

There are structural differences between the ICD-O and the ICD-10, even though they are both designed to promote international comparability in the collection, processing, classification, and presentation of health conditions by defining the universe of diseases, disorders, injuries and other related health conditions listed in a hierarchical fashion. In case of mortality statistics these coding rules improve the usefulness by giving preference to certain categories instead of others and, by forcing the selection of a single underlying cause of death while setting the other reported causes as non-underlying causes of death. The combination of underlying and non-underlying causes is considered to be the multiple causes of death.

The topography codes listed in the second chapter of the ICD, which is centered on Neoplasms, contain the information and description on the behavior of the neo-

Table 3.1: 2021 ICD-10-CM Codes.[19]

Chapter	ICD-10 Code	Description
I	A00 – B99	Certain infectious and parasitic diseases
II	C00 – D49	Neoplasms
III	D50 – D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00 – E89	Endocrine, nutritional and metabolic diseases
V	F01 – F99	Mental, Behavioral and Neuro-developmental disorders
VI	G00 – G99	Diseases of the nervous system
VII	H00 – H59	Diseases of the eye and adnexa
VIII	H60 – H95	Diseases of the ear and mastoid process
IX	I00 – I99	Diseases of the circulatory system
X	J00 – J99	Diseases of the respiratory system
XI	K00 – K95	Diseases of the digestive system
XII	L00 – L59	Diseases of the skin and subcutaneous tissue
XIII	M00 – M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00 – N99	Diseases of the genitourinary system
XV	O00 – O9A	Pregnancy, childbirth and the puerperium
XVI	P00 – P96	Certain conditions originating in the perinatal period
XVII	Q00 – Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00 – R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00 – T88	Injury, poisoning and certain other consequences of external causes
XX	U00 – U85	Codes for special purposes
XXI	V00 – Y99	External causes of morbidity
XXII	Z00 – Z99	Factors influencing health status and contact with health services

plasm, and this is achieved by assigning a specific range of codes identifying these behaviours. In particular the following behaviors have been identified:

- Malignant,
- Secondary or Metastatic,
- In situ neoplasms,
- Benign neoplasms,
- Neoplasms of Uncertain or Unknown behavior.

Appropriate codes for each site of the body are then listed in alphabetical order.

Let us see a small example that we hope will facilitate the reader's understanding: the specific codes for breast and lung (here presented for comparison and illustrative reasons) neoplasms are:

Table 3.3: ICD-10 Alphabetic Index Entry for Breast and Lung Neoplasms

Site	Malignant	Secondary or Metastatic	In situ	Benign	Uncertain or Unknown
Breast	C50	-	D05	D24	D48.6
Lungs	C34.9	C78.0	D02.2	D14.3	D38.1

To describe the topography of a neoplasm, the ICD-O instead uses a set of four characters, taking up the malignant Neoplasms section of the ICD-10. Therefore, the topographic code, C50 with reference to breast cancer, C34.9 for lung cancer, remains the same for each neoplasm that affects the relative part of the body.

Instead, to indicate the behavior of the neoplasm (malignant, benign, etc.) a fifth digit is used in the Morphology field. The ICD-0 also describes the type or morphology of the neoplasm. As for what concerns the Morphology codes deployed by the ICD-O-3 we refer the readers to consult the manual released by the World Health Organization containing all the official indexes and relative descriptions, and we will limit ourselves to indicating those of interest for our analysis (breast tumors), using if necessary for comparison, those relating to lung cancer, again, to achieve a greater clarity.

Table 3.5: 5th Digit Behaviour Code for Neoplasms

Behavior Code	Description
/0	Benign
/1	Uncertain whether benign or malignant: • Borderline malignancy, • Low malignant potential, • Uncertain malignant potential.
/2	Carcinoma in situ: • Intraepithelial, • Non-infiltrating, • Non-invasive.
/3	Malignant, primary site
/6	• Malignant, metastatic site, • Malignant, secondary site.
/9	Malignant, uncertain whether primary or metastatic site.

Table 3.7: ICD-0-3 coding for breast neoplasms

Topography code	Morphology Code	Description
C50	8500/3	Invasive breast carcinoma of no special type
C50	8504/2	Encapsulated papillary carcinoma
C50	8504/3	Encapsulated papillary carcinoma with invasion
C50	8507/3	Invasive micropapillary carcinoma of breast
C50	8509/2	Solid papillary carcinoma in situ
C50	8509/3	Solid papillary carcinoma with invasion
C50	8519/2	Lobular carcinoma in situ, pleomorphic
C50	9715/3	Anaplastic large cell lymphoma, ALK negative
C50	-	Breast implant-associated anaplastic large cell lymphoma

3.3 Main Risk Assessment Models for breast cancer

As we have already seen, several factors impact on the risk of developing breast cancer but also on the outcome of the disease. As other forms of cancer, also breast cancer depends both on environmental and hereditary risk factors and although many risk factors have been identified the it remains unknown the triggering cause remains unknown. What epidemiologists and scholars in general have been able to do is identify common patterns of incidence on some specific populations. Therefore, the analysis of these factors, suitably tested on various, numerous and constantly updated patient databases, can lead to the development of increasingly accurate predictive models. These models can be subsequently implemented in tools useful both to clinicians and to individuals for enabling health predictions for individuals. The aforementioned tools therefore require an interface designed to simplify their use especially by clinicians. However, the main problems remain the presence of various predictive models, each using a different set of risk factors, hence the question of which of these is the most correct or accurate.

There are numerous models at the base of the tools and below we propose to illustrate the main ones.

3.3.1 The Gail model for breast cancer risk assessment

The Gail model was first developed in 1989 by dr. Mitchell Gail and his collegues of the department of Biostatistics of the National Canacer Institute's Division of cancer Epidemiology and Genetics, and it is the result of a massive screening program which involved 280.000 women with an age between 35 and 74.

This tool calculates a woman's risk of developing breast cancer within the next 5 years and up to the age of 90, therefore it is unable to say for sure whether a person will develop the cancer or not. This model focused on the woman's personal medical history, her family's and also her reproductive history, in particular it takes into account 7 key factors for the prediction: age, age at the first menstrual cycle, age at birth of the first child, family history of breast cancer, number of past breast biopsies, number of breast biopsies showing atypical hyperplasia and the ethnicity.

Women with a 5-year risk exceeding 1,67% were considered as high risk and the FDA guideline for these subjects suggests taking a risk-lowering drug to reduce the likelihood of cancer occurrence.[20]

Over time it was found that the Gail model underestimates the risk of prediction regarding the family situation but also the woman's life-style behaviours and since the original model did not take into account ethnicity differences, it had to be subsequently adapted (Gail Model 2), making the model one of the most used in the clinical setting, but above all the basis for the development of the following models.[21] In fact, the Gail Model 2 also presents some problems, such as greater accuracy in predictions for non-Hispanic white women but a lower predictive capacity for subjects with demographic values other than those on which the model was calibrated. The latest update of the Gail model is now known as the Breast cancer risk assessment Tool, available on the National Cancer Institute website.[22]

3.3.2 The Tyrer-Cuzick model (IBIS tool) for breast cancer risk assessment

The second most known model is the Tyrer-Cuzick one, also known as the Ibis tool, which is used to calculate the person's probability of being a carrier of BRCA1 or BRCA2 mutations. Specifically, this model estimates a woman's 10-year probability of developing breast cancer and this estimate is based on factors similar to those already considered by Gail such as the woman's age, age at the first menarche but also, the body mass index, obstetric history, age at menopause, previous ovarian cancers, hormone replacement therapy and family history.[23]

3.3.3 Adjuvant! Online

An adjuvant therapy is a treatment that is often combined with the most drastic breast operation in order to bring greater benefit to the patient and decrease the possibility of a relapse. Such treatment may include one or more of the following therapies:

- antibody therapy, in which proteins that stimulate the growth of cancer cells attack it;

- chemotherapy;
- hormone therapy;
- radiotherapy

Adjuvant! Online is software that allows medical personnel to predict treatment benefits while also providing a 10-year survival estimate for patients in the early stages of breast cancer. However, it is important to underline that although the program is open source and there is a site to refer to, at the time of writing this section the site is under maintenance. Consequently, an indirect approach is necessary for the description of the operation of the software, through the publication of scientific papers in which the search for them using Adjuvant is compared with those determined by similar programs from differential models.

The estimates provided by Adjuvant! are based on information gathered by the Surveillance, Epidemiology and End Results (SEER) registry.

The tool was developed by analyzing a population sample aged between 20 and 79, who had undergone a surgical operation between 1988 and 1992 and has been the subject of numerous subsequent validations which seem to confirm the robustness of the estimates obtained via Adjuvant!, although some adaptation is required when the population of interest is Asian.[24] Clinicians in particular can enter data such as the ER and PR test restul, tumor staging and size and other information into the program. The program then reports the mortality rate which corresponds to the combination of the values entered and searched for in the database.

3.4 Instruments for the analysis

Two languages were mainly used for our analysis, R and Python, which we will briefly illustrate below, along with the main advantages and disadvantages of both.

3.4.1 R

R is a programming language accompanied by a free software developed by Bell Laboratories by Kohn Chambers and some of his colleagues. The environment provides numerous tools that allow the user to perform complex statistical analysis

and view the results, it is also an open-source and free project which makes it flexible, extensible (through the creation of new functions and extensions), intuitive and affordable to most. An interface is also made available to users that facilitates viewing both the code and its output at the same time. A common feature of both R and Python is that both are interpreted languages. Each program is made up of a set of instructions and compilers and interpreters are tools that allows the conversion of the language used by humans to communicate with machines into machine language, which is the one used, written and read by machines. In a complied language, the machine translates the program directly, while in an interpreted language the translation takes place through a different program, that is the interpreter that reads and executes the code. [25]

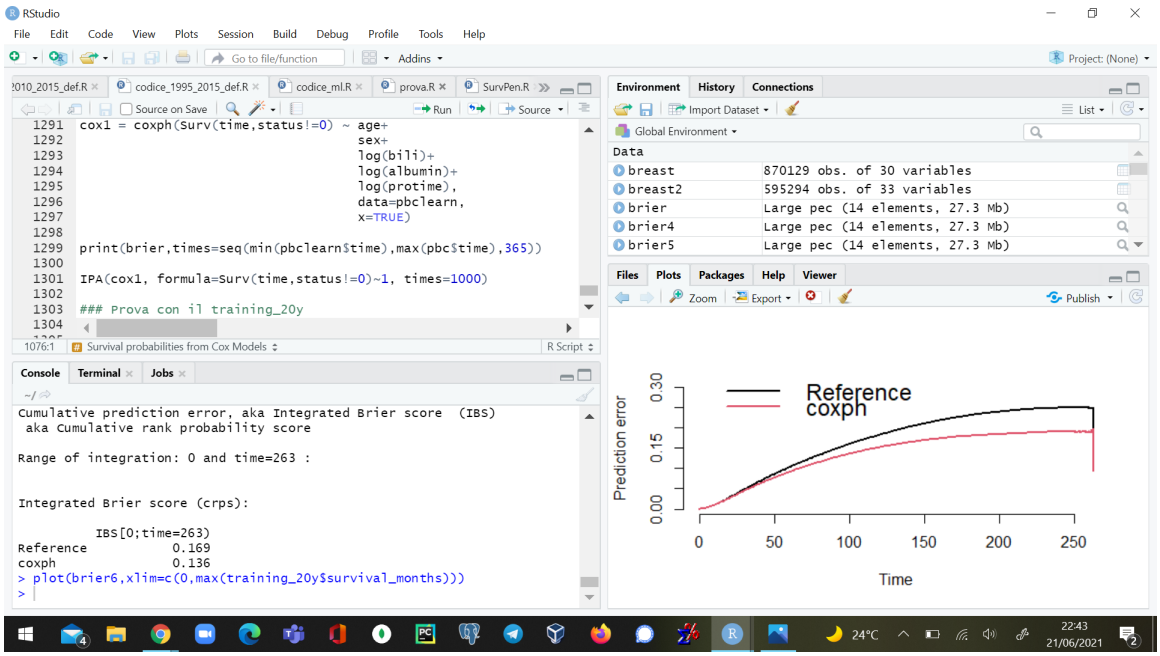


Figure 3.1: RStudio interface

3.4.2 Python

Python is an interpreted programming language (just like R), object oriented and a higher level than other languages that lends itself very well to Rapid Application development, coding and numerical computation. With object-oriented programming (OOP) we indicate a programming style that allows the user to define software objects (generally a region of the allocated memory), to interact with each other.[26] The higher abstraction, on the other hand, indicates that the language is charac-

terized by a higher level of abstention from the operating details of a computer and the classic features of machine language. The language is also sufficiently intuitive and easy to understand and supports a series of modules and libraries that allow you to perform numerous types of analysis.

Anaconda Navigator and Jupyter Notebooks

As reported by the official website of *Anaconda Navigator*, it is a desktop graphical user interface (GUI) part of the free and open source distributed software *Anaconda* of the programming languages Python and R, aimed at simplifying the management and the usage of the libraries available for both languages. Among the applications available by default in *Anaconda Navigator* we find:

- *Jupyter Notebook*, is an interactive and open source web application that allows you to create and share documents containing code, equations and functions, thus providing an ideal environment for computations;
- *Rstudio*, an integrated Development Environment; (IDE) for the R programming language;
- *Spyder*;
- *JupyterLab.*[27]

Google Colaboratory

Google Colaboratory, also called Colab for simplicity, is a free Jupyter notebook environment produced by Google Research running entirely in cloud that allows to run python code in the browser. It does not require any type of setup to be used, it provides a series of libraries and datasets that the user can utilize for their own analyzes, access to computational resources, understood as space and memory (including the GPU) are free although not unlimited.

Precisely for these reasons we decided to use this tool to carry out the analysis concerning the application of machine learning algorithms. Furthermore, just as it is possible to execute code of type Python, SQL or JavaScript in RStudio, also in Google Colab it is possible to execute some commands in R language (although there are some limitations), making both tools very comfortable and flexible.

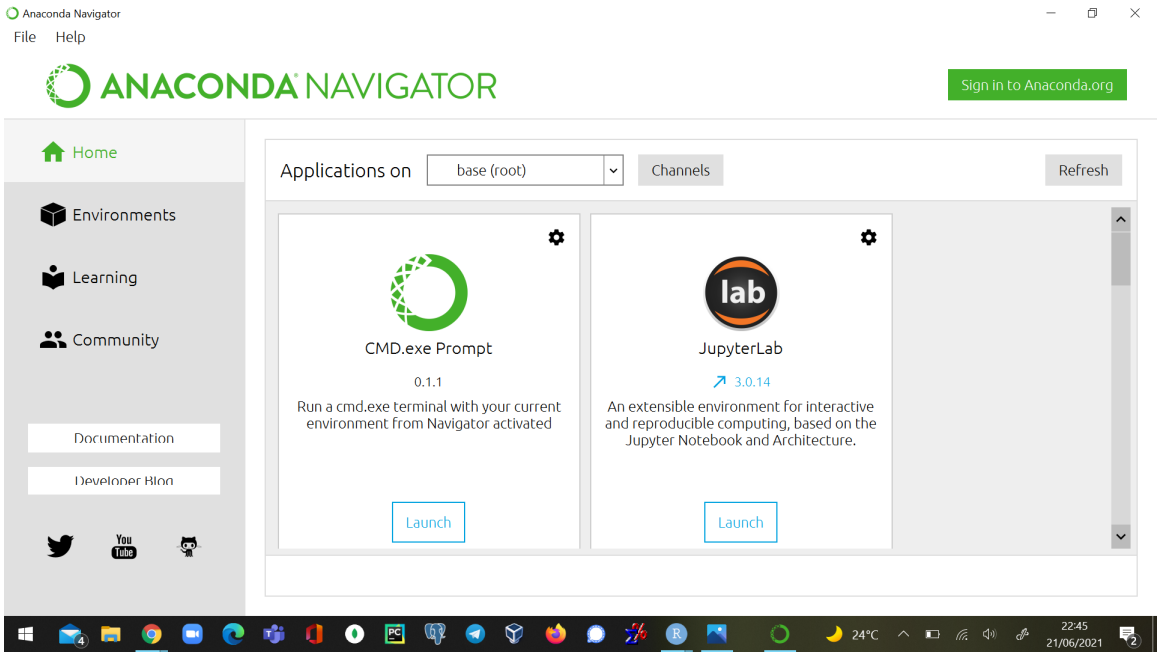


Figure 3.2: Anaconda Navigator interface

3.4.3 Main packages used

In the following section the main packages used for the analysis are listed, for both R and Pytohn.

R packages

R has been used extensively with regard to the first part of the analysis, i.e. the one relating to the preparation of the dataset and the application of the more classic models of survival analysis, i.e. Kaplan-Meier, Cox and the Accelerated Failure Time models, through the use of three libraries or packages: *survival*, *survminer* and *pec*.

In 2019 Lang et al.[28] have also developed an R package called *mlr3*[29] containing, in addition to the aforementioned models used by the survival analysis, also some machine learning models implemented from Python, adapted to this particular type of statistical analysis. However, the package is still under development and due to some incompatibilities between libraries previously installed on the PC used for this research, it was necessary to resort to Python in order to implement these models.

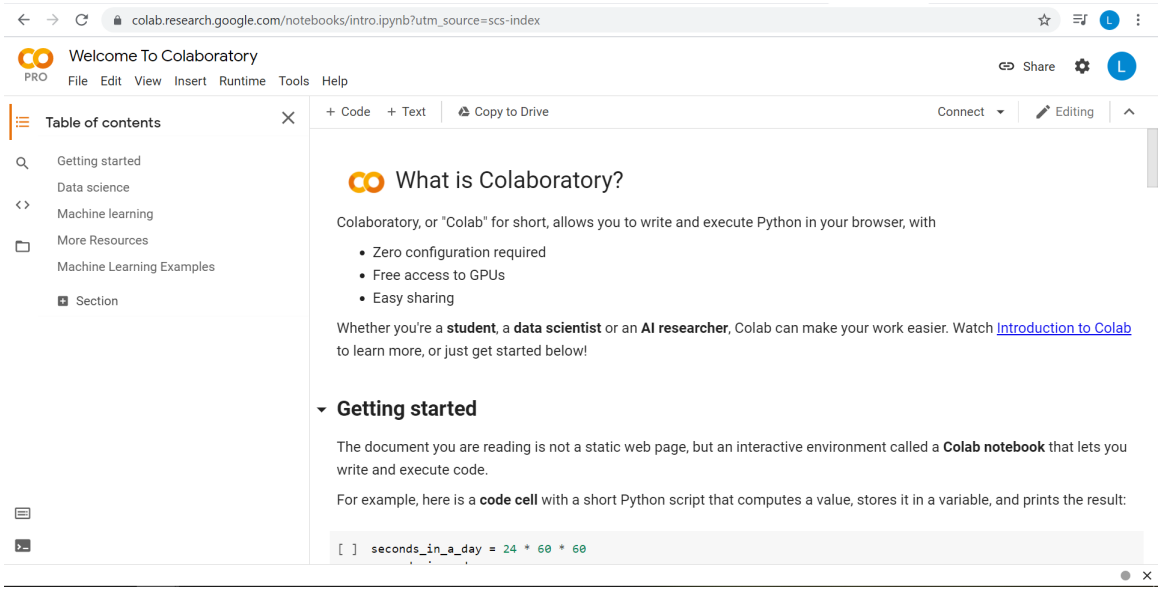


Figure 3.3: Google Colaboratory interface

Python packages

For the implementation of the machine learning algorithms mainly PyTorch, pycox and scikit-survival were used. *PyTorch* is an open source library based on another library, Torch, developed by Facebook's AI Research lab. The library offers two high-level computation tools:

- Tensor computing and
- neural networks.

pycox is a python package for survival analysis and time-to-event prediction with PyTorch. It is built on the torchtuples package for training PyTorch models.

scikit-survival is a Python module for survival analysis built on top of scikit-learn, a free software machine learning library for the Python programming language which features various classification, regression and clustering algorithms such as support vector machines, random forests, gradient boosted models, all algorithms that were also used during the development of this project.

Chapter 4

Survival Analysis

As previously anticipated, survival analysis is a branch of statistics used to study the mortality of biological organisms and failures in mechanical systems. In economics it is often referred to as "duration analysis" or "duration model", but the most representative name is probably "time-to-event analysis", as these statistical approaches try to investigate the time required for an event of interest to occur. The event of interest not necessarily is the death (mechanical or biological) of the system, thus opening up to numerous applications such as for example in the field of marketing and customer analytics.

Before delving into the methodologies and tools of survival analysis, we provide the reader with the definition of the main features and some fundamental terms of this type of analysis.

4.1 Main features of survival analysis

The main feature of survival analysis is that it is not necessary that all data have exactly the same starting point and not even the same ending point. A subject could potentially enter the study at any time.

4.1.1 Censoring

It is important to underline that the event may not occur, i.e. a machine may not fail, the patient may not die or have a heart attack or develop a recurrence, and a customer may not subscribe to an insurance policy. We do not know when and if

these subjects will experience the event of interest, all we know is that during the follow up period they did not experience it, and this phenomenon and possibility is called censoring. Censoring is a type of missing data problem, characteristic of survival analysis. In fact, the censored subjects cannot be removed from the study as their removal would completely bias the results towards those for whom the event of interest has occurred.[30]

In particular, it is possible to distinguish three types of censoring:

- right censoring, in this type of censoring, the most common, the subject enters the study at $t = 0$ but "leaves" it before the event of interest occurs. This can happen because the subject for various reasons such as in the case of clinical trials, where the patient could leave the study of his own free will, or the subject is lost in the follow up, or alternatively an event other than that of interest occurs that makes it impossible to continue follow up.

- left censoring, where the birth event has not been observed.

- interval censoring, which happens when the follow-up period is expressed as intervals of time.

- left truncation, this type of censoring occurs when the event of interest might have already happened.

4.1.2 Survival function

The purpose of the statistical methods used in and by survival analysis is that of estimating the survival function form the survival data.

This function represents the probability that the event of interest has not occurred at time t but it can also be interpreted as the probability that the subject will survive pas any specified time.

Let T be a continuous random variable representing the random lifetime taken from the population under study, we can define the survival function as follows:

$$S(t) = Pr(T > t) \tag{4.1}$$

The output of the survival function will obviously be values between 0 and 1 and it

will be a non-increasing function of t. Some main features of this function are that at time $t = 0$ the probability of surviving beyond a certain time t will be equal to 1, while for t which goes to infinity this probability inevitably reaches zero.

Another relevant aspect is that in theory, or rather, in case of numerous observations, the function is smooth, but in reality, since time is discrete, the function is step-wise as it is possible to observe in Figure 4.1.

