POLITECNICO DI TORINO

Master of Science Degree in Computer Engineering

Master of Science Thesis

Data-Driven Analysis to Improve Oncological Processes in Hospital

Supervisors
Prof. Silvia Anna Chiusano
Prof. Ernestina Menasalvas Ruiz

Candidate
Manuel Scurti

October 2019

"Someone is sitting in the shade today
because someone planted a tree a long time ago."
— Warren Buffett

Abstract

Big Data technologies are becoming a pervasive technology in day-to-day life. Marketing, financial, automotive are some examples of sectors in which already we can see examples of how big data technologies can have a fundamental impact on society. The Healthcare sector, however, still has not benefitted from the wide scale use of Big Data technologies. This is because, the sector has been traditionally slow in adopting ICT and until recently, most clinical data was not stored digitally but on paper.

Clinical decision making for specific diseases cannot be made using current hospital IT systems as information is either not structured or has been collected for operational purposes without having in mind to reuse them for further analysis. Thus, there is an unmet need for smart analytics capabilities to help report and measure key quality indicators.

Cancer is the uncontrolled growth and spread of cells that arises from a change in one single cell. With more than 3.7 million new cases and 1.9 million deaths each year, cancer represents the second most important cause of death and morbidity in Europe.

In this thesis, we focus on lung cancer from which we count with the data of the anonymized data of patients being diagnosed of a lung cancer in the last 10 year and we present a data-driven approach to help clinicians measuring two key performance indicators: i) length of stay and ii) patients at risk of developing lung cancer. The main challenge is to deal with unstructured informations in order to extract the knowledge.

In order to achieve the goal this thesis presents a method to be able, first of all, to extract processes of the patient from unstructured data sets. To structure the project and make it replicable, CRISP-DM has been adopted as a methodology to fulfill the goals. The thesis also presents methods to analyze, clean and prepare data to obtain structured datasets from which the mentioned KPIs can be measured. The thesis also presents results that have been already discussed with the healthcare professionals.

Acknowledgements

Desde estas líneas me gustaría expresar mi más sincero agradecimiento a todos los compañeros de trabajo en el CTB, desde mi tutor, la profesora Ernestina Menasalvas Ruiz, a quien agradezco su labor de ayuda, de orientación y de motivación durante todo el desarrollo del proyecto; a Johnny, para sus comentarios y ayuda cuando se lo pedía, siempre dispuesto a ayudar; y a los demas por hacerme sentir parte de un equipo.
Finalmente quiero agradecer los que conocí en mi experiencia en Madrid, con su alma alegre y festiva.
Vorrei poi ringraziare la professoressa Silvia Anna Chiusano, per il suo lavoro di supervisione del progetto di tesi e orientamento alla carriera; Matteo, Valerio, Antonio e i miei compagni di università con cui ho condiviso difficoltà, scadenze, successi e la vita fuori dal Politecnico, sin dai primi anni di triennale; Tutte le persone che sono state parte del mio Erasmus perché siete parte di una esperienza indimenticabile, in particolare Maria, Giulia ed Eleonora per le avventure di Madrid (e Avila!) passate insieme; Davide, Emiliano, Jobba e tutti i miei amici di Pescara con cui sono cresciuto e mi hanno supportato a distanza; Veg, amico nonché supporter (o sopporter), e per le sempre ispiranti chiacchierate su progetti e futuro; Gaetano, collega di lavoro e amico, con cui ho lavorato fianco a fianco per un intero semestre e con cui a Madrid ho condiviso tanti momenti e tante risate; Martina, per aver dato nuove forme e colori non solo ai grafici di questa tesi, ma anche ai mesi passati insieme, condividendo viaggi, pensieri ed emozioni contrastanti; Alex, mio fratello, quasi mentore spirituale nonché critico d'eccellenza (da buon architetto) che mi ha sempre spinto a dare il meglio (anche grazie alle sue scommesse sulla mia carriera universitaria).
Questo lavoro di tesi chiude un percorso, iniziato 5 anni fa, e segna anche un nuovo inizio. A questo percorso ci preparava già il professor Paolo Cortese, durante le sue sempre interessanti lezioni e digressioni di Analisi Matematica, il quale ringrazio per la forma mentis impartitami per raggiungere questo traguardo.

Infine, un grazie speciale va a loro, ai miei genitori, che mi hanno dato la possibilità di seguire la strada che volevo e mi hanno sempre sostenuto. A loro dedico questo importante successo. Grazie.

Table of Contents

Figures

List of Tables

1 Introduction and Motivation

1.1 Introduction

Everyday, an enormous mole of data is being generated. A report of IDC (4), dated back in 2018, estimated that the global traffic of data will reach 33 zettabytes (ZBs) in 2018 and 175 ZBs by 2025. This trend is confirmed also by IBM, who states that users are generating 2.5 quintillion bytes of data each day (5).

IDC report also stated that amongst all industries, healthcare is facing the fastest growth of data generated and it is expected to rise by 36% by 2025. In fact, healthcare organizations accumulated 85% more data in 2019 than they did just two years before. This is due to the growing integration of information systems inside hospitals and the digitalization of Electronic Health Record (EHR). The clinical history of a patient is often composed by clinical notes, images from radiography, reports of health status generated along its life. Consequently, its clinical history can be reconstructed by combining data generated after each visit. At this scope, EHR contains detailed information of all the activities recorded of contacts between the patient, doctors and healthcare professionals in general, from consultations to hospitalization, medical tests, emergencies and so on.

The amount and speed of data generation in the health sector has created new challenges for analysis, management and extraction of knowledge from data. Big data techniques can answer these needs. Potential benefits of the application of data analysis techniques to healthcare would pave the way towards improvements in diagnosis and personalized medicine, as well as to carry out research which will allow the knowledge generated to be translated into new hospital policies to reduce costs and waste of resources.

In particular, cancer is one of the diseases with major impact in the public health sector. After cardiovascular diseases, it is one of the first causes of death and morbidity in Europe, with more than 3.7 million new cases and 1.7 million deaths per year (6).

Among all the different characterizations of cancer, lung cancer is the most common type of cancer affecting population, with around the 85% of the cases ending fatally. That is the highest probability among all types of cancer. According to AIRC data (7), in Italy there are around 373.300 new cancer patients, having 52% male subjects and 48% of

female subjects. Considering the entire Italian population, excluding skin related cancer, lung cancer is the third most frequent cancer.
Whether the data is structured or not, massive data processing techniques, such as data mining and machine learning make it possible to discover associations between the values of variables (association problems), groups of subjects that behave in a similar manner (segmentation or clustering) or even, based on historical data, are capable of constructing a model that would allow predictions or estimations of the future (classification).
While these techniques are widely used in some industrial sectors, they have been applied to a much lesser extent in the health sector for many reasons, highlighted among which is the lack of computerization in healthcare processes and the fact that the majority of clinical data is unstructured (free text) and difficult to exploit.
A major challenge derived from these data, as we have already mentioned, is the ability to track all processes to which the patient has been subjected to. This is extremely important for different objectives, such as knowing reasons of early re-admissions, emergencies, length of stay optimization as well as identification of people at risk of developing lung cancer. However, it is not possible to infer these processes from unstructured clinical notes.
With that said, the aim of this thesis work is then to structure clinical notes into clinical processes, in order to reconstruct the activities performed by the patient in a hospital and thus enabling the detection of problems and possible optimizations in hospital policies. This master thesis has been developed as part of the BigMedilytics EU funded project.

1.2 State of The Art

Several approaches have focused in the past in the use of different sources of information to improve the knowledge available about a certain specific disease. Some of the most relevant works, closely related with the main aspects to be discussed in this paper, are referred to Length of Stay (LOS); it is one of the main variables that affect both the economic impact of a disease in a hospital, as well as the quality of life of the patients suffering from a specific disease. In (8), an analysis of how different variables correlate with LOS by using Electronic Health Record (EHR) admission data is presented. In this case, data is collected from all the hospital services and the main goal of the analysis is to improve the management of LOS among patients. Another similar approach is followed in

(9) to create a model for predicting survival and length of stay in critically ill patients using sequential organ failure scores, or in (10) with older adults after cardiac surgery or in (11) about the determining factors associated with prolonged LOS in patients following cardiac surgery.

1.3 Goals

The main aim of this master thesis is to apply data analysis techniques to unstructured medical data, more specifically, to data of patients suffering lung cancer, in order to extract patterns to help oncology health professionals to understand possible ways to improve:

1. Length of hospitalization for oncology patients
2. Identification of people at risk of developing lung cancer

In order to fulfil these goals, the following sub-goals have to be achieved:

- O1 - Explore patients database
- O2 - Extract patients processes from the database
- O3 - Analyze processes to extract patterns and useful insights

1.4 Structure of Contents

The description of the work has been divided in chapters:

- Chapter 2 - Theoretical Framework: Introduces and describes the theory that explains why the research problem in this thesis work exists
- Chapter 3 - Materials and Methods: Describes all data involved in the process of structuring and analysis, together with an overview of the main tools being used
- Chapter 4 - Framework Design and Development: In this part we describe the problem and the proposed solution. Starting by the business problem to be solved we will go deep in analyzing the data produced in the hospital. Once data is explored, we present our proposal to group notes into clinical processes and we will see how to

extract the information required to fulfil the goals. We end this chapter by presenting the results and insights obtained so far.

- Chapter 5 - Discussion of Results: This chapter presents the discussions of the results.
- Chapter 6 - Conclusions and Perspectives: In this chapter we present the main conclusions and we will analyze possible lines of work that the development of this works has opened.

2 Theoretical Framework

2.1 Lung cancer

Amongst all types of cancer, lung cancer is the most common one, both in terms of incidence and mortality. It registers 1.5 million deaths per year worldwide, accounting for 18% of cancer deaths(12).

Tumor stage is one of the first factors in preliminary diagnosis. According to the stage, different decisions for developing treatment can be planned. The American Joint Committee on Cancer (AJCC) defines two standard systems for measuring cancer stage(13): i) Stage Grouping ii) TNM. Stage Grouping system consists on encoding tumor stage using roman numerals, mixed with alphabets and numbers. TNM system, instead, uses three parameters to define the stage: a) Size of the tumor (T) b) Number of lymph nodes (N) c) Presence of metastasis (M). The majority of patients with lung cancer are diagnosed at an advanced stage(14). In these advanced cases, where cancer spreads to other organs, the five-year survival is about 5% only (15). In fact, lung cancer has lower survival rates (18.6%) than other types of cancer such as colorectal, breast and prostate (15). The main risk factor to develop a lung cancer is the smoking habit (16). The risk of lung cancer increases with the number of cigarettes smoked and with the time of exposure to tobacco. Also passive smokers have a high risk of lung cancer. Furthermore, radon is considered another important cause of lung cancer after tobacco.

2.2 Hospital Processes

In an hospital management context, a process is a set of interrelated or interacting healthcare activities, which are performed for a subject of care with one or more health issues. The primary input and output to a clinical process is the health state. Consequently, a clinical process includes all visits, clinical judgments, reports resuming medical care progress and state of health of the patient at each stage of the hospitalization. Each patient usually experiences more than one process while suffering a disease like lung cancer. See figure 2.1 for an example clinical history of a patient suffering from lung cancer.

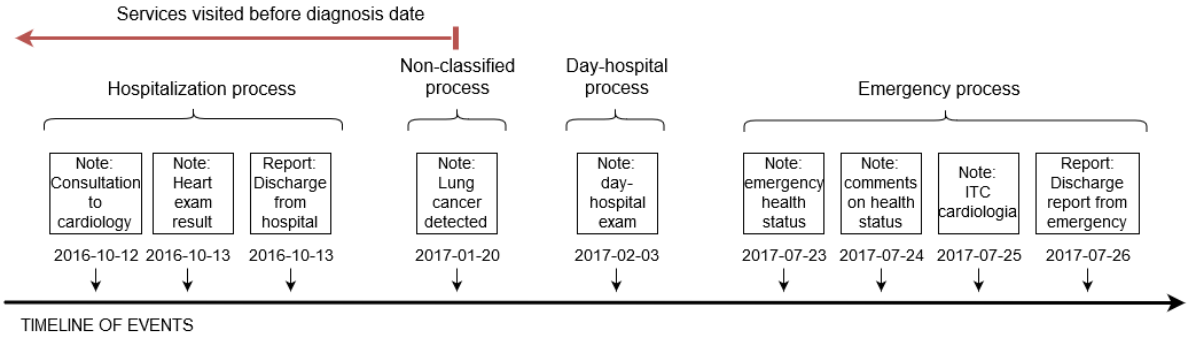

Figure 2.1: An example timeline of events of a single patient

For an hospital, a business process management approach allows for identification and management of processes that are interrelated(17); in order to analyze performances and make continuous improvements of the results, eliminating errors, and redundant processes in the institution. Process extraction allows for understanding how processes are executed in the system. Its application helps to identify bottlenecks, anticipate problems, record policy violations, recommend countermeasures and simplify processes for improving business performance.

An hospital process begins when a patient goes either to emergencies, consultations, hospital or day-hospital. Emergencies is the hospital service in charge of treating illnesses and injuries that require an urgent medical response, providing out-of-hospital treatment and transport to definitive care. Once the emergency has ended, and health status is judged as stable, depending on the status, the patient can go back home or being hospitalized for some days, where he will be monitored each day. Consultations with doctors can take place in doctors' offices or clinics, in hospital outpatient departments or, in some cases, in patients' own homes (18). Often patients are required to first consult a general practitioner (GP) about any new episode of illness. The GP may then refer them on to a specialist, if indicated. Under general conditions of admissions, a patient can be hospitalized for some days, to do exams, be monitored on the evolution of its illness and follow a treatment. In cases in which the treatment is short, the patient could be hospitalized for a single day in a so called, day-hospital. In figure 2.2, there is a resume between the previously explained types of processes and their relationships:

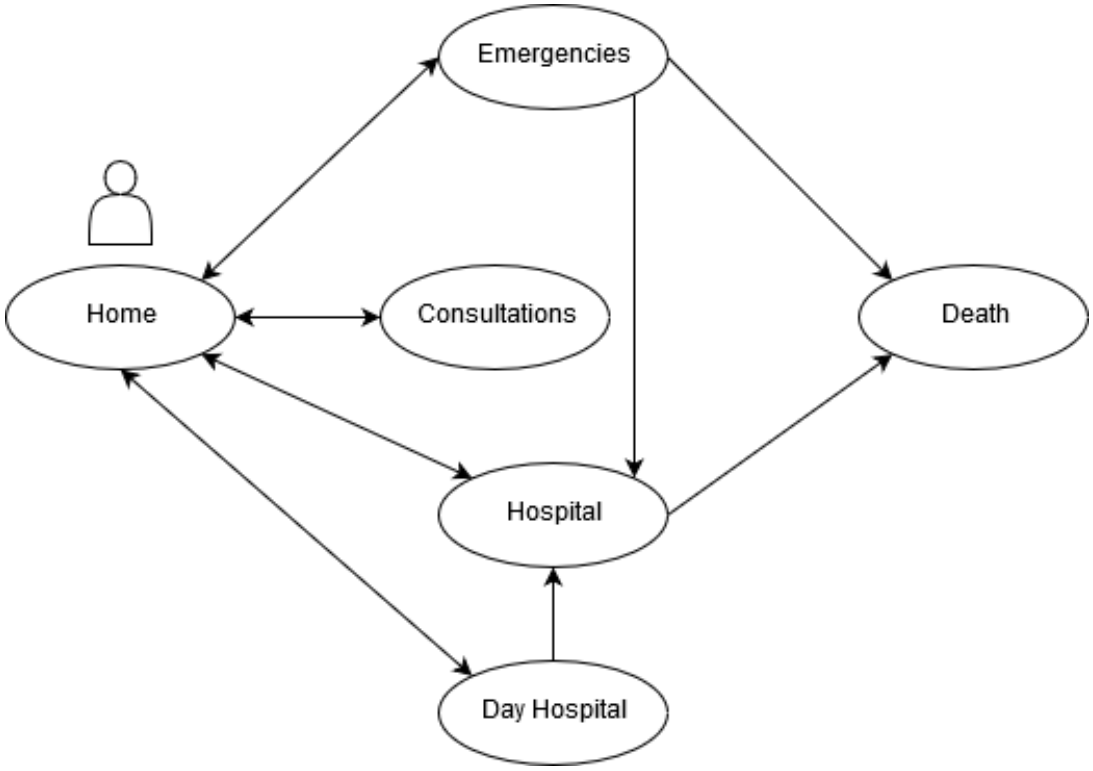

Figure 2.2: Processes relationships. The arrows indicate all possible paths for a patient

2.3 Key Performance Indicator

"What gets measured gets done." P. Drucker

A Key Performance Indicator (KPI) is a measurable value that reflects critical goals of a company to achieve success. It measures the trend of a business process and helps making decisions to improve the outcomes(19). Each company can use a set of metrics to track the status of a specific process. With KPIs, the company focuses the attention to one or more metric to become a point of reference to identify progresses.
The administration is in charge of defining business goals for its company, then, a so called business analyst will analyze company processes to define KPIs. However, not all processes can be analyzed with KPIs and in general, the opportunity to analyze a process using them is measured with a robustness scale, which considers(20):

- Ease of understanding - KPIs must explain what it is measuring at the minimum possible grain level
- Cost of data - Collecting data to enable analysis on a KPI costs. These costs can

include collecting data, structuring data and experts on interpret results

- Meaningfulness - KPIs must reflect only critical goals i.e. the ones which give the company an improvement on quality of processes
- Frequency of data change - the more informative the indicator is with time, the more robust will be

Once KPIs are defined, a clear snapshot of trends, state of objectives and space for improvements is given to the company.
In healthcare, performance indicators are used to increase patient's satisfaction, reduce costs and improve services quality. The most common KPIs defined in an hospital context are(21):

- Average Hospital Stay - Evaluate the amount of time patients are staying
- Treatment Costs - Compute costs per patient (e.g. on a daily basis)
- Hospital Readmission Rates - Track how many patients are coming back and why
- ER Wait Time - Identify rush hours in your emergency room
- Patient Wait Time - Monitor waiting times to increase patient satisfaction
- Number of unscheduled visits - Patients that needs assistance without having scheduled any visit

Such objectives are key factors of healthcare quality and patient's satisfaction. During the development of this thesis there will be a discussion on some performance indicators that are the ones we are giving a solution to manage and analyze data in order to be able to quantify the KPI such that the hospital can make strategic decisions.

2.4 Data Science

Data Science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data Science unifies statistics, data analysis, machine learning.

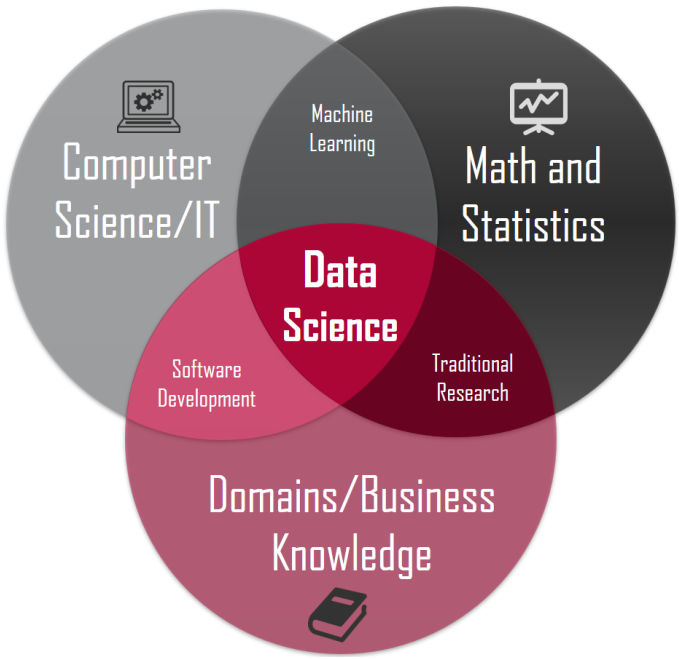

Figure 2.3: Disciplines influencing Data Science. Source: Barber 2018(1)

Before data science, the term data mining became popular in an article called From Data Mining to Knowledge Discovery in Databases in 1996(22), in which it referred to the overall process of discovering useful information from data. In 2001, William S. Cleveland wanted to bring data mining to a new level by combining computer science with data mining. Practically, he made statistics a lot more technical which he believed would expand the possibilities of data mining and produce a powerful force for innovation. Now you can take advantage of compute power for statistics and he called this combination Data Science. Around this time, this is also when web 2.0 emerged where websites are no longer just a digital pamphlet, but a medium for a shared experience amongst millions and millions of users, these are websites like MySpace (2003), Facebook (2004), YouTube (2005). We can now interact with these websites, meaning we can contribute, post, comment, like, upload, share, leaving our footprint in the digital landscape we call Internet, and help create and shape the ecosystem we now know and use today. This data became too much to handle using traditional statistical methods. This trend of growing data, is known as the «Big Data» phenomenon and more and more business activities are getting interested in exploiting these new analytics tools to plan their growing strategies. In fact, the ability to make data-driven decisions is crucial to any business nowadays. With each transaction, note, call or message, a world of valuable information is created.

