 Corso di Laurea Magistrale in Data Science and Economics

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

**PERFORMING A SURVIVAL ANALYSIS ON THE SEER BREAST CANCER DATASET: When Does Machine Learning Become Useful?**

Relatore:

Chiar.mo Prof. Federico Ambrogi

Tesi di Laurea Magistrale

Lirida Papallazi

Matricola n. 938943

Anno Accademico 2020/2021

Indice

[Abstract 1](#_Toc63898831)

[Introduction 2](#_Toc63898832)

[Principles of Oncology 7](#_Toc63898833)

[Data Sources and Literature 10](#_Toc63898834)

[The SEER Program 10](#_Toc63898835)

[ICD-10 and ICD-O-3 11](#_Toc63898836)

[Analytical Backgroud 15](#_Toc63898837)

[Analysis 16](#_Toc63898838)

[Conclusion 18](#_Toc63898839)

[References 19](#_Toc63898840)

# Abstract

Introduction

Cancer is still the second leading cause of death after cardiovascular diseases (ischaemic heart disease, stroke), respiratory (chronic obstructuve pulmonary disease and lower respiratory infections) and neonatal conditions - which include birth trauma, neonatal sepsis, preterm birth complications ((WHO), The top 10 causes of death, 2020). 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: Cumulative percentage of total global deaths (year 2019)

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 2* and from table ?) could be considered somewhat reassuring and a confirmation of the negative trend of death rate from cancer (relative tu the United States) as reported by the American Cancer Society (Staff, 2021). 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 2: Pie chart of the percentage of global deaths by top 20 leading causes of death (year 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 diseases | 5,2 |
| 7 | Tuberculosis | 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 |

Table 1: Global Health Estimates 2019, Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. Geneva, World 2020

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 countires have paritally 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 concers 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. (P. Armaroli, s.d.)

In July 2020 the situation pushed the Board of Directors of the European Cancer Organisation - a federation of 31 Memebr 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”. (Organization, 2020) 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 dealyed.

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 to study mortality in biological organisms and therefore for cancer studies, but also failures in mechanical systems and sociological events taking the name of *event history analysis*. Another definition wants survival analysis to corresponds to a set of statistical approaches used to investigate the time it takes for an event of interest to occur. The 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. (Zahid, 2019)

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

The initial purpose of this analysis is that of trying to conduct a complete survival analysis 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 acutally useful to improve the performance and when these algorithms really outperform the classical metodologies, by taking into consideration variables that in survival analysis are not considered and also changing our dataset.

Secondo dataset: https://www.kaggle.com/gunesevitan/breast-cancer-metabric

# Principles of Oncology

La costante crescita di incidenza dei tumori o se si vuole vedere il lato positivo ha aumentato l’importanza che l’oncologia assume in ambito medico. Questa crescita costante nei numeri di pazienti con cancro può essere dovuta anche ad un miglioramento progressivo dei programmi di screening ed in generale delle tecnologie mediche che hanno permesso nel tempo di diagnosticare la malattia già dalle prime fasi di avanzamento della stessa.

La malattia è diventata il centro di numerosi studi volti al miglioramento delle cure, diventate nel tempo sempre più complesse e specifiche.

Come detto anche precedentemente, le neoplasie con più alta incidenza nell’uomo e nella donna sono rispettivamente, quelle della prostata e quelle della mammella. Rimane elevata comunque anche l’incidenza del carcinoma del polmone che, come riportato anche nella Tabella 1 dell’Introduzione, è stata la nona causa di morte e rimane il più “mortale” tra i carcinomi.

INSERIRE TABELLE DI INCIDENZA E MORTALITA’ PER TIPOLOGIE DI TUMORI DAL SEER O DAL DATASET DELLA WHO.

Sottotitolo: Che cosa è il cancro

Secondo la WHO , cancro è il termine generico con il quale si indicano un vasto gruppo di malattie che possono intaccare qualsiasi parte del corpo. Alcuni tra i sinonimi più comuni, che abbiamo fra l’altro già avuto l’occasione di vedere anche all’interno di questo paper, sono tumore e neoplasma.

Il National Institue of Healht definisce il neoplasma o tumore come “an abnormal mass of tissue that forms when cells grow and divide more than they should, or do not die when they should”, i.e.: una crescita anormale di cellule che potenzialmente possono diffondersi in altre parti del corpo. In particolare è possibile distinguere due tipologie di neoplasmi:

* Benigni, che in quanto manchevoli della capacità di diffondersi ai tessuti circostanti, possono essere rimossi senza solitamente temere in recidive.
* Maligni, i quali invece presentano la capacità di intaccare i tessuti circostanti e una volta rimossi, se possibile, rimane comunque la possibilità che possano ritornare.

Ciò non significa comunque che anche i tumori benigni non causino danni sulla salute: la dimensione e la crescita di alcuni di questi possono infatti generare il cosiddetto “mass effect”, ossia comprimere con la propria massa i tessuti circostanti causando danni ai nervi, riduzione del flusso di sangue che può comportare anche ad una ischemia o alla necrosi dei tessuti. I tumori benigni al cervello ad esempio, rimangono comunque tra i più pericolosi/life threatening. Inoltre rimane comunque la possibilità che i tumori benigni, diventino più aggressivi ed eventualmente si trasformino in tumori maligni (ciò avviene tipicamente mediante un processo noto come tumor progression).