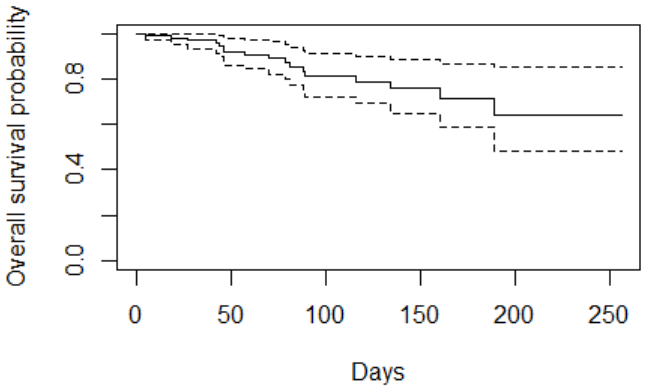


Figure 4.1: Kaplan-Meier with subset of 100 observations

It can also be very useful to plot the graph of the cumulative distribution function (CDF), representing the probability that the survival time is less than or equal to the specific time.

4.1.3 Hazard function

In order to understand the concept of hazard function it is necessary to speak first of failure rate.

The failure rate is the frequency with which an engineering component fails and is consequently a measure that depends on time. The hazard function results from the calculation of the failure rate for small time intervals and is indicated with $h(t)$. It follows that this function can be interpreted as the instantaneous failure rate and is considered a constant, i.e. the hazard rate for $\delta(t)$ that goes to zero. More practically, the hazard function is the probability that a subject will verify the event

of interest in a small time interval, provided that the individual has survived until the beginning of that time interval.

Mathematically, the hazard function can be written as follows:

$$h(t) = \lim_{\delta(t)\to 0} \frac{Pr(t \leq T \leq t + \delta(t) | T > t)}{\delta(t)} \tag{4.2}$$

4.2 Methodologies and tools of survival analysis

Depending on the purpose for which survival analysis is used, different analytical tools can be distinguished. Specifically:

- to describe the survival times of the members of a group via:
  - life tables
  - Kaplan-Meier curves
- survival times comparison of two or more groups, mainly via Log-rank test;
- to describe how covariates impact survival, via:
  - Cox proportional hazards regression
  - parametric survival models.

There are two main methods used to compute the survival function, life tables (or actuarial) and the Kaplan-Meier method.

4.2.1 Life tables and the actuarial method

The actuarial method is probably known to most people for its wide use in the insurance sector, where insurance policies are defined on the basis of mortality analyses, conducted by constructing life tables or mortality tables. For each age, these tables show the probability of a person's survival to their next birthday. From these tables it is also possible to understand the remaining life expectancy for people of different ages. This type of analysis is used when the exact time of occurrence of the event is not known, rather the time interval in which it occurred is known.

Obviously, the survival tables can now also include information regarding the lifestyle of people and are usually separate for men and women.

An important aspect in the construction of the life tables is the consideration also of the censored subjects and the consequent assumption that the subjects who withdraw from the study do so randomly during the time interval.[31]

Life tables typically contain the following information, and in Figure 4.2 we report an example:

- time interval t;
- number of patients at the beginning of the time interval n_t;
- number of patients who died during the time interval d_t;
- number of patients lost to follow-up during the time interval l_t;
- number of patients who withdrew alive during the time interval w_t;
- number of patients exposed to the risk of death $e_t = (w_t + l_t)/2$;
- conditional probability of death $q_t = \frac{d_t}{e_t}$;
- conditional probability of survival $p_t = 1 - \frac{d_t}{e_t} = 1 - q_t$;
- Survival Rate $P_t = p_1 * p_2 * ... * p_t$

4.2.2 Kaplan-Meier curves

The Kaplan-Meier or product-limit method, unlike the actuarial one, is favored when:

- the exact time of occurrence of the event of interest is known;
- the subjects lost during the follow-up participate in the determination of the survival function until their withdrawal.

The Kaplan-Meier estimator is a non-parametric statistic proposed by Kaplan and Meier in their 1958 paper, published in the Journal of the American Statistical Association, to estimate the survival function $S(t)$ even in presence of censoring and it is based on conditional probabilities.

Recall that a non-parametric statistic is not based on the assumption of an underlying probability distribution, which is also sensible since survival data typically

TAB. 38 *Rigolato 1775-1799: tavola di mortalità calcolata col metodo dei decessi generalizzato (r=-0,16‰)*

Classi d'età	*Età centrale*	$(1-r)^{\bar{x}}$	*N.ro dei decessi*	*Decessi rettificati*	*Decessi rettificati cumulati*	l_x	q_x	d_x	L_x	T_x	e_x
$(x, x+a)$	$\bar{x}$		D_x	$\frac{D_x}{(1-r)^{\bar{x}}}$							
0-1	0,5	0,9985	151	151	557	1000	271,1	271	865	33727	33,7
1-2	1,5	0,9955	36	36	406	729	88,7	65	697	32862	45,1
2-3	2,5	0,9925	18	18	370	664	48,6	32	648	32166	48,4
3-4	3,5	0,9895	13	13	352	632	36,9	23	621	31518	49,9
4-5	4,5	0,9866	5	5	339	609	14,7	9	605	30897	50,7
5-6	5,5	0,9836	4	4	334	600	12,0	8	596	30293	50,5
6-7	6,5	0,9807	4	4	330	592	12,1	7	589	29697	50,2
7-8	7,5	0,9777	6	6	326	585	18,4	10	580	29108	49,8
8-9	8,5	0,9748	6	6	320	575	18,8	11	570	28528	49,6
9-10	9,5	0,9719	2	2	314	564	6,4	4	562	27959	49,6
10-11	10,5	0,9689	1	1	312	560	3,2	2	559	27397	48,9
11-12	11,5	0,9660	1	1	311	558	3,2	1	558	26838	48,1
12-13	12,5	0,9631	3	3	310	557	9,7	6	554	26280	47,2

Figure 4.2: Example of life table [32]

have a right-skewed distribution and therefore applying directly normal distributions is not correct.

Mathematically, the estimator can be defined as the number of subjects who have not experienced the event yet, divided by the number of subjects at the beginning of the study, thus we can write:

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{4.3}$$

where n_i represents the number of subjects at risk prior to time t, whilst d_i represents the number of occurred events of interest at time t.

As for the actuarial method, the Kaplan-Meier also determines the survival function by constructing tables, where the occurrence of the event or the censoring of a subject determines the time, and subjects are no longer delineated according to predetermined time intervals. At the beginning of each time interval both the number of patients at risk and the number of deaths by the end of it are defined, and the new survival function is computed as the product of the survival functions of the previous intervals.

In addition to the survival table, it is also possible to obtain a graph representing

the Kaplan-Meier curves. These curves are characterized by being a series of declining horizontal steps, which for a sufficiently large sample of subjects approximates the real survival curve. The graph of the K-M suvival curve will therefore report time on the x axis, and on the y axis the probability that the subject has not yet verified the event of interest. Furthermore, as previously mentioned, the Kaplan Meier curves are particularly useful for comparing the survival curves of different groups of patients, i.e. the survival for patients with a different tumor stage, or that of patients who have been subjected to different types of medical treatment or interventions.[33]

In absence of censoring, the Kaplan-Meier estimate coincides with the empirical survival function.

4.2.3 The Nelson-Aalen Estimator

The Nelson-Aalen Estimator can be used in order to estimate the cumulative hazard function $\Lambda(t)$ and it can be defined as:

$$\hat{\Lambda}(t|X_i) = \sum_{j=1}^{i} \frac{d_j}{n_j} \tag{4.4}$$

where we recall d_i being the number of deaths at t_i and n_i be the number of subjects alive before t_i.

4.2.4 Log-rank test

The log-rank test is the method typically used in clinical trials to compare the survival curves of two or more groups. It is a non-parametric hypothesis test which is appropriate in presence of non normally distributed data as in the case of survival analysis where the data are right skewed and censored. In absence of censoring, the Wilcoxon rank sum test could be used.

Similarly to other hypothesis tests, the null hypothesis log-rank test assumes that there is no difference in the survival probabilities of the individuals recruited at different times during the study, and that the events happened at the times specified.

The log-rank test is a very useful way of comparing the survivals of two or more groups but is based on the strong assumption that within each level of the treatment

variable (or covariate) the populations are homogeneous in their survival

4.2.5 Proportional hazards regression

Proportional hazards models are a family of survival models in statistics that focuses on the hazard function by relating time that passes to one or more covariates, in particular the impact of the covariate on the hazard rate is multiplicative, hence the concept of proportional hazards.

As Wikipedia reports, survival models are characterized by the presence of two components: the baseline hazard function, denoted as $\lambda_0(t)$, which describes the way in which the risk of the event for instant or time changes over time at baseline levels of covariates, and the effect of parameters, which describe how the **hazard** (and not the lifetime) changes in response to covariates.

The European Agency for the Evaluation of Medicinal Products (EMEA) defines a baseline covariate as a "qualitative factor or a quantitative variable measured or observed before a subject starts taking study medication (usually before randomisation) and expected to influence the primary variable to be analysed".[34] It is also possible to distinguish many types of baseline covariates as they typically depend on the study.

4.2.6 Cox proportional hazard model

Cox Proportional Hazards Regression Model, also known as Cox partial likelihood, is a semi-parametric model developed by Sir David Cox in 1972 and it determines the relationship between the survival time and the impact of the covariates on the survival distribution.

Cox observed that if the proportional hazard assumption holds then it is possible to estimate the effect parameters without any consideration of the hazard function.

The model describing the hazard function at time t for the i-th subject with covariate vector X_i can be written as follows:

$$\lambda(t|X_i) = \lambda_0(t) \exp(\beta_1 X_{i1} + ... + \beta_p X_{ip}) = \lambda_0(t) \exp(X_i \beta) \tag{4.5}$$

Notice that in the Cox model, it is βX and not X that directly determines the

survival distribution of a subject and this quantity is defined *prognostic index*.

Cox model fits the data by maximizing the likelihood of the event of interest occurring for subject i at time t_i. In particular, let R_i be the risk set at time t_i, namely the set of subjects that are still alive at time t_i and let us also suppose that there are no ties in the observation times, namely only one subject can experience the event at time t_i. Therefore the likelihood that the event will occur only for this particular subject at time t_i given the risk set R_i will be:

$$L_i(\beta) = \frac{\lambda_0(t_i)e^{x_i\beta}}{\sum_{i \in R_i} \lambda_0(t_i)e^{x_i\beta}} = \frac{e^{x_i\beta}}{\sum_{i \in R_i} e^{x_i\beta}} \tag{4.6}$$

with $0 < L_i(\beta) \leq 1$ and it is a partial likelihood since only part of the parameters occur in it and in this case the likelihood is independent of the baseline hazard $\lambda_0(t)$. Assuming the subjects as statistically independent of each other the joint probability of all realized events can be written as the following partial likelihood:

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{x_i\beta}}{\sum_{i \in R_i} e^{x_i\beta}} \tag{4.7}$$

The corresponding log of Cox's partial likelihood is:

$$ln(\beta) = \sum_{i=1} (x_i\beta - \ln \sum_{i \in R_i} e^{x_i\beta}) \tag{4.8}$$

This function can be maximized by taking the derivatives with respect to β obtaining the partial score function:

$$ln'(\beta) = \sum_{i=1} (x_i - \frac{\sum_{i \in R_i} e^{x_i\beta} x_i}{\sum_{i \in R_i} e^{x_i\beta}}) \tag{4.9}$$

The maximization of the partial likelihood can be obtained taking both the score function and the Hessian matrix and using the Newton-Raphson algorithm.

To sum up, the Cox model is then fitted in two steps:

- At first, the parametric part is fitted via maximization of the Cox partial likelihood,
- Then, the non-parametric baseline hazard is estimated as a function of the parametric results.

4.2.7 Breslow Estimator

In order to estimate the baseline hazard function different approaches have been proposed and the main ones are:

- Breslow estimator (the most commonly used)
- Kalbfleisch/Prentice estimator, who use an argument similar to the derivation of the Kaplan-Meier estimate

Breslow suggested estimating the survival function as:

$$\hat{S}(t) = \exp\left(-\hat{\Lambda}(t)\right) \tag{4.10}$$

where $\hat{\Lambda}(t)$ is the Nelson-Aalen estimator of the integrated hazard. An important aspect that needs to be kept in mind is that when the number of deaths is small with respect to the number of subjects exposed the Breslow estimator and the Kaplan-Meier are asymptotically equivalent.

4.2.8 Accelerated Failure Time Regression Models

Accelerated Failure Time Regression Models (AFT) models are alternative parametric models to proportional hazards, which can be used when the PH assumption is violated. In fact, they assume the effect of the covariate on the hazard rate accelerates or decelerates the life course of the disease by some constant.

More formally, let $\boldsymbol{T}$ be a random variable of survival times and $\boldsymbol{X}$ be instead the vector of covariates. The AFT model defines the relationship between the survival function $S(t|X)$, for every $t \in T$ as:

$$S(t|X) = S_0(te^{(\beta^t X)}) \tag{4.11}$$

where:

- S_0 is the baseline survival function
- β^t is the vector of the regression coefficients
- $exp(\beta^t X)$ is the accelerated factor.

Or in alternative:

$$S(t|\theta) = S_0(\theta t) \tag{4.12}$$

where $\theta = \exp(\beta_1 X_1 + ..+)\beta_p X_p$. The relationship between the covariates and survival time can also be illustrated as a linear relation by simply taking the natural logarithm of (4.21), from which we can obtain the following:

$$Y = \ln T = \mu + \theta^t X + \sigma\epsilon \tag{4.13}$$

where:

- μ is the slope
- $\sigma > 0$ is the scale parameter
- θ^t is the vector of regression coefficients with $\theta = -\beta$
- ϵ is the error which is assumed to follow a specific parametric distribution

In particular for each distribution of ϵ there is a parametric distribution for T and the commonly used parametric distribution, and also the ones that were tried in this dissertation are the Weibull, exponential, log-normal and log-logistic. The results will be illustrated in the following chapter.

How can we interpret an Accelerated Failure Time model? Suppose that theta is equal to two: for the proportional hazard assumption the hazard function should always be twice as high, while for AFT models a $\theta = 2$ means that the relevant elements in a subject's history occur twice as fast.[35]

4.2.9 Performance and accuracy metrics in survival analysis

There are numerous metrics that have been devised to measure the performance of survival analysis models. Below we illustrate the most common ones that coincide with those used in our survey.

There are two important aspects that a metric for evaluating the performance of prediction models in survival analysis cannot ignore: calibration and discrimination. Discrimination represents the ability to separate observations with different labels

in a distinct and above all correct way. Calibration instead calculates how close the estimates are to the corresponding real value.

ROC curve and AUC

A classic performance measure used in machine learning is the AUC-ROC curve. ROC curve stands for *Receiver Operating Characteristic* curve and is a graph that shows the performance of the classification model and can be obtained by representing the true positive rate vs the false positive rate at classification settings.

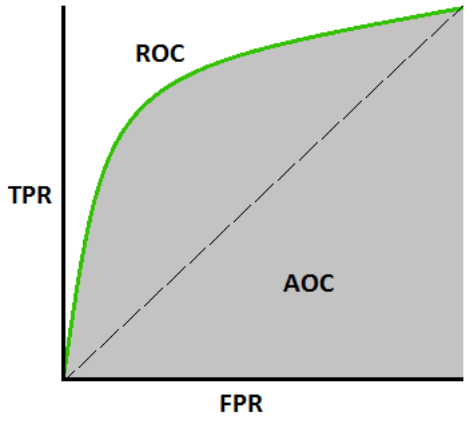


Figure 4.3: AUC-ROC Curve [36]

AUC instead stands for *Area Under the ROC* curve and therefore measures the area below the ROC curve, providing a measure of the performance across all possible classification thresholds and will have a value between 0 and 1.

Harrell's Concordance Index

As the name also suggests, the c-index is meant to measure the agreement between two variables and in particular it a measure of discrimination. This index is a generalization of the concept of area under the curve (AUC), applied to data that can be censored and consequently to the survival analysis and indicates the correctness of the model based on the individual risk scores in defining the ranking of survival and can be defined as the ratio of all pairs of subjects whose predicted survival times are correctly ordered among all subjects that can actually be ordered.[37]

The correct interpretation of the index is as follows:

- concordance-index = 0.5: random predictions;

- concordance-index = 1: perfect concordance
- concordance-index = 0.0: in this case it is necessary to multiply the predictions by -1 to get perfect concordance.

Brier Score

The Brier score is always used to measure the accuracy of the prediction made by the survival model, at a specific t period. The Brier score prediction error is the squared distance between the indicator function of the event having taken place before time t and the predicted event probability, or, equivalently, the average squared distances between the observed survival status and the predicted survival probability. It can take values between 0 and 1 and usually a model is considered useful if its Brier score is less than 0.25, consequently the lower the Brier Score, the greater the goodness of the model.

Index of Performance accuracy

The Index of Prediction Accuracy (IPA) was proposed in 2018 by Kattan and Gerds as an alternative to the more commonly used accuracy indexes of survival analysis models.[38] As a matter of fact, since the c-index is just a discrimination measure it requires further calibration procedures, whilst the IPA results from the re-scaling of the Brier score and does not require further adaptations as well as providing a more simplified interpretation, particularly useful especially for clinical staff. Another feature of the IPA is that unlike the C-index, it also considers whether the model is dangerous and/or useless. The IPA is then obtained by the following formula:

$$\text{IPA} = 1 - \frac{\text{model Brier score}}{\text{null model Brier score}} \tag{4.14}$$

where the null model does not contain any covariate (typically estimated through a Kaplan-Meier model, in the absence of competing risks or an Aalen-Johansen method, in presence of competing risks). As far as the interpretation is concerned, an IPA equal to 100% represents a perfect model, a negative or zero value represents a useless model and instead a harmful model will be represented by a negative value. Consequently in general, the higher the IPA value, the greater the accuracy,

usefulness and precision of the model.

4.3 Machine learning algorithms

Several machine learning algorithms have been implemented in order to understand if these more "complex" models bring real added value to the more classic models of survival analysis. In this chapter we take care to introduce the basic mathematical concepts while in the next chapter the logic used and the relative results will be illustrated.

4.3.1 Introduction

As Tom Mitchell (former chair of the machine learning department at the Carnegie Mellon University (CMU)), stated in his book entitled "Machine Learning" (1997), Machine learning is the study of computer algorithms that improve automatically through experience and by progressive learning on the data given as input.[39]

In general, it is possible to distinguish three types of machine learning:

- supervised learning, the algorithms belonging to this class predict the dependent variable/target/label from a given set of predictors by generating a function that maps the inputs to the desired output. Some examples of supervised learning algorithms are linear and logistic regression, decision trees and their evolution, i.e. random forests, k-nearest neighbors (KNN).

- unsupervised learning, in this type of learning the algorithm is provided with a dataset without explaining what you want to achieve or what the correct result is. The algorithm then automatically tries to find a pattern among the data; a classic example of an algorithm belonging to this class is clustering, which allows to identify similar groups of observations/data points.

- reinforcement learning, in this case what we want to achieve is an optimal solution to achieve a certain goal, and every time we get close to the final goal, you receive a reward.

In this dissertation we will mainly focus on supervised algorithms, where the goal is that of learning functions that map x a points to y labels and once known,

these functions can be used to classify images, documents or to stay in the medical field, if a patient has or will develop in future a certain disease. This is done by providing the machine learning algorithm with a dataset containing the observations and the related labels, on which the algorithm will train and which for this reason is called training set and subsequently the goodness of the model is tested on a dataset without the labels and called test set. Depending on the type of labels y, two types of problems can be distinguished:

- classification problem if the labels are categories, for example spam or non-spam emails,

- regression problem when the label is instead numeric, such as the demand for a particular product.

The identification of the error in the case of classification problems is immediate and binary: if the predicted label does not coincide with the real one, an error has been made. In the case of regression problems, on the other hand, it is necessary to delineate a measure of the distance between the predicted label and the real one. Therefore, to measure the goodness of a prediction in both types of problems, a non negative loss function l is used, so if y represents the correct label of an observation x, the goodness of the prediction $\hat{y}$ is measured by $l(y, \hat{y}) \geq 0$.

The zero-one loss is the most popular type of loss function with regard to classification problems and in particular:

$$l(y, \hat{y}) = \begin{cases} 0 & if \quad y = \hat{y}, \\ 1 & \text{otherwise.} \end{cases} \tag{4.15}$$

. As for regression problems, two typical loss functions are the absolute loss $l(y, \hat{y}) = |y - \hat{y}|$ function and the quadratic loss function $l(y, \hat{y}) = (y - \hat{y})^2$. We can therefore define the function $f : X \to Y$ as a predictor, where X is the set of data points while Y represents the set of possible labels.

4.3.2 ERM, Overfitting and Underfitting

As previously anticipated, the input given to a learning algorithm is the training set so the first idea would be that of computing the training error

$$\hat{l}(f) = \frac{1}{m} \sum_{t=1}^{m} l(y_t, f(x_t)) \tag{4.16}$$

The Empirical Risk Minimizer (ERM) is the algorithm that outputs the predictor f in the set F that minimizes the training error, i.e.:

$$\hat{f} = \underset{f \in F}{\operatorname{argmin}}\, \hat{l}(f) \tag{4.17}$$

Using this technique can lead to some problems such as that of overfitting (low training error, but high test error) or underfitting (high training error and high test error). Furthermore, it is necessary the assumption that the examples (x, y) are randomly generated from a joint probability distribution D thus implying that both the training and the test sets are random samples.

4.3.3 Linear Prediction and Perceptron

In order to understand some features of neural networks, we introduce a few important and fundamental concepts in this section.

Linear Classifiers

Linear classifiers predict a label for an observation given in input, based on the value assumed by the linear combination of the characteristics or features of the observation.

From a geometric point of view a linear classifier (or predictor) is a function $h : \mathbb{R}^d \to \{-1; +1\}$ with the partition $\{S^+, S^-\}$ of $\mathbb{R}^d$ is such that:

$$h(x) = \begin{cases} +1 \quad if \quad x \in S^+, \\ -1 \quad if \quad x \in S^-. \end{cases} \tag{4.18}$$

where S^+, S^- are the half-spaces defined by a S hyperplane in $\mathbb{R}^d$. Algebraically, a

hyperplane is the locus of the points $x \in \mathbb{R}^d$ which satisfy the equation $a_1x_1 + .. + a_dx_d = b$ with $a_1, .., a_d, b$ real coefficients. Using the notation: $u^T a = \sum_{i=1}^{d} u_i a_i$ for the inner product, you can rewrite the hyperplane S as $S(a, b) = \{x \in \mathbb{R}^d : a^T x = b\}$.