Over the past ten years, data is becoming a game changer of virtually any field of science, engineering and services. Companies, public administrations, public health collected huge amounts of data. It is estimated that in 2025, we will exchange on the Internet an amount of 463 exabytes of data (23). Furthermore, over the last two years alone 90% of the whole amount of data was generated (24), that means an exponential growing is happening. However, even if we have so much data, they are often captured and stored without planning any further analysis on them, leaving their potential knowledge hidden and requiring a great effort to extract their real value. Furthermore, data is mostly temporal, and its information is hidden not in a single sample but in a sequence of events, this makes the extraction process even more complicated. Let us think for example to clinical notes, despite containing lot of knowledge regarding the patient status, this temporal information, written in natural language, cannot be processed by a computer and cannot be analyzed directly by machine learning techniques. However, a process of data wrangling can prepare the data in order to extract knowledge. This is exactly the focus and challenge of this thesis.

2.5 Structuring a Data Science Project: CRISP-DM

As every scientific method, a rigorous process must be followed to reach knowledge in an objective, reliable, repeatable and shareable way.

Data Science needed a standard methodology in the process of translating business questions into data science goals, together with standardized data transformations, data analysis techniques, and metrics to evaluate the effectiveness of the results.

A common problem in data science projects, is the repeatability of the results. In fact, it can happen that an algorithm with a known result on certain data, will not give back the same result in a different environment. This means that the success or failure of a data mining project relies on the person or team carrying it out.

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a standard process model that describes common approaches used by data mining experts and aims to make large data mining projects, less expensive, more reliable, more repeatable, more manageable and faster (25). It was published in 1999, but the second version of the model came only between 2006 and 2008. Now it is still widely adopted for data science projects.

CRISP-DM methodology divides the data mining process into six phases:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

In figure 2.4, there is a resume of the phases. They stand in a cycle which represent that while each phase is implemented, one can revisit and refine previous steps. This cyclic nature results in a loop that can keep working even though the solution has been achieved, to improve the given solution or give answers to new future questions. In what follows we will describe in deep each of the phases of the methodology as this is the methodology we have followed for the development of this project. We will see, however, that the work of this thesis belongs mainly to the first 3 phases: business understanding, data understanding and data preparation. In fact, in any data science project these phases account for the 50-80% of the efforts of the whole project.

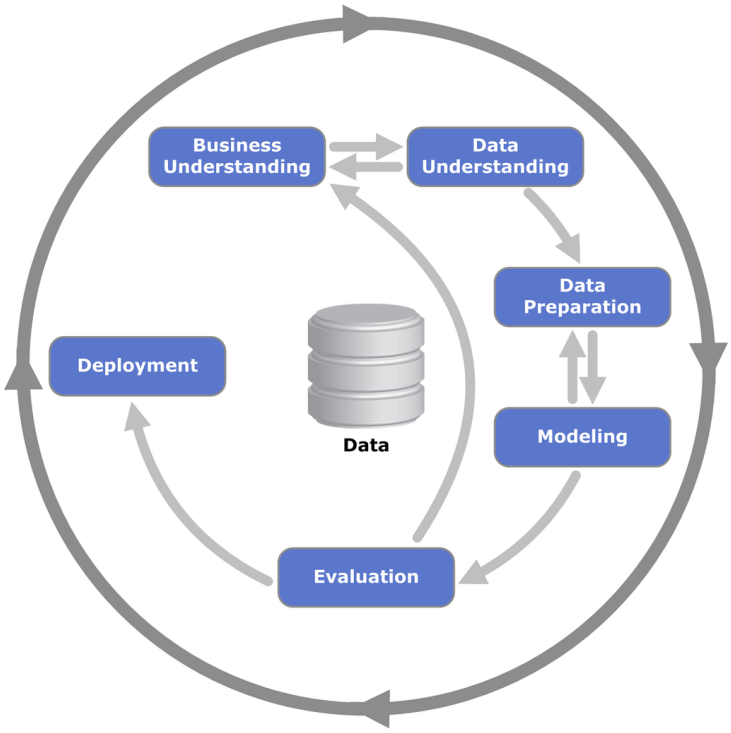

Figure 2.4: Process diagram showing CRISP-DM phases and their relationships. Source: Data Science Central (2)

2.5.1 Business Understanding

It is the initial phases of each data mining project. It focuses on understanding the project objectives and requirements from a business perspective, and then converting questions into a data mining problem definition, and a preliminary project design to achieve answers.

2.5.2 Data Understanding

This phase includes all the steps needed to get from initial data to a general understanding of the population, identifying data quality problems, discovering insights into data, or to detect interesting subsets to be used to discover information.

At this scope, a preliminary descriptive analysis is needed, to extract features and test significant variables. Testing significant variables is often performed with correlation analysis. The term feature"identifies characteristics of a data sample. For example, Name", "Age", Gender"are typical features of patients data set. Finally, data visualizations will help identifying significant patterns and trends in data. Humans are naturally less familiar

with numbers, and thus by using charts we increase the comprehension of the whole data set.
A wide range of tools is available for data scientists with the aim of helping them in data visualization and exploration phase, it is worth mentioning: Ggplot2, Matplotlib and Dplyr.

2.5.3 Data Preparation

Once a general understanding of the available data is obtained, it is time to adapt data for our objectives. This step covers all the tasks carried out to build the final data set, that will be fed into a modelling tool, from the initial raw data. This phase is more likely to be repeated multiple times. Tasks include data cleaning, attribute selection, filtering, construction of new attributes, transformation of data. This stage usually goes through two steps:

- Data Retrieval
- Cleaning, Filtering and Transformation

Data Retrieval
It is the very first step in a data science project, but it is reported inside data preparation instead of data understanding because here we work massively on data transformations while for preliminary exploratory analysis it is not needed to retrieve all the data into an offline machine but instead could be enough to execute just some queries.
Data Retrieval consists on obtaining data from company's data lakes, storage drives, web APIs or any kind of available data sources needed to address project objectives.
In this phase, typically querying a database will be the main activity, using technical skills like MySQL to process the data. However, databases are not the only form of input data that can be retrieved, even other data formats may be ingested into the data pipeline, like CSV (Comma-Separated Values) files. For addressing multiple types of input data, two popular programming languages are particularly powerful for these tasks: R and Python, which they read any type of file using specific packages. Another type of database, where data can be stored and retrieved in an efficient way, are non-relational databases (NoSQL), like MongoDB, which they become popular for managing huge quantities of

data. Furthermore, Web APIs are more and more adopted by companies that want to make public part of their data, and thus they become another possible source of data. Finally, especially for the Internet of Things (IoT), data can be retrieved from streaming sources, and ingested into processing pipelines like Apache Hadoop, Apache Spark and Apache Flink for real-time insights or create a collection of data to be further analyzed.

Cleaning, Filtering and Transformation

After obtaining data, a cleaning and filtering process is needed. This is because data often arrives unstructured, contains errors and missing values and this leads to false conclusions or even worse make it impossible to produce analysis at a first glance. And so, cleaning and understanding the data has a high impact on the quality of the results.

Data cleaning is the process of identifying and correct, or eventually delete, corrupt or inaccurate records from a data set. Data cleaning may be performed using automated tools or as batch processing through scripts(26).

The result of this process should make data consistent with other similar data sets in the storage. Errors may be caused by user entry errors, by corruption in transmission or storage, by different data dictionary definitions, or by unavailability of information. The process of data cleaning may involve removing typographical errors or validating and correcting values using a known list of possible values. Validation may be strict or fuzzy. In a strict validation a record can be rejected if any field is missing or if contains any error, while in a fuzzy validation a record can be adjusted, with statistical techniques, to be filled with all the needed values or corrected. Validation, however, must not be confused with data validation, because in that validation often means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

Finally, data transformation can be applied after cleaning and filtering processes were applied. Data transformation is used to convert data from one structure into another one. Applying such process has the benefit of giving the data set the right shape in the context of the needed analysis, but it can happen that this phase is not needed.

2.5.4 Modelling

In this stage, various modelling techniques are selected and applied. Then, parameters are fine-tuned to the optimal values. Typically, for the same data mining problem there are

different techniques available to be tried. There is no single solution that fits every data mining problem; thus it is a good approach to try out a reasonable number of distinct algorithms(27).
Typically Machine Learning techniques are applied at this stage and they can be applied both in an unsupervised or supervised fashion, depending on the objective. Generally, objectives vary from prediction of new data points, classification or clustering. After having defined the objective, the stage is usually composed by an iterative process:

- Algorithm implementation
- Training - Data is fed into the algorithm, to eventually learn patterns in data
- Validation - The trained algorithm is checked against a validation data set to tune parameters that best fit our data
- Testing - The algorithm is tested against new data to evaluate the abstraction power of the model

Once the model has obtained enough accuracy, it can be used to produce insights.

2.5.5 Evaluation

After one or more models are built, it is essentially to evaluate them using some performance metrics. Performance metrics usually include accuracy, precision and in some cases computational effort, but metrics also depends on the objective of the analysis. A key objective is to determine if there is some important business issue that has not been sufficiently considered, in order to avoid situations in which the results are not realistic. At the end of this stage, a decision on the use of results should be reached.

2.5.6 Deployment

Creation of the model could not coincide with the end of the project. Usually, knowledge gained will need to be organized and presented in a way that customer can request and use it. Depending on the requirements, the deployment could be as simple as generating a report or as complex as developing a repeatable data mining process to supply APIs to make data available on-demand or even to make them available on a dashboard.

3 Materials and Methods

3.1 Datasets

In what follows we are going to describe the main sources of data that have been used in the development of this thesis. These are: i) Raw dataset of clinical documents coming from the hospital and ii) Database of patients containing the clinical information extracted for each patient once the document database has been pre-processed in a NLP pipeline.

3.1.1 Documents File

The first source of data used in this project has been retrieved from the computer-based patient record system used in the oncological service of an hospital, only from patients affected by lung cancer. Prior to receiving this data from the hospital, all documents have been anonymized, such that in no way it is possible to obtain personal data of the patients, such as name or surname. Consequently, all documents regarding the patients that have been diagnosed with lung cancer are available. This includes notes generated by the oncological service and all the notes prior or after diagnose of the patient in any service (oncological or not) of the patient in the hospital.
This database is a comprehensive longitudinal database that contains documents (clinical notes and reports) starting from 2009. The data collected include therapies, nursery notes, reports, laboratory tests. Its data structure uses a proprietary relational model which stores all patient's information in different structures.
An extraction procedure captures the information from the database and generates one -CSV file.
This database is a live database having new patients and documents each day. Consequently each 6 months the data is updated with the information generated in the period since the last extraction.
The CSV file presents the following structure:

- **ID** - Unique identifier of the document
- **EHR** - Unique identifier of the patient

- **birthDate** - Date of birth of the patient
- **gender** - Gender of the patient
- **category** - Classify the document between Report and Note
- **subcategory** - Label assigned by the doctors to recognize service which produced the note or report and the purpouse of its creation
- **date** - Creation date of the document
- **hospital** - Hospital where the document was created
- **language** - Language of the text
- **text** - Textual content of the record, containing natural written text
- **city** - City where the patient lives

The focus of the analysis will be mainly put on: ID, EHR, category, subcategory, creation date and text. In order to understand the main challenges of this work, it is crucial to have clear the contents of these notes and the difference between a Report and a Note.
Clinical notes and reports mainly differ on the moment when documents are generated. A clinical note is any note the health professional writes during his contact with the patient. While a report is generated by a physician at the end of a hospitalization or once the patient request for a report. Consequently, normally reports are longer as they summarize the contents of many notes.
Typically, clinical notes contain sections about the following information: history of present illness, allergy, medication, past medical history, past surgical history, family history, social history, physical exam, studies, laboratory tests and assessment. There are two main challenges in section classification for clinical notes. First, terms that physicians use to designate sections are ambiguous and various, for example, «history of present illness» might appear as «HPI», «history», «history of current illness». Second, physicians often omit section headers when they author clinical notes. For both these reasons it is hard to infer a section type given the text in the section (28).
A **report**, instead, gives a detailed description of the symptoms, signs, diagnosis, treatment, and follow-up of an individual patient. In figure 3.1, a sample of a possible report is presented.

Possible part of a report
SERVICIO: Oncología Radioterápica *FECHA INGRESO: 07/03/2012 15:06* *FECHA ALTA: 14/03/2012* *MÉDICO RESPONSABLE INFORME: Dr. XXX (Médico Adjunto). Dra. YYY (Médico Residente)* *Motivo de Ingreso:* *Disnea y fiebre.* *Antecedentes Personales* *NO alergias medicamentosas conocidas.* *- No HTA, no DM, no DL.* *Antecedentes Familiares:* *-* *HISTORIA ONCOLÓGICA:* *.....* *Pruebas complementarias: Analíticas y Radiodiagnóstico* *- ECG:* *- Rx Tórax:* *- .....* *Analítica al alta:* *- BIOQUÍMICA:* *-.....*

Figure 3.1: An example of report

On the other hand, a **note** is usually a short text updating other nurses about the status of the patient, its cause of hospitalization, consultations and so on, according to the specified subcategory. In figure 3.2, an example note is presented.

Sample Text in Clinical Note
Paciente que acude sin cita para solicitar control analítico. *AP: ex ADVP hace 10a. No ttos. F. 30cig/d. No B.* *AF: esposo VHC tratado con respuesta viral sostenida.* *MC: VHC+ conocido hace 10a. Seguida por su médico de primaria. Siempre transas y ECO normales. Solicito analítica con carga viral y genotipo y ECO.*

Figure 3.2: An example of note

Each time a patient goes either to external consultation, inter-consultation, surgery or any service of the hospital, a clinical note is produced.

A patient can enter either from emergencies or general admissions, then each step of the

process is documented either by a note or a report. The process can last more than one consecutive day, as a regular hospitalization, or can occur during different days with in-day hospital, where visits, exams, or surgery can be done on a single day hospitalization. Amongst all possible subcategories of notes and reports, a subset of them is crucial for this project, because they characterize some temporal references for extracting processes or they help in the classification task to assign a label to notes and thus, to processes. These subcategories are:

- ITC (Inter-consultation), CEX (External consultation) - these kind of notes refers to consultations happened between clinicians. It contains information regarding the health status of a patient. ITC notes are often created in consultations processes, but they can also appear inside hospitalization processes, after surgery treatments. Its grade of presence inside a process could infer that the process is a consultation process.

- Discharge report - these reports are produced at the end of a hospitalization or emergency process. A crucial information to extract from here is the date of hospitalization and discharge of the patient. It can also contain the service and the doctor which produced the report. Discharge reports can be divided into two main categories, as specified in the title of the subcategory: hospitalization discharge report and emergency discharge report. Its presence inside a process could determine the process class. If there are both types of discharge reports inside a process, then it means that the patient moved from emergencies to hospitalization.

- «Traslado de Cadaver» (Death Report) - It is a note which informs of the patient's death. The presence of this note inside a process describes a process that started in emergencies or hospital and ended with the death. This class of processes is crucial to analyze the ending treatments received by the patient. Furthermore, its creation date could mark the end of a process, in case it was not possible to retrieve it in any other way.

- «Motivo de Ingreso» (Cause of Hospitalization) - It is a note which informs the reason of admission of the patient. This could describe the beginning of a process, but it is crucial for other applications, such as analyzing reasons of early re-admissions. By combining its creation date with the creation date of the temporal nearest corpse

transfer notes, one could infer a new process.

- HDIA (Day Hospital) - These notes can be of different types: infirmary comments, treatments, medical follow-up. In all cases, if found inside a process, they form a day-hospital process.

Figure 3.3 depicts the main notes and reports, of relevance for this work, that could be produced after each stage, as a consequence of the main flow of process that was depicted in figure 2.1.

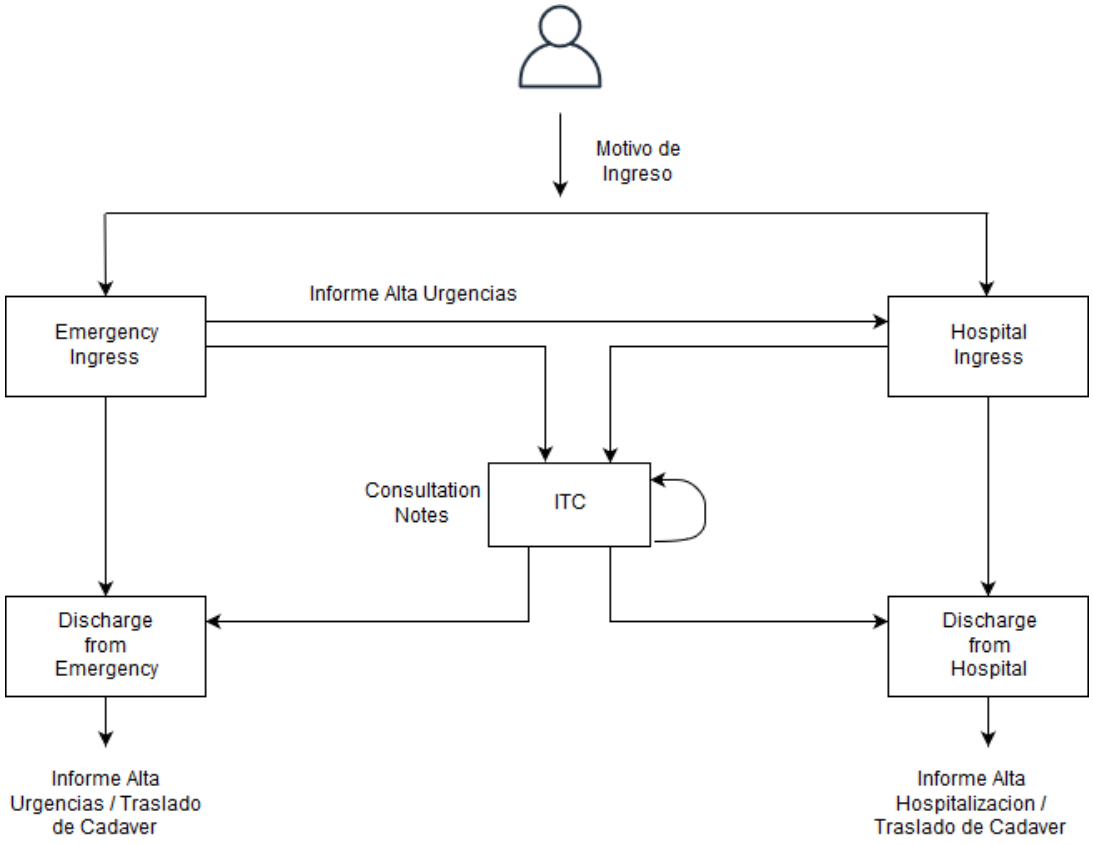

Figure 3.3: Hospital workflow, as described by domain experts. Titles showed outside boxes represent the possible category of notes produced after each stage

3.1.2 NLP Pipeline

The raw data coming from hospital goes through a Natural Language Processing pipeline in order to extract mainly:

- Available dates from text
- Diagnoses

- Treatments
- Antecedents
- UMLS (Unified Medical Language System) concepts

Furthermore, each text has been decomposed into sections, paragraphs and sentences, to give users the maximum granularity level and thus understand the context of the different entities. The results of this pipeline are saved into a central database managed by MySQL. The schema of this database is out of the scope for this thesis. However, in this database the main entity is the patient and not any longer the document as the information has been extracted from notes and integrated for each patient.

3.1.3 Patients File

As a consequence of the data pipeline, a table is generated which contains information of patients (967 at the time of this study) being described by 122 attributes. These attributes are filled by information found inside texts, after an NLP process is applied. Amongst all these attributes, the main ones, used for the scope of this thesis are being described:

- **EHR** - It is the unique identifier of a patient. This field appears as a foreign key inside the csv file of clinical notes, that is needed to join the two tables.
- **birthDate** - Contains the birth date of the patient
- **gender** - Gender of the patient
- **startAge** - Age of the patient at the start of his therapies
- **deathDate** - Death date of the patient
- **city** - City where the patient lives
- **tumorType** - Describes the class of its lung cancer, between Squamous and Non-Squamous carcinoma.
- **diagnosisDate** - It is the date in which the patient was diagnosed with lung cancer.
- **stage** - Describes the stage of the cancer at diagnosis time

These fields are being used to make correlations between the processes to which a patient has been subjected to and its personal profile.
Despite the structured information extracted there is still crucial data that has not been extracted such us: some metadata grouping together temporal related notes and reports (i.e. processes), notes and reports classes, processes classes. Besides as we have said, we have dates in which notes has been generated and in which diagnoses have happened but notes and events are not grouped, thus the structured database does not contain information on clinical processes of each patient and it is not possible to know whether a note is from a consultation, hospitalization or emergency, and from which service it is coming or reasons of admissions.