E’ interessante notare come la frequenza e la tipologia di tumori maggiormente diffusi cambi nei vari Paesi, in quanto numerosi studi hanno confermato che essi siano legati alle condizioni ambientali, abitudini, tradizioni e stili di vita. La WHO riporta che all’incirca un terzo delle morti per cancro sono dovute principalmente a 5 tipologie di comportamenti ritenuti rischiosi:

* High body mass index,
* Unhealthy diet with low fruit and vegetable intake,
* Lack of physical activity,
* Tobacco use, active or passive, which is also the most important risk factor accounting for approximately 22% of cancer deaths,
* Alchohol usage.

In un contesto del genere diviene chiara l’importanza che la prevenzione può avere, prima ancora dei complicati trattamenti farmacologici e chirurgici volti ad appunto curare la malattia già sviluppatasi; la stessa WHO stima che una percentuale tra il 30 ed il 35% dei tumori odierni possa essere prevented by avoiding risk factors and implementing existing and collaudate prevention strategies.

Negli ultimi decenni si sono ottenuti sostanziali progressi nel trattamento di alcuni tumori sino ad ottenere, in alcuni casi, complete guarigioni che fino a trenta anni fa erano solo una speranza, come nel caso del tumore alla mammella. Ed in altri casi invece la mortalità è purtroppo variata di poco, come nel caso del tumore al pancreas.

Questi progressi sono stati ottenuti grazie ad una combinazione di progresso farmacologico e tecnologico, ad un miglioramento delle strategie terapeutiche ma anche ad una maggiore sensibilizzazione dell’opinione pubblica e delle strutture sanitarie, ma è soprattutto una early detection a ricoprire un ruolo rilevante. Le principali componenti di questa sono

# Data Sources and Literature

As previously stated, for our analysis we will 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” ((NCI), s.d.). It therefore seems appropriate to give a description of the main characteristics 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. This becomes necessary also to define a common frame of reference that we will use in this report. For more and more accurate information on the program, however, we invite the readers to consult the official website of the same.

The SEER Program

In particular this program covers approximately 34.6% of the U.S. population throught the contribution of the so called authorised SEER registries, and 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.

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.

ICD-10 and ICD-O-3

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.

|  |  |  |
| --- | --- | --- |
| Chapter | ICD-10 CODES | 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 Neurodevelopmental 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 |

Table 2: 2021 ICD-10-CM Codes ((WHO), International Statistical Classification of Diseases and Related Health Problems 10th Revision, 2019)

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 neoplasm, 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's see a small example also to facilitate the reader's understanding: the specific codes for breast and lung (here presented for comparison and illustrative reasons) neoplasms are:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 |

Table 3: ICD-10 Alphabetic Index Entry for Breast and Lung Neoplasms

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 with respect to the lungs 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 all the official indexes and relative descriptions, and we will limit ourselves to indicating those of interest for our analysis (breast tumors), comparing them, as previously done, with those relating to lung cancer, in order to achieve a greater clarity.

|  |  |
| --- | --- |
| Behaviour 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, metastiatic site, * Malignant, secondary site. |
| /9 | Malignant, uncertain whether primary or metastatic site. |

Table 4: 5th Digit Behaviour Code for 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 implan-associated anaplastic large cell lymphoma |

Table 5: ICD-0-3 coding for breast neoplasms

# Analytical Backgroud

Analysis

Conclusion

# References

(CDC), C. f. (2020, February 13). *ICD-10 (Mortality)*. Retrieved from National Center for Health Statistics: https://www.cdc.gov/nchs/icd/icd10.htm

(NCI), N. C. (n.d.). *Surveillance, Epidemiology, and End Results Program (SEER)*. Retrieved from Surveillance, Epidemiology, and End Results Program (SEER): https://seer.cancer.gov/

(WHO), W. H. (2019). *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. Retrieved from ICD-10 Version: 2019: https://icd.who.int/browse10/2019/en

(WHO), W. H. (2020, December 9). *The top 10 causes of death*. Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Organization, E. C. (2020). *The Impact of Covid-19 on Cancer in Europe: The 7-Point Plan to Address the Urgency and Build Back Better*. Retrieved from The Impact of Covid-19 on Cancer in Europe: The 7-Point Plan to Address the Urgency and Build Back Better: https://www.cirse.org/wp-content/uploads/2020/12/Impact-of-COVID-19-on-Cancer\_7-Point-Plan\_Final-1.pdf

P. Armaroli, J. B. (n.d.). *Rapporto sulla ripartenza degli Screening - Maggio 2020*. Retrieved from Osservatorio Nazionale Screening: https://www.osservatorionazionalescreening.it/content/rapporto-ripartenza-screening-maggio-2020

Staff, A. M. (2021, January 12). *Facts & Figures 2021 Reports Another Record-Breaking 1-Year Drop in Cancer Deaths*. Retrieved from American Cancer Society: https://www.cancer.org/latest-news/facts-and-figures-2021.html

Zahid, T. (2019, March 18). *Survival Analysis — Part A: An introduction to the concepts of Survival Analysis and its implementation in lifelines package for Python*. Retrieved from towards data science: https://towardsdatascience.com/survival-analysis-part-a-70213df21c2e