Let now θ be the angle between two vectors v and x such that the length of the projection of x on a multiplied by $||a||$ is given by $a^T x = ||a|| \, ||x|| \cos\theta$, where $||x||$ represents the Euclidean norm of a vector $x = (x_1, .., x_d)$. Consequently, the vector a is perpendicular to the hyperplane $S = S(a, b)$, which cuts v at the distance $b/||a||$ from the origin. It follows that the two half-spaces defined by S can be rewritten as follows: $S^+ = \{x : a^T > b\}$ e $S^- = \{x : a^T \leq b\}$. The linear classifier h associated with the $S(a, b)$ hyperplane can be represented as $h(x) = sgn(a^T x - b)$ where the sgn function is defined as follows:

$$sgn(x) = \begin{cases} +1 \quad if \quad x > 0, \\ -1 \quad \text{otherwise.} \end{cases} \tag{4.19}$$

For simplicity, it is possible to reduce any hyperplane to a so-called homogeneous hyperplane, in which $b = 0$: a non-homogeneous hyperplane $S(a, b)$ in d dimensions is in fact equivalent to a homogeneous hyperplane $S(\tilde{a}, 0)$ in $d + 1$ dimension with $\tilde{a} = (a_1, .., a_d, -b)$ with the points $x \in \mathbb{R}^d$ mapped to the points $\tilde{x} = (x_1, .., x_d, 1) \in \mathbb{R}^{d+1}$.

Perceptron

The perceptron is a binary classifier, but it is also a type of neural network model as it consists of a single node or neuron, predicting the labels by taking only a row of data in input. In particular, binary classifiers are a type of linear classifiers, and therefore functions $f : X \to Y$ where in particular Y contains only two labels, typically $\{+1, -1\}$.

The Perceptron computes the weighted sum of the inputs and a bias (typically the first weight corresponds to the bias, which will be also updated at each iteration). The linear combination of the weighted sum and the bias defines the *activation* which is later transformed into an output value using a *transfer function*, such as a unit step activation function. It follows that the algorithm will predict 1 if the *activation*¿0, 0 otherwise. Figure 4.7 can provide valuable help for less experienced

readers.

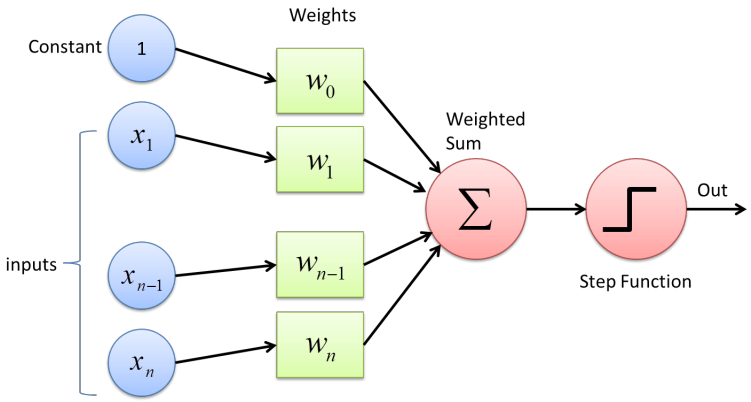


Figure 4.4: Perceptron[40]

Algorithm 1: Perceptron

Input: Training set $(x_1,y_1),\ldots,(x_m,y_m)$
Initialization: of the weights w $= (0,\ldots,0)$ and of the thresholds
Repeat: Read next training example (x_t, y_t);
1 **while** $y_t w^T x_t > 0$ **do**
2 **end**
3 **if** $y_t w^T x_t \leq 0$ **then**
4 $\quad w \leftarrow w + y_t x_t$
5 **end**
Output: w

The combination of many perceptrons forms an artificial neural network.

Statistical vs Sequential Risk

In order to be able to analyze a learning algorithm and to be able to understand and consider reliable the results obtained, it is necessary to set up a mathematical model that defines how the examples (x, y) are generated. In the statistical learning framework, it is assumed that the examples are extracted randomly and independently from an established but unknown probability distribution D on $X \times Y$ and consequently every observation x is equally probable. Likewise, this implies that each dataset, including the training set and the test set, is a random sample. In statistical learning a problem is fully specified by the pair (D, l) where l is the loss function and the performance of a predictor h is calculated by measuring the statistical risk $l_d(h) = \mathbb{E}[l(Y, h(X))]$, which corresponds to the expected value of the loss function on a random example (X, Y) extracted from D.

The best possible predictor is the *Bayes optimal predictor*, defined as $f^*(x) = \min_{\hat{y} \in Y} \mathbb{E}[l(Y, \hat{y})|X = x]$ and where $f^*(x)$ is the prediction that minimizes the conditional risk, in turn corresponding to the expected loss of the prediction with reference to the distribution of the Y conditional label at x. It therefore follows that $\mathbb{E}[l(Y, h(x))|X = x] > \mathbb{E}[l(Y, f^*(x))|X = x]$ for each predictor $h : X \rightarrow Y$ and for each $x \in X$ and since the conditional risk average is equal to the risk itself, it follows that $l_D(h) > l_D(f^*)$ for each h editor, where $l_D(f^*)$ is defined as Bayes error, typically greater than zero.

Online Learning

In the statistical learning model the assumption made is that the data are generated by a probabilistic model, not known and the evaluation of the goodness of the predictor takes place through the statistical risk. The online learning model, on the other hand, was designed to evaluate a predictor when data is provided in the form of any sequence of data. The perceptron is a special case of this model. The protocol of the online learning model is defined by the following:

Algorithm 2: Online Learning

1 The algorithm generates a starting model w_1
2 For $t = 1, 2, ...$

1. The initial model w_1 is tested on the next example (x_t, y_t)
2. The algorithm then updates the model w_t generating a new model w_{t+1}

The sequence of models generated for each t are evaluated by measuring the sequential risk (which replaces the statistical risk), that is $\frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(y_t w_t^T x_t \leq 0)$ which measures as T varies, the fraction of prediction errors made by the sequence of models generated on the first T examples. Therefore, if in statistical risk what we want to measure is how quickly it decreases as the size of the training set increases, sequential risk intends to provide a measurement of how quickly the risk decreases as T increases.

The sequential learning model learns through progressive optimizations of the original model, as opposed to the statistical model which solves a global optimization problem. The sequential model is therefore particularly advantageous in all

those contexts in which data are constantly provided and the predictive model must therefore be adapted to the new input data.

Online Gradient Descent (OGD) algorithm

The *Online Gradient Descent* is a sequential algorithm capable of optimizing any convex loss function l. Remember, what we want to obtain is a minimization of the convex and differentiable loss function $l : \mathbb{R}^d \rightarrow \mathbb{R}$. Below we illustrate the OGD with projection assuming that $l_1, l_2, ..$ are convex and twice differentiable functions, followed by a brief description of the arguments:

Algorithm 3:

Parameters: constant ρ, radius $r > 0$
Initialization: $w_0 = 0$

1 For $t = 1, 2, ...$

1. $w'_{t+1} = w_t - \frac{\rho}{\sqrt{t}\nabla l_t(w_t)}$
2. $arg\,min_{w:||w||\leq r}||w - w'_{t+1}||$

The optimization of the loss function takes place starting from an arbitrary point w_0 to which the operation in the first step of the algorithm is applied, where $\rho_t > 0$ is a parameter. If the current point is not a minimum then the gradient $\nabla l_t(w_t)$ will be greater than 0 and consequently w_{t+1} will move towards the minimum of the function. In the second step w'_{t+1} is projected into a Euclidean sphere of radius r so if $||w - w'_{t+1}|| \leq r$ then $w_{t+1} = w'_{t+1}$. The ultimate goal is to minimize the difference between the sequential risk of the algorithm and that of any v model such that $||v|| \leq r$ and therefore v^*_T is the best predictor for the first T steps and consequently $arg\,min_{v:||v||\leq r}\frac{1}{T}\sum_{t=1}^{T} l_t(v)$.

4.3.4 Survival Support Vector Machines (SVM)

Readers who are already sufficiently familiar with machine learning algorithms topics will probably have already heard of Support Vector Machines (SVM) and it is consequently immediate to understand that Survival SVM are simply an extension to the field of survival analysis. However, we provide below a brief explanation regarding the logic and operation of this algorithm, always referring the reader to the

sources cited to deepen the concepts illustrated.

It is a supervised machine learning algorithm developed at Bell Laboratories by Vladimir Vapnik extremely effective in binary classification problems but also widely used in regression problems. The basic idea is to find a hyperplane that best divides the observations of training set given as input to the model, so that a data point will be assigned a label depending on whether it is on a side rather than from another of the hyperplane thus defined.

SVM

Support Vector Machines is a learning algorithm for linear calculators which, given a linearly separable training set $(x_1, y_1), ..., (x_n, y_n) \in \mathbb{R}^d \times \{-1; 1\}$, generates the linear classifier corresponding to the unique solution $w^* \in \mathbb{R}^d$ of the following problem of convex optimization with linear constraints:

$$\begin{cases} \min_{w \in \mathbb{R}^d} \frac{1}{2}||w||^2, \\ \text{s.t.} \quad y_t w^T x_t \geq 1 \quad t = 1, .., m \end{cases} \tag{4.20}$$

where $w*$ represents the maximum margin separator hyperplane.

It is possible to prove that for each $(x_1, y_1, .., x_m, y_m) \in \mathbb{R}^d \times \{-1, +1\}$ linearly separable, the vector u^* which realizes the maximum margin

$$\lambda^* = \max_{u:||u||=1} \min_{t=1,..,m} y_t u^T x_t \tag{4.21}$$

satisfies $u^* = \lambda * *w^*$, where ω^* is a solution of 4.34.

Put simply, solving the problem by maximizing the u margin while keeping the $||u||$ norm constant is equivalent to solving the $||\omega||$ norm by minimizing the $||\omega||$ norm while maintaining the ω margin constant. The solution of the problem corresponds to: $\omega^* = \sum_{t \in I} \alpha_t y_t x_t$ where I represents the set of data points of the training set (x_t, y_t) such that $y_t(\omega^*)^T x_t = 1$. In particular the x_t arguments are the so-called support vectors, i.e. those observations on which ω^* has margin exactly equal to 1 and consequently removing from the training set the examples different from those of support the SVM solution would not suffer variations. In the case of a non-linearly separable training set, assumption made at the beginning of this

section, quantities called slack variables are added to equation 4.21 which provide a measure of how much each margin constraint is violated by a potential solution ω.

4.3.5 Random survival forests

Random survival forests were developed in 2008 by Ishwaran et al. in 2008 as an extension of the random forests method, for the analysis of right-censored survival data.[41]

In order to fully understand how the algorithms works it is necessary to start by introducing the concept of decision trees to then move on to the random forest algorithm.

Decision Trees

Decision trees are one of the simplest and most intuitive classification and regression algorithms as they are often used unconsciously to make decisions in everyday life. They belong to the family of supervised learning algorithms. Let $\chi = \chi_1 x ... x \chi_d$ tree predictor $h_T : \chi \rightarrow Y$ is structured as ordered and rooted tree T, where at each internal node corresponds a test whilst to each leave correspond a label in Y. The test for the internal nodes with k children is a function $f : \chi_i \rightarrow (1, .., k)$ mapping each element of χ_i to a node children.

Simply put, a decision tree defines the prediction following a set of if-else conditions.

Figure 4.5 shows a very trivial example of a decision tree that could help the reader understand the structure of the algorithm.[42]

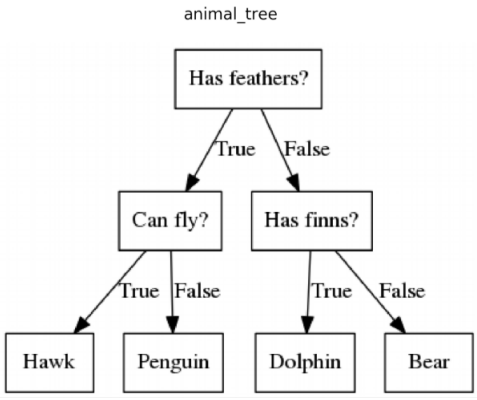


Figure 4.5: Decision tree example

Building a tree usually follows two steps:

- growth, obtained by binary subdivision, for which the variability within the node is sought to break the nodes (usually using the Gini index or evaluating the cross-entropy)
- pruning, which involves removing nodes.

In order to improve predictive accuracy, multiple trees can be combined through defined ensemble solutions such as bagging, boosting or random forests.

Random Forests

Random forests were introduced by Breinman in 2001 and they involve the construction of numerous decision trees that operate as an ensemble. Each tree predicts a certain label and the final prediction of the model will coincide with the one that was most frequently predicted.

Random Survival Forests

The survival random trees algorithm is based on the Breiman algorithm but adapted for the survival analysis features, in particular we summarize the main points of interest below:

- the survival period (or follow up) and the information regarding censoring in the splitting criterion used during the growth phase of the tree are taken into consideration.
- the measurement of the effectiveness/cost of the split must provide a measurement of the difference based on the difference in survival.
- Finally, the predicted value for a terminal node and the measurement of the accuracy of the prediction must incorporate survival information.

Note: for the next part concerning machine learning algorithms some parts have been adapted from a previous work (on the classification of Turkish banknotes) by the author published on Github for a university project.

4.3.6 Artificial Neural Networks

An artificial Neural Network (ANN) is composed by nodes interconnected between them via links: each node represents an artificial neuron and each link has a weight representing the strength of one neuron's influence on another. The learning process is very similar to that of the perceptron: examples are processed by computing the weighted associations between input and relative output. The difference between the prediction and the real value of the observation is then computed and the network adjusts the weights according to a learning rate and progressive adjustments will help the network to make more accurate predictions.

Artificial Neuron

An artificial neuron is a mathematical function: it receives one or more inputs that can be the features of the object in analysis and computes the weighted sum of them in order to produce an output.

Activation function

The weighted sum previously obtained is then passed into a non-linear function known as activation function.

Among the main examples of activation functions are:

- *Threshold function* - With this function the neuron remains non active up until a certain threshold is reached but it is quite difficult from a mathematical point of view.

- *Sigmoid function* - In this function the highest the hyperparamenter β, the steepest the graph around the zero.

$$\sigma(x) = \frac{1}{1 + \exp(-\beta x)} \tag{4.22}$$

- *ReLu activation function* - This is a non linear function that gives an output x if x is positive, otherwise it will output 0.

$$f(x) = max(0, x) \tag{4.23}$$

- *Softmax activation function* - The softmax function is a generalization of the logistic function and it takes as input a vector m of Z real numbers and normalizes it into a probability distribution that consists of Z probabilities. By assigning it to the output layer of the neural network for categorical target variables, the outputs can be interpreted as posterior probabilities. We briefly remind the reader that in Bayesian statistics the posterior probability is the conditional probability that derives once an important fact of the past has been taken into account.[43]

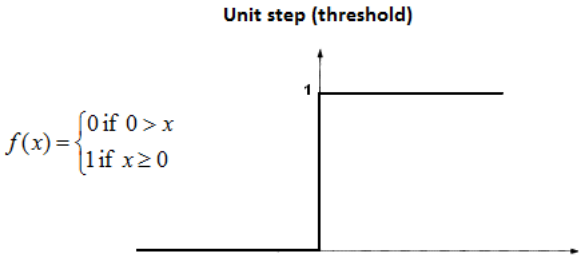


Figure 4.6: Threshold function

The choice of the activation function depends on the learning problem and it is possible to choose a different activation function for each layer.

Hyperparameters

A hyperparameter is a parameter whose value is used to control the learning process while the other parameters (such as the weights of the nodes) are derived through the training. The most known hyperparameters are probably:

- *Learning rate* - It determines the step size at each iteration that the model takes to adjust the errors at each observation. The highest the learning rate, the shorter the training time but the lower the accuracy.

- *Batch size* - It controls the number of training samples that will be propagated through the network.

- *Hidden layers* - Are those layers between the input and the output ones and are called "hidden" because they are not visible outside of the neural network.

- *Epochs* - It controls the number of complete passes through the training dataset, namely it is a measure of the number of times all the traing vectors are used once to update the weights.

- *Dropout* - The dropout hyperparameter can be interpreted as the probability during the training phase of a node in a layer of being discarded or equivalently the probability of not being discarded. It is comprised between 0.0 and 1.0 where the former means that the layer does not give any output and the latter means that no dropout is used. A value between 0.5 and 0.8 is usually considered a good dropout rate and it is used to prevent overfitting.

Accuracy metrics

Once the loss function that takes as inputs the weights, the biases and the training examples is fixed, it is necessary to define a metric that measures the "goodness" of the model. Ideally, the lower the output of the loss function, the better the neuron is at predicting the data. The most known metric are probably the *Accuracy* and the *Kullback-Leibler divergence* or *Cross-entropy*.

Accuracy is typically defined as the ratio between the number of correct prediction with the number of total predictions. For binary classification in particular we can write:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.24}$$

Where:

- TP = true positives;

- TN = true negatives;

- FP = false positive;

- FN = false negatives.

The *Kullback-Leibler divergence* originates from the mathematical definition of entropy of a discrete probability distribution and it measures how similar a probability distribution p is to a candidate distribution q.[44]

$$D_{KL}(p|q) = \sum_{i=1} p_i log \frac{p_i}{q_i} \tag{4.25}$$

4.3.7 Feed-Forward Neural Network

A feed-forward neural network is an artificial neural network where the term "forward" refers to the fact that information moves in only one direction, forward, from the input nodes, through the hidden nodes, ending in the output nodes, without employing cycles or loops.

This type of neural network computes a function $f : \mathbb{R} \rightarrow \mathbb{R}$. Each node i (but the input nodes) solves the function $g(v)$.

Single-Layer Perceptron

A single layer perceptron is the simplest kind of feed-forward neural network.

Multilayer Perceptron

MLP is a class of feed-forward artificial neural network and it consists of at least three layers of nodes: an input, a hidden and an output layer.

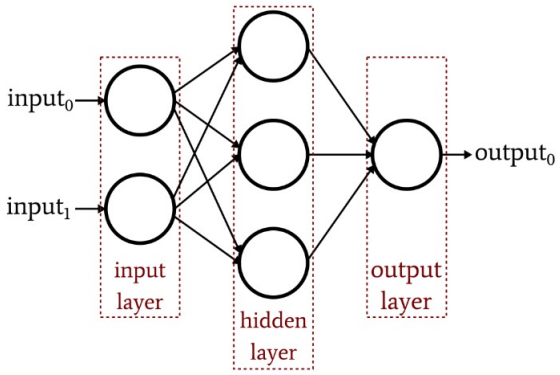


Figure 4.7: Multi-layer Perceptron

The nodes can be partitioned into a sequence of layers such that each node of a layer has incoming edges only from nodes in the previous layer and the outgoing edges go only to nodes in the next layers.

4.3.8 CoxTime

The *CoxTime*, introduced in 2019 by H. Kvamme et al. [45] model provides and extension of Cox Regression beyond the proporional hazards assumption through the use of neural networks. The main difference between the classic Cox model and the CoxTime, consists in the fact that while in the former the partial likelihood is usually minimized through the Newton-Raphson method, in this new algorithm the optimization occurs through stochastic gradient descent in order for the method to scale well especially with large datasets.

Assuming a fixed dimension of the Risk set R_i the authors are able to reduce the loss function in the following formula

$$loss = \frac{1}{n} \sum_{i:D_i=1} \log \left(\sum_{j \in \tilde{R}_i} exp[g(x_j - g(x_i))] \right) \tag{4.26}$$

where is it:

- $f(t)$, probability density function
- $F(t)$, cumulative distribution function
- $S(t)$, survival function
- $h(t) = \frac{f(t)}{S(t)}$, hazard rate
- $H(t)$, cumulative hazard
- T^*, True event time
- $D = \mathbb{I}(T = T^*)$, indicator function that labels the event observed at time T as an event or alternatively a censored observation
- C^*, censoring time
- x_i, covariates
- $exp[g(x)]$ is the relative risk function
- $g(x) = \beta^T x$
- β is the parameter vector

- n, number of events in the data set
- R_i, risk set
- $\tilde{R}_i$, sampled risk set

The function can now be optimized using SGD by replacing the original $g(x)$ formula with one parametrized by a neural network. As a matter of fact another change to Cox's original model is to make the relative risk function dependent on time.

4.3.9 DeepSurv

DeepSurv is a Cox proportional hazards deep neural network aiming at modeling the interactions between a patient's covariates and the effectiveness of a specific treatment. [46] The model of Katzman et al., unlike the previous one, remains anchored to the assumption of proportional hazards on which the classic Cox model is also based.

Both the CoxTime and the DeepSurv models exploit the ability of neural networks to understand and learn complex and non-linear relationships between covariates and the probability of survival of individuals.

The objective function to optimize is the average negative log partial likelihood with regularization, using the previous notation:

$$l(\theta) = -\frac{1}{n} \sum_{i:D_i=1} \left(\hat{h_\theta}(x_i) - \log \sum_{j \in R(T_i)} e^{\hat{h_\theta}(x_j)} \right) + \lambda ||\theta_2^2|| \tag{4.27}$$

where:

- λ is the l_2 regularization parameter
- $R(t) = \{i : T_i \geq t\}$, is the set of patients that are still at risk of failure at time t
- θ is the network

$\hat{h_\theta}(x)$ is the output of the network and is a single node with linear activation function, estimating the log-risk function in the Cox model.

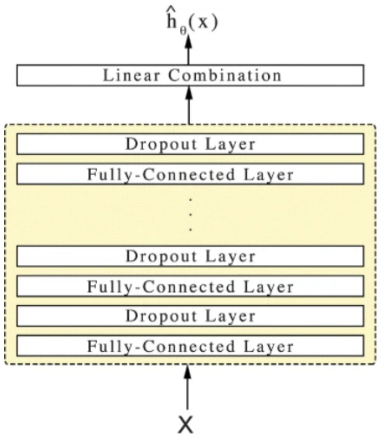


Diagram of DeepSurv. DeepSurv is a configurable feed-forward deep neural network. The input to the network is the baseline data x. The network propagates the inputs through a number of hidden layers with weights θ. The hidden layers consist of fully-connected nonlinear activation functions followed by dropout. The final layer is a single node which performs a linear combination of the hidden features. The output of the network is taken as the predicted log-risk function $\hat{h}_\theta(x)$. The hyper-parameters of the network (e.g. number of hidden layers, number of nodes in each layer, dropout probability, etc.) were determined from a random hyper-parameter search and are detailed in Table 3

Figure 4.8: Diagram of the DeepSurv [46]

Again, given the vastness of these topics it is not possible to go into every single algorithm in detail here. We therefore advise the reader to deepen the topics further by consulting the aforementioned papers.