3.1.4 Sections File

Another consequence of the NLP pipeline, is the disaggregation of texts into paragraphs. As mentioned in section 3.1.1, reports and notes can contain labels indicating the beginning of a paragraph. The NLP pipeline detects these labels, using an internal dictionary, and divides the text as many pieces as the number of labels it detects. As a result, the following table called Sections is derived:

- **document** - It is the unique identifier of the document where the section was extracted
- **id** - Unique identifier of the section inside a document
- **begin** - Specifies the initial character position inside the text where the section begins
- **end** - Specifies the ending character position inside the text where the section ends
- **name** - It is the detected label indicating the begin of the section
- **kind** - It is the corresponding label inside the dictionary used to detect the beginning of the section
- **text** - Text of the paragraph

This dataset is being used for extracting services and dates of hospitalization and discharge inside texts, but requires to be cleaned and formatted.

3.2 UMLS - Unified Medical Language System

The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records(29). The purpose of UMLS is to facilitate the development of computer systems that behave as if they "understandthe meaning of the language of biomedicine and health. Often clinicians use terms abbreviations, that make impossible for a non-informed automated system like a simple entity recognition to extract knowledge from medical texts.
UMLS once applied to texts, produces a table with all the recognized concepts inside texts. Some interesting fields, for our scope, are presented.

- **document** - Unique identifier of the document where the concept was extracted
- **section** - Unique identifier of the section inside the document
- **id** - Unique identifier of the concept inside the UMLS dictionary
- **begin** - Specifies the initial character position inside the text where the concept was detected
- **end** - Specifies the ending character position inside the text where the concept was detected
- **sty** - Category of the concept found
- **concept** - Concept found in the text
- **negated** - Flag that specifies if the concept is found to be negated or not (e.g. patient has no fever)

For this thesis, UMLS is being used to extract services name and causes of hospitalization from Spanish texts. It has been also massively used inside the preliminary data pipeline to disaggregate texts and extract patients' information.

3.3 MySQL - Database Management System

All data being used in this project has been retrieved from a MySQL relational database. MySQL is an open source relational database management system, owned by Oracle. It runs virtually on all platforms, including Linux, Unix and Windows.
MySQL is based on a client-server model. The core of MySQL is a MySQL server, which handles all database instructions (or commands). MySQL server is available as a separate program for use in a client-server networked environment and as a library that can be embedded (or linked) into separate applications. Clients of MySQL can be developed under different forms. In this thesis the client used to communicate to the server is an R package called RMySQL.
MySQL enables data to be stored and accessed across multiple storage engines, including InnoDB, CSV, and NDB. MySQL is also capable of replicating data and partitioning tables for better performance and durability. MySQL users aren't required to learn new commands; they can access their data using standard SQL commands.

3.4 R (Programming Language)

The programming language R is generally used for statistical computing and graphics. It represents a GNU Project and shows similarities to the S language, an environment created at the Bell Laboratories. Overall, R can be perceived as a different implementation of S.
R is a tool which plays an important part in the area of statistics. Being a open source system, many different packages containing a variety of tools, methods and techniques are freely available. (The R Foundation, 2016) Packages are also available for the sector of machine learning, data mining and multivariate statistics. Especially the package e1071 created by the Department of Statistics - Probability Theory Group (Formerly: E1071) - placed at the Technical University of Vienna is one of the most used ones. This, according to Geethika Bhavya, who analyzed the most downloaded R packages from January to May 2015.
Also, in the area of data mining and knowledge discovery R offers a vast amount of packages and implementations of different algorithms. In this project, this programming

language made possible to get from data to results without focusing too much on implementation details, thanks to its natural orientation to data science tasks. Around 10 R packages were used within the scope of this thesis, the most used ones are presented below.

3.5 Visualization Packages in R

ggplot2

It is a popular package nowadays for R users, to make powerful and elegant plots, particularly taught for data science tasks. Users create plots by writing code in a declarative style, by simply providing to ggplot2(30) the data, how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. This package has been widely used for all the plots you will find in this thesis.

3.6 Data Manipulation Packages in R

stringr

Strings play a big role in our data cleaning and preparation tasks. Stringr package(31) provides a cohesive set of functions designed to make working with strings as easy as possible.

dplyr

dplyr(32) is a powerful R-package to transform and summarize tabular data with rows and columns. The package contains a set of functions (or «verbs») that perform common data manipulation operations.

It is easy to build pipelines of operations on data with this package, and it was the main reason of its adoption in this project.

RMySQL

RMySQL package(33) contains all the basic functions to connect, authenticate and retrieve data from a MySQL database. It has been used to retrieve all the data used in this project.

4 Framework Design and Development

Before diving into the execution of the required steps to achieve the thesis objectives, it has to be clarified what follows. There is a common misperception that data analysis is mostly a process of running statistical algorithms on high-performance data engines. In practice, this is just the final step of a longer and more complex process; 50 to 80 percent of an analyst's time is spent wrangling data to get it to the point at which this kind of analysis is possible.
Not only does data wrangling consume most of an analyst's workday, it also represents much of the analyst's professional process: it captures activities like understanding what data is available; choosing what data to use and at what level of detail; understanding how to meaningfully combine multiple sources of data; and deciding how to distil the results to a size and shape that can drive downstream analysis. These activities represent the hard work that goes into both traditional data «curation» and modern data analysis (34).

4.1 Methodology

In order to achieve our objectives, we follow a methodology based on CRISP-DM with special emphasis on data understanding and preparation.

4.2 Business Understanding

In this project two KPIs have been given by the hospital with the aim to be understood and improved: Length of hospital stay for oncological processes and identification of people at risk of developing lung cancer. We further explain them.

4.2.1 Length of hospital stay for Oncological Processes

Objective: Reducing length of hospital stay, by detecting patients with high probability of consumption of hospital care resources (e.g. emergency visit, admission, readmission) and proactive action to avoid hospitalization in safety conditions, improving patient

satisfaction, their outcomes, and reducing costs for the healthcare system. Furthermore, detection of end-stage (end-of-life) patients to promptly react to maximize patient and family well-being and avoid the excessive and useless consumption of resources. Finally, finding factors in distance from the hospital affecting severity of patient's cancer at the time of admission.

Success Criteria: Process Outcomes KPI 1 - Length of hospital stay.

This key performance indicator is composed by the following metrics:

1. Average hospitalization time
2. Number of responsible doctors during hospitalization
3. Estimated distance from patient's home to hospital
4. Main reasons of early re-hospitalization
5. Number of ER (emergency) admissions

4.2.2 Identification of People at Risk of Developing Lung Cancer

Objective: As stated before, identifying people with high probability of developing lung cancer. Moreover, detection of the frequently used services and most common symptoms in patients before their diagnosis date of cancer. Thus, enabling the improvement of such services and identify a set of diseases correlated with the danger of developing lung cancer, in order to improve the treatment effectiveness in such cases.

Success Criteria: Process Outcomes KPI 2 - Identification of people at risk of developing lung cancer

This key performance indicator is composed by the following metrics:

1. Main services visited by patients before lung cancer diagnosis
2. Main reasons of visiting services

Consequently, in this master thesis we will focus on these KPIs and will face the challenge of extracting values to measure KPI by analyzing data sources of the hospital as in 4.3. When analyzing the steps of a Data Science project we have mentioned that first stages are time demanding. It is also true that the first stages will highlight first insight for

further analysis. In fact, in this thesis we will focus on data understanding and preparation and we will analyze the first insights obtained.

4.2.3 Technical Goals

After business goals have been defined, it is now time to translate them into technical objectives, as showed in the following. We will refer to these with the term Data Preparation Goals (DPG).

General Preparation Goals

- DPG 0.1 - Cleaning and standardization of available data
- DPG 0.2 - Extraction of dates of admission and discharge from reports
- DPG 0.3 - Extraction of processes from the information contained in the clinical documents

KPI 1 Specific Goals In order to obtain insights to fulfil the KPI 1 the following tasks are required:

- DPG 1.1 - Classification of processes into a fixed set of categories, to enable more precise analysis
- DPG 1.2 - Compute length of stay per type of process
- DPG 1.3 - Compare length of stay per type of process: before vs after diagnosis date of cancer
- DPG 1.4 - Compute length of stay per stage of cancer
- DPG 1.5 - Correlation between patients' cancer data and their hospitalization processes
- DPG 1.6 - Correlation between patients' cancer data and their hospitalization processes that end with their death

KPI 2 Specific Goals In order to obtain insights to fulfil the KPI 2 the following tasks are required:

- DPG 2.1 - Extract diagnosis date for each patient

- DPG 2.2 - Extract services names from notes and reports
- DPG 2.3 - Extract causes of hospitalization from notes and reports
- DPG 2.4 - Compute top 10 most common causes of hospitalization before diagnosis date
- DPG 2.5 - Compute top 10 most used services before diagnosis date
- DPG 2.6 - Compare services usage between Cardiology, Pneumology and Emergencies by period
- DPG 2.7 - Compute top 3 causes of hospitalization for Emergencies, Pneumology, Cardiology
- DPG 2.8 - Compute main causes of hospitalization for Emergencies, Pneumology, Cardiology
- DPG 2.9 - Compute how most common services and causes of hospitalization vary with patients data like gender, age, stage and smoking habit

Once the goals are clear the next step is to analyze available data to make a first evaluation of whether the goals will be feasible with the given data.

4.2.4 Situation Assessment

Apart from the data sources we have already described, the hospital has provided to us some general data regarding the hospital that will be later used for validation in the extraction process. In particular the following data is given regarding the use of resources by patients.
Independently of the service, a patient has an average of 1.61 emergency visit, while the length of hospital stay on average is 7.72 days.
Besides, the hospital has also given to us information, assumptions and considerations to be able to understand the data, that must be considered for our goals. In particular:

- The time window to look for services and causes of hospitalization before lung cancer diagnosis must be set. Doctors have established to be from 6 months before diagnosis

date to 24 months before.

- Types of processes are defined by doctors. They will be illustrated in section 4.4.7

4.3 Data Understanding

In this section, a first preliminary exploration on data is described. As a consequence, we will produce a report of the data understanding.

4.3.1 Description of Data Files

4.3.1.1 Documents File

Data source with 296003 records and 13 features containing all notes and reports of an hospital, written in Spanish. These notes belong to 989 distinct lung cancer patients.

	id	EHR	birthDate	gender	category	subcategory	date	hospital	language	begin	end	text	city
1	1	[redacted]	1968-01-17	Female	Report	Informe Alta de Hospitalización	2012-03-14	[redacted]	es	0	11963	SERVICIO: Oncologia Radioterapica FECHA INGRESO: 07...	NA
2	2	[redacted]	1968-01-17	Female	Report	Informe de Seguimiento Consulta Externa	2011-10-28	[redacted]	es	0	6611	SERVICIO: Oncologia Medica FECHA CREACION INFORM...	NA
3	3	[redacted]	1968-01-17	Female	Report	Informe de Seguimiento Consulta Externa	2012-03-05	[redacted]	es	0	4842	SERVICIO: Oncologia Radioterapica FECHA CREACION IN...	NA

Figure 4.1: A first look on clinical notes data set

Documents File - Overview		
Variable name	Type	Subtype
ID	Numeric	Continuous
EHR	Numeric	Continuous
birthDate	Categorical	Ordinal
gender	Categorical	Ordinal
category	Categorical	Nominal
subcategory	Categorical	Nominal
date	Categorical	Ordinal
hospital	Categorical	Ordinal
language	Categorical	Nominal
begin	Numeric	Continuous
end	Numeric	Continuous
text	Categorical	Nominal
city	Categorical	Nominal

Table 4.1: Overview of documents file

A detailed description of the meaning of each field can be found in section 3.1.1

4.3.1.2 Patients File

Data source with 967 records with 122 attributes describing patients who have been diagnosed with lung cancer.
The information on this data file contains the resulting records after processing the clinical documents with a NLP process, later grouped with all the clinical documents of a patient and having as final result a database of patients.
Out of 22 patients from the initial 989 patients, for which we have clinical notes associated, were removed due to lack of essential information.
An overview of the main fields used for the scope of the thesis is given in table 4.2.

Patients File - Overview		
Variable name	Type	Subtype
EHR	Numeric	Continuous
birthDate	Categorical	Ordinal
gender	Categorical	Ordinal
startAge	Numeric	Continuous
deathDate	Categorical	Ordinal
city	Categorical	Nominal
tumorType	Categorical	Nominal
diagnosisDate	Categorical	Ordinal
stage	Categorical	Ordinal

Table 4.2: Overview of patients file

A detailed description of the meaning of each field can be found in section 3.1.3.

4.3.1.3 Sections File

Data source with 755792 records with 7 attributes identifying each section of each text in the documents file. Each record describes the type of section, its position in the text, and the document from where it has been extracted.
An overview of the fields is given in table 4.3.

Sections File - Overview		
Variable name	Type	Subtype
document	Numeric	Continuous
id	Numeric	Continuous
begin	Numeric	Continuous
end	Numeric	Continuous
name	Categorical	Nominal
kind	Categorical	Nominal
text	Categorical	Nominal

Table 4.3: Overview of sections file

4.3.2 Exploratory Analysis

A general overview of each data source is given in this section.

First, we will start the exploration of the raw dataset (clinical documents), then we will explore the information regarding patients.

4.3.2.1 Distribution of Documents: by Year

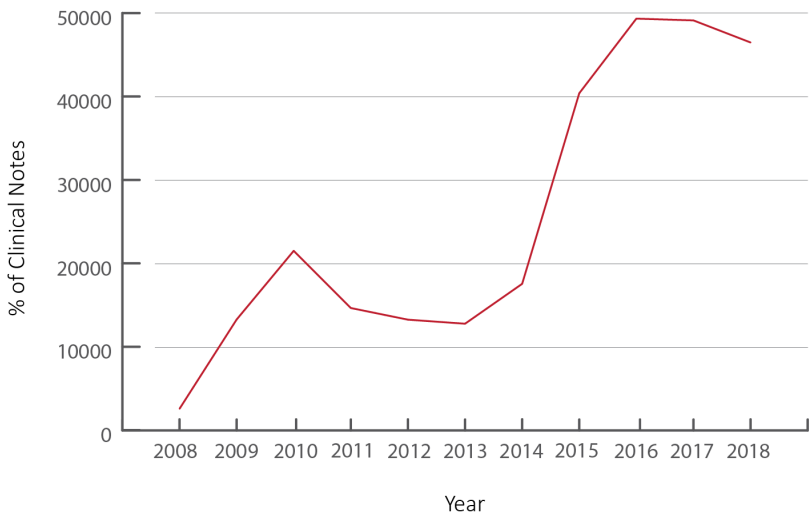

Figure 4.2: Number of documents produced per year

Data Source: Documents file.

Figure 4.2 presents the number of documents produced per year.

The analyzed period ranges from year 2008 to 2018, as they are respectively, the first year for which we have memory of any clinical document for this hospital and the last year (at

the time of development of the project).

For the generation of this graph, documents created in 2019 were excluded, because at the time of writing this thesis, the available data does not represent the whole year.

A first evidence is that the number of documents in the hospital is increasing since the adoption of the system.

4.3.2.2 Number of Documents per Patient

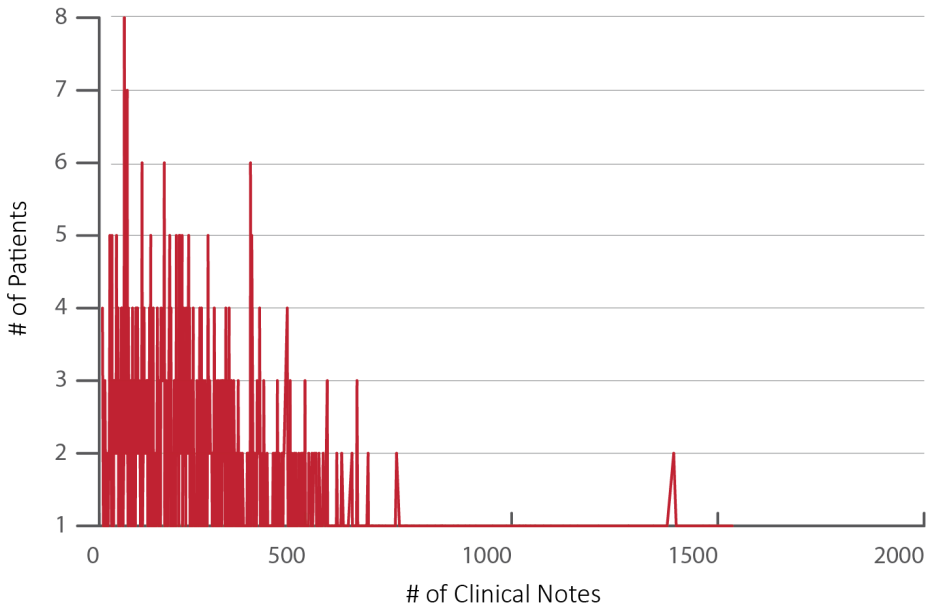

Figure 4.3: Number of documents per patient

Data Source: Documents file.

Figure 4.3 presents the number of documents per patient, having on the x-axis the number of documents, and on the y-axis the number of patients with such number of documents. On average, each patient has around 292 notes or reports in the data set, with a minimum of 8 documents per patient to a maximum of 1535 documents per patient.

This result suggests that a reasonable quality on processes extraction can be achieved for most of patients, because a sufficient number of reports and notes to extract processes is available.

4.3.2.3 Distribution of Documents: Reports vs Notes

Category	Count	%
Reports	14547	5%
Notes	269445	95%

Table 4.4: Distribution of documents: Reports vs Notes

Data Source: Documents file.

Table 4.4, shows the number of reports versus the number of notes.

As evidenced, most of documents are notes. This could affect the number of processes that will be extracted, as the hospitalization and discharge dates of them are often defined by dates contained in reports. However, the great number of notes is an index of how much confusion it is hided inside the dataset, and thus, it highlights the need of a semantic grouping to give rich meaning to notes.

4.3.2.4 Distribution of Documents: Reports per Patient

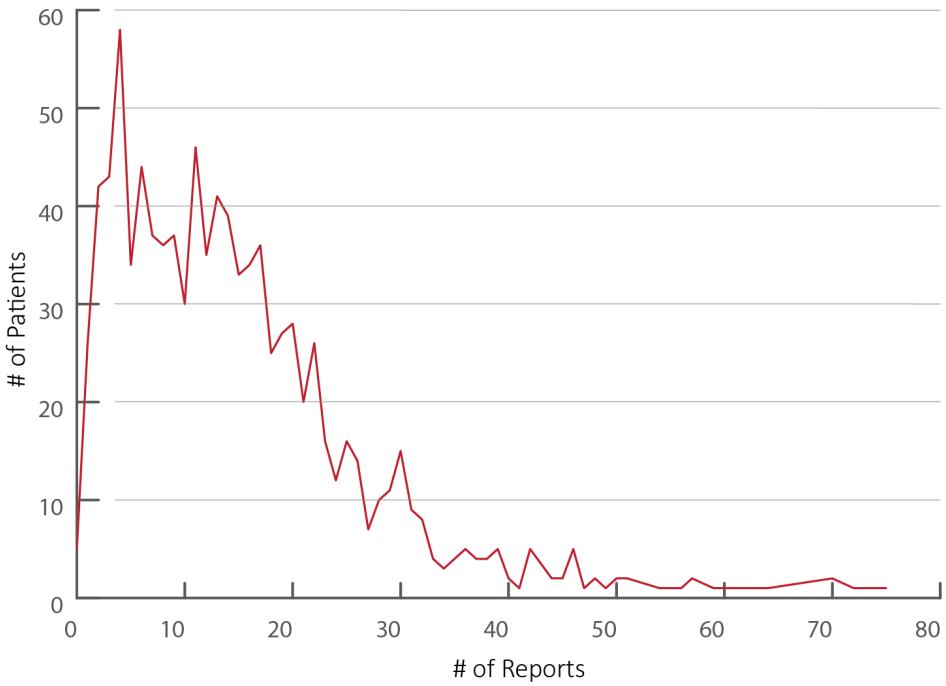

Figure 4.4: Distribution of number reports per patient

Data Source: Documents file.

Figure 4.4 shows the distribution of number of reports per patient.

On average, each patient has 13 reports, with a minimum of 0 reports per patient to a maximum of 75 reports per patient.
For those patients having 0 reports records it will be difficult to have processes extracted relying only on hospitalization and discharge dates, due to this lack.

4.3.2.5 Distribution of Documents: by Main Subcategory of Interest

Main Subcategories Types - Frequencies		
Subcategory	Count	%
Informe Alta	8368	2.92
ITC	12377	4.32
Motivo Ingreso	1901	0.66
Traslado Cadaver	211	0.07

Table 4.5: Main Subcategories of Interest

Data Source: Documents file.
Table 4.5, represents the number of documents by subcategory, together with the percentage relative to the total number of documents independently of the category.
The set of subcategories has been restricted to only the 4 most important subcategories for the scope of the project: «Informe Alta» (Discharge Report), ITC, «Motivo de Ingreso» (Cause of Hospitalization), «Traslado Cadaver» (Death Report). Hence, excluding notes belonging to other categories not listed.
As seen, ITC is the category with the major number of documents.
Notice that, each single note or report is useful for the analysis, but as already discussed, the categories being analyzed here have implicitly more information than usual, and thus they are frequently exploited for processes extraction. The list of categories has been already discusses in section 3.1.1.
Other interesting insight on clinical documents can be achieved only further to a pre-processing stage.