Chapter 5

Analysis

The first part of this chapter focuses on the construction of the dataset used for our investigation. Secondly, the analytical methods implemented will be illustrated.

5.1 Building of the data set

As previously said the dataset was built using the Seer*Stat software made available by the SEER program. The focus of the analysis was female patients who were diagnosed with breast cancer in the years between 1995 and 2015, thus obtaining a time span of 20 years. Specifically, we opted for those patients for whom breast cancer was the only one or at least the first of two or more tumors; the final sample, without taking into account the Nan Values, includes 870.129 subjects.

In appendix B the reader will find the sequence of steps to implement in order to replicate the creation of the dataset.

5.1.1 Variable selection and description

In this section we intend to describe the variables extracted from the SEER program, to understand their typology and conduct an exploratory analysis on them. For a more precise and meaningful list of values, the reader is advised to always consult the website of the SEER program which provides various pages and also dictionaries describing the variables. However, since the publications are numerous and for many of them the site re-sends to other links[47], we have proposed to collect the descriptions relating to the variables used in this project, with the aim of providing

the reader with a more "fluent" reading.

Patient ID

It is a unique categorical variable representing the identifier per each patient/subject.

Race recode

Categorical variable describing the ethnicities which can assume the following values:

- White
- Black
- Other (American Indian/AK Native, Asian/Pacific Islander)

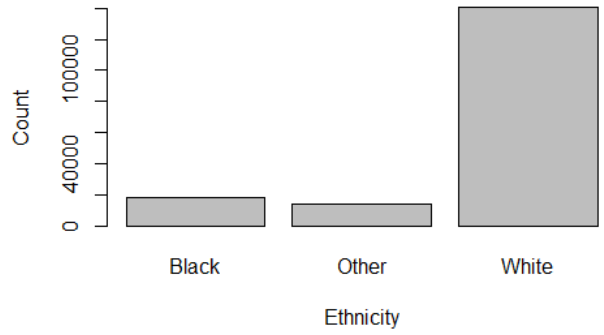


Figure 5.1: Ethnicity count in dataset previous to pre-preprocessing

As it is possible to observe from Figure 5.1 patients of white ethnicity highly outnumbers the other two ethnicities.

State-County

Categorical variable describing both the State and County at diagnosis.

Primary Site – labeled

Categorical variable describing the site of the tumor:

The **C50.9** was recoded as NaN and therefore the observations having this value were not considered in the analysis.

Table 5.1: Primary Site – labeled

Code	Description
C500	Nipple
C50.1	Central portion of breast
C50.2	Upper inner quadrant of breast
C50.3	Lower inner quadrant of breast
C50.4	Upper outer quadrant of breast
C50.5	Lower outer quadrant of breast
C50.6	Axillary tail of breast
C50.8	Overlapping lesion of breast
C50.9	Breast NOS (DO NOT CODE)

Grade

Categorical variable describing the grade of the tumor:

- Moderately differentiated; Grade II
- Poorly differentiated; Grade III
- Undifferentiated; anaplastic; Grade IV
- Well differentiated; Grade I

This variable is equivalent to the *Breast - Adjusted AJCC 6th Stage (1988-2015)* variable.

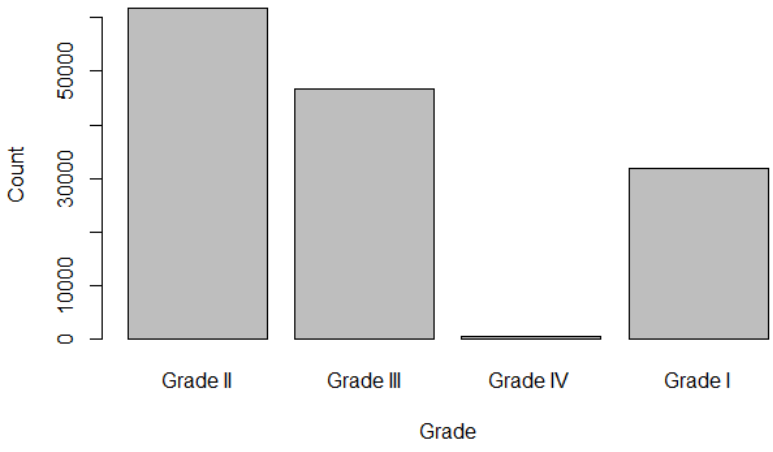


Figure 5.2: Grade distribution in dataset previous to pre-preprocessing

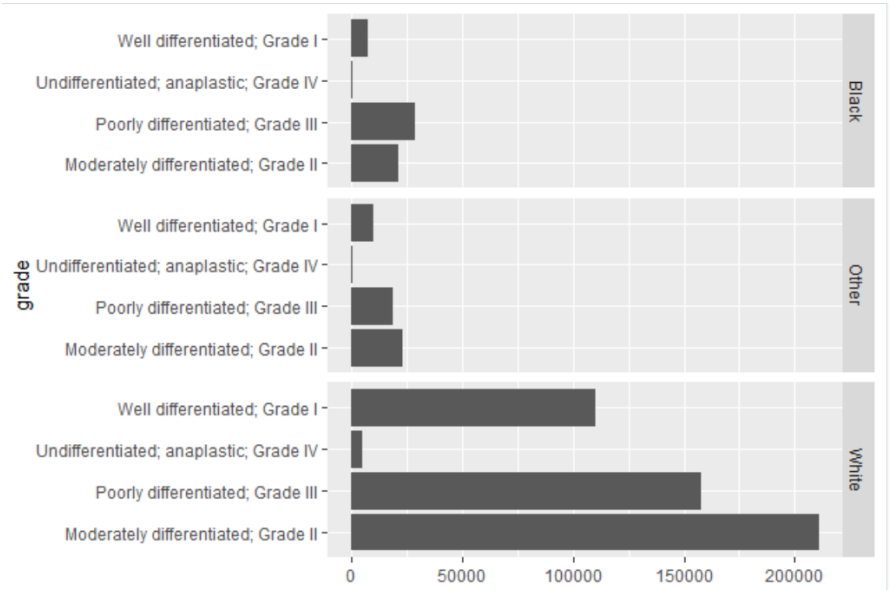


Figure 5.3: Grade distribution per ethnicity

Laterality

Laterality is a categorical variable as well although expressed with a numerical value, in particular:

Table 5.3: Laterality

Code	Description
0	Not a paired site
1	Right: origin of primary
2	Left: origin of primary
3	Only one side involved, right or left origin unspecified
4	Bilateral involvement at time of diagnosis, lateral origin unknown for a single primary; or both ovaries involved simultaneously, single histology; bilateral retinoblastomas; bilateral Wilms tumors
5	Paired site: midline tumor (effective with 01/01/2010 dx)
9	Paired site, but no information concerning laterality

Codes from the **95** to the **98**, as well as the **0** and the **3** because of the low frequency of these two values, were recoded as NaN and therefore the observations having these values were not take into account in the analysis.

ICD-O-3 Hist/behav, malignant

Categorical variable that reports the ICD-O-3 code (illustrated in chapter 3 *Data Sources and Literature*) of the type of malignant tumor:

Table 5.5: ICD-O-3 Hist/behav, malignant

Code	Description
8000/3	Neoplasm, malignant
8010/3	Carcinoma, NOS
8032/3	Spindle cell carcinoma, NOS
8046/3	Non-small cell carcinoma
8050/3	Papillary carcinoma, NOS
8070/3	Squamous cell carcinoma, NOS
8072/3	Squamous cell carcinoma, large cell, nonkeratinizing, NOS
8140/3	Adenocarcinoma, NOS
8200/3	Adenoid cystic carcinoma
8201/3	Cribriform carcinoma, NOS
8211/3	Tubular adenocarcinoma
8230/3	Solid carcinoma, NOS
8255/3	Adenocarcinoma with mixed subtypes
8260/3	Papillary adenocarcinoma, NOS
8343/3	Papillary carcinoma, encapsulated
8401/3	Apocrine adenocarcinoma
8480/3	Mucinous adenocarcinoma
8500/3	Infiltrating duct carcinoma, NOS
8501/3	Comedocarcinoma, NOS
8502/3	Secretory carcinoma of breast
8503/3	Intraductal papillary adenocarcinoma with invasion
8504/3	Intracystic carcinoma, NOS
8507/3	Ductal carcinoma, micropapillary
8510/3	Medullary carcinoma, NOS
8520/3	Lobular carcinoma, NOS
8521/3	Infiltrating ductular carcinoma
8522/3	Infiltrating duct and lobular carcinoma
8523/3	Infiltrating duct mixed with other types of carcinoma
8524/3	Infiltrating lobular mixed with other types of carcinoma
8530/3	Inflammatory carcinoma
8540/3	Paget disease, mammary
8541/3	Paget disease and infiltrating ductal carcinoma of breast
8543/3	Paget disease and intraductal carcinoma
8560/3	Adenosquamous carcinoma
8572/3	Adenocarcinoma with spindle cell metaplasia
8573/3	Adenocarcinoma with apocrine metaplasia
8575/3	Metaplastic carcinoma, NOS
9020/3	Phyllodes tumor, malignant

RX Summ - Surg Prim Site (1998+)

It is a categorical variable describing whether or not a surgery was performed on the primary site:

Table 5.7: RX Summ - Surg Prim Site (1998+)

Code	Description
00	None; no surgical procedure of primary site; diagnosed at autopsy only
10-19	Site-specific codes. Tumor destruction; no pathologic specimen or unknown whether there is a pathologic specimen
20-80	Site-specific codes. Resection; pathologic specimen
90	Surgery, NOS. A surgical procedure to the primary site was done, but no information on the type of surgical procedure is provided
98	Special codes for hematopoietic neoplasms; ill-defined sites; and unknown primaries (See sitespecific codes for the sites and histologies), except death certificate only
99	Unknown if surgery performed

The **99** code was recoded as NaN and therefore the observations having this value were not considered in the analysis.

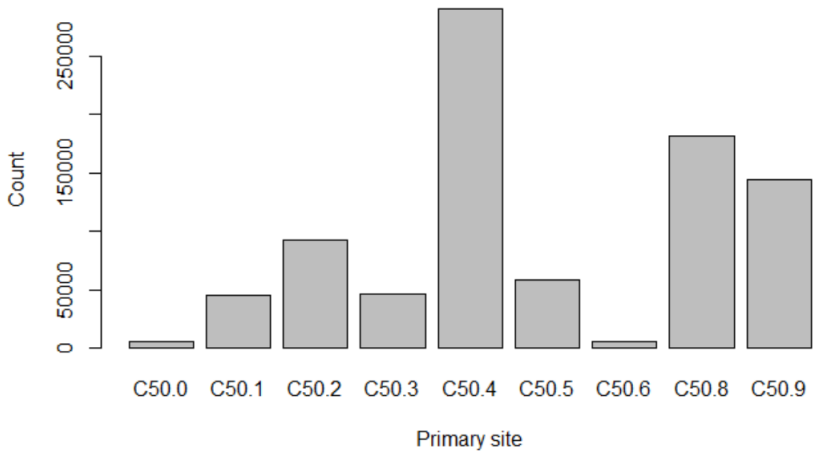


Figure 5.4: Primary site distribution

Regional nodes examined (1988+)

This is both a categorical and numerical variable describing if regional nodes were examined, and if so, the number of nodes examined:

Table 5.9: Regional nodes examined (1988+)

Code	Description
00	No nodes were examined
01-89	Exact number of nodes examined
90	90 or more nodes were examined
95	No regional nodes were removed, but aspiration of regional nodes was performed
96	Regional lymph node removal was documented as a sampling, and the number of nodes is unknown/not stated
97	Regional lymph node removal was documented as a dissection, and the number of nodes is unknown/not stated
98	Regional lymph nodes were surgically removed, but the number of lymph nodes is unknown/not stated and not documented as a sampling or dissection; nodes were examined, but the number is unknown

Codes from the **95** to the **98** were recoded as NaN and therefore the observations having these values were not take into account in the analysis.

Regional nodes positive (1988+)

As the previous one, this is both a categorical and numerical variable:

Table 5.11: Regional nodes positive (1988+)

Code	Description
00	All nodes examined are negative
01-89	Exact number of nodes positive
90	90 or more nodes are positive
95	Positive aspiration of lymph node(s) was performed
97	Positive nodes are documented, but number is unspecified
98	No nodes were examined

Codes from the **95** to the **98** were recoded as NaN and therefore the observations having these values were not take into account in the analysis.

ER Status recode Breast Cancer (1990+)

Estrogen receptor (ER) positivity is a prognostic indicator and a powerful predictor of endocrine therapy response in breast cancer. In most tumors, ER is either clearly expressed or completely absent, however some of them, the borderline ones, might have weak ER-positivity. For positive ER breast tumors, as also explained in the second chapter *Principles of Oncology*, endocrine therapy is recommended.

This is therefore a categorical variable:

- 1 = Positive
- 2 = Negative
- 3 = Borderline
- 4 = Unknown
- 9 = Not 1900+ Breast

Both codes **4** and **9** were not considered in the analysis and were therefore recoded as Nan values.

PR Status recode Breast Cancer (1990+)

For progesteron testing (PR) holds the same interpretation as per the Estrogen testing. This is again, a categorical variable and in particular the possible observable values are:

- 1 = Positive
- 2 = Negative
- 3 = Borderline
- 4 = Unknown
- 9 = Not 1900+ Breast

As per the *ER Status recode* variable both the codes **4** and **9** were not considered in the analysis and were therefore recoded as Nan values.

COD to site recode

Categorical variable describing whether the patient is still alive or dead, and in the latter case the cause of death is reported.

Survival months

Numerical variable describing the follow up time.

Type of follow up expected

Categorical variable describing the type of follow-up expected for a SEER case:

- 1 = "Autopsy Only" or "Death Certificate Only" case
- 2 = Active follow up case
- 3 = In situ cancer of the cervix uteri only
- 4 = Case not originally in active follow up, but in active follow up now (San Francisco-Oakland only)

Sequence number

Categorical variable that in this case will assume only two values based on the filters applied in the SEER*Stat software and they would be: "One primary only" and "1st of 2 or more primaries".

Primary by international rules

This is a categorical and specifically, binary variable that will report "Yes" if the tumor is a primary tumor. Again, by construction of the dataset, this condition is verified for all subjects and was therefore inserted as a control variable to verify the correctness of the selections.

Total number of in situ malignant tumors for patient

This is a numerical variable describing the count of a subject's total reported in situ/malignant cancers.

The valid values are the ones ranging from **00** to **98**, whilst the **99** code stands for *Unknown*: this value was recoded into a Nan value and not considered in the analysis.

Total number of benign borderline tumors for patient

This is a numerical variable describing the count of a subject's total reported in benign/borderline cancers.

The valid values are the ones ranging from **00** to **98**, whilst the **99** code stands for *Unknown*: this value was recoded into a Nan value and not considered in the analysis.

Year of birth

Categorical variable describing the year of birth of the patient.

Month of diagnosis

Categorical variable describing the month of tumor diagnosis of the patient.

Year of diagnosis

Categorical variable describing the year of tumor diagnosis of the patient, by construction the years reported range from 1995 to 2015.

Breast - Adjusted AJCC 6th Stage (1988-2015)

Categorical variable created from the EOD 3rd Edition and Collaborative Stage disease information. The 99 value stands again for *Unknown stage* and it was therefore recoded into a Nan value and then excluded from the analysis. This variable is equivalent to the *stage* variable.

Breast - Adjusted AJCC 6th Stage T (1988-2015)

Categorical variable created from the EOD 3rd Edition and Collaborative Stage disease information, which focuses on the T stage.

Breast - Adjusted AJCC 6th Stage N (1988-2015)

Categorical variable created from the EOD 3rd Edition and Collaborative Stage disease information, which focuses on the N stage.

Breast - Adjusted AJCC 6th Stage M (1988-2015)

Categorical variable created from the EOD 3rd Edition and Collaborative Stage disease information, which focuses on the M stage.

SEER historic stage A (1973-2015)

Categorical variable derived from Collaborative Stage (CS), however the variable has been simplified:

Table 5.13: SEER historic stage A (1973-2015)

Code	Description
0	In situ
1	Localized
2	Regional
4	Distant
8	Localized/Regional
9	Unstaged

The **9** value means that the variable is *Unstaged*: it was therefore recoded into a Nan value and then excluded from the analysis.

Marital status at diagnosis

Categorical variable that identifies the patient's marital status at the time of diagnosis for the reportable tumor.

CS Tumor size

This is both a numerical and categorical variable describing the size for the breast cancers diagnosed between 2004 and 2015. The **888**, the **990** together with the values from **996** to **996** were recoded as Nan, and the reason behind this solution

Table 5.15: Marital status at diagnosis

Code	Description
1	Single (never married)
2	Married (including common law)
3	Separated
4	Divorced
5	Widowed
6	Unmarried or domestic partner (same sex or opposite sex or unregistered)
9	Unknown

can be understood by looking at Table 5.17 that reports the description of the values that this variable can assume:

Table 5.17: CS Tumor size

Code	Description
000	No mass or tumor was found
001-988	Exact size in millimeters
989	989 millimeters or larger
990	Microscopic focus or foci only; no size of focus is given
991	Less than 1 cm
992	Less than 2 cm
993	Less than 3 cm
994	Less than 4 cm
995	Less than 5 cm
996-998	Site-specific codes where needed
999	Unknown
888	Not applicable

EOD (Extent of Disease) - Tumor size

This is both a numerical and categorical variable describing the size for the breast cancers diagnosed between 1988 and 2003. It was necessary to include this variable as well as our dataset collects diagnoses from 1995 to 2015, but the previous variable only describes the size of the tumor since 2004. In particular, the code **999** represents a tumor of unknown size and has therefore been replaced by the wording Nan.

A similar set of encodings to the previous one applies to this variable.

Age at diagnosis

This is a numerical discrete (or categorical, depending on its view), derived from the difference between the *Year of diagnosis* and *Year of birth* per each subject.

Since the 9 value means that the value of the variable for the subject is *Unknown*, it was recoded into a Nan value and then excluded from the analysis.

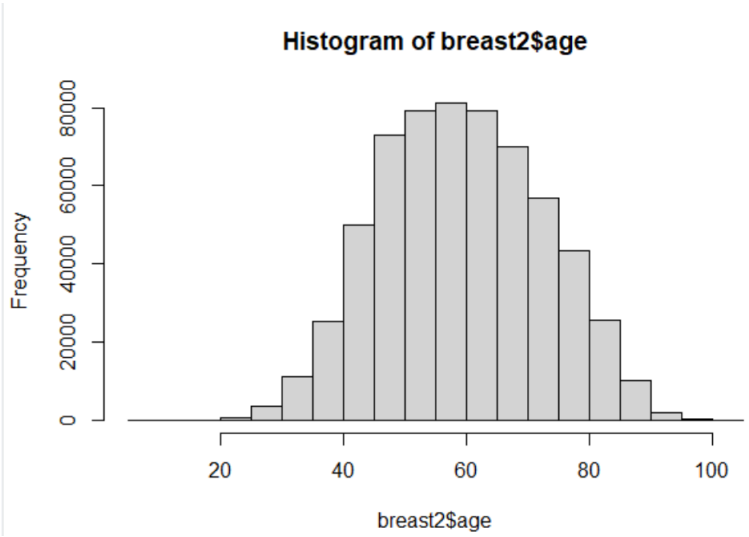


Figure 5.5: Age distribution before selection of range for analysis

State

This is a categorical variable derived from the *state_county* variable, from which the first two letters (representing the state acronym) of each observation where taken and inserted into a new column.

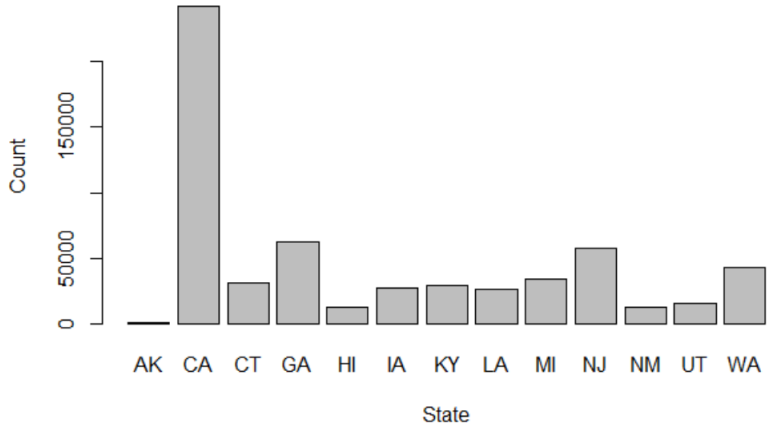


Figure 5.6: State distribution

5.1.2 Pre-processing and Exploratory Data Analysis

Each dataset, in order to be used for its own analyzes, requires some "cleaning" and data adaptation processes. This procedure is called pre-processing and is the first step for the application not only of machine learning algorithms but also for statistical analysis.

Nan values

The first phase of the pre-processing process requires the removal of the nan values, which can take place in various ways: removal of rows containing at least one missing value, elimination of columns containing at least one nan value, or defining a threshold of nan value beyond which the row (or column) must be removed. It is clear that in the third option some missing values therefore remain a fairly recurrent solution is to replace the missing values with the average of the values of the variable, a solution that works if the variable is of a numerical type.

Categorical variables

In this same chapter we have described for each selected variable its type, and as it is possible to notice many of the chosen variables are categorical. With categorical variable, we remember, we mean a variable whose values can be divided into groups or categories, as the name also indicates, some examples are gender, ethnicity...

It is important to highlight some further selection parameters consistent with the type of analysis to be conducted, specifically:

- only the subjects with an age ranging from 18 to 80 were selected;
- values of the *survival time* (namely the follow up) equal to 0 were replaced with 0.01.
- since the *survival time* is expressed in months we have decided to create a variable that expresses the follow up in days simply by multiplying each observation by **30.436875**, namely the coefficient of conversion of months to days.
- the *surgery of primary site* variable has been recoded in such a way as to replace the numerical values with their description.
- for the *tumor size* variable the values from **991** to **995** were recoded as respectively, 10, 20, 30, 40 and 50 in order to solve the problem deriving from having some categorical values in a variable that wants to be numeric instead.
- Latly, in order for the code to understand the type of variable it receives as input in the functions, it is necessary to use the *factor* or *as.factor* functions in the case of categorical variables, *as.numeric* in the case of numeric variables instead.

Moreover, as we have seen in the previous paragraph, not all observations have been preserved due to unacceptable values for some variables, such as for example in the number of nodes examined the observations that reported the words "No nodes were examined" or for estrogen and progesteron receptors the subjects for which the test was not known. The exclusion took place by casting the unacceptable values to nan values and then eliminating the corresponding observations from the dataset finally used for the analyzes. Removing observations is one of the possible solutions for handling nan values, but not the only one.

Some alternatives involve:

- changing the nan values with a specific value, in case of numerical data the replacement could be the zero but this solution deeply biases the results;
- an alternative could be replacing the nan value with the average or the median;
- using K-nearest neighbor for the imputation of the value;

- using the Multiple Imputatio by Chained Equations (MICE), an algorithm which fits a linear regression and predicts a value for the nan values, based on the the values that are not missing.

We opted for removing the rows containing nan values on the basis that the amount of data available was more than sufficient to obtain a good dataset for our analyzes and therefore of the reliable estimates. To be more precise, therefore, the analysis conducted can be classified as a complete case analysis, valid under missing completely at random assumption.

Variable Correlation

Negativity due to the presence of one of the two receptors, ER or PR, determines a lower therapeutic efficacy. In particular, the typical response rates, as reported by the *Lab Tests Online* [48] site are:

- ER-positive, PR-positive: 75-80%
- ER-positive, PR-negative: 40-50%
- ER-negative, PR-positive: 25-30%
- ER-negative, PR-negative: 10% or less.

```
> # Number of concordant pairs
> length(which(breast2$estrogen_receptor == breast2$progesteron_receptor))
[1] 660077
> # Number of discordant pairs
> length(which(breast2$estrogen_receptor != breast2$progesteron_receptor))
[1] 107567
```

Figure 5.7: Count of concordant and discordant pairs of the estrogen/progesteron receptors test

In Figure 5.7 we report the number of observations for which the two tests are concordant and discordant and it is possible to note that the first are much more frequent. Another doubt consists in the fact that the two variables can also be, given their correlated similarity. In order to exclude a correlation between them, we performed the chi-squared test in R using the function chisq.test() of which we show the result in Figure 5.8. Being the p-value of less than the significance level

```
> chisq.test(final$estrogen_receptor, final$progesteron_receptor)

        Pearson's Chi-squared test

data:  final$estrogen_receptor and final$progesteron_receptor
X-squared = 382010, df = 4, p-value < 2.2e-16
```

Figure 5.8: Output of the Chi square test for correlation between the ER and PR variables

of 0.05, we can reject the null hypothesis and conclude that the two variables are independent.

The final dataset comprises a grand total of over 350.000 subjects and the variables that were selected for conducting the analysis are the following:

- etnhicity;
- primary site labeled;
- grade;
- laterality;
- regional nodes examined;
- regional nodes positive;
- estrogen receptor;
- progesteron receptor;
- survival months;
- number of in situ malignant tumors;
- number of benign tumors;
- tumor size;
- age;
- status;
- state.

5.2 Methods

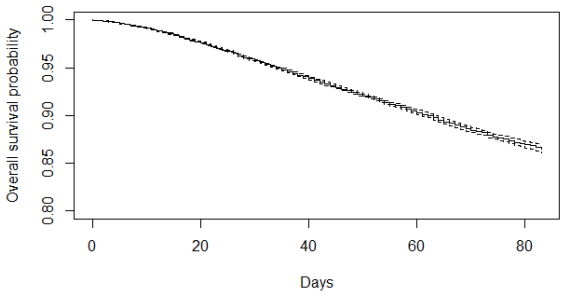
After having seen the various steps of preparing and cleaning the dataset, in this section we will focus on the methods actually implemented in our analysis.

5.2.1 Kaplan-Meier

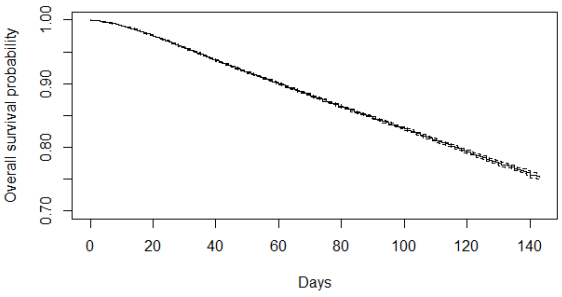
The first step of our analysis was to use the Kaplan meier method to estimate survival curves and probabilities. Specifically, we decided to fit the model on the data covering the following time intervals:

- 2010 to 2015 = 5 years;
- 2005 to 2015 = 10 years;
- 2000 to 2015 = 15 years;
- 2010 to 2015 = 20 years;

From the graphs below, which are the basic or default graphs obtained with R, it is possible to notice some important differences.

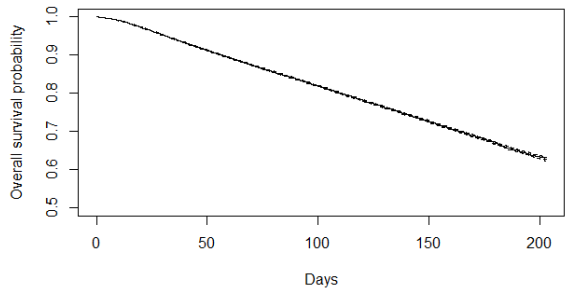


:
fig:Kaplan-Meier with data from 2010 to 2015

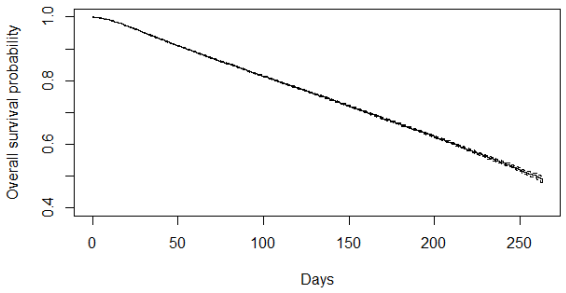


:
fig:Kaplan-Meier with data from 2005 to 2015

Let's start with a consideration: in the chapter dedicated to the illustration of the main methods used in survival analysis we mentioned that the graph of the survival curves is characterized by a "continuous" trend when the sample size is sufficiently large, exactly as in our case, otherwise the characteristic appearance will be that of a step-wise function as it is possible to observe these features of the survival curve in Figure 4.1, where we fitted the Kaplan-Meier model only on one hundred



:
fig:Kaplan-Meier with data from 2000 to 2015



:
fig:Kaplan-Meier with data from 1995 to 2015

observations extracted from the original dataset. Furthermore, the occurrence of the event under analysis determines the conclusion of a time interval while the height of the steps represents the consequent variation in the cumulative probability.

The other aspect that can be noticed is that from which it is possible to notice that only with data collected over a time interval equal to or greater than 20 years does the function reach the 50th percentile. This is one of the reasons we decided to extend the follow up time by considering an interval of 20 years, since even in this time laps the survival curve barely manages to reach the 50% percentile. To prove this registration from the SEER we extracted the observations relating to the diagnosis of lung cancer always in the period between 1995 and 2015. In Figure 5.9 where we reported the Kaplan-Meier model mounted on this second dataset it is possible to note that the survival curve is in fact much more similar to what we imagine: this in fact surpasses the fiftieth percentile and approaches the 0.

Estimating the probability of survival past the first and fifth year

Subsequently, again through the use of a simple function in R, we estimated the probability of survival at one year and at 5 years.

Subsequently, again through the use of a simple function in R, we estimated the probability of survival at one year, at 5 years and at 10 years. One-year survival is 98.70% (95% CI: 98.60-98.70%), 5-year survival is about 88.90% (95% CI: 88.80-89.00%), while 10-year survival is 77.6% (95% CI: 77.40-77.80%) and these results seem to be consistent with our graphs.

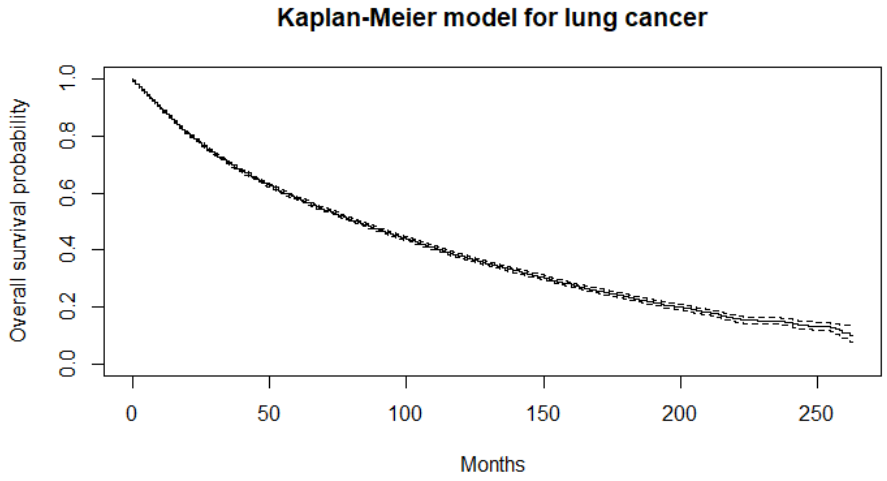


Figure 5.9: Kaplan-Meier model fitted on the Lung cancer patients

Mean and Median lifetime

Two other important statistics are precisely the mean and median survival time and as already mentioned in the previous paragraph, only for the survival curve fitted on the data ranging from 1995 to 2015, since it is possible to note that only this reaches the fiftieth percentile, representing the median.

We recall that the median lifetime asks answers to the question "how long does it take until half the sample population experiences the event?". We start off by fitting a simple curve that does not consider any difference between groups and therefore we will specify just the intercept ($\tilde{1}$).

```
> km_fit_months <- survfit(Surv(survival_months, status) ~ 1,
+                          data=training_20y)
> km_fit_months
Call: survfit(formula = Surv(survival_months, status) ~ 1, data = training_20y)

      n  events  median 0.95LCL 0.95UCL
 369586   78392     250     247     252
> median(survival_months) # 83 months
[1] 83
> delta.followup <- 1-training_20y$status
> survfit(Surv(survival_months, delta.followup) ~ 1, data=training_20y)
Call: survfit(formula = Surv(survival_months, delta.followup) ~ 1,
    data = training_20y)

      n  events  median 0.95LCL 0.95UCL
 369586  291194     102     102     103
```

Figure 5.10: Median survival time

The median in this case provides a measure of how well the subjects were followed up on average. Considering all survival times without asking the problem of censored

subjects leads to a bias in the calculation of the median. One measure that can be thought about is the so-called "potential" median survival which is obtained simply by treating censored observations as events. It therefore follows that the median of the follow-up is 250 months, while the potential follow up time is 120.

As it is possible to see from Figure 5.10, the median survival time is of 83 months and in particular it is possible to observe the over 350.000 people in our dataset, 69.955 people were uncensored (followed for the entire time, until occurrence of event) and among this set there was a median survival time of 260 months (the median is used because of the skewed distribution of the data). The 95% confidence interval for the median survival time for the 18 uncensored individuals is (256, 263).

Let us now look at the fit of the Kaplan-Meier model on first the age divided into quartiles and then on the dimension divided into quartiles. It is possible to see from the images below that for the age intervals [19,59] neither the median nor the mean of survival are determined. We also observe the same thing for tumor sizes in the range of [0-17].

```
> km_fit_months <- survfit(Surv(survival_months, status) ~ 1,
+                          data=training_20y)
> summary(km_fit_months)
Call: survfit(formula = Surv(survival_months, status) ~ 1, data = training_20y)

   time n.risk n.event survival  std.err lower 95% CI upper 95% CI
   0.01 357176      77    1.000 2.46e-05        1.000        1.000
   1.00 356850     195    0.999 4.62e-05        0.999        0.999
   2.00 356380     273    0.998 6.54e-05        0.998        0.999
   3.00 355836     318    0.998 8.23e-05        0.997        0.998
   4.00 355218     388    0.996 9.90e-05        0.996        0.997
   5.00 354425     392    0.995 1.14e-04        0.995        0.996
   6.00 353550     384    0.994 1.26e-04        0.994        0.995
   7.00 352677     434    0.993 1.39e-04        0.993        0.993
   8.00 351831     458    0.992 1.51e-04        0.991        0.992
   9.00 350981     476    0.990 1.63e-04        0.990        0.991
  10.00 350137     493    0.989 1.75e-04        0.989        0.989
  11.00 349255     572    0.987 1.87e-04        0.987        0.988
  12.00 348185     595    0.986 1.99e-04        0.985        0.986
  13.00 345467     646    0.984 2.12e-04        0.983        0.984
  14.00 342669     625    0.982 2.23e-04        0.982        0.983
  15.00 339849     690    0.980 2.35e-04        0.980        0.981
```

Figure 5.11: Summary of the Kaplan-Meier fit

The summary of the Kaplan-Meier fit in Figure 5.11 goes through each time point in the study in which an individual was lost to follow up or died and re-computes the total number of people still at risk (n.risk), the number of events at that time point (n.event), the proportion of individuals who survived up until that point (survival) and the standard error (std.err) and 95% confidence interval (lower 95% CI, upper 95% CI) for the proportion of individuals who survived at that point.

```
> fit_quantile
Call: survfit(formula = Surv(survival_days, status) ~ training_20y$fage,
    data = training_20y, type = "kaplan-meier")

                              n events median 0.95LCL 0.95UCL
training_20y$fage=[19,50] 97850  14374     NA      NA      NA
training_20y$fage=(50,59] 86433  13101     NA      NA      NA
training_20y$fage=(59,68] 83707  15850   7305    7183    7396
training_20y$fage=(68,84] 89186  34583   4231    4200    4261
```

Figure 5.12: Kaplan-Meier fit with age expressed in quantiles

```
> fit_quantile_size
Call: survfit(formula = Surv(survival_days, status) ~ training_20y$fsize,
    data = training_20y, type = "kaplan-meier")

                                n events median 0.95LCL 0.95UCL
training_20y$fsize=[0,11]   99798  14588     NA    8005      NA
training_20y$fsize=(11,17]  85535  15651     NA    7974      NA
training_20y$fsize=(17,26]  85369  19608   7457    7335    7548
training_20y$fsize=(26,990] 86474  28061   5479    5387    5570
```

Figure 5.13: Kaplan-Meier fit with size expressed in quantiles

5.2.2 AFT models

Accelerated failure time (AFT) models are parametric models that provides an alternative to the commonly used proportional hazards models. The assumption made by these models is that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. AFT models are typically robust to omitted covariates.

For our analysis we decided to test four AFT models with the following distributions:

- Weibull;
- Exponential;
- log-normal;
- log-logistic.

We first compare the four models fitted using the ANOVA statistical test which estimates how a quantitative dependent variable changes according to the levels of one or more categorical independent variables. ANOVA tests whether there is a difference in means of the groups at each level of the independent variable. The null hypothesis in this case is that there is no difference in means.

```
> anova(fit_weibull, fit_exp, fit_lnorm, fit_llogis)

 Terms
1 ethnicity + grade + estrogen_receptor + progesteron
e, 3)
2 ethnicity + grade + estrogen_receptor + progesteron
e, 3)
3 ethnicity + grade + estrogen_receptor + progesteron
e, 3)
4 ethnicity + grade + estrogen_receptor + progesteron
e, 3)
  Resid. Df   -2*LL Test Df   Deviance Pr(>Chi)
1    416686 1210998      NA         NA       NA
2    416687 1221009    = -1 -10010.999        0
3    416686 1215648    =  1   5360.936        0
4    416686 1209256    =  0   6392.707       NA
```

Figure 5.14: ANOVA test for the AFT models

The AIC and BIC of the four models were also calculated and it is possible to observe the results in Figure 5.15. It is emphasized that usually only one of these indices is more than enough.

```
%%R
print(glance(fit_exp) %>%
      dplyr::select(AIC, BIC))

print(glance(fit_weibull) %>%
      dplyr::select(AIC, BIC))

print(glance(fit_lnorm) %>%
      dplyr::select(AIC, BIC))

print(glance(fit_llogis) %>%
      dplyr::select(AIC, BIC))
```

```
# A tibble: 1 x 2
    AIC   BIC
  <dbl> <dbl>
1 4298. 4394.
# A tibble: 1 x 2
    AIC   BIC
  <dbl> <dbl>
1 4274. 4374.
# A tibble: 1 x 2
    AIC   BIC
  <dbl> <dbl>
1 4318. 4419.
# A tibble: 1 x 2
    AIC   BIC
  <dbl> <dbl>
1 4261. 4362.
```

Figure 5.15: AIC and BIC for the AFT models

5.2.3 Comparing survival times between groups of observations: the log-rank test

After estimating the survival curves on our dataset, the log-rank test allows us to compare the curves of two or more population groups. A classic example of

application of the log-rank test that can be easily found in the literature is that which compares the survival curves between female and male subjects. In our case, because we decided to focus on breast cancer, cases of breast cancer among men were not taken into consideration as the relative incidence of breast cancer is less than 1% for all diagnosed cases of this tumor.[49]

When the test is performed with the intent of comparing two groups the null hypothesis is that there are no significant differences between the groups, similarly when more than two groups are being compared, the null hypothesis can be interpreted as the lack of significant differences between all groups under analysis.

```
> survdiff(Surv(survival_days, status) ~ grade, data=training_20y)
Call:
survdiff(formula = Surv(survival_days, status) ~ grade, data = training_20y)

                                            N Observed Expected (O-E)^2/E (O-E)^2/V
grade=Moderately differentiated; Grade II  152818    30989    33573     198.9     350.4
grade=Poorly differentiated; Grade III     123865    33775    26174    2207.3    3332.6
grade=Undifferentiated; anaplastic; Grade IV  3841     1377     1170      36.5      37.2
grade=Well differentiated; Grade I          76652    11767    16990    1605.9    2059.3

 Chisq= 4060  on 3 degrees of freedom, p= <2e-16
```

Figure 5.16: Log-rank test for the grade variable

In Figure 5.16 we investigate if there is a significant difference between the survival curves of subjects with a different grade of tumor and as it is possible to imagine there is a substantial difference between the survival of a person whose tumor is diagnosed in the first stage rather than the third or fourth and the p_value much lower than the 0.05 threshold confirms this intuition.

5.2.4 Cox proportional hazards regression model

The methods seen so far are examples of univariate analyzes, in the sense that they describe survival as a function of a single variable, ignoring the impact of the other covariates. Furthermore, another aspect to note is that these models are particularly effective when the variables are categorical, such as the gender, ethnicity or grade of the tumor, and less effective when we insert variables such as the number of malignant and benign tumors, or the size of the tumor.

The aim of Cox's method, we recall, is to determine how certain factors impact the survival of a subject as also described by his equation (which we report again

here), where the dependent variable is expressed as a hazard function:

$$h(t|x) = b_0(t) \exp \sum_{i=1}^{n} b_i x_i \tag{5.1}$$

where:

- t is the survival time
- $h(t|x)$ is the hazard function
- $b_0(t)$ is the baseline function, namely the probability of experiencing the event of interest when all the other covariates are equal to zero
- $x_1, ..., x_n$ are the covariates
- $b_1, ..., b_n$ are the coefficients measuring the impact of the covariates

In general, therefore, the cox model estimates the hazard function as a linear combination of the covariates (independent over time and estimated in turn by partial likelihood) and the baseline hazard, hence the concept of proportional hazards.

The value of interest for the cox model is the hazard ratio, that is the value of $\exp(b_i)$, representing the multiplicative coefficient of the covariates on the hazard rate (HR).

To correctly interpret the meaning of the hazard rate, just remember the following rule:

- if $HR = 1$: No effects
- if $HR < 1$ or $b_i < 1$: Reduction in Hazard
- if $HR > 1$ or $b_i > 1$: Increase in Hazard

Checking the Proportional Hazard assumption

As previously mentioned, the Cox model is based on the assumption that the hazards are proportional but how do we verify this assumption? The cox.zph function, always belonging to the survival package of R, allows this verification to be carried out by analyzing the Schoenfeld residuals. The null hypothesis of this test is that

the hazards are proportional or alternatively that the hazard ratio is constant over time; the output of this test will provide a result for each of the covariates with the relative p-value: in the event that the p-value should be less than the usual value of 0.05, then we will reject the null hypothesis and we will find ourselves in that case of facing a time-dependent variable that therefore needs to be stratified. In Figure 5.17 it is possible to observe that all the p_values are much lower than the reference value.

```
> prop_hazards6 <- cox.zph(fit_breast_cox6)
> print(prop_hazards6)
                          chisq df      p
ethnicity                   460  2 <2e-16
grade                      4261  3 <2e-16
estrogen_receptor          5033  2 <2e-16
progesteron_receptor       3666  2 <2e-16
ns(age, 3)                 4697  3 <2e-16
regional_nodes_examined     367  1 <2e-16
regional_nodes_positive     562  1 <2e-16
ns(tumor_size, 3)          2758  3 <2e-16
GLOBAL                    10824 17 <2e-16
```

Figure 5.17: Proportional Hazard test results for model fitted on the whole training set

Cox regression model

The covariates selected for our models are therefore the following:

- ethnicity
- grade (or breast_ajcc)
- estrogen receptor
- progesteron receptor
- regional nodes examined
- regional nodes positive
- cubic spline of age
- cubic spline of tumor size

The addition to the Cox model of powers of the predictor of interest by means of third degree polynomials, allows a gain in terms of flexibility of the function.

Cubic regression splines are in fact tools that allow us to identify hidden relationships and at the same time to produce smooth and easy to interpret curves. So what is a *spline*? According to what Wikipedia reports, "In mathematical analysis, a spline is a function, consisting of a set of connected polynomials". Much more simply, a spline adds multiple curves together in order to make the irregular curve that tightens the observations continuous.