4.3.2.6 Distribution of Patients: by Age

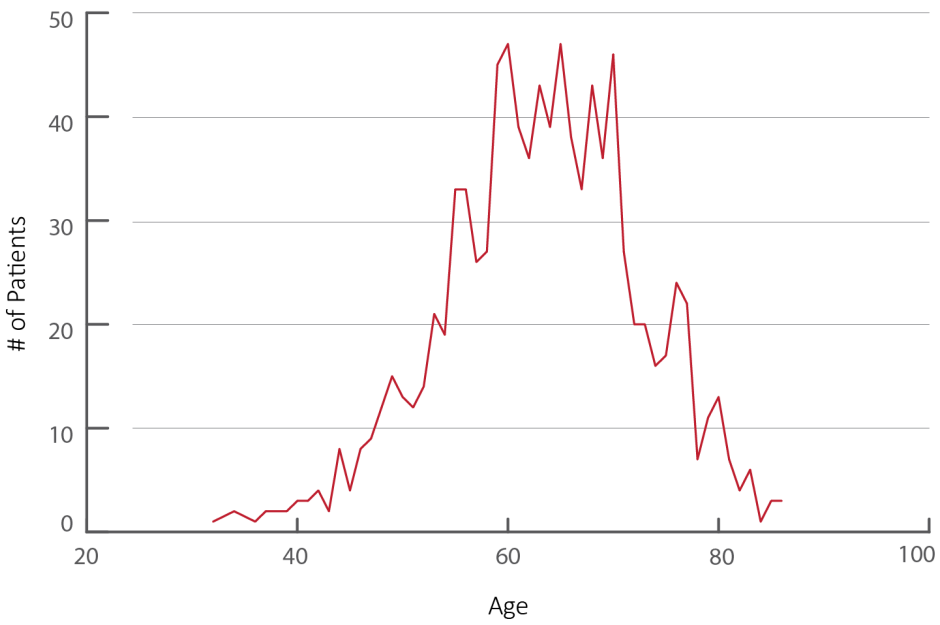

Figure 4.5: Distribution of patients by age

Data Source: Patients file.

In figure 4.5, there is the distribution of age of patients.

The figure shows ages at which patients were diagnosed with cancer.

Observe that the mean age of patients is 63 years old with a minimum of 32 and a maximum of 86 years old.

4.3.2.7 Distribution of Patients: by Gender

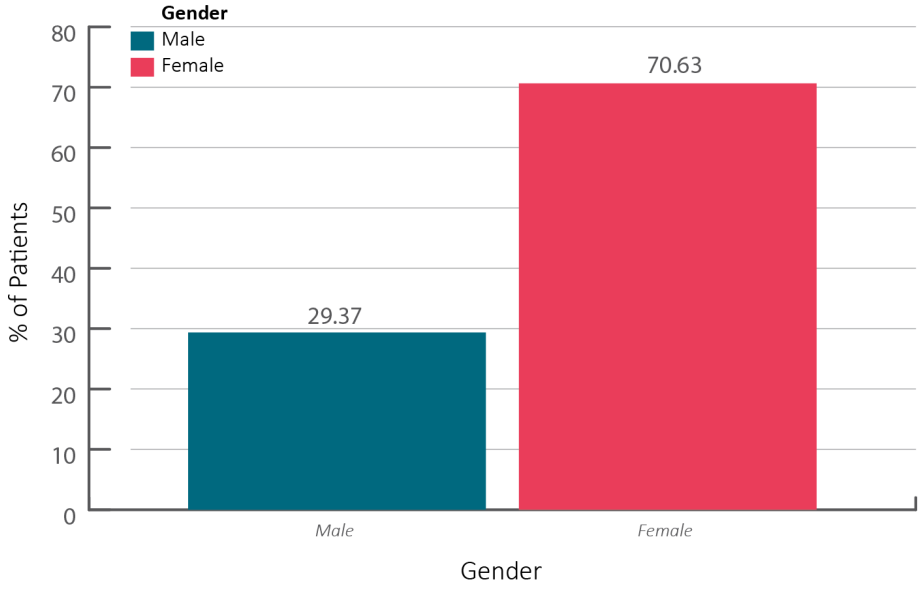

Figure 4.6: Distribution of patients by gender

Data Source: Patients file.

Figure 4.6 shows the distribution of patients by gender.

Observe that the majority of patients are men, with a 70.63% of the total, while women are only the 29.37% of the population.

4.3.2.8 Distribution of Patients: by City

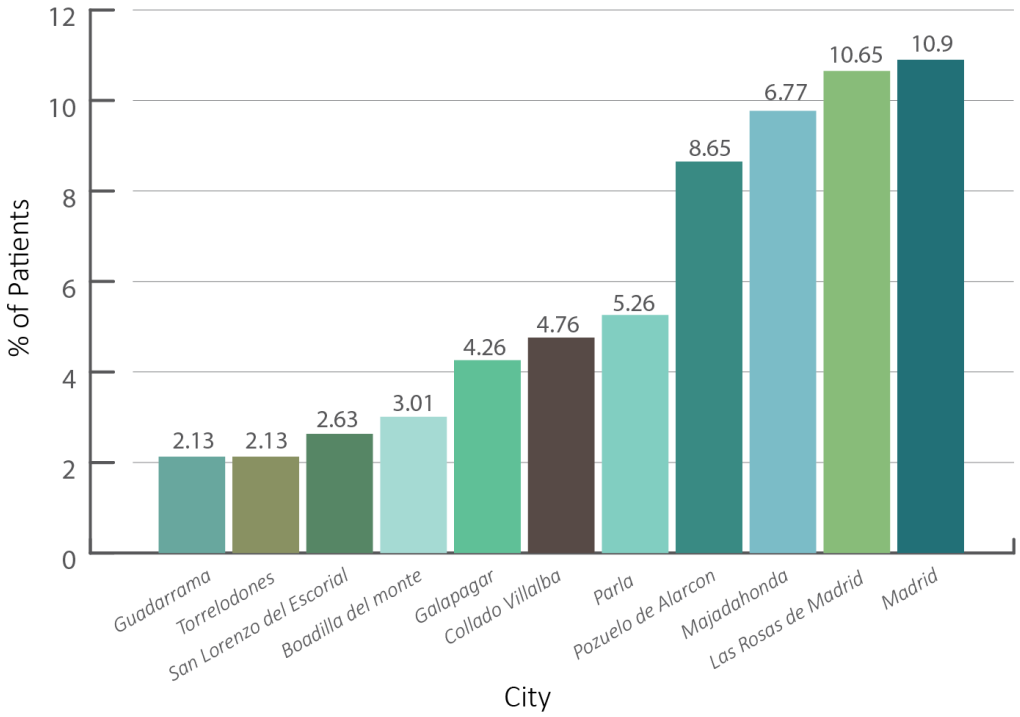

Figure 4.7: Top 10 most frequent cities where patients come from

Data Source: Patients file.

Figure 4.7, shows the top 10 most frequent cities of origin where patients come from. Patients of this data set live in 137 distinct cities of Spain, in the metropolitan area of Madrid. However, for the 17.48% of patients we have missing data about their city.

We can observe that most of the patients come from Madrid. No more useful information can be gained from this graph.

4.3.2.9 Distribution of Patients: by Stages

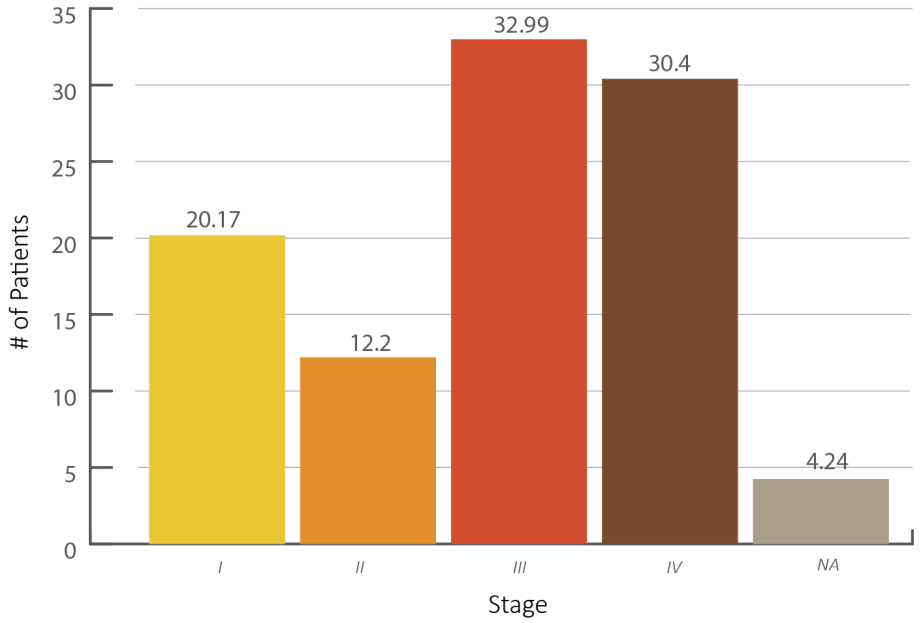

Figure 4.8: Distribution of patients by stage of cancer

Data Source: Patients file.

Figure 4.8, shows the number of patients of the different stages.

The missing data for stage of the tumor accounts for the 4.24% of patients.

The majority of patients stays between stage III and stage IV, which are the stages with the highest risk of mortality. This could suggest a problem in timeliness of lung cancer diagnosis which can be improved by analyzing services and causes of hospitalization months before of the diagnosis date, to detect those patients that already had lung cancer and needed an oncological consultation instead of other services.

4.3.2.10 Distribution of Patients: by Tumor Type

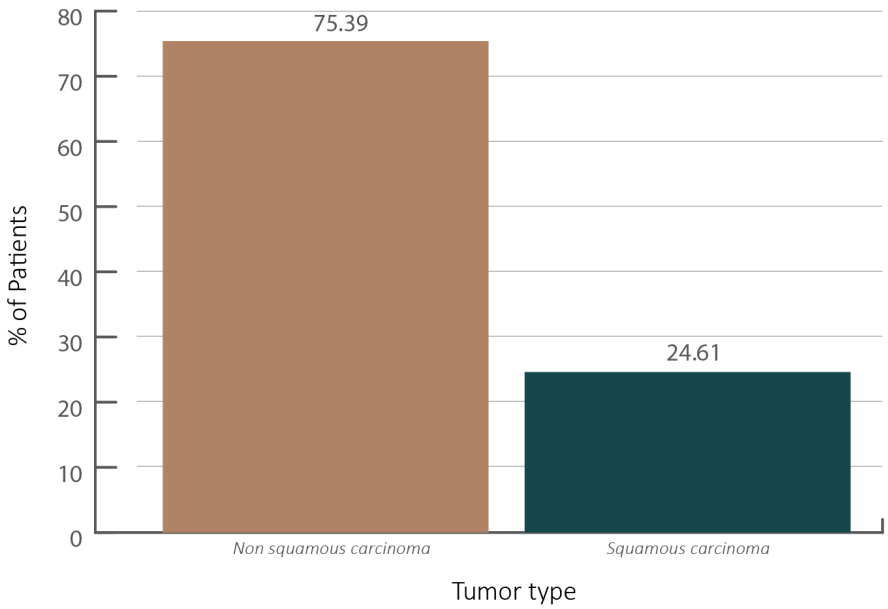

Figure 4.9: Distribution of patients by type of lung cancer

Data Source: Patients file.

Figure 4.9, presents the number of patients having a tumor of type Non Squamous Carcinoma versus patients with Squamous Carcinoma.

The majority of patients is affected, as it can be seen, by lung cancer of type Non-Squamous Carcinoma. No more insights are available at this time.

4.3.2.11 Frequency of Main Types of Sections

Section Types - Frequencies		
Section Type	Count	%
FECHA ALTA	7094	0.86
FECHA EXITUS	61	0.01
FECHA INGRESO	7061	0.86
MOTIVO INGRESO	4929	0.6
SERVICIO	13725	1.67

Table 4.6: Frequencies of Main Types of Sections

Data Source: Sections file.

Table 4.6, shows the frequency of a set of sections type inside texts. The numbers upside columns are the percentage of the frequency with respect to the total number of sections extracted from texts.

In particular, these 5 sections are the main ones that will be used with the aim to extract information from texts: hospitalization date («fecha ingreso»), discharge date («fecha alta»), death date («fecha exitus»), cause of hospitalization («motivo ingreso»), service name («servicio»). These are all keywords who mark the beginning of a particular section. As evidenced by the graph, the number of these sections in which we have interest is little compared to the total number of sections inside texts. However, without performing further investigations we can't say if this will result in missing data at the extraction phase.

4.4 Data Preparation

4.4.1 Cleaning and Formatting Data

Since patients' data comes from a NLP process, it does not need any further data cleaning operations. Thus, the whole stage will be conducted to clean the documents file and the sections file.
In order to enable operations like aggregations, filtering and search on a specific variable a data cleaning stage is needed. All these operations do, at their core, comparisons between strings of texts. For categorical variables (e.g. subcategory of a document), each possible category must have only one possible name representing it, otherwise the results of comparisons will be misleading, even if the meaning of the word is the same. This results in the need of having uniformity in the format of names, texts, and whatever we are interested on performing comparison operations on it (e.g. class «apple» must be referred always with the term «apple», and not with «apples» or «Apple»). Data cleaning is also helpful in making the whole pipeline of data processing replicable for possible new incoming data and it is essential for extracting additional information from text that is still not available.

4.4.1.1 Documents File Cleaning

Input: Documents File.
Objectives:

- Cleaning subcategory field
  - Transforming all letters into their lower-case
  - Special characters removal (e.g. punctuation)
  - Spanish accents substitution (e.g. «à» becomes «a»)
  - Fix words spacing, leaving single space between words
  - Spanish stop words removal, in particular: «de», «la», «las», «del»

- Formatting date field - Chosen final format: YYYY-MM-DD
- Text duplicates removal: to detect pairs of duplicated documents, the following fields are exploited: EHR, subcategory, text, date.

Results:
Subcategory field contains initially 251 distinct strings, with no missing value. After cleaning, the distinct strings are 250.
Date field has been formatted to the following format: YYYY-MM-DD, in order to be coherent with other data sources.
A total of 12011 documents were duplicated, and thus only one copy of them leaved in the database. The final number of documents being used for analysis is 283992. Please notice that full texts of all documents are excluded from any cleaning treatment, as it is not needed to have the whole text processed.

4.4.1.2 Sections File Cleaning

Input: Sections File.
Objectives:

- Cleaning text field
  - Filtering sections by type, selecting only the ones used for our scope: service, hospitalization date, discharge date, death date, cause of hospitalization
  - Transforming all letters into their lower-case
  - Special characters removal excluding characters involved in the dates format (e.g. «
    »)
  - Spanish accents substitution (e.g. «à» becomes «a»)
  - Fix words spacing, leaving single space between words
  - Spanish stop words removal, in particular: «de», «la», «las», «del»

Results:
All texts correctly cleaned and formatted. They are now ready to be used for information

extraction.

4.4.2 Dates Extraction

Input: Sections file.

Output: Enriched Sections file.

The desired output will be a new table having the following fields:

- id - Unique identifier of the document where dates were extracted - format YYYY-MM-DD
- hospitalization - String containing the extracted hospitalization date - format YYYY-MM-DD
- discharge - String containing the extracted discharge date - format YYYY-MM-DD
- exitus - String containing the extracted death date - format YYYY-MM-DD

Notice that discharge date and death date never show together in the same section, as a process can end only in one direction, death or discharge.

This structure will enable the table to be joined to the documents file, using the id field, in order to directly have extracted dates available for use.

Algorithm Design:

For dates extraction, there is a subset of sections of our interest: hospitalization date («fecha ingreso»), discharge date («fecha alta»), death date («fecha exitus»). Hospitalization dates will be helpful as a marker of the beginning of a process while instead, based on the presence of one or the other, death date or discharge date will mark the end of a process. Since we are interested in using only these 3 types of sections, the input data has been filtered to include only those mentioned types of sections. After filtering, a total of 14216 sections are available for the analysis.

The common format for these sections is in most of the cases as follows. The sentence begins with the marker and it is followed by the date. An example for a generic hospitalization date is given. Notice that the date format needs to be uniformed to dates inside the documents file.

fecha ingreso DD/MM/YYYY

In some cases, it could happen that the date is not present and thus a blank space is

found, or the date is semantically wrong, i.e. it is a date that has no sense for the context in which it appears. For all the other cases, all words are separated by one white space and thus it is straightforward to get the date, thanks to the cleaning process happened before.

Algorithm:

A description of the algorithm to extract dates from documents is described in the pseudocode 1. For the sake of brevity, only the case of hospitalization date extraction is reported.

Algorithm 1 Extraction of hospitalization date from section text, here represented by the variable hosp_date

```
hosp_date =  foreach section in filtered_sections do
    document_id = section.document_id tokens = tokenizer(section.text)  for token in
    tokens do
        if token == "ingreso" then
            hosp_date = tokens.next()  break
        end
    end
    if hosp_date is not null then
        addHospDateToReport(document_id, hospitalization_date)  break
    end
end
```

After dates are being extracted, their format is changed to YYYY-MM-DD, in order to make them comparable to other dates formats inside documents file and patients file.

Evaluation:

Now that dates are extracted, it is now time to evaluate their quality. Two metrics will be computed:

1. Number of typos in dates
2. Number of missing dates

Number of Typos in Dates - Metric Definition

Typos in dates are errors made at writing time, due to human error. These errors can include: hospitalization date reported to be later of the discharge date, dates for which it is impossible to have a record (i.e. dates before year 2008) or date ranges whose duration is longer than a certain threshold.

A particular case, that is not counted in this metric, is when creation date of the report is

further of the discharge date reported, or creation date is before hospitalization date. In such cases, the hospitalization and discharge dates are fixed accordingly, e.g. if creation date is before hospitalization date then creation date becomes hospitalization date. This will help to avoid situations in which reports define a process but the system will consider it as a report that does not belong to that process.

For a basic detection of meaningless dates, it is enough to perform basic comparisons with a reference date. In fact, meaningful dates will be past of year 2008, since it is the first year of adoption of the information system of the hospital. Documents having typos in dates will be marked with a boolean flag in a new field. The number of typos in dates will be then, the number of flags marked as TRUE in this new field.

Number of Typos in Dates - Result:

The number of documents containing typos in dates are 23, that is the 0.8% of the total. Since they're only a few, they will be manually checked and eventually corrected, because they could still be helpful in reaching high coverage of documents having a process in the processes extraction phase.

Number of Missing Dates - Metric Definition

On the other hand, dates can be missing, and in such cases the related document will be marked according to which date is missing:

- OK - means that both hospitalization and discharge (or death) dates are available
- MALL (missing all) - means that no date is available from text
- MI (missing hospitalization) - hospitalization date is missing
- MD (missing discharge) - discharge date and death date are missing

Number of Missing Dates - Result:

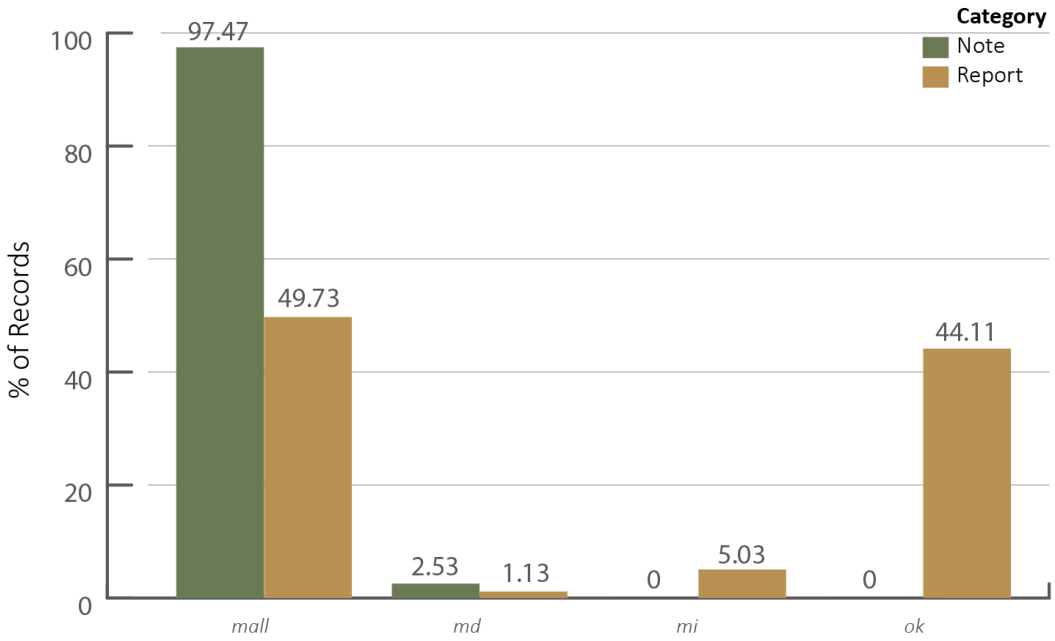

Figure 4.10: Distribution of dates completeness

Notes - Dates Assessment		
Category	Count	%
MALL	262628	97.47
MD	6814	2.53
MI	2	0
OK	1	0

Table 4.7: Distribution of dates completeness extracted from Notes

Reports - Dates Assessment		
Category	Count	%
MALL	7234	49.73
OK	6417	44.11
MI	732	5.03
MD	164	1.13

Table 4.8: Distribution of dates completeness extracted from Reports

Tables 4.8, 4.7 give detailed counts of dates types for Reports and Notes. Figure 4.10, shows the distribution of documents according to their grade of dates missing, as defined

in the metric definition. To obtain the graph, a join between documents file and enriched sections, on the id field has been performed.
As seen in the graph, almost the whole number of notes have no dates inside texts, while half of the reports have them. This suggests that even 23 more reports (i.e. the ones affected by typos) can improve the quality of processes.
Another possibility to increase the number of processes extracted, despite having a low number of dates, is to match creation dates of notes regarding the death of the patient, and thus the end of a process, with their corresponding i.e. nearest occurrence of notes regarding the hospitalization of the patient (emergency note or cause of hospitalization). The assumption here is that every note written in dates between hospitalization notes and death notes are belonging to the same process.

4.4.3 Extracting Services and Causes of Hospitalization

Input: Sections file.
Output: Services and Causes Table.
The desired output is a single table having for each document id, its corresponding service who produced the document and cause of hospitalization of the patient. Thus, the table will have the following schema:

- id - Unique identifier of document
- service - Extracted service name
- cause - Extracted cause of hospitalization

Even this table can be joined with the documents file using the id field. This will be helpful when analyzing services and causes of hospitalization related to the patient's profile.

Algorithm Design:
The extraction of services and causes of hospitalization will be splitted into two separate procedures.
Services names appear, when available, in the service section of a report. Thus, it is straightforward to extract them, since it is only a matter of tokens scan, as already seen

in dates extraction (section 4.4.2).

On the other hand, causes of hospitalization requires more attention to details, even if it is based on the same data source. In fact, what we want to extract is a set of 1 to 3 words at most, resuming the cause of hospitalization, that is in general the name of a disease or a symptom. However, those sections often present more than 3 words, giving useless details for our scope. Thus, the problem becomes recognizing names of diseases from a bag of words.

For this purpose, UMLS can be exploited. UMLS table, as seen in section 3.2, have a field specifying the category of the concept. By selecting only concepts related to diseases or symptoms, we will be able to extract only terms of interest from section's text, with precision.

Some extra cases must be considered while extracting causes. These cases are: i) More than one cause, but with different meaning, is extracted from text ii) More than one cause, with same meaning, is extracted from text.

For the first case, the chosen policy is to take the most frequent one in the set, and, in case all words appear with the same frequency, take the one which appears first in the text. For the second case, having the same disease, the chosen policy is to take the most detailed among the options (e.g. between «pain» and «abdominal pain», the second one will be selected).

Algorithm:

A brief overview of the processing to obtain causes of hospitalization is given.

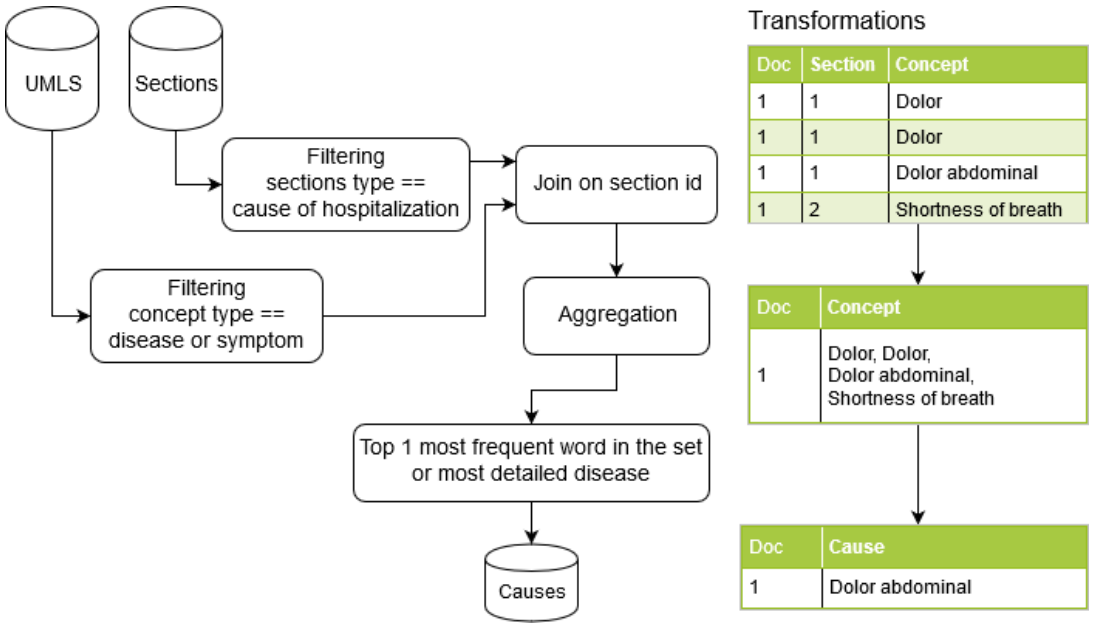

Figure 4.11: Extraction of causes of hospitalization. On the right there is an example of how data is manipulated

After UMLS table has been filtered by selecting only concepts of type «Disease or Symptom» and which are not negated (field negated equals to zero), a join is performed with the sections file on the section id field. This results in a table having for each document, all concepts regarding diseases or symptoms detected in sentences texts. Then, a simple aggregation is performed to have for each document a string containing all the detected concepts. Finally, a scan is performed to the whole table to select, for each document, a single cause of hospitalization, using the already discussed policies.

A general description of the algorithm for services extraction is given in the pseudocode 2. The algorithm works using a filtered version of the sections file which selects only sections regarding the service which produced the document.

Algorithm 2 Extraction of service name from section text

```
service_name =  foreach section in filtered_sections do
    document_id = section.document_id tokens = tokenizer(section.text)  for token in
    tokens do
        if token == "servicio" then
            service_name = tokens.next()  break
        end
    end
    if service_name is not null then
        addServiceNameToDocument(document_id, service_name)  break
    end
end
```

Evaluation:

From a total of 14864 sections regarding causes of hospitalization, a final set of 8780 causes of hospitalization were assigned to their respective documents. On the other hand, a total of 11944 services names were extracted from a total of 13725 sections regarding services. The resulting table obtained by joining causes of hospitalization and services name for each document have 4898 records. This means that only 4898 documents have both cause of hospitalization and service name. However, this is still a good sample to understand the main causes and services used before the diagnosis date of cancer of patients.

4.4.4 Classifying Documents

Input: Documents file.

Output: Enriched Documents file.

The desired output is a final table, which includes all the previous extracted information of dates, as discussed in section 4.4.2, and an additional field describing the class to which the document belongs to. The final structure will be as follows:

- **id** - Unique identifier of the document
- **EHR** - Identifier of the patient
- **cat** - Category of the document (Report or Note)
- **subc** - Subcategory of the document

- **text** - Original text contents
- **creation** - Creation date of the document
- **hospitalization** - Extracted hospitalization date of the patient
- **discharge** - Extracted discharge date of the patient
- **exitus** - Extracted death date of the patient
- **note_type** - Class of the document

Algorithm Design:
Documents classification consists on giving to a document d a class c from the following set of classes:

- **h** - Document regarding hospitalization
- **u** - Document regarding emergency
- **c** - Document regarding consultation
- **d** - Document regarding day hospital
- **nc** - Non-classified document

This classification will be helpful to understand the composition of a group of documents when extracting processes, and thus, classifying a process by its content. For example, a process containing for the majority documents of type consultation, it will result as a consultation process.
Since there is no labelled data for this task, we cannot apply any supervised algorithm in this classification task. However, there is a more efficient way to classify documents that does not need any labelled data. In fact, each documents comes, as seen in section 3.1.1, with a subcategory. This subcategory classifies the document according to its purpose. Hospitalization documents often have a subcategory containing words or abbreviations starting with «hos», while emergency notes often start with «urg» and day hospital notes are indicated with «hdia». Finally, consultation notes often have their subcategory starting with either «cex», «itc», «cons». For any other note we will simply label it with a non-classified label.

Algorithm:

The algorithm works by scanning all the documents and checking the presence of certain letters inside words of their subcategory e.g. documents having a subcategory containing the string «urg» are classified as documents created at the emergency.

Evaluation:

These classification rules led to a coverage of 94% of notes classified, with the remaining 6% without a class. The distribution of documents by note type is given.

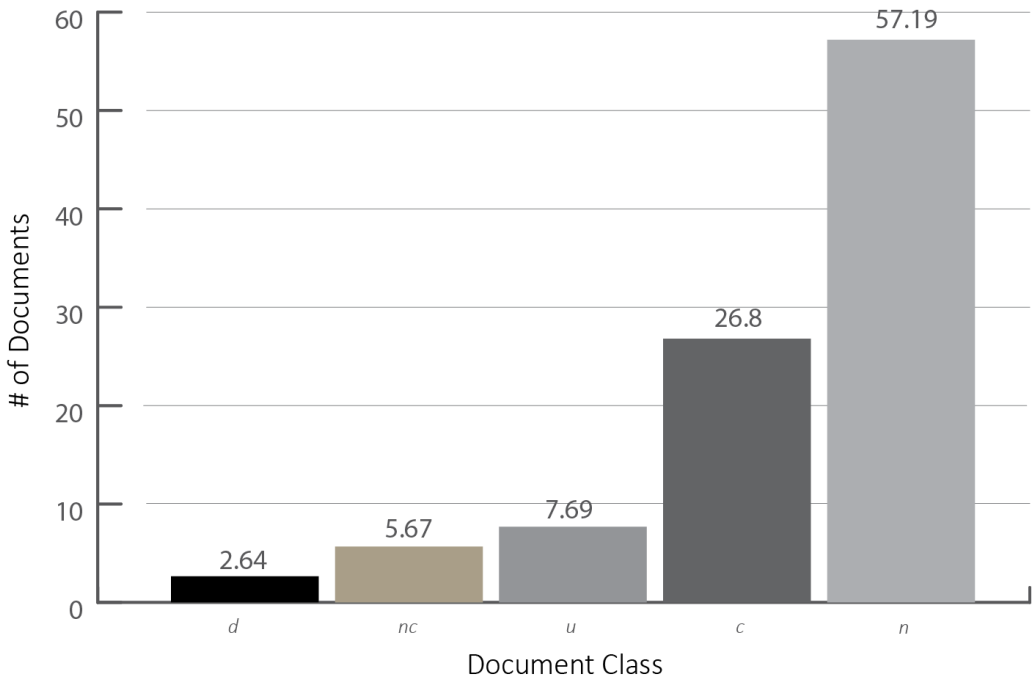

Figure 4.12: Distribution of documents by note type

Figure 4.12 shows that the majority of documents are of type hospitalization.

4.4.5 Processes Extraction

Until now, we have been enriching our available information on documents in order to make it easier to extract processes. Extracting processes from documents in fact, represents a key-objective of this project, and their quality will be crucial for the reliability of further analysis using them. However, as we will see later, quality of these processes can't be evaluated a priori or in an automated way.
Different strategies can be outlined for this task, but they all rely on dates, which are the markers of the begin and the end of a process.
These strategies are:

- Clustering Techniques
- Processes Extraction using Date Ranges

4.4.5.1 Clustering Techniques

An unsupervised learning approach has been evaluated. The idea is to use the creation date of the document as a feature, to create clusters from data. However, applying clustering techniques for this task is an unnecessary overhead for two main reasons:

1. Clustering must be applied to each patient separately, because each patient can have more than one process. This results in a high computational cost, as well as the efforts to fine-tune all the parameters of the chosen algorithm.
2. The number of clusters (i.e. processes) is not known a priori, and thus the chosen algorithm must not require any parameter asking for this information.

One could observe that there are families of algorithms with the characteristic of not asking the number of clusters to extract before processing. However, since the chosen algorithm will work only with one feature (the creation date of clinical notes), the quality of the result will be inappropriate. In fact, the distance metric should take into account two opposite cases to create clusters, the first one in which clinical notes are written in consecutive days, and in the second one, there could be processes described by hospitalization and discharge dates inside some text but without notes written consecutively. Thus, clustering is not applicable, and even if it could be, it is an avoidable

computational cost.

4.4.5.2 Processes Extraction using Date Ranges

Input: Enriched Documents File. Obtained in section 4.4.4

Output: Processes Table.

The desired output is a table describing processes, having the following structure:

- EHR - Unique identifier of the patient
- hospitalization - Start date of the process
- discharge - End date of the process
- pid - Process ID, i.e. unique identifier of the process

Algorithm Design:

Hospitalization and discharge dates form a date range that can be used to identify processes. The main idea is that a process is a date range, and so, all documents whose creation date is in between any date range will be considered as part of that process. Thus, the number of processes extracted is directly dependent on the number of date ranges available in the documents file.

Since a date range requires both hospitalization and discharge date to be associated as a process, it means that the number of date ranges extracted equals the number of processes extracted only using existing dates. The number of complete date ranges and thus processes is equal to the number of dates classified as «ok», as seen in section 4.4.2.

An enhancement of date ranges has been obtained by associating death reports to first occurrence of its corresponding hospitalization note. The idea is to match creation dates of notes regarding the death of the patient, and thus the end of a process, with their nearest occurrence of notes regarding the hospitalization of the patient, i.e. the beginning of the process. These kind of notes were discussed in section 3.1.1.

Furthermore, date ranges can overlap each other, thus some notes can belong to more than one process. Overlapping processes may refer, in most of the cases, to the same main process. Main source of this situation is that the patient changes the service where is hospitalized and so a new report is open, but if this happens the first date range should

end at the beginning of the second date range. For situations in which the patient goes, for example, from emergencies to hospitalization, the wanted behaviour is to create two processes, having the second process starting right at the end of the first process. However, it could be interesting to analyze only the main processes of each patient. Thus, there is the need to solve these conflicts, in order to enable these analysis. Three possible approaches have been evaluated:

- Merging dates - joining two date ranges as one, taking as beginning date the minimum start date between the two, and as ending date the maximum end date between the two.
- Intersecting dates - Using the maximum start date between the two as beginning date and the minimum end date between the two.
- Flag intersecting processes - Create groups of related processes to further analyze them if needed, by marking them with a Main PID.

Intersecting dates does not seem to be an applicable idea, in fact, this approach could lead to clinical notes left without process because they were excluded from their original process, and thus they are not joinable to any other process. Merging dates, instead, gives no information on two separate processes, but only for the main process. Thus, the chosen approach to solve conflicts is to flag overlapping date ranges.

Algorithm:

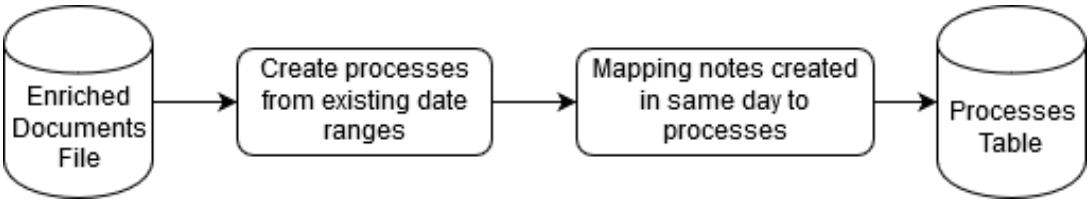

Figure 4.13: Processes Extraction Overview

Figure 4.13, shows the main steps of the algorithm used to extract processes.
The algorithm consists of two steps:

1. Creating processes using existing date ranges - Documents are associated to processes

according to date ranges extracted from texts.

2. Mapping same day documents to new processes - Sets of documents of the same patient created in the same day are associated to the same process.

In the following figures, we will follow an example to understand the basic steps.
A generic enriched documents file is presented in figure 4.14.

Documents File

EHR	document	creation	ingress	discharge
34	27	07/07/2016		
34	19	08/10/2018		
34	84	08/07/2016		
34	54	09/08/2013		09/08/2013
34	36	01/03/2017		
34	58	15/10/2018	08/10/2018	15/10/2018
34	71	01/03/2017		
34	5	09/07/2016		
34	3	13/10/2018		
34	78	05/08/2013	05/08/2013	

Figure 4.14: Processes Extraction - Initial Step

The algorithm starts by collecting all date ranges where both hospitalization date and discharge date are available, thus the ones classified as «ok», as explained in section 4.4.2. In this case, only the document 58 has both hospitalization and discharge date, and document 78 and 54 can be considered as a process. Thus all documents whose creation date is inside those processes are flagged as grouped (rows in yellow, figure 4.15). All retrieved date ranges are saved into the processes table, as shown in figure 4.13.

Documents File

EHR	document	creation	ingress	discharge
34	27	07/07/2016		
34	19	08/10/2018		
34	84	08/07/2016		
34	54	09/08/2013		09/08/2013
34	36	01/03/2017		
34	58	15/10/2018	08/10/2018	15/10/2018
34	71	01/03/2017		
34	5	09/07/2016		
34	3	13/10/2018		
34	78	05/08/2013	05/08/2013	

Processes Table

EHR	pid	ingress	discharge
34	1	08/10/2018	15/10/2018
34	2	05/08/2013	09/08/2013

Figure 4.15: Processes Extraction - Step 1

In the second step, a scan is performed to identify all documents created in the same day for the same patient. Figure 4.16 shows the new state of the processes table.

Documents File

EHR	document	creation	ingress	discharge
34	27	07/07/2016		
34	19	08/10/2018		
34	84	08/07/2016		
34	54	09/08/2013		09/08/2013
34	36	01/03/2017		
34	58	15/10/2018	08/10/2018	15/10/2018
34	71	01/03/2017		
34	5	09/07/2016		
34	3	13/10/2018		
34	78	05/08/2013	05/08/2013	

Processes Table

EHR	pid	ingress	discharge
34	1	08/10/2018	15/10/2018
34	2	05/08/2013	09/08/2013
34	3	01/03/2017	01/03/2017

Figure 4.16: Processes Extraction - Step 2

There will be still documents without any process assigned.

For these documents it has been decided to create from them processes. Each document is then mapped to a new process.
This has been done in order to not exclude them from further analysis.

Evaluation:
The evaluation method to determine how effective the algorithm was on data is to use coverage.
Coverage is defined as the ratio between the number of documents belonging to a process over the total number of documents. No precise metric has been defined to evaluate instead, the quality of processes extracted, i.e. the accuracy to which each document belongs to the right process. This is not a feasible metric, since there is no labelled data to perform checks. However, we will define some approximated quality metric with the help of clinicians.

Processes Extraction - Coverage at each step	
Step	Coverage
Step 1	65.4%
Step 2	89.5%

Table 4.9: Processes Extraction Performance

This first approach has brought interesting results, achieving a final 89.5% of coverage. However, some observations must be done:

- Existing date ranges come with noise, due to human errors or incomplete reports, leaving notes or reports outside their corresponding process, and thus worsening the coverage
- Enhancing the number of date ranges by matching death reports to hospitalization notes could increase noise, by joining together notes which instead they do not belong to the same process
- There are some notes created in consecutive days which are not included in any existing date range. However, they could still be part of a process, and thus grouped together.

4.4.6 Improving Processes Extraction

Input: Enriched Documents File. Obtained in section 4.4.4

Output: Processes Table.

More details were already explained in section 4.4.5.2.

Algorithm Design:

An improvement of the previously explained approach in section 4.4.5.2, has been tested. The aim of the new approach is to rely less on existing date ranges, and thus postponing that step later. Consequently, it has been chosen also to not enhance the number of date ranges extracted from text.

This approach focuses more on the assumption that notes created in consecutive dates belong to the same process.

Algorithm:

This attempt is divided in three steps:

1. Mapping notes created in consecutive days to new processes - notes created in consecutive days have an high probability to be correlated
2. Creating processes using existing date ranges
3. Mapping same day notes to new processes

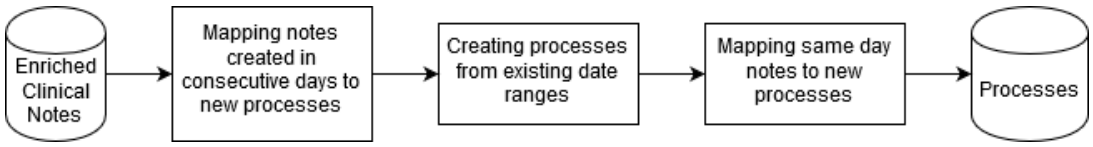

Figure 4.17: Processes Extraction Improvement

An example is again given, using the same table of figure 4.14.