For instance, Durrleman and Simon use splines to identify non-linear convariate-response relationships in the Cox model.[50]

```
Call:
coxph(formula = Surv(survival_months, status) ~ ethnicity + grade +
    estrogen_receptor + progesteron_receptor + ns(age, 3) + regional_nodes_examined +
    regional_nodes_positive + ns(tumor_size, 3), data = training_20y,
    x = TRUE)

  n= 416705, number of events= 90652

                                              coef  exp(coef)   se(coef)       z Pr(>|z|)
ethnicityOther                          -5.113e-01  5.997e-01  1.625e-02 -31.472  < 2e-16 ***
ethnicityWhite                          -3.442e-01  7.088e-01  1.035e-02 -33.243  < 2e-16 ***
gradePoorly differentiated; Grade III    2.127e-01  1.237e+00  8.068e-03  26.362  < 2e-16 ***
gradeUndifferentiated; anaplastic; Grade IV  1.771e-01  1.194e+00  2.595e-02   6.826 8.76e-12 ***
gradeWell differentiated; Grade I       -1.600e-01  8.521e-01  1.018e-02 -15.713  < 2e-16 ***
estrogen_receptorER_Negative             1.052e-01  1.111e+00  7.242e-02   1.453 0.146264
estrogen_receptorER_Positive            -5.857e-02  9.431e-01  7.222e-02  -0.811 0.417369
progesteron_receptorPR_Negative         -1.739e-03  9.983e-01  4.157e-02  -0.042 0.966633
progesteron_receptorPR_Positive         -1.519e-01  8.591e-01  4.126e-02  -3.681 0.000232 ***
ns(age, 3)1                              4.350e-01  1.545e+00  2.192e-02  19.840  < 2e-16 ***
ns(age, 3)2                             -6.472e-02  9.373e-01  8.289e-02  -0.781 0.434953
ns(age, 3)3                              2.119e+00  8.321e+00  1.692e-02 125.232  < 2e-16 ***
regional_nodes_examined                  1.827e-04  1.000e+00  4.928e-04   0.371 0.710785
regional_nodes_positive                  6.366e-02  1.066e+00  6.994e-04  91.032  < 2e-16 ***
ns(tumor_size, 3)1                       5.553e+00  2.581e+02  8.215e-02  67.597  < 2e-16 ***
ns(tumor_size, 3)2                      -2.437e-01  7.837e-01  4.288e-01  -0.568 0.569778
ns(tumor_size, 3)3                      -7.666e+00  4.684e-04  8.930e-01  -8.585  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5.18: Cox regression model output (part 1)

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                          exp(coef) exp(-coef) lower .95 upper .95
ethnicityOther                            5.997e-01  1.667e+00 5.809e-01 6.191e-01
ethnicityWhite                            7.088e-01  1.411e+00 6.945e-01 7.233e-01
gradePoorly differentiated; Grade III     1.237e+00  8.084e-01 1.218e+00 1.257e+00
gradeUndifferentiated; anaplastic; Grade IV 1.194e+00  8.377e-01 1.135e+00 1.256e+00
gradeWell differentiated; Grade I         8.521e-01  1.174e+00 8.353e-01 8.693e-01
estrogen_receptorER_Negative              1.111e+00  9.001e-01 9.639e-01 1.280e+00
estrogen_receptorER_Positive              9.431e-01  1.060e+00 8.186e-01 1.087e+00
progesteron_receptorPR_Negative           9.983e-01  1.002e+00 9.202e-01 1.083e+00
progesteron_receptorPR_Positive           8.591e-01  1.164e+00 7.924e-01 9.314e-01
ns(age, 3)1                               1.545e+00  6.473e-01 1.480e+00 1.613e+00
ns(age, 3)2                               9.373e-01  1.067e+00 7.968e-01 1.103e+00
ns(age, 3)3                               8.321e+00  1.202e-01 8.050e+00 8.601e+00
regional_nodes_examined                   1.000e+00  9.998e-01 9.992e-01 1.001e+00
regional_nodes_positive                   1.066e+00  9.383e-01 1.064e+00 1.067e+00
ns(tumor_size, 3)1                        2.581e+02  3.874e-03 2.197e+02 3.032e+02
ns(tumor_size, 3)2                        7.837e-01  1.276e+00 3.382e-01 1.816e+00
ns(tumor_size, 3)3                        4.684e-04  2.135e+03 8.137e-05 2.696e-03

Concordance= 0.742  (se = 0.001 )
Likelihood ratio test= 64525  on 17 df,   p=<2e-16
Wald test            = 77206  on 17 df,   p=<2e-16
Score (logrank) test = 85873  on 17 df,   p=<2e-16
```

Figure 5.19: Cox regression model output (part 2)

We look at the figure to understand the fit output of the Cox model. In addition to dwelling on the values of the second column, named *exp (coef)* and representing the hazard ratio, we also look at the p_value; the general rule in the absence of other specifications is that a p-value lower than 0.05 is to be considered statistically significant, which occurs for almost all the variables except the *estrogen receptor, the progesteron receptor* and *the number of nodes examined*. It should also be noted that the parameter estimates are estimated taking the other predictors into account. Moreover in absence of further specification the coefficients for one value of the categorical variables are estimated taking as reference one of them. Hence, for example, for the ethnicity variable the coefficients are computed with respect to the "black" values as reference.

Continuing with this variable, it is possible to observe a hazard ratio for Caucasian ethnicity equal to 0.7151, therefore it is possible to deduce that, cetera paribus, belonging to this ethnicity decreases the hazard by a factor of 0.7151 or alternatively by 28.49% and instead belonging to an ethnic group other than white and black instead reduces the hazard by a percentage equal to about 40%.

The issue of overfitting

We now come to the model implemented in our project and briefly describe the method adopted. The first step was to divide the entire dataset into three subsamples called training, equal to 60% of the original dataset, and the validation and test set, each equal to 20% of the original dataset. The next step was to fit the Cox model on increasing dimensions of the training sample in order to first verify the accuracy of the model, and the error made by the model. To understand the logic of this reasoning it is necessary to recall the concept of overfitting. As already said in the previous chapters, in data science a model learns the relationship between the inputs or features and the outputs or the labels during the training phase, where the model is given a dataset containing both the features and the labels. Then the model is evaluated on the test set, that is a dataset containing only the input values and then verifies the correctness or accuracy of the labels assigned by the model.

The simplest model we can refer to is that of linear regression, namely:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{5.2}$$

Training a linear regression model trivially means minimizing the distance between the line we are trying to fit into our data and the data itself. A model is considered a good model when it is capable of generalizing, that is, when it is endowed with such a sensitivity as to be able to correctly classify even data never seen previously.

Therefore, when the model perfectly follows the trend of the observations of the training set we have minimized the total distance of the points from the curve that we are trying to fit to the data but the model will hardly be able to correctly classify new data as it is said, has learned too well those of the training set and is therefore unable to generalize and capture the dominant trend.

In this case we will talk about *overfitting* or *high variance*.

Instead, we will talk about *underfitting* or *high bias* when the model generalizes too much and is therefore unable to correctly classify the observations of the training set.[51]

Having said this, it is generally known that the greater the number of observations given to the model for training, the better the fit and consequently the less the overfitting will be. The question that now arises is how many observations are enough? We now understand the logic explained at the beginning of this paragraph, which is to train our cox model on increasing dimensions of the training set.

5.2.5 Tuning of the Cox model

The next step consisted of fitting the Cox model to increasing size random sample of the original dataset in order to validate its robustness.

To evaluate the goodness of the fit three metrics were used: the C-index, Brier score (computed at distances of 1 year, namely at 1, 2, 3,.. years up until the last year considered) and the IPA.

Figure 5.20 shows the concordance indices for our increasing sample sizes. As we can observe, the value for the sample with a size equal to 10% of the original one is missing because the model fitted on this sample of observations converges before the

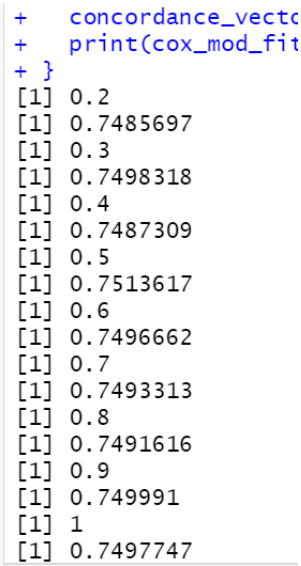


Figure 5.20: Concordance Index for increasing sample sizes

analysis of all the variables and it follows that so that the model does not converge in advance the dataset on which the model is fitted must have a size greater than 20% of the original dataset, i.e. approximately 71.435 observations.

We also note that the concordance rates for all subsamples are very similar (standard deviation equal to 0.0007) and that they can be considered good results normally considered the concordance rate has a value between 0.55 and 0.7 due to the noise present in the data.

In the following two pictures, specifically Figure 5.22 and Figure 5.23 we show instead the output for a sample size equivalent to the 60% of the original training set.

How do we interpret the brier score and IPA results? Recall that both these indexes, together with the Concordance Index, provide a measure of the accuracy of the survival function prediction but at a specific time t. In the case of the Brier score this was calculated for 5 periods: at 0.01 (in the pre-processing phase year 0 was changed to 0.01 to not eliminate zero survival), 5, 10, 15 and 20 years. This index typically assumes values between 0 and 1 and contrary to what one can guess, values close to zero are preferred. On the basis of what has been said, therefore, our classic Cox model provides a good estimate of the survival function even if it is possible to observe how the index worsens as the period t increases.

```
sample = seq(0.2 , 1, by = 0.1)
concordance_vector <- c()

### Da fare singolarmente

for (n in sample){
  print(n)
  data1 <- sample_frac(training_20y,n)
  cox_mod_fit <- coxph(Surv(survival_months, status) ~ ethnicity +
                         grade +
                         estrogen_receptor +
                         progesteron_receptor +
                         ns(age, 3) +
                         #breast_ajcc +
                         regional_nodes_examined +
                         regional_nodes_positive +
                         ns(tumor_size,3),
                         data = data1,
                         x = TRUE)
  concordance_vector[n*10-1] <- cox_mod_fit$concordance[[6]]
  print(cox_mod_fit$concordance[[6]])
  brier <- pec(object=cox_mod_fit,
               formula=Surv(survival_months, status) ~ ethnicity+
                                                       grade+
                                                       estrogen_receptor+
                                                       progesteron_receptor+
                                                       ns(age,3)+
                                                       regional_nodes_examined+
                                                       regional_nodes_positive+
                                                       ns(tumor_size,3),
               data=data1,
               exact=TRUE,
               cens.model="marginal",
               splitMethod="none",
               B=0,
               verbose=TRUE)
  print(brier6,times=seq(min(data1$survival_months),max(data1$survival_months),60))
  ipa <- IPA(cox_mod_fit, formula=Surv(survival_months,status!=0)~1,times=60)
  print(ipa)}
```

Figure 5.21: Function for getting the C-index, brier index and IPA for increasing sample size of the training set

Testing the model

Lastly we fit the Cox model on the test set and output the concordance index first and then the IPA and the results are reported in Figure 5.24 and Figure 5.25.

```
[1] 0.6
[1] 0.7410487

Prediction error curves

Prediction models:

Reference     coxph
Reference     coxph

Right-censored response of a survival model

No.Observations: 416705

Pattern:
                Freq
 event          90652
 right.censored 326053

IPCW: marginal model

No data splitting: either apparent or independent test sample performance

Cumulative prediction error, aka Integrated Brier score  (IBS)
 aka Cumulative rank probability score

Range of integration: 0 and time=0.01 time=60 time=120 time=180 time=240 :
```

Figure 5.22: Output of the function for obtaining the C-index. Brier score and IPA for a sample size of 60% of the original training set (part 1)

5.2.6 The next step

To verify therefore whether the machine learning algorithms bring real added value compared to the more classic models, in the second phase of our analysis we acted as follows. The first step was to split the pre-processed dataset into two equal parts, which we named *original training* and *external validation*. Subsequently, the algorithms, including the classic Cox one, were trained on increasing dimensions of the original training, specifically portions equivalent to the 0.1%, 0.5%, 1% and 5%. The models trained in this way were then tested on the external validation set, a dataset completely untouched during the learning phase, in order to verify the performance of the selected algorithms.

In the next section we will take care to very briefly describe the selected algorithms and especially the process of selecting their parameters.

5.3 Application of the Machine Learning Algorithms

In this section we will illustrate the machine learning algorithms used and the main tuning steps of the related topics.

```
Range of integration: 0 and time=0.01 time=60 time=120 time=180 time=240 :


Integrated Brier score (crps):

          IBS[0;time=0.01) IBS[0;time=60) IBS[0;time=120) IBS[0;time=180) IBS[0;time=240)
Reference                0          0.049           0.097           0.135           0.162
coxph                    0          0.044           0.085           0.113           0.131
                  Variable times Brier  IPA IPA.drop
 1:              Null model    60  10.4  0.0     11.2
 2:              Full model    60   9.3 11.2      0.0
 3:               ethnicity    60   9.3 10.9      0.3
 4:                   grade    60   9.3 10.9      0.3
 5:       estrogen_receptor    60   9.3 11.0      0.2
 6:    progesteron_receptor    60   9.3 11.1      0.1
 7:             ns(age, 3)    60   9.6  8.2      3.0
 8: regional_nodes_examined    60   9.3 11.2      0.0
 9: regional_nodes_positive    60   9.5  9.1      2.1
10:       ns(tumor_size, 3)    60   9.5  8.9      2.3
```

Figure 5.23: Output of the function for obtaining the C-index. Brier score and IPA for a sample size of 60% of the original training set (part 2)

```
> concordance(fit_breast_cox6, newdata = test_20y)
Call:
concordance.coxph(object = fit_breast_cox6, newdata = test_20y)

n= 79197
Concordance= 0.738 se= 0.002019
concordant discordant     tied.x     tied.y    tied.xy
 562003156  199531220       3163     841850          4
```

Figure 5.24: Concordance Index of the Cox model fit on the test set

5.3.1 Survival Random Forests

As the first machine learning algorithm we opted for random survival forests, an adaptation of the random forests algorithm for survival analysis implemented by the scikit-survival package. Random forests is a supervised learning algorithm based on the creation of decision trees on random data selections. The predictions of each tree are subsequently collected and the best is selected. In the version implemented by scikit-survival the sample size is always the same as the original dataset given in input.

First model

For the first model built the arguments were selected as follows:

- number of trees = *1.000*;
- minimal number of samples required to split an internal node = *10*;
- minimal number of samples required to be a leaf node = *15*;

```
> IPA(fit_breast_cox6,
+     formula=Surv(survival_months,status!=0)~1,
+     times=60, newdata=test_20y)
                    Variable times Brier  IPA IPA.drop
 1:                Null model    60  10.2  0.0     10.3
 2:                Full model    60   9.1 10.3      0.0
 3:                 ethnicity    60   9.1  9.9      0.3
 4:                     grade    60   9.1  9.9      0.3
 5:        estrogen_receptor    60   9.1 10.1      0.1
 6:    progesteron_receptor    60   9.1 10.2      0.1
 7:                ns(age, 3)    60   9.4  7.3      2.9
 8: regional_nodes_examined    60   9.1 10.3     -0.0
 9: regional_nodes_positive    60   9.3  8.4      1.9
10:       ns(tumor_size, 3)    60   9.3  8.2      2.1
```

Figure 5.25: Index of Prediction Accuracy computed fitting the Cox model on the test set

- number of features that have to be considered when looking at the best split = *sqrt* namely square root of the the maximal number of features.
- random number generator = *20*
- bootstrap if set to *True* then the samples are drawn with replacement, therefore the *False* option was selected

Cross validation

In order to obtain a better result, the arguments and consequently the performance of the random survival forest model were optimized by means of a process known as *hyperparameter tuning*. This technique involves testing different combinations of parameters in order to find the best one, but by evaluating each model on the same training set you risk running into an overfitting problem, a problem that is usually tackled via *cross validation*. The most common cross validation technique is the *K-fold CV* in which the training set is further split into K subsets and consequently the model is fitted k times, training it on K-1 subset and using the K-th one as a validation sample, in this case we opted for a 3-fold CV with 30 iterations (due to the high computational time and space that this technique requires). At the end of the hyperparameter tuning process all models are analyzed and the best of them is selected. For this step we decided to use the *RandomizedSearchCV* function provided by Scikit-Learn.

The algorithm was then computed again using the best parameters found in this way.

5.3.2 Gradient Boosted Models

Gradient boosting represents a set of techniques aimed at optimizing loss functions, based on the logic of combining multiple bases or weak learners (typically decision trees). The predictions are additively combined by improving or "boosting" the overall model, but if in the random forest the decision trees are computed independently and then averaging the prediction, gradient boosted models are constructed sequentially and in sequence. The starting model is very simple and the following parameters have been selected:

- loss function = *coxph*
- number of trees = *200*;
- learning rate = *0.1*;
- maximal depth = *1*;

Cross validation

Also for this model we have tried to improve the performance of the model by tuning the arguments or parameters of the same. The technique function used is again the K-fold cross validation, in particular k has been selected equal to 3, via *RandomizedSearchC* function. A maximal of 30 iterations was selected.

5.3.3 Support Vector Machines

Support vector machines are another example of a supervised learning algorithm, used mainly for classification problems but it can also be extended to regression problems. The typical use in classification problems is a consequence of the fact that the basic idea of the model is to find the hyperplane in an n-dimensional space that allows to separate as definitively and clearly as possible the points belonging to a class rather than another, as it is possible to observe from Figure 5.26

The SVM model (as also the Random forest model) was implemented using first the scikit-surv and the PyTorch packages, however in Table 6.1 the results reported are those relating to the scikit-learn version. As a matter of fact, testing the PyTorch

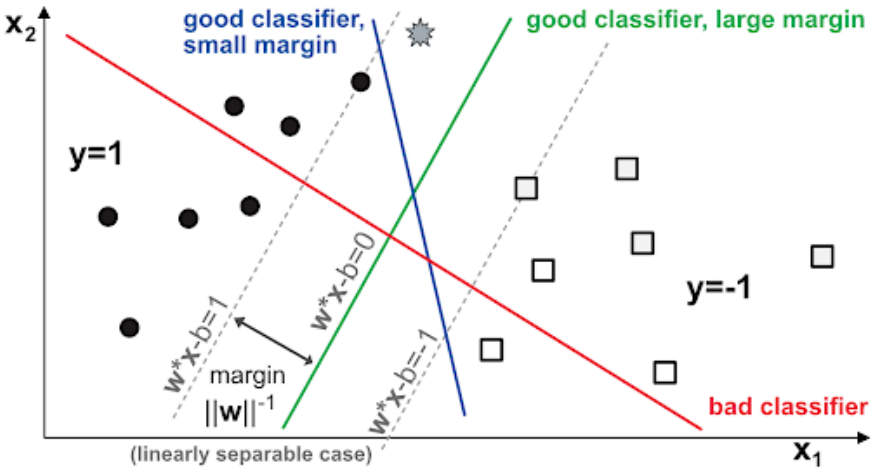


Figure 5.26: Support vector machine [52]

version on the external validation set involved a continuous crashing of the program and a bias in the results that were consequently obtained on different samples.

The scikit-surv model was implemented using the FastSurvivalSVM() function and the arguments selected for the first fit were the following:

- alpha, representing the penalizing weight for the hinge loss function = *1.0* by default;
- rank ratio representing a mixing parameter between the regression and ranking objective = *0.0*;
- tolerance parameter = 1.0 exp −3;
- optimizer = *avltree*

Cross Validation

The next phase was related to the optimization of the parameters through cross validation. In addition to those previously mentioned, the following parameters have also been subject to tuning:

- fit intercept, boolean variable indicating whether or not the model had to compute the intercept of the regression model;
- maximal number of iterations.

Once again we used a 3-fold cross validation, with a maximum number of iterations equal to 30.

5.3.4 CoxTime

As previously anticipated, the main differences between the classic Cox model and the CoxTime model of Kvamme et al. are the following:

- the optimization of the partial likelihood underlying the Cox model takes place by means of stochastic gradient descent, which allows to overcome the proportional hazards assumption.
- the $g(x)$ of the relative risk function $exp[g(x)]$ is replaced by a parameterized by the neural network.
- addition of a time dependent term to the risk function that is why in this case the concordance index calculated is that of Antolini, Boracchi and Biganzoli which is precisely time dependent.

For non-proportional models, such as DeepHit, CoxTime or LogisticHazard, it defining the "estimated risk" from the model is not obvious. The Antolini Concordance has been proposed as an attempt to address this issue, as it considers the full survival function and not just a single number per individual. This index uses the predicted survival function as outcome prediction, and the ability to discriminate among subjects having different outcome is summarized over time. In particular, the index compares the predicted survival function of a subject who experienced the event of interest to:

- the survival function of the people who experienced the event before his/her survival time and
- the survival function of the people who developed the event or were censored after his/her survival time

Censored subjects are considered together with those subjects that experienced the event before their observed times. In addition to this the Ctd is equivalent to Harrell's index in case of separation between survival curves on the whole follow-up period.[53]

In the following paragraphs we specify the arguments inserted for the training the models.

Alternative 1

- two hidden layers of respectively 64 and 32 nodes;
- learning rate is equal to the best one found with the $learn_rate_finder method; dropout = 0.03$;
- batch size = epochs = 256;
- optimizer = Adam;
- activation function = ReLu.

Alternative 2

- three hidden layers with respectively 64, 64 and 32 nodes;
- learning rate is equal to the best one found with the $learn_rate_finder method; dropout = 0.03$;
- batch size = epochs = 256;
- optimizer = Adam;
- activation function = ReLu.

Alternative 3

- two hidden layers with respectively 64 and 32 nodes;
- learning rate with exponential decay starting from a 0.1 value;
- dropout = 0.03;
- batch size = epochs = 256;
- optimizer = Adam;
- activation function = ReLu.

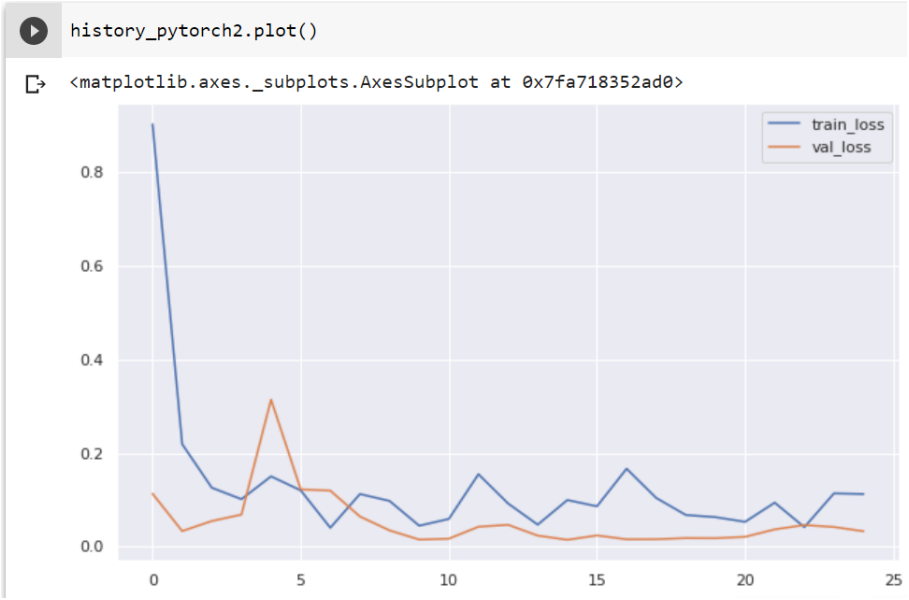


Figure 5.27: Train-validation loss plot for the CoxTime model

5.3.5 DeepSurv

Lastly, the model developed by Katzman et al. was fitted, which unlike the CoxTime models is still based on assumption of proportional hazards but it exploits neural networks to learn complex relationship between the covariate. Another peculiarity of DeepSurv is that it optimizes the average negative log partial likelihood with regularization.