Step one scans the list of documents and finds all documents created in consecutive days (rows in yellow). The new state of the processes table is shown in figure 4.18.

Documents File

EHR	document	creation	ingress	discharge
34	27	07/07/2016		
34	19	08/10/2018		
34	84	08/07/2016		
34	54	09/08/2013		09/08/2013
34	36	01/03/2017		
34	58	15/10/2018	08/10/2018	15/10/2018
34	71	01/03/2017		
34	5	09/07/2016		
34	3	13/10/2018		
34	78	05/08/2013	05/08/2013	

Processes Table

EHR	pid	ingress	discharge
34	1	08/07/2016	09/07/2016

Figure 4.18: Processes Extraction Improvement - Step 1

Then, step two uses existing date ranges inside documents to create processes.

Documents File

EHR	document	creation	ingress	discharge
34	27	07/07/2016		
34	19	08/10/2018		
34	84	08/07/2016		
34	54	09/08/2013		09/08/2013
34	36	01/03/2017		
34	58	15/10/2018	08/10/2018	15/10/2018
34	71	01/03/2017		
34	5	09/07/2016		
34	3	13/10/2018		
34	78	05/08/2013	05/08/2013	

Processes Table

EHR	pid	ingress	discharge
34	1	08/07/2016	09/07/2016
34	2	08/10/2018	15/10/2018

Figure 4.19: Processes Extraction Improvement - Step 2

Finally, each set of notes created in the same day is mapped to a new process.

Documents File

EHR	document	creation	ingress	discharge
34	27	07/07/2016		
34	19	08/10/2018		
34	84	08/07/2016		
34	54	09/08/2013		09/08/2013
34	36	01/03/2017		
34	58	15/10/2018	08/10/2018	15/10/2018
34	71	01/03/2017		
34	5	09/07/2016		
34	3	13/10/2018		
34	78	05/08/2013	05/08/2013	

Processes Table

EHR	pid	ingress	discharge
34	1	08/07/2016	09/07/2016
34	2	08/10/2018	15/10/2018
34	3	01/03/2017	01/03/2017

Figure 4.20: Processes Extraction Improvement - Step 3

Evaluation:

The coverage obtained at each step is shown in table 4.10.

Processes Extraction Improvement - Coverage at each step	
Step	Coverage
Step 1	76.3%
Step 2	78.5%
Step 3	91.8%

Table 4.10: Processes Extraction Improvement

As highlighted by results, using existing date ranges does not have a strong impact on the coverage at step two, while instead using consecutive dates at first step achieved an interesting coverage, which compared to the first step of the other approach is higher. Furthermore, by using this method a higher coverage of 91.5% has been reached, while the remaining 8.2% could be analyzed to be further included.

In this second approach, results reached a higher coverage than before, however to choose between the two methods, a quality metric should be defined together with domain experts. This metric is the average number of days spent in each type of process. This will be explained after the classification phase.

Remaining notes without any assigned PID are eventually assigned to one process, resulting in a one-to-one mapping, giving a virtual coverage of 100%. This has been done both for the first approach and for the second approach of processes extraction. These single note processes are classified as consultation processes.

4.4.7 Classifying Processes

Input: Processes Table. Obtained in section 4.4.5.2

Output: Processes Table.

The desired output is a table describing processes, having the following structure:

- EHR - Unique identifier of the patient
- hospitalization - Start date of the process
- discharge - End date of the process
- pid - Process ID, i.e. unique identifier of the process
- cat - Category of the process

The main objective is to fill the category field.

Algorithm Design:

Classifying processes means to assign at each process one label, giving the user a quick description of what happened during the execution of that process. The following classes of processes has been defined and validated by the health professionals:

- Consultations: Correspond to visits to a hospital to see a doctor; normally they are scheduled visits.
- Day Hospital-Home: A patient goes to a hospital normally for treatment and returns home. In the same date normally, the patient will also have a consultation.
- Day Hospital-Hospitalization. As the previous process, but the patient must be hospitalized after or during treatment.
- Emergency-Hospitalization. This process starts when a patient visits the emergency, and he/she is not discharged and must stay at the hospital.

- Emergency-Home. As in the previous process, but the patient is discharged to home.
- Home-Hospitalization. Normally, it corresponds to processes of scheduled surgeries and a patient comes directly to be hospitalized.
- Hospitalization-Death. While the patient is hospitalized, he/she dies.
- Emergency-Death: While being in emergencies the patient dies.
- Non-Classified: This category groups clinical notes that could not be classified in either of the previous categories.

Thus, the objective is to assign each process to the most representing class. No supervised learning approach could be used at this task due to lack of labelled data. Again, this task has been performed by extracting features from documents inside each process and then applying a set of manually defined if-else rules. To do that we can combine the following characteristics inside processes:

- Majority of documents' category inside PID - Each document has its own category. Thus PID category could be inferred by looking to which category belongs the majority of documents inside a PID.
- Looking for the category of discharge reports - Discharge reports come with a fixed nomenclature inside its title. e.g. «informe alta hospitalizacion» (discharge report), «informe alta urgencias» (discharge from emergency report), «informe alta consulta» (discharge from consultation report). Thus, they could describe the whole process if found inside a PID.

By combining these information, the category could be assigned by looking at the majority of documents category and presence of particular «alta» (discharge) note, with weighted biases.

Algorithm:

A total of 12 features were extracted from processes to be used in the classification task. These have been used to describe the distribution of categories of notes inside PIDs to infer the class of the process. Amongst all, these are the main features used:

- Number of clinical notes in the cluster

- Types of reports contained (e.g. discharge report)
- Number of clinical notes per category of the note (e.g., "five notes are related to emergency notes, one is related to death").
- Presence of hospitalization or emergency process prior to the currently analyzed process.

Evaluation:
By only using the majority of documents' categories inside processes we classified the 93% of processes. Then, by combining also the presence of particular documents inside processes we reached a coverage of 99% of classified processes.

4.4.8 Validation of Results

In this chapter two main methods for extracting processes have been shown. However, in order to build a data pipeline to make the analysis repeatable on new incoming data, a decision on which method to put in production stage must be taken.
Two questions arise at this scope:

- How do we check that clinical notes in the same process are semantically related?
- How do we know if the category assigned to the process is the right one?

To evaluate the consistency of the grouping and categorization process, the idea is to use statistical measures and check manually, together with the domain-specialists if the values obtained are reasonably near to the expected values.
First approach:
The first method of processes extraction reached a coverage of 89% and a total of 96.5% classified processes. The average number of days per category of the process is shown below.

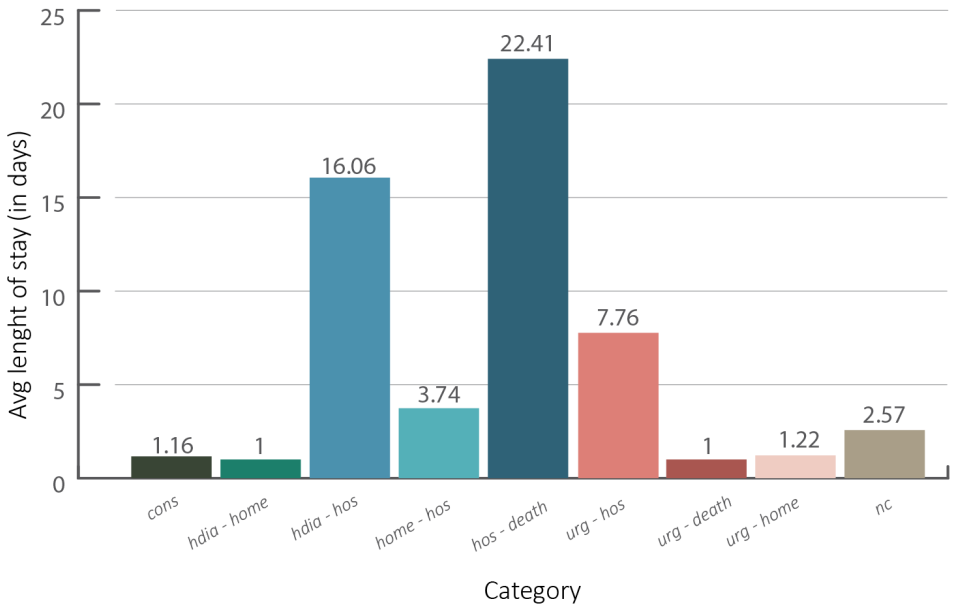

Figure 4.21: Method 1 - Length of stay per type of process

Second approach:

The second method of processes extraction reached a coverage of 92% and a total of 99% classified processes. The average number of days per category of the process is shown below.

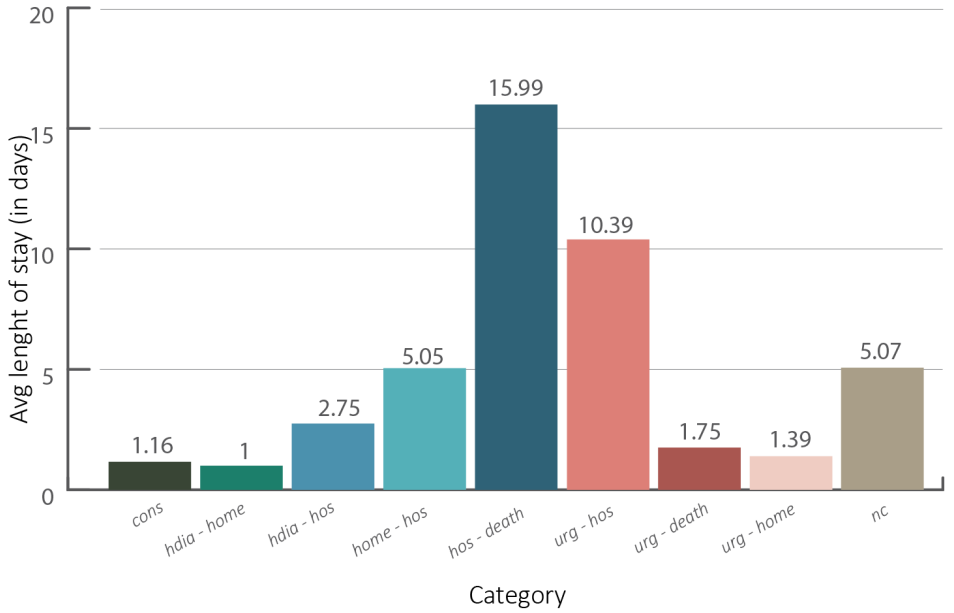

Figure 4.22: Method 2 - Length of stay per type of process

Together with the numerical average, standard deviation, minimum and maximum values,

results show that the second method produces shorter processes, thus they are more reliable than processes extracted using the first method. Thus, the second approach is chosen to be integrated in the final data pipeline.

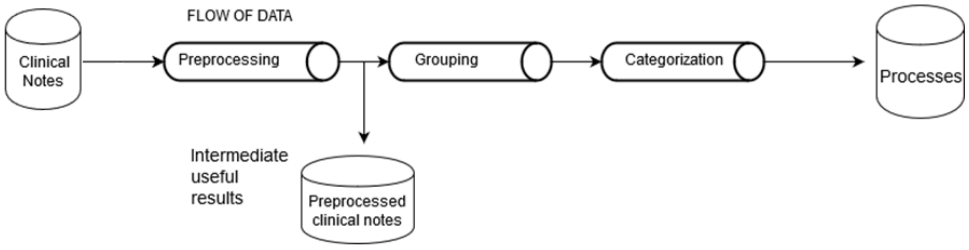

Figure 4.23: Final data pipeline

After data passes through the entire pipeline, it is ready to be analyzed. In the results sections, the outcomes of analysis performed on processes will be presented and discussed.

4.5 Implementation

Before implementing the data pipeline shown at figure 4.23, an evaluation of the needs has been carried out.
The implementation needs to be:

- Repeatable to any data change and to any level of expertise
- Understandable and readable at code level
- Reusable for other experiments

Since it is not a primary objective to be efficient and fast, the chosen programming language is R. Thus, the choice has been oriented towards scripting languages, since there is no need to compile or optimize code. R was born for data wrangling and analysis tasks, and this makes it preferable to other scripting languages. Libraries like Ggplot and Dplyr made the whole work straight to the objectives, while maintaining clear code.
To guarantee reusability and readability, each component of the data pipeline is implemented in a separate R script. Each component is divided in functions. Each function makes a single data manipulation action (e.g. extracting dates) from an input data frame and then outputs a new data frame.

Using this approach, intermediate results can be exported from the outputs of any function of interest with a single line of code. This is helpful for other team members in need for particular data generated by the processing pipeline (e.g. hospitalization and discharge dates of each patient).

5 Discussion of Results

5.1 Exploratory Analysis of Output Data

Once we processed the initial data to extract relevant information from documents to achieve for our objectives, it is time to explore the now available processes table and services and causes of hospitalization table, as mentioned in sections 4.4.5.2 and 4.4.3.

5.1.1 Exploring Processes Table

Table Structure:

A total of 48849 processes were extracted from documents file. They belong to 989 distinct patients.

The final structure of the table have the following fields.

Processes Table - Overview		
Variable name	Type	Subtype
EHR	Numeric	Continuous
ingress	Categorical	Ordinal
discharge	Categorical	Ordinal
pid	Numeric	Continuous
cat	Categorical	Nominal

Table 5.1: Overview of processes table

The detailed description of each field can be found at section 4.4.5.2.

5.1.1.1 Number of Processes per Patient

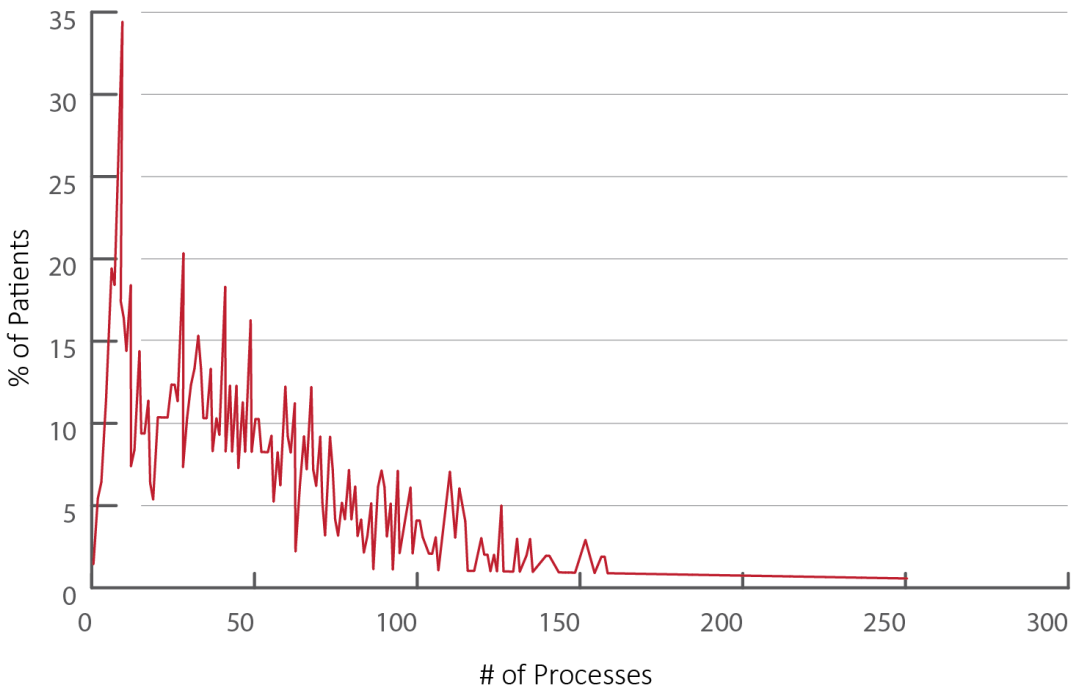

Figure 5.1: Number of processes per patient

Data Source: Processes Table.

Figure 5.1 shows the distribution of number of processes per patient.

On average, each patient has been subjected to 81 processes, with a minimum of 1 process per patient to a maximum of 251 processes per patient.

5.1.1.2 Distribution of Processes: by Category

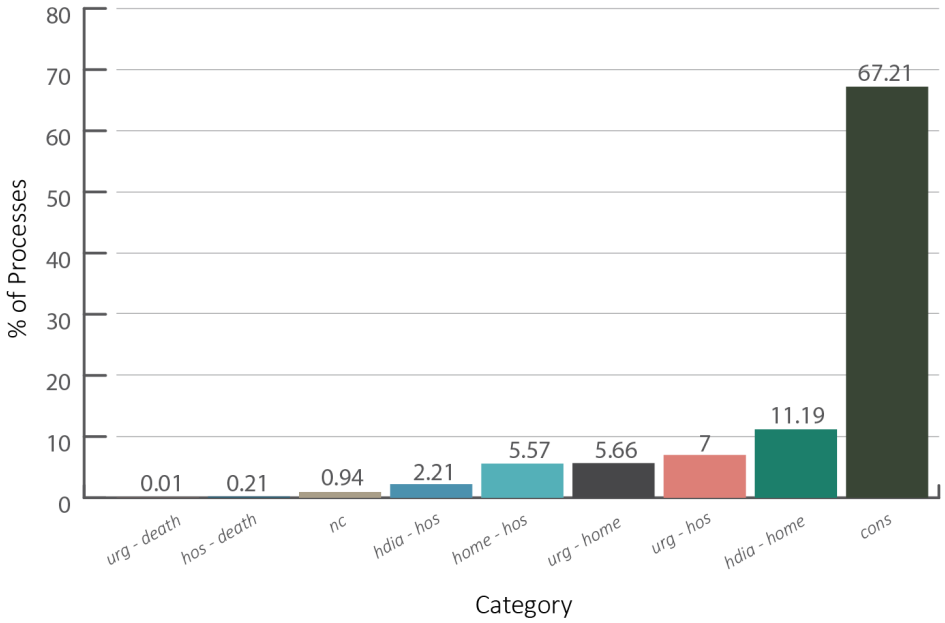

Figure 5.2: Percentage of processes per category

Data Source: Processes Table.

Figure 5.2 shows the distribution of number of processes per category of the process. As it can be seen, the majority of processes are consultation processes, followed by in-day hospitalizations and Emergencies.

5.1.1.3 Distribution of Processes: by Cancer Stage of Patients

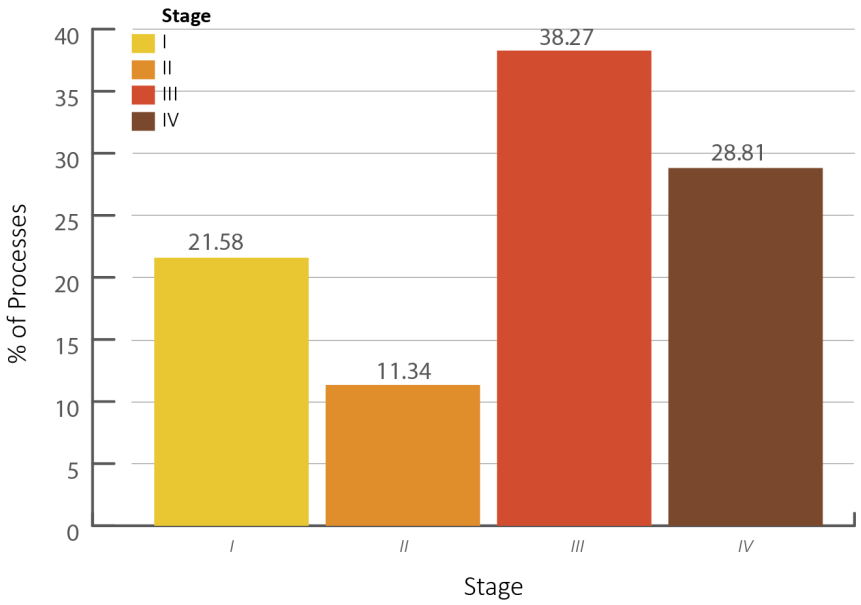

Figure 5.3: Percentage of processes per stage

Data Source: Processes Table.

Figure 5.3 shows the distribution of processes by cancer stage of the patients.

The third stage has the majority of processes, followed by the fourth stage.