In order to fit this model we use once again both with pycox and PyTorch: the C-index of the first library's models is time dependent, whilst the second one is the "simple" Harrell's index.

Pycox

For the models implemented with the pycox package three alternatives were evaluated as for CoxTime, and the same parameters were actually selected, with the exception of a dropout value equal to 0.05 and therefore higher than the previous one. Consequently, the selected parameters are not specifically reported in this paragraph in order to avoid providing a copy of the previous paragraph.

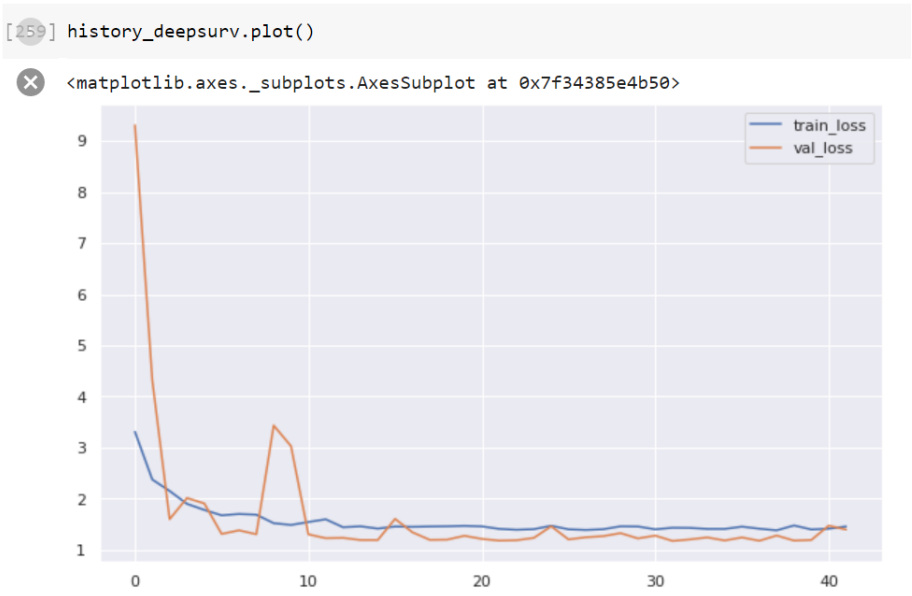


Figure 5.28: Train-validation loss plot for the DeepSurv model

PyTorch

As for the DeepSurv model implemented with the PyTorch package, the selected parameters are the following:

- three hidden layers with respectively 64, 64 and 32 nodes;
- optimizer = Adam;
- learning rate = 0.01;
- epochs = 256;
- dropout = 0.035;
- activation function = glorot uniform

Moreover, using the *integrated_brier_score* function we can plot the Brier score over time (Figure 5.29) and being the value under the usual 0.25 threshold, it seems that the model will yield good performances.

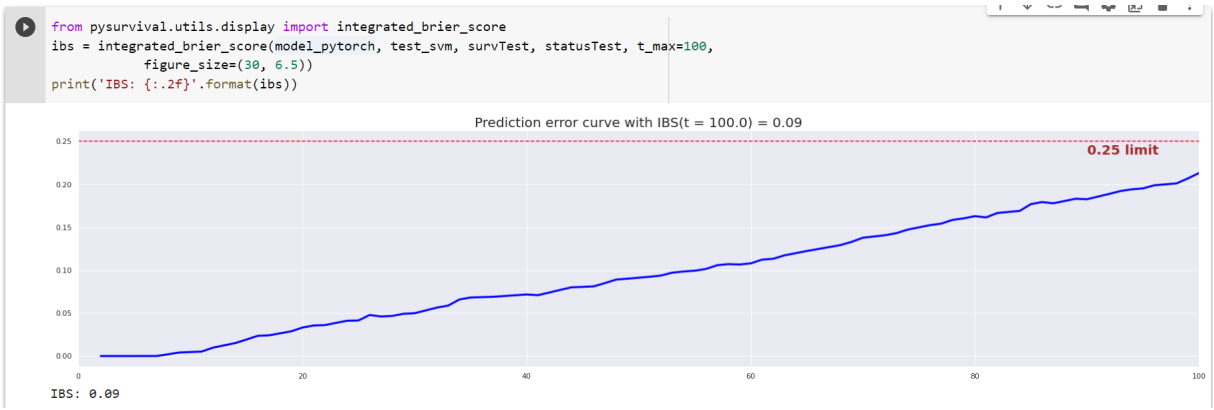


Figure 5.29: Integrated Brier Score for the DeepSurv model in PyTorch

Chapter 6

Conclusion and final considerations

6.1 Results

Table 6.1 shows the final and best results obtained by fitting the models trained on the input training sets of different data dimensions, on the external validation set. It is also reported, as already illustrated in the previous chapter, that, as regards the DeepSurv, two concordance indexes are reported, resulting from the implementation of the model using two different packages, respectively PyTorch and Pycox. Specifically, the concordance index determined by Pycox is the time dependent concordance index also known as Antolini concordance index.

Table 6.1: Concordance Index of the external validation dataset for the different training sample size taken into consideration

Model	0.1%	0.5%	1%	5%
Cox	68.81%	69.37%	73.62%	74.10%
Random Forest	70.31%	72.90%	73.01%	74.20%
Gradient Boosted	66.77%	72.00%	72.04%	72.72%
SVM (Sckit-survival)	67.86%	68.04%	68.10%	68.22%
CoxTime Alternative 2	97.71%	98.31%	98.40%	99.27%
DeepSurv-td Alternative 1	87.23%	88.00%	88.63%	91.76%
DeepSurv (PyTorch)	60.10%	62.45%	65.05%	66.22%

Observing the results, it is possible to state that there is a real added value

brought by machine learning algorithms compared to the more classic Cox model, although it is necessary to bear in mind that the performances of both the CoxTime and the DeepSurv are evalutated using a time dependent concordance index and not Harrell's index. As it was natural to think as the size of the training dataset increases the performance of these improves considerably but it is also necessary to note some problems that the application of the Cox model does not involve, namely the computational power, time and memory necessary for these can define good estimates. The Cox model trained on the entire training dataset not only does not cause the forced closure of the program, but it even takes a few seconds to define the output, something that does not happen with the other models, which instead require minutes for both the learning and the testing. It should also be noted that the Cox model remains rather constant, an aspect which is also confirmed by the verification of its robustness on increasing dimensions of the training set (phase 1 of this analysis).

6.2 Main issues encountered

The development of this analysis required the tackling of several and numerous problems.

Novelty of the topics

First of all, the author's lack of knowledge in the medical, epidemiological and biostatistic fields which required an in-depth study of the subject through the material provided by the supervisor as well as attending courses on the topics offered by sites such as Coursera and edX.

The SEER program

The next problem was to understand the use of the software provided by the SEER program for data extraction, but above all to understand which variables to select. In fact, the reader is reminded that the SEER has been collecting data relating to cancer diagnostics since 1973 and the literature as well as the classifications have been updated over time. The SEER database therefore also contains the data re-

lating to the previous classifications which have fallen into disuse but which the new observations possess as variables, albeit empty or with a value reported as nan. Another problem of the variables is, as is well known, the possible correlation but even of equality between them, and consequently many of the variables that were initially extracted were excluded from the analysis for this very reason. Examples of these two aspects are provided respectively by the variables *estrogen_receptor* and *progesteron_receptor*, and *grade* and *breast_ajcc* (where ajcc stands for American Joint Committee on Cancer). The first two as explained also in the previous chapters, could seem related to the characteristics possessed, but we have always previously demonstrated through the Chi-square test that they are not, but it was important to investigate the possible correlation in order to exclude this possibility or act accordingly in the event of an actual correlation. As for the second pair, both represent the same variable, ie the tumor staging, yet they are contained in the seer as two distinct variables.

Tools utilised

In the third instance, as already anticipated for our analysis, numerous "tools" were used: the first part of the analysis was developed in R, while the second in Python but using the Google Colaboratory service. This was basically mandatory due to two fundamental problems: first of all the power and the computational space to train and subsequently test in particular the machine learning algorithms. It is well known to most that for these the performances in general improve when the dataset given as input in the learning phase is sufficiently large, this also to avoid the typical problem of overfitting. Training these complex models on the entire training dataset would have required computing power and time that unfortunately we did not have available. The parameter tuning phase through Cross validation on a portion of the dataset equal to 10% of the original one is forcibly interrupted by the program due to lack of space and power.

The second problem, as mentioned above, is due to the fact that the machine learning models used in this dissertation, even if developed for Python, were actually collected by the mlr3 package which due to incompatibility with previous package installations and probably also due to the fact that the Rstudio IDE was not installed

by the author on Anaconda, but separately as the installation on Anaconda caused another series of problems.

6.2.1 Different packages

The implementation of the algorithms through different packages (sklearn, pytorch, survival, survminer) necessarily required a different data preparation phase for each function. The adaptation of the data was not always obvious, above all due to the request for the input and output format of the various functions and this too required a considerable effort.

6.3 Final considerations

The added advantage of the most modern machine learning models over the more classic models of survival analysis has been proven. This said, the Cox, Kaplan-Meier and Accelerated Failure Time models remain excellent survival data analysis models and are also the basis of the most complex algorithms. But above all, in their simplicity, intuitiveness and strong assumptions (such as that of proportional hazards) they are extremely performing, efficient and fast.

As already anticipated, despite the effort and time employed for the present analysis, we were able to cover only a small part of the much more complex survival analysis. There are still many areas of extension and application of survival analysis. Here are just a few ideas for future analysis.

Future propositions and works

Survival analysis is such a large topic which can be applied to many other fields, not just the medical one, and that despite the hours and effort spent on a thorough and timely analysis there are many other topics that would be extremely interesting to develop in the future.

The first objective of all is certainly that of trying to optimize even more the algorithms implemented even more by means of cloud solutions that allow a computation if not faster, certainly more stable.

Certainly the SEER program is a source of data of the utmost importance and from which you can still learn a lot. It would also be interesting and certainly useful to add to the variables that are already collected by the program other aspects that monitor the personal life of patients such as the drug therapies followed, hypertension or hypotension problems, pathologies, diabetes, type of diet followed, etc.. This is why the figure of the general practitioner becomes extremely important, the only one who can have this data which, if correctly collected, can be an enormous source of information. To this end, in the absence of datasets or programs that do such a thing, it could be interesting to simulate the values of these variables for the subjects present in the dataset from datest similar to the SEER but which also collect this type of information.

It would also be very useful to extend the Brier Score and especially the IPA to all the algorithms used in this analysis in order to define and therefore compare the models based on a unique metric. In fact, we recall that the Brier score and the IPA, although less commonly used than the C-index, are both discrimination and calibration measures. The C-index instead measures only the first aspect. As regards the Brier score, this has actually been implemented but the results have not been reported because the application of the algorithms on the test set causes a continuous crash of the program.

Lastly, another point of interest could be to compare the results obtained in Python using the different libraries available to us, with those obtained using the mlr3 package for R.

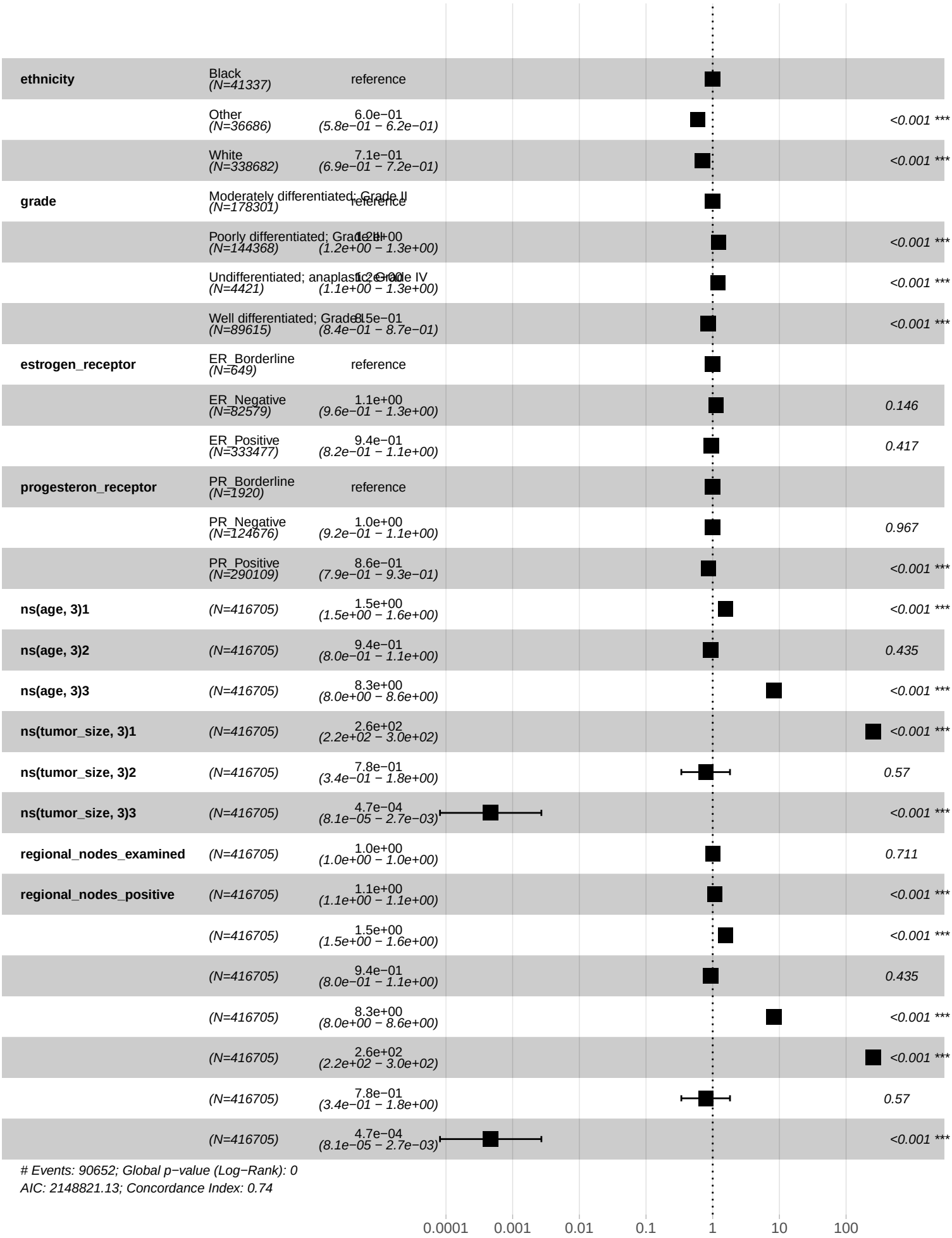
Chapter 7

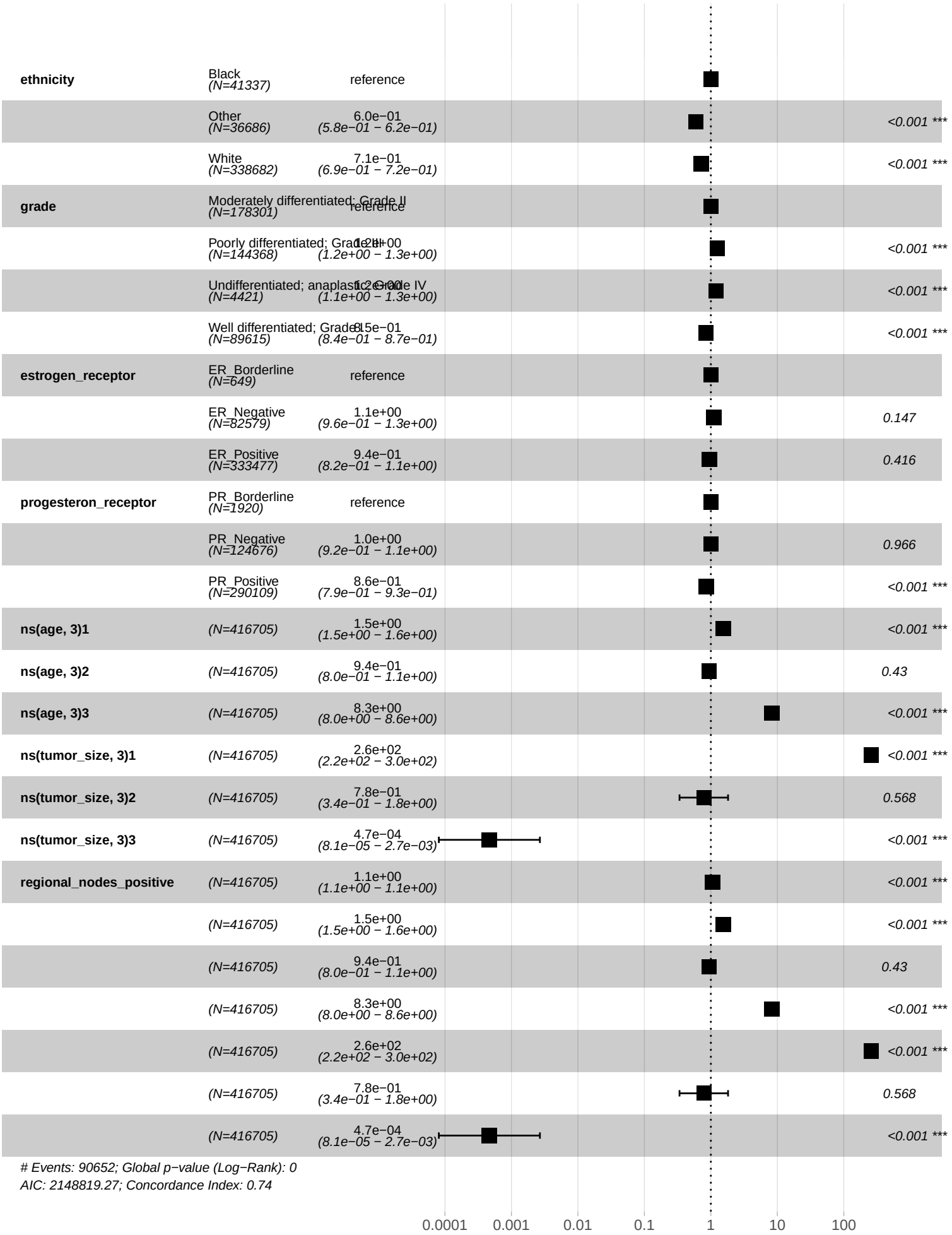
Appendix A

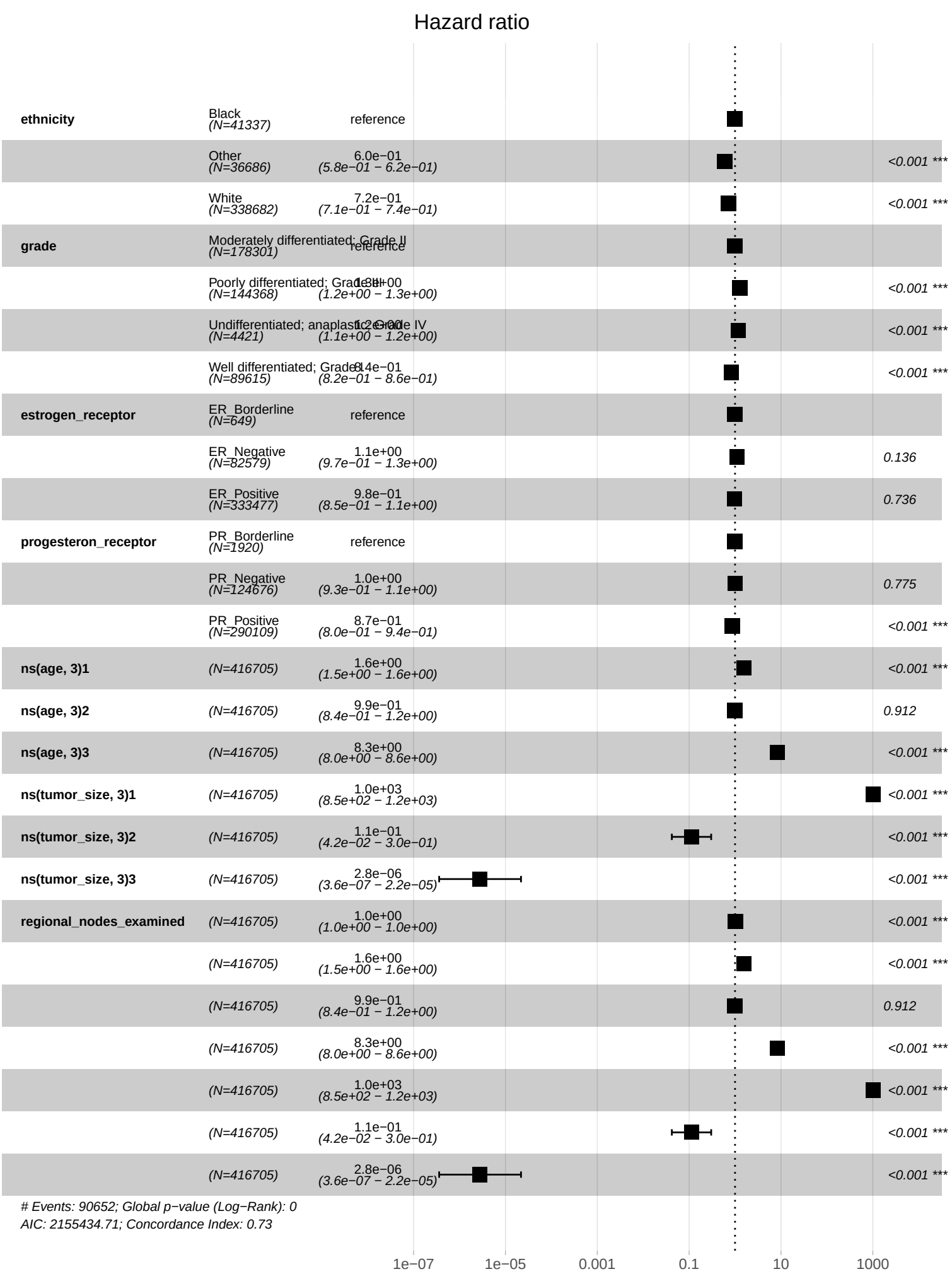
In this appendix we report some very interesting graphs on the Cox model obtained in R through the ggforest function. The graphs in question are Forest plots, or Blobbograms, graphical representations used to compare the results of different studies relating to the same issue. The dashed vertical line represents the value one.