5.1.1.4 Distribution of Processes: by Category and Cancer Stage

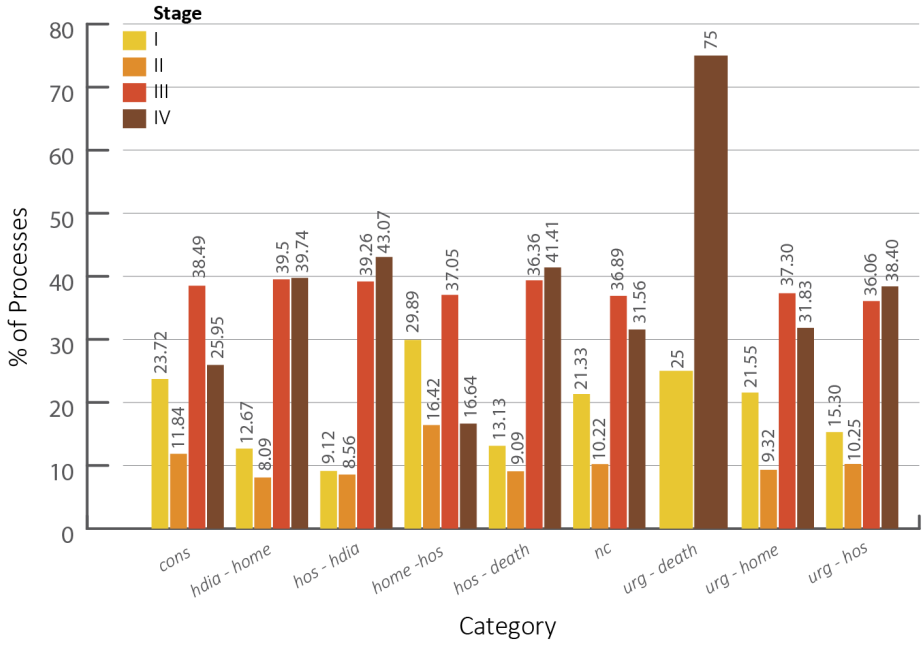

Figure 5.4: Percentage of processes per category and stage

Data Source: Processes Table.

Figure 5.4 shows the distribution of processes by cancer stage of the patients.

The third stage has the majority of processes, followed by the fourth stage.

5.1.2 Exploring Services and Causes of Hospitalization Table

Table Structure:

A total of 8780 causes of hospitalization and 11944 services names were extracted from documents file. By joining them using the document ID we obtain a table having 4898 reports for which both service and cause of hospitalization are available.

The final structure of the table have the following fields.

Services and Causes Table - Overview		
Variable name	Type	Subtype
id	Numeric	Continuous
service	Categorical	Nominal
cause	Categorical	Nominal

Table 5.2: Overview of services and causes table

The detailed description of each field can be found at section 4.4.3.

This table presents 32 distinct services names and 103 distinct causes of hospitalization.

More insights will be given in the discussion of KPI 2 at section 5.3.

5.2 KPI 1 - Length of Hospital Stay

As previously explained, the KPI Length of Hospital Stay aims at reducing the average number of days that patients stay at the hospital. However, the focus here is to show that the whole work enabled the extraction of such insights and data that will help clinicians to reduce length of stay rather than showing how to reduce such length of stay. This is because reducing length of stay is something that depends on the hospital and their interventions.

5.2.1 Length of Stay per Type of Process

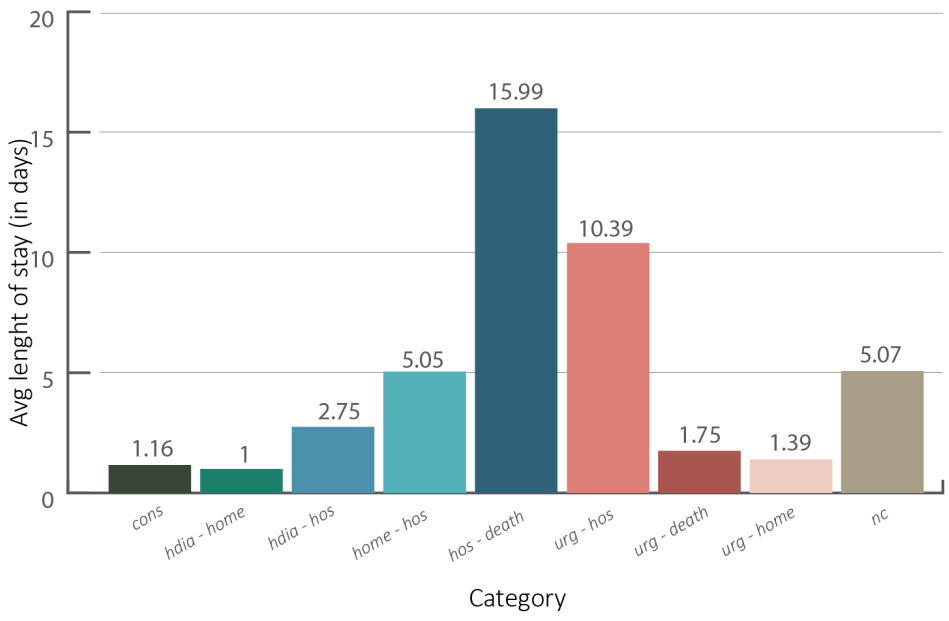

Figure 5.5: Average length of stay per category of the process

Data Source: Processes Table.

Figure 5.5, represents the length of hospital stay for the different categories of processes.

The processes considered for this graph are the ones occurred after diagnosis date of patients' cancer.

It is interesting to notice that patients near to death, here grouped as «hos-death», are the ones to stay longer at the hospital. It could be interesting to see if there is any specific profile for patients which are subjected to hos-death processes. Furthermore, this result shows where clinicians can optimize the length of stay based on the type of process, thus differentiating policies by type of process.

Detailed results are shown in table 5.3, together with their standard deviation.

Standard deviation shows that «hos-death» and «urg-hos» have an highly variable average length, while «hdia» processes, due to the classification rules given by clinicians (as mentioned in section 4.4.7), are lasting exactly one day.

Category	Avg Length (Days)	Std Deviation
cons	1.16	1.63
hdia-home	1.00	0.00
hdia-hos	2.76	4.17
home-hos	4.91	7.37
hos-death	15.05	20.82
urg-death	1.75	0.50
urg-home	1.42	0.87
urg-hos	9.90	9.49
nc	3.61	12.24

Table 5.3: Length of Stay per Type of Process

5.2.2 Length of Stay per Type of Process: Before vs After Diagnosis Date

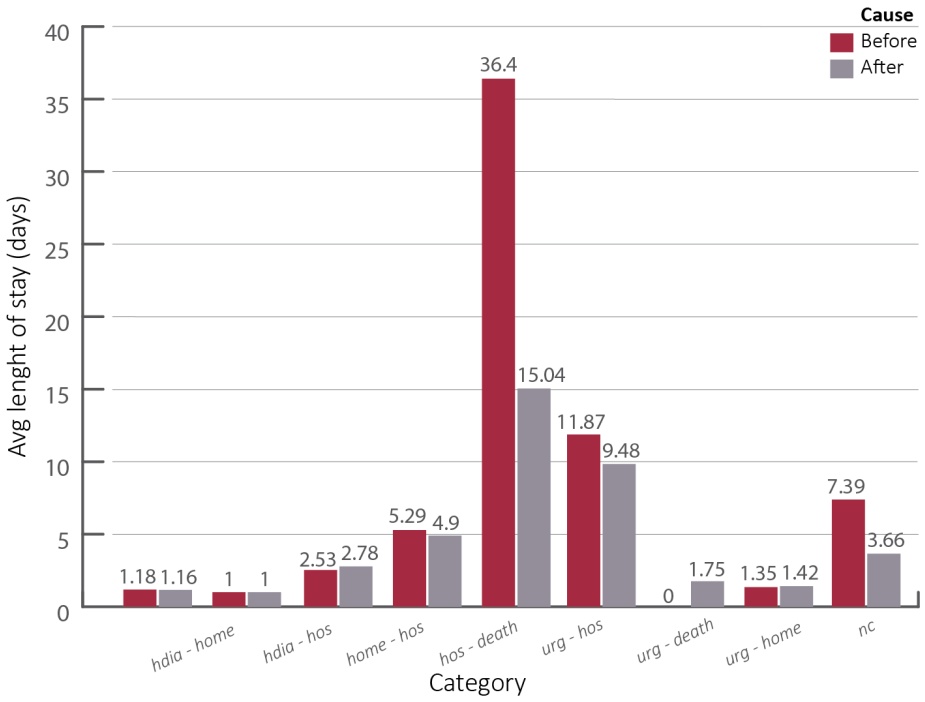

Figure 5.6: Average length of stay per category of the process compared before and after diagnosis date

Data Source: Processes Table joined with Patients file.

Figure 5.5, compares the length of hospital stay for the different categories of processes before and after diagnosis date of patients. To divide between the two categories the hospitalization date and diagnosis date have been compared. Processes are considered to be happened after diagnosis date if the hospitalization date is later than the diagnosis date.

It can be noticed that processes happened before diagnosis date are slightly longer than the ones happened later. «hos-death» processes have a great difference on their duration, however, the number of samples for the «hos-death» processes happened before diagnosis date make this insight not reliable. In fact only 5 processes happened before diagnosis date belong to the «hos-death» category.

For the reader, it can be confusing the fact that processes ending with the death of the patient are happening before diagnosis date, however, even if this is a rare condition (only 5 cases for «hos-death» processes), it could happen that the patient dies when waiting for

the results of an oncological consultation.

To complete the comparison of processes categories before and after diagnosis date, their relevance by number of samples is available in figure 5.7.

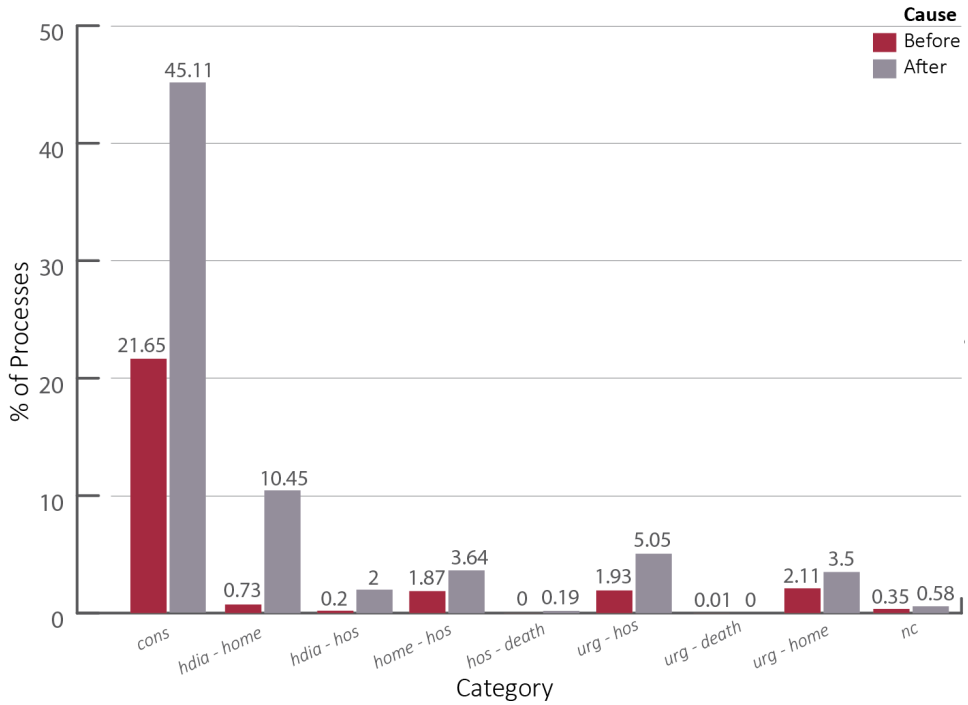

Figure 5.7: Number of processes per category: before vs after diagnosis date

5.2.3 Length of Stay per Stage of Cancer

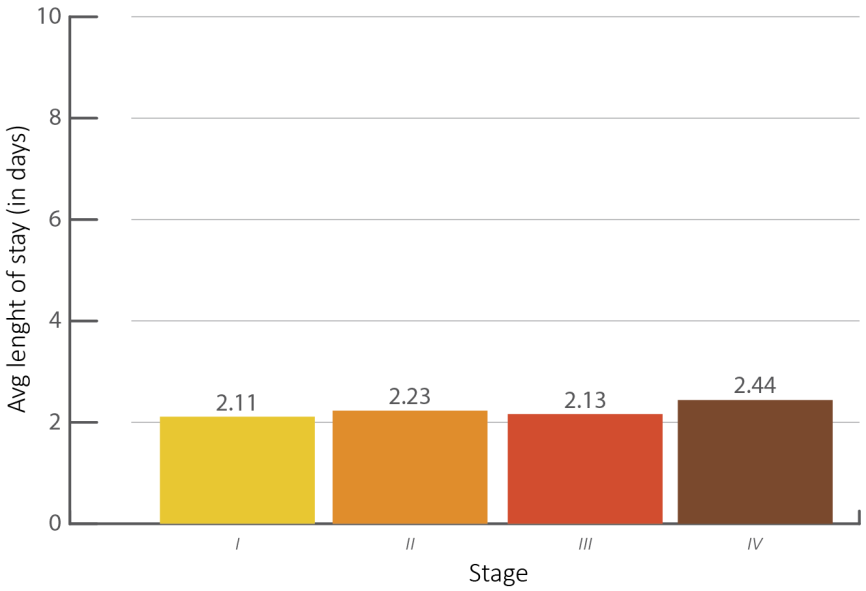

Figure 5.8: Average length of stay per stage of cancer

Data Source: Processes table joined with Patients file.

Figure 5.8, represents the length of hospital stay for the different stages of cancer of patients at diagnosis date. The processes considered for this graph are the ones occurred after diagnosis date of patients' cancer. Patients having no defined stage (classified as «NA») are excluded from this graph.

No interesting insight can be extracted from this view, as it can be seen, length of stay doesn't affect the duration of processes.

More details are given in table 5.4.

Stage	Avg Length (Days)	Std Deviation
I	2.11	5.54
II	2.23	3.96
III	2.13	4.32
IV	2.44	5.33

Table 5.4: Length of Stay per Stage of Cancer

5.2.4 Correlation Between Patients Data and HOS-DEATH Processes

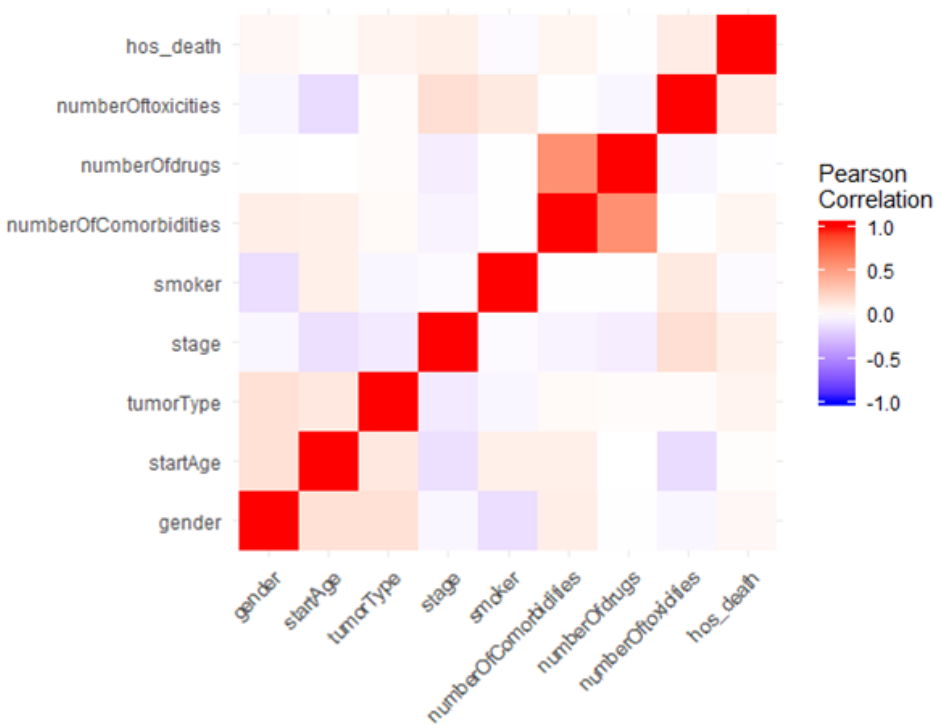

Figure 5.9: Pearson correlation(3) between patients data and hos-death processes

Data Source: Processes table joined with Patients file.

Figure 5.10 shows how a subset of variables of our interest from patients file are correlated with the presence or absence of processes classified as «home-hos» in the clinical history of the patient.

The heatmap shows in red scale the positive correlations and in blue scale the negative ones. Our objective is to see if the variable called «home-hos», indicating the presence or not of at least one process of this kind in the clinical history of the patient, is correlated with other data about patients. Thus, we are looking only to the first row of the heatmap. No relevant correlation is found from this point of view. This could be also due to lack of data, since patients subjected to «hos-death» are only 83 out of a total of 989 patients.

5.2.5 Correlation Between Patients Data and HOME-HOS Processes

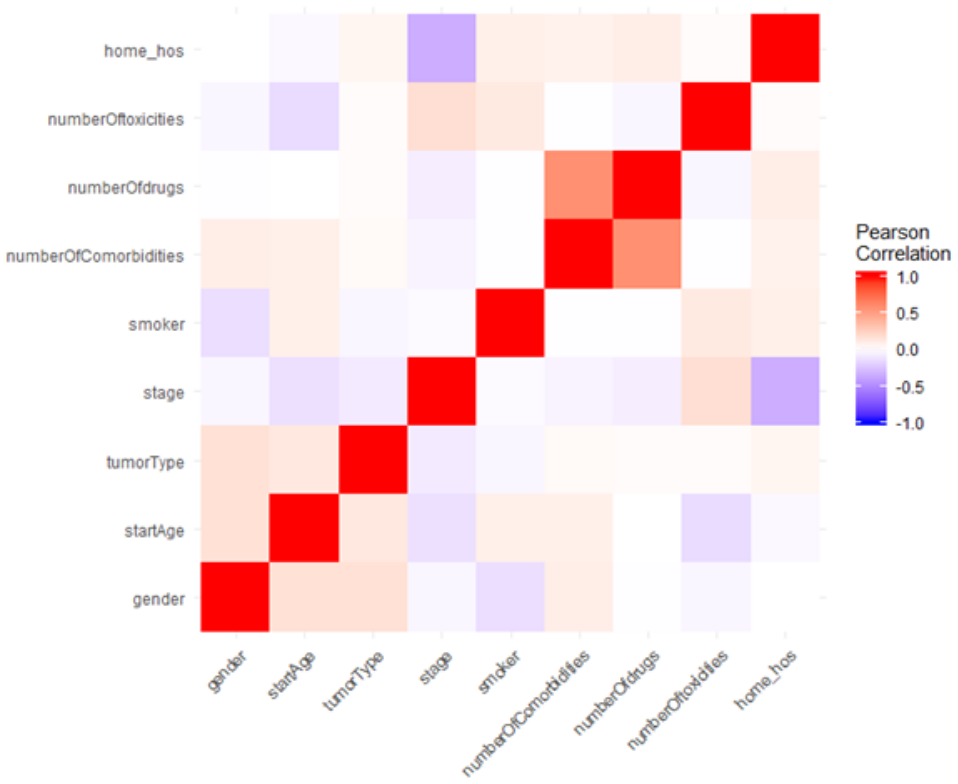

Figure 5.10: Pearson correlation between patients data and home-hos processes

Data Source: Processes table joined with Patients file.

Figure 5.10 shows how a subset of variables of our interest from patients file are correlated with the presence or absence of processes classified as «home-hos» in the clinical history of the patient.

The heatmap shows, again, in red scale the positive correlations and in blue scale the negative ones. Again, the variable «home-hos» is indicating the presence or not of such process in the clinical history of patients.

The interesting insight here is that as the stage of cancer advances, the less likely a patient is subjected to a «home-hos» process.

5.3 KPI 2 - Identification of People at Risk of Developing Lung Cancer

The KPI Identification of people at risk of developing lung cancer aims at recognizing the patients that might have lung cancer but have not been diagnosed yet. Having the diagnosis date available, the goal is to identify if there are evidence before the diagnose date that can lead the physicians to think that a specific patient might already have lung cancer, being thus necessary an oncology consultation as soon as possible. Again, the results here shows only useful insights for clinicians to assess the situation and improve hospital policies rather than showing a way to improve this KPI.

5.3.1 Top 10 Most Common Causes of Hospitalization Before Diagnosis Date

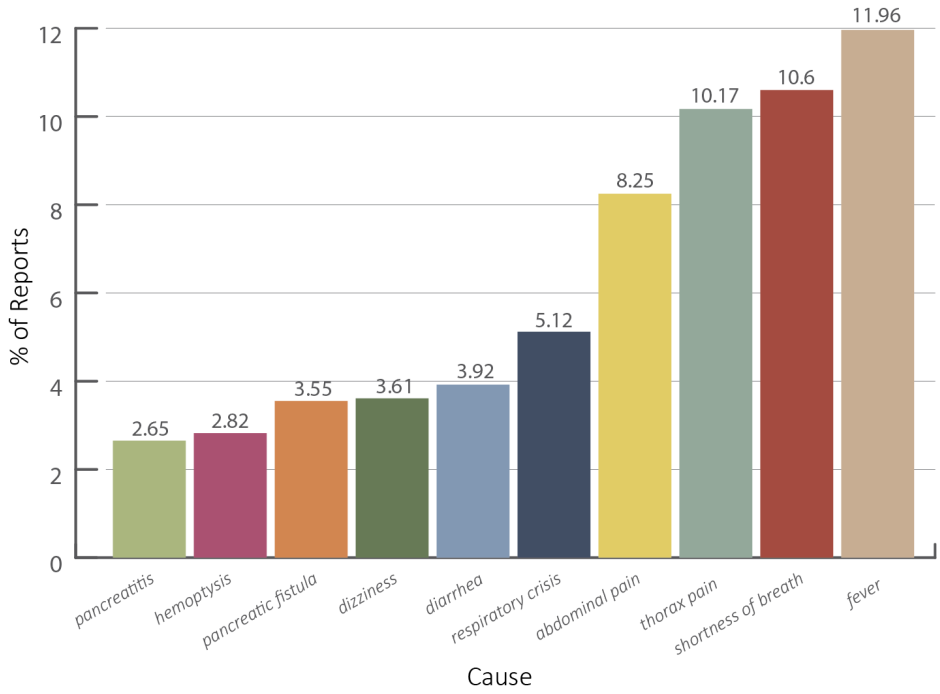

Figure 5.11: Top 10 most common reasons of ingress before diagnosis date

Data Source: Services and Causes of Hospitalization Table.

Figure 5.11, shows the top 10 causes of hospitalization for patients hospitalized in periods

prior to their diagnosis date.

The most common causes are fever (fiebre), shortness of breath (disnea), abdominal pain and thorax pain. These are diseases that often relates with lung cancer patients.

5.3.2 Top 10 Most Used Services Before Diagnosis Date

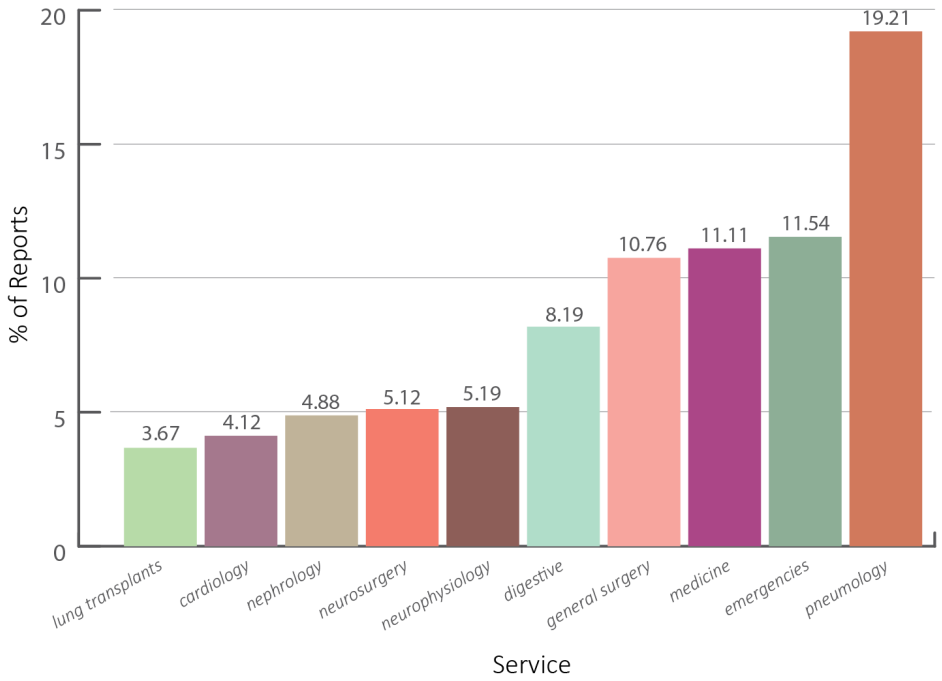

Figure 5.12: Top 10 most used services before diagnosis date

Data Source: Services and Causes of Hospitalization Table.

Figure 5.12, shows the top 10 services where patients go before diagnosis date. This chart can help visualize where to find services to improve in detecting patients at risk faster than before.

As seen, the most used services is Pneumology (neumologia) followed by Emergencies (urgencias generales) and medicine (medicina interna). Also Cardiology (cardiologia) has a relevant position in this chart. In the following results, we will focus the attention on Pneumology, Emergencies and Cardiology, because we think that they are more likely to be related with lung cancer.

5.3.3 Time Passed From Last Visit to Cardiology, Pneumology or Emergencies to Diagnosis Date

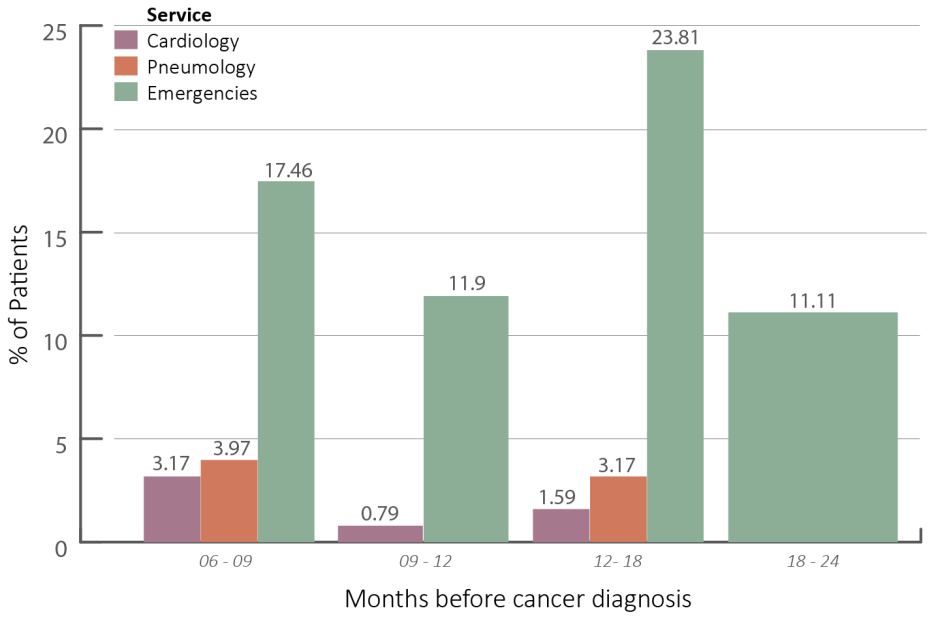

Figure 5.13: Number of months passed between last visit to Cardiology, Pneumology and Emergencies

Data Source: Services and Causes of Hospitalization Table.

Figure 5.13, compares services usage between Cardiology, Pneumology and Emergencies before diagnosis date, considering only the most recent visit of each patient. The time window for this analysis has been limited from a minimum of 6 months before diagnosis date to a maximum of 24 months.

As it can be noticed, both Cardiology and Pneumology have more and more visits as the date of diagnosis is getting closer.

5.3.4 Top 3 Causes of Hospitalization for Emergencies, Pneumology and Cardiology

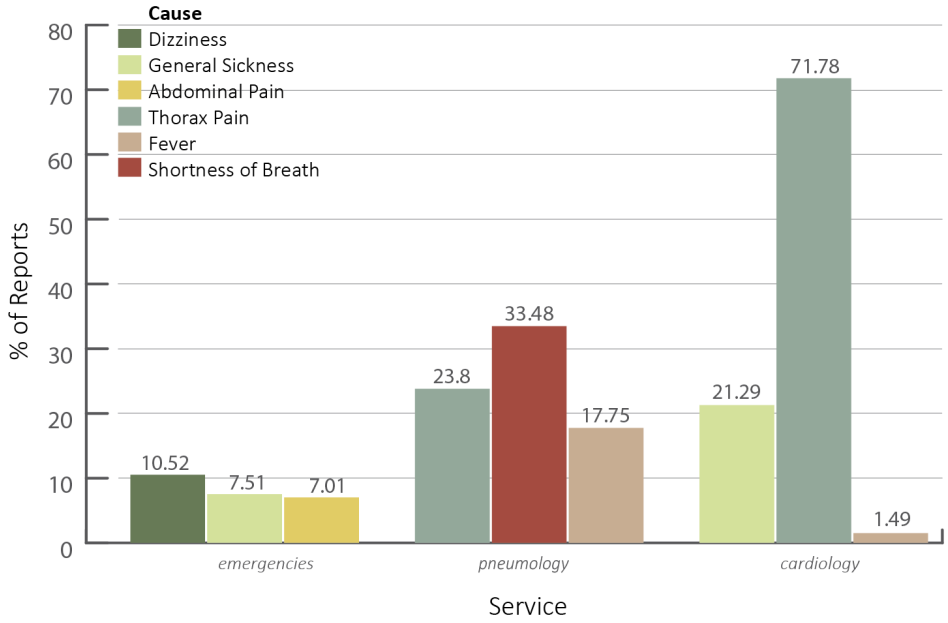

Figure 5.14: Top 3 causes of hospitalization for Emergencies, Pneumology and Cardiology

Data Source: Services and Causes of Hospitalization Table.

Figure 5.14, compares top three causes of hospitalization from Cardiology, Pneumology and emergency services before diagnosis date.

Thorax pain appears to be a frequent cause of hospitalization registered in patients going to Pneumology and the most frequent one for Cardiology. Patients going to Pneumology instead, often present shortness of breath, that is a typical symptom of lung cancer development. For Emergencies instead, it is interesting to notice that abdominal pain is in the top three causes, since it is another symptom of lung cancer development.

5.3.5 Causes of Hospitalization for Cardiology, Pneumology and Emergencies by Period

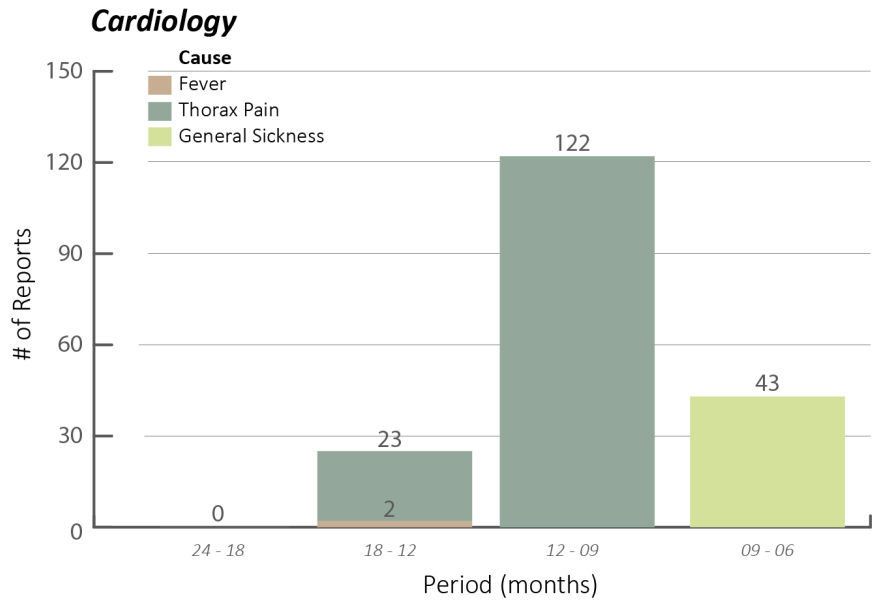

Figure 5.15: Top 3 Causes of Hospitalization for Cardiology by Period

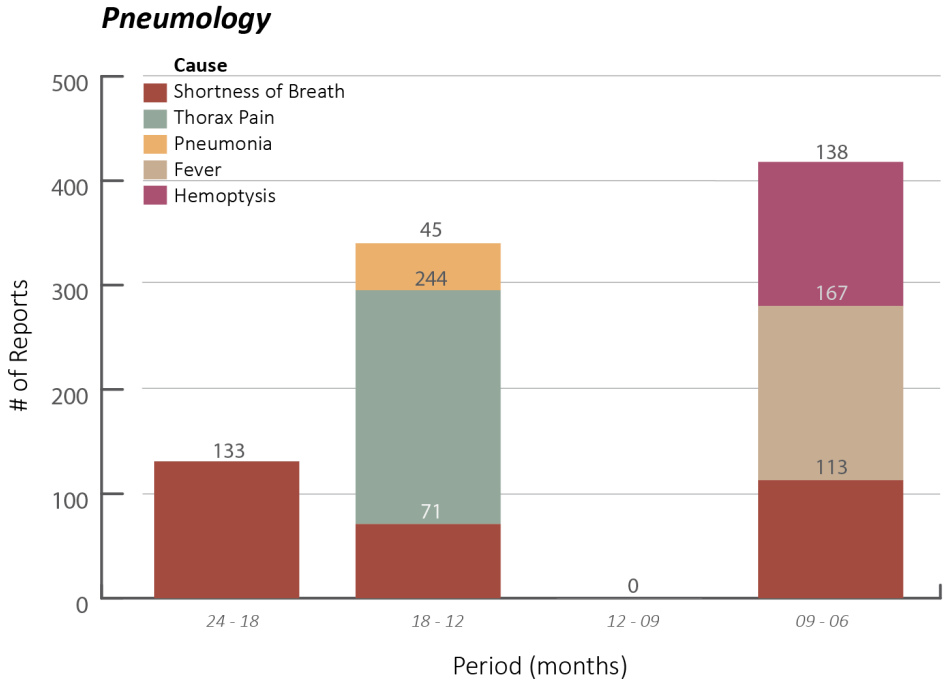

Figure 5.16: Top 3 Causes of Hospitalization for Pneumology by Period

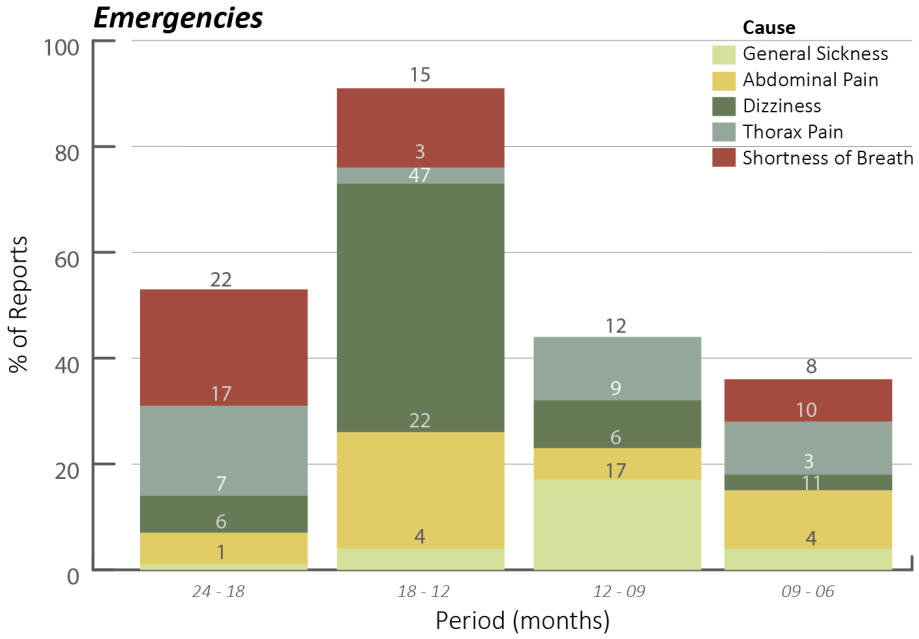

Figure 5.17: Top 3 Causes of Hospitalization for Emergencies by Period

Data Source: Services and Causes of Hospitalization Table.

Figures 5.15, 5.16, 5.17 show the most frequent causes for cardiology, Pneumology and emergencies. To obtain this graph, reports have been divided into 4 periods, which are the number of months prior to diagnosis date: i) 24 to 18 months ii) 18 to 12 months iii) 12 to 9 months iv) 9 to 6 months. To compare the results shown in this graph, it has to be taken into account the total number of reports for each service: i) 941 reports for Pneumology ii) 599 reports for emergencies iii) 202 reports for cardiology.

For cardiology, it has been registered an high number of thorax pain cases from 18 to 9 months prior to diagnosis date. However, the lack of a reasonable number of cases does not help in achieving evidences.

For Pneumology, it is interesting to notice a constant presence of shortness of breath cases, and the presence of hemoptysis only in the last few months.

While for emergencies, it is interesting that abdominal and thorax pain are frequent causes of hospitalization. These patients may have needed an oncological consultation instead of a general consultation.

5.3.6 Services and Causes of Hospitalization by Age of Patients

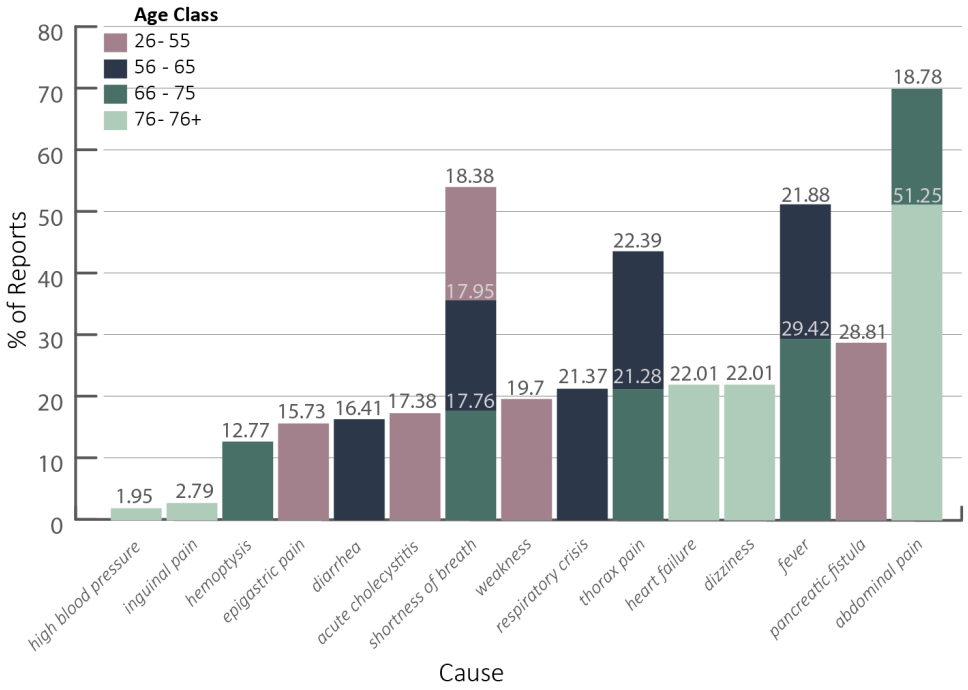

Figure 5.18: Top 5 causes of hospitalization by age of patients

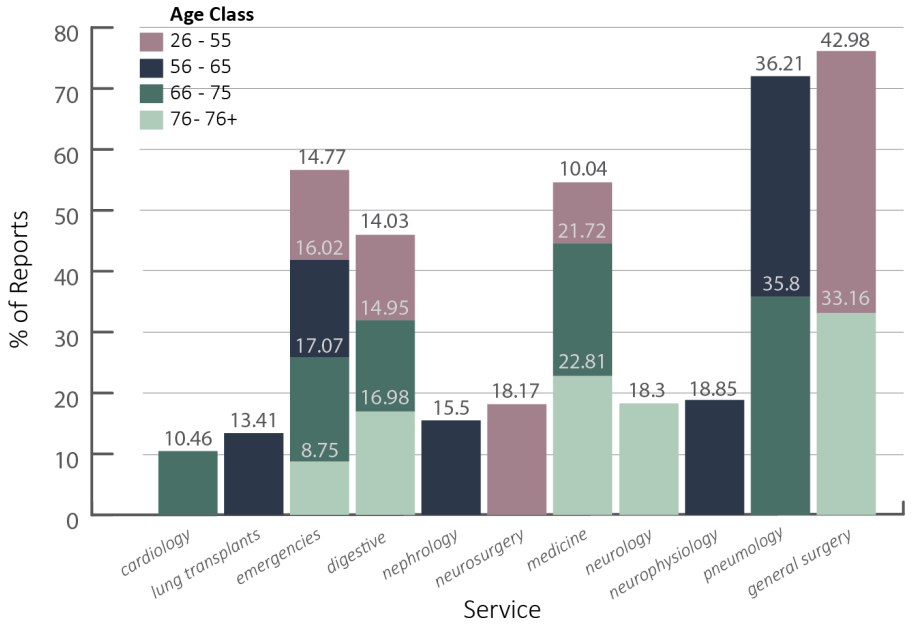

Figure 5.19: Top 5 consulted services by age of patients

Data Source: Services and Causes of Hospitalization Table.

Patients were divided according to their age, in 4 classes: i) 26 to 55 years old ii) 56-65

years old iii) 66-75 years old iv) 76 and more than 76 years old.

Figure 5.18 reports the most common causes of hospitalization by age class. It is interesting to notice that for classes 56-65 and 76-76+ the most common cause is related to lung cancer symptoms. This is also true for the age class 66-75 where shortness of breath, thorax pain and abdominal pain are in the top 5. It is also interesting to notice that for youngest age class, 26-55, there are no relevant causes of hospitalization related to lung cancer symptoms, apart from shortness of breath.

Figure 5.19, instead, shows the most used services by age class.

It is interesting that Pneumology and cardiology, which are related to lung cancer, are being used only by age classes 66-75 and 56-65.

5.3.7 Services and Causes of Hospitalization by Gender