The *ggforest* function outputs the hazard ratios resulting from the fitted Cox model, giving a direct view of how much and how the covariates affect the hazard. Each hazard ratio represents the relative risk of death with reference to a value taken as a reference. The graph also shows the confidence interval of the statistics (always at 95%) and the relative p-value in order to understand if the result is statistically significant.

We can observe other extremely significant results for the *grade* variable where the diagnosis of cancer in the fourth stage determines an extremely high probability (we are around 1300%) the risk of death, which is actually quite predictable.


Hazard ratio
ethnicity
Black (N=41337) reference
Other (N=36686) 6.0e−01 (5.8e−01 − 6.2e−01) <0.001 ***
White (N=338682) 7.1e−01 (6.9e−01 − 7.2e−01) <0.001 ***
grade
Moderately differentiated; Grade II (N=178301) reference
Poorly differentiated; Grade III (N=144368) 1.2e+00 (1.2e+00 − 1.3e+00) <0.001 ***
Undifferentiated; anaplastic; Grade IV (N=4421) 1.2e+00 (1.1e+00 − 1.3e+00) <0.001 ***
Well differentiated; Grade I (N=89615) 8.5e−01 (8.4e−01 − 8.7e−01) <0.001 ***
estrogen_receptor
ER_Borderline (N=649) reference
ER_Negative (N=82579) 1.1e+00 (9.6e−01 − 1.3e+00) 0.146
ER_Positive (N=333477) 9.4e−01 (8.2e−01 − 1.1e+00) 0.417
progesteron_receptor
PR_Borderline (N=1920) reference
PR_Negative (N=124676) 1.0e+00 (9.2e−01 − 1.1e+00) 0.967
PR_Positive (N=290109) 8.6e−01 (7.9e−01 − 9.3e−01) <0.001 ***
ns(age, 3)1 (N=416705) 1.5e+00 (1.5e+00 − 1.6e+00) <0.001 ***
ns(age, 3)2 (N=416705) 9.4e−01 (8.0e−01 − 1.1e+00) 0.435
ns(age, 3)3 (N=416705) 8.3e+00 (8.0e+00 − 8.6e+00) <0.001 ***
ns(tumor_size, 3)1 (N=416705) 2.6e+02 (2.2e+02 − 3.0e+02) <0.001 ***
ns(tumor_size, 3)2 (N=416705) 7.8e−01 (3.4e−01 − 1.8e+00) 0.57
ns(tumor_size, 3)3 (N=416705) 4.7e−04 (8.1e−05 − 2.7e−03) <0.001 ***
regional_nodes_examined (N=416705) 1.0e+00 (1.0e+00 − 1.0e+00) 0.711
regional_nodes_positive (N=416705) 1.1e+00 (1.1e+00 − 1.1e+00) <0.001 ***
(N=416705) 1.5e+00 (1.5e+00 − 1.6e+00) <0.001 ***
(N=416705) 9.4e−01 (8.0e−01 − 1.1e+00) 0.435
(N=416705) 8.3e+00 (8.0e+00 − 8.6e+00) <0.001 ***
(N=416705) 2.6e+02 (2.2e+02 − 3.0e+02) <0.001 ***
(N=416705) 7.8e−01 (3.4e−01 − 1.8e+00) 0.57
(N=416705) 4.7e−04 (8.1e−05 − 2.7e−03) <0.001 ***
Events: 90652; Global p−value (Log−Rank): 0
AIC: 2148821.13; Concordance Index: 0.74
0.0001
0.001
0.01
0.1
1
10
100


Hazard ratio
ethnicity
Black (N=41337) reference
Other (N=36686) 6.0e−01 (5.8e−01 − 6.2e−01) <0.001 ***
White (N=338682) 7.1e−01 (6.9e−01 − 7.2e−01) <0.001 ***
grade
Moderately differentiated; Grade II (N=178301) reference
Poorly differentiated; Grade III (N=144368) 1.2e+00 (1.2e+00 − 1.3e+00) <0.001 ***
Undifferentiated; anaplastic; Grade IV (N=4421) 1.2e+00 (1.1e+00 − 1.3e+00) <0.001 ***
Well differentiated; Grade I (N=89615) 8.5e−01 (8.4e−01 − 8.7e−01) <0.001 ***
estrogen_receptor
ER_Borderline (N=649) reference
ER_Negative (N=82579) 1.1e+00 (9.6e−01 − 1.3e+00) 0.147
ER_Positive (N=333477) 9.4e−01 (8.2e−01 − 1.1e+00) 0.416
progesteron_receptor
PR_Borderline (N=1920) reference
PR_Negative (N=124676) 1.0e+00 (9.2e−01 − 1.1e+00) 0.966
PR_Positive (N=290109) 8.6e−01 (7.9e−01 − 9.3e−01) <0.001 ***
ns(age, 3)1 (N=416705) 1.5e+00 (1.5e+00 − 1.6e+00) <0.001 ***
ns(age, 3)2 (N=416705) 9.4e−01 (8.0e−01 − 1.1e+00) 0.43
ns(age, 3)3 (N=416705) 8.3e+00 (8.0e+00 − 8.6e+00) <0.001 ***
ns(tumor_size, 3)1 (N=416705) 2.6e+02 (2.2e+02 − 3.0e+02) <0.001 ***
ns(tumor_size, 3)2 (N=416705) 7.8e−01 (3.4e−01 − 1.8e+00) 0.568
ns(tumor_size, 3)3 (N=416705) 4.7e−04 (8.1e−05 − 2.7e−03) <0.001 ***
regional_nodes_positive (N=416705) 1.1e+00 (1.1e+00 − 1.1e+00) <0.001 ***
(N=416705) 1.5e+00 (1.5e+00 − 1.6e+00) <0.001 ***
(N=416705) 9.4e−01 (8.0e−01 − 1.1e+00) 0.43
(N=416705) 8.3e+00 (8.0e+00 − 8.6e+00) <0.001 ***
(N=416705) 2.6e+02 (2.2e+02 − 3.0e+02) <0.001 ***
(N=416705) 7.8e−01 (3.4e−01 − 1.8e+00) 0.568
(N=416705) 4.7e−04 (8.1e−05 − 2.7e−03) <0.001 ***
Events: 90652; Global p−value (Log−Rank): 0
AIC: 2148819.27; Concordance Index: 0.74
0.0001
0.001
0.01
0.1
1
10
100


Hazard ratio
ethnicity
Black (N=41337) reference
Other (N=36686) 6.0e−01 (5.8e−01 − 6.2e−01) <0.001 ***
White (N=338682) 7.2e−01 (7.1e−01 − 7.4e−01) <0.001 ***
grade
Moderately differentiated; Grade II (N=178301) reference
Poorly differentiated; Grade III (N=144368) 1.3e+00 (1.2e+00 − 1.3e+00) <0.001 ***
Undifferentiated; anaplastic; Grade IV (N=4421) 1.2e+00 (1.1e+00 − 1.2e+00) <0.001 ***
Well differentiated; Grade I (N=89615) 8.4e−01 (8.2e−01 − 8.6e−01) <0.001 ***
estrogen_receptor
ER_Borderline (N=649) reference
ER_Negative (N=82579) 1.1e+00 (9.7e−01 − 1.3e+00) 0.136
ER_Positive (N=333477) 9.8e−01 (8.5e−01 − 1.1e+00) 0.736
progesteron_receptor
PR_Borderline (N=1920) reference
PR_Negative (N=124676) 1.0e+00 (9.3e−01 − 1.1e+00) 0.775
PR_Positive (N=290109) 8.7e−01 (8.0e−01 − 9.4e−01) <0.001 ***
ns(age, 3)1 (N=416705) 1.6e+00 (1.5e+00 − 1.6e+00) <0.001 ***
ns(age, 3)2 (N=416705) 9.9e−01 (8.4e−01 − 1.2e+00) 0.912
ns(age, 3)3 (N=416705) 8.3e+00 (8.0e+00 − 8.6e+00) <0.001 ***
ns(tumor_size, 3)1 (N=416705) 1.0e+03 (8.5e+02 − 1.2e+03) <0.001 ***
ns(tumor_size, 3)2 (N=416705) 1.1e−01 (4.2e−02 − 3.0e−01) <0.001 ***
ns(tumor_size, 3)3 (N=416705) 2.8e−06 (3.6e−07 − 2.2e−05) <0.001 ***
regional_nodes_examined (N=416705) 1.0e+00 (1.0e+00 − 1.0e+00) <0.001 ***
(N=416705) 1.6e+00 (1.5e+00 − 1.6e+00) <0.001 ***
(N=416705) 9.9e−01 (8.4e−01 − 1.2e+00) 0.912
(N=416705) 8.3e+00 (8.0e+00 − 8.6e+00) <0.001 ***
(N=416705) 1.0e+03 (8.5e+02 − 1.2e+03) <0.001 ***
(N=416705) 1.1e−01 (4.2e−02 − 3.0e−01) <0.001 ***
(N=416705) 2.8e−06 (3.6e−07 − 2.2e−05) <0.001 ***
Events: 90652; Global p−value (Log−Rank): 0
AIC: 2155434.71; Concordance Index: 0.73
1e−07
1e−05
0.001
0.1
10
1000

Chapter 8

Appendix B

8.1 Preamble

Before illustrating the steps necessary to replicate the creation of the dataset, it is important to inform the reader that access to the data provided by the SEER is subject to the request by the interested party and the signature of the so-called SEER Research Data Use Agreement, which must be signed and sent by email whether you decide to register through an institutional account or in the case of non-institutional users. Following the sending of the data agreement, the request will be processed and the user will be sent a username and a temporary password. For further information, the reader is invited to consult the SEER web page, the link to which will be inserted in the final references.[54]

Upon reviving the credentials it will be possible to download and use the SEER*Stat software.

8.2 Steps for the building of the dataset

We now focus on the steps to be implemented using the SEER software.

Selecting the session

As previously said the extraction of data is possible via different sessions, in this case the reader is invited to select the *Case Listing Session*.

8.2.1 Database Name

Selecting this session will automatically open a dialog box called *Database Name* in which the registers made available by the SEER are listed. For this analysis, the register selected is the one called "Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Lousiana Cases, Nov 2018 Sub (1975-2016 varying)".

A note on this registry: due to the Hurricanes Katrina and Rita, Louisiana cases diagnosed for the second half of 2005 (July – December 2005) were excluded from SEER 18 databases through the 1975-2016 SEER Research Data. However, July-December Louisiana cases were treated differently by session type and in particular for case listings, SEER excludes Louisiana cases diagnosed for the six month period, July-December 2005, by default.

As also the SEER site reports, it is important to highlight this point since the exclusion of Louisiana cases diagnosed in this time-frame may be have an impact on incidence rates computed at county or state level.

8.2.2 Selection

The next step requires clicking on the *Selection* tab in which two sections appear. In the *Select Only* the items:

- Malignant Behavior
- Known Age
- Cases in Research Database

should already be checked.

Below this small window there should be a much larger and empty one with three options beside it: *Edit*, *Copy* and *Clear*, the last two being grayed out.

Clicking on *Edit* opens a new window called *Case Selection* which is divided into three sections:

- Variable;
- Values;
- Selection Statement.

In the *Variable* and in the *values* section select the following:

- "Race, Sex, Year Dx, Registry, County" → Sex → Female
- "Race, Sex, Year Dx, Registry, County" → Year of diagnosis → from 1995 to 2015
- "Site and Morphology" → Site recode ICD-O-3/WHO 2008 → Breast
- "Multiple Primary Fields" → Sequence number → One primary only + 1st of 2 or more primaries

Clicking on OK this window will be closed and the selected filters will be applied.

8.2.3 Table

By moving on the *Table* tab it is now possible to select the variables that will make up the dataset, in particular the user should see three sections *Column*, *Sort*, *Available Variables*.

The last section is the one that will be used to select variables. Specifically for this project the following variables were chosen by selecting the variable and then clicking on *Column*:

- "Other" → Patient ID
- "Race, Sex, Year Dx, Registry, County" → Race recode (White, Black, Other)
- "Race, Sex, Year Dx, Registry, County" → State-County
- "Site and Morphology" → Primary Site – labeled
- "Site and Morphology" → Grade
- "Site and Morphology" → Laterality
- "Site and Morphology" → ICD-O-3 Hist/behav, malignant
- "Therapy" → RX Summ - Surg Prim Site (1998+)
- "Extent of Disease" → Regional nodes examined (1988+)
- "Extent of Disease" → Regional nodes positive (1988+)

- "Extent of Disease" → ER Status recode Breast Cancer (1990+)
- "Extent of Disease" → PR Status recode Breast Cancer (1990+)
- "Cause of Death (COD) and Follow-up" → COD to site recode
- "Cause of Death (COD) and Follow-up" → Survival months
- "Cause of Death (COD) and Follow-up" → Type of follow up expected
- "Multiple Primary Fields" → Sequence number
- "Multiple Primary Fields" → Primary by international rules
- "Multiple Primary Fields" → Total number of in situ malignant tumors for patient
- "Multiple Primary Fields" → Total number of benign borderline tumors for patient
- "Dates" → Year of birth
- "Dates" → Month of diagnosis
- "Race, Sex, Year Dx, Registry, County" → Year of diagnosis
- "Stage - 6th edition" → Breast - Adjusted AJCC 6th Stage (1088-2015)
- "Stage - 6th edition" → Breast - Adjusted AJCC 6th Stage T (1088-2015)
- "Stage - 6th edition" → Breast - Adjusted AJCC 6th Stage N (1088-2015)
- "Stage - 6th edition" → Breast - Adjusted AJCC 6th Stage M (1088-2015)
- "Stage - Summary/Historic" → SEER historic stage A (1973-2015)
- "Other" → Marital status at diagnosis
- "Extent of Disease" → CS tumor size (2004-2015)
- "Extent of Disease" → EOD 10 - size (1988-2003)

8.2.4 Output

Moving to the *Output* section it will be possible to give a name to the dataset containing the variables selected in the previous step.

8.2.5 Execute, save and use

Finally, to extract the dataset it is necessary to click on the lightning bolt symbol in the session bar and in order to use the data it is necessary to select the columns, copy and paste them in a txt file (less heavy) or Excel, and then open it with the chosen software to conduct the analysis (in our case R).

Bibliography

[1] American Cancer Society, *How Common Is Breast Cancer?*, https://www.cancer.org/

[2] World Health Organization (WHO), *The top 10 causes of death* (09-12-2020), https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] Staff, ACS Medical Content and News, *Facts Figures 2021 Reports Another Record-Breaking 1-Year Drop in Cancer Deaths* (12-01-2021), https://www.cancer.org/latest-news/facts-and-figures-2021.html

[4] P. Armaroli, J. Battagello, F. Battisti, P. Giubilato, P. Mantellini, P. S. de Bianchi, C. Senore, L. Ventura, M. Zappa, M. Zorzi, *Rapporto sulla ripartenza degli Screening - Maggio 2020*, Osservatorio Nazionale Screening, https://www.osservatorionazionalescreening.it/content/rapporto-ripartenza-screening-maggio-2020

[5] European Cancer Organization, 2020, https://www.cirse.org/wp-content/uploads/2020/12/Impact-of-COVID-19-on-Cancer_7-Point-Plan_Final-1.pdf

[6] Z. Taimur, *Survival Analysis — Part A: An introduction to the concepts of Survival Analysis and its implementation in lifelines package for Python* (18/03/2019), towards data science, https://towardsdatascience.com/survival-analysis-part-a-70213df21c2e

[7] Institute for Health Metrics and Evaluation, *GBD Results Tool*, http://ghdx.healthdata.org/gbd-results-tool

[8] G.M. Poltieri, *Patologia Generale Fisiopatologia Generale* III Edizione (2012), Piccin Nuova Libreria

[9] D. Shier, J. Butler, R. Lewis, *Hole's Human Anatomy Physiology* XIV Edition (2016), McGraw-Hill Education

[10] R. Villa, *Tumori benigni* (04-06-2018), AIRC, https://www.airc.it/cancro/informazioni-tumori/cose-il-cancro/tumori-benigni

[11] M. Roser, H. Ritchie, *Cancer* (2015), Our World in Data, https://ourworldindata.org/cancer

[12] AIRC, *Radioterapia* (21-10-2020), https://www.airc.it/cancro/affronta-la-malattia/guida-alle-terapie/radioterapia

[13] M. Ehrgott, A. Holder, *Operations Research Methods for Optimization in Radiation Oncology*, Journal of Radiation Oncology Informatics (2014)

[14] T. DeVita, S. Hellman, S.A. Rosenberg, Cancer: Principles and Practice of Oncology Review (2012), Lippincott Williams Wilkins

[15] Wikipedia, *La Mammella*, https://it.wikipedia.org/wiki/Mammella

[16] A. Marchet, F. Meggiolaro, E. Goldin, *Mammella* (2017), Piccin Nuova Libreria

[17] Roche, *Tumore al seno* (2018), https://peripazienti.roche.it/it/trials/cancer/bc.html

[18] Surveillance, Epidemiology and End Results (SEER), https://seer.cancer.gov/analysis/

[19] World Health Organization (WHO), ICD-10 Version: 2019, *International Statistical Classification of Diseases and Related Health Problems 10th Revision*

[20] S.G. Komen, *My Family Health History Tool, Estimating Breast Cancer Risk* (02/24/2021), https://www.komen.org/breast-cancer/risk-factor/understanding-risks/gail-method/

[21] National Cancer Institute, The Breast Cancer Risk Assessment Tool, https://bcrisktool.cancer.gov/

[22] National Cancer Institute, Breast Cancer Risk Assessment Tool: in practice- https://bcrisktool.cancer.gov/calculator.html

[23] J. Cuzick, *IBIS Breast Cancer Risk Evaluation Tool*, https://ems-trials.org/riskevaluator/

[24] I.A. Olivotto, C.D., *Population-based validation of the prognostic model ADJUVANT! for early breast cancer* (25/04/2005), Pubmed, https://pubmed.ncbi.nlm.nih.gov/15837986/

[25] The R Project, https://www.r-project.org/about.html

[26] *What is Python? Executive Summary* (2021), https://www.python.org/doc/essays/blurb/

[27] T. Osterbuhtr, *What is Anaconda and how does it relate to Python?*, https://www.venturelessons.com/what-is-anaconda/

[28] M. Lang, B. Bischl, J. Richter, P. Schratz, G. Casalicchio, S. Coors, Q. Au, M. Binder, M. Becker, *mlr3: Machine Learning in R - Next Generation* (2019), https://cloud.r-project.org/web/packages/mlr3/index.html

[29] M. Becker, M. Binder, B. Bischl, M. Land, F. Pfisterer, N.G. Reich, J. Richter, P. Schratz, R. Sonaben, *mlr3 book* (2021), https://mlr3book.mlr-org.com/

[30] Faculty of Mathematics and Physics of Charles University, Survival data analysis, https://www2.karlin.mff.cuni.cz/ pesta/NMFM404/survival.html

[31] NCSS Statistical Software, Chapter 570: Life-Table Analysis, https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Life-Table_Analysis.pdf

[32] Metodi Ricerche, *Rigolato 1741. La canonica del reverendo Nicolò Vuezil e le anime delle ville* (2009), https://www.alteraltogorto.org/alto-gorto/rigolato/rigolato-1741.html

[33] T. Zahid, Survival Analysis - Part A (16/03/2019), https://towardsdatascience.com/survival-analysis-part-a-70213df21c2e

[34] Committee for Proprietary Medicinal Products (CPMP) - EMEA, *Points to consider on adjustment for baseline covariates* (22/05/2003), https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-adjustment-baseline-covariates_en.pdf

[35] A. Faruk, *The comparison of proportional hazards and accelerated failure time models in analyzing the first birth interval survival data* (2018), J. Phys.: Conf. Ser. 974 012008

[36] S. Narkhede, *Understanding AUC - ROC Curve* (26/06/2018), towards data science, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[37] Statistical odds and ends, *What is Harrell's C-index?* (26/10/2019), https://statisticaloddsandends.wordpress.com/2019/10/26/what-is-harrells-c-index/

[38] M.W. Kattan, T.A. Gerds, *The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models*, Diagnostic and Prognostic Research (2018), https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0029-2

[39] T. Mitchell, *Machine Learning*, McGraw Hill (1997), http://www.cs.cmu.edu/ tom/mlbook.html

[40] S. Sharma, *What the Hell is Perceptron* (2017), https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53

[41] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, *Random Survival Forests*, The Annuals of Applied Statistics (2008), Vol.2, N.3, 841-860, https://arxiv.org/pdf/0811.1645.pdf

[42] D.H. Chowdary, *Decision Trees Explained With a Practical Example* (2020), https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53

[43] V. Avinash Sharma, *Understanding Activation Functions in Neural Networks* (2017), //medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0

[44] tdhopper.com, *Cross Entropy and KL Divergence*, https://tdhopper.com/blog/cross-entropy-and-kl-divergence

[45] H. Kvamme, Ø. Borgan, I. Scheel, *Time-to-Event Prediction with Neural Networks and Cox Regression*, Journal of Machine Learning Research 20 (2019), 1-30, https://jmlr.org/papers/volume20/18-424/18-424.pdf

[46] J.L. Katzman, U. Shaham, A. Cloninger et al., *DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network*, BMC Med Res Methodol 18, 24 (2018). https://doi.org/10.1186/s12874-018-0482-1

[47] Breast, Collaborative Stage Data Set (08/07/2013), http://web2.facs.org/cstage0205/breast/Breastschema.html

[48] Dr. A. Fortunato, *Recettore per gli Estrogeni/Progesterone* (07/12/2019), Lab Tests Online, https://labtestsonline.it/tests/recettore-gli-estrogeniprogesterone

[49] M. Yalaza, A. Inan, M. Bozer, *Male Breast Cancer*, Journal of Breast Health (01/01/2016), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5351429/

[50] S. Durrleman, R. Simon, *Flexible regression models with cubic splines*, Stat. Med. 8 (1989) 551-561

[51] A. Al-Masri, *What Are Overfitting and Underfitting in Machine Learning?* (22/06/2019), Towards Data Science, https://towardsdatascience.com/what-are-overfitting-and-underfitting-in-machine-learning-a96b30864690

[52] G. Coqueret, *Support Vector machines*, Machine Learning for Factor Investing (25/05/2021), http://www.mlfactor.com/svm.html

[53] L. Antolini, P. Boracchi, E. Biganzoli, *A time-dependent discrimination index for survival data* (04/12/2005), https://doi.org/10.1002/sim.2427

[54] NIH, How to Request Access to SEER Data, https://seer.cancer.gov/data/access.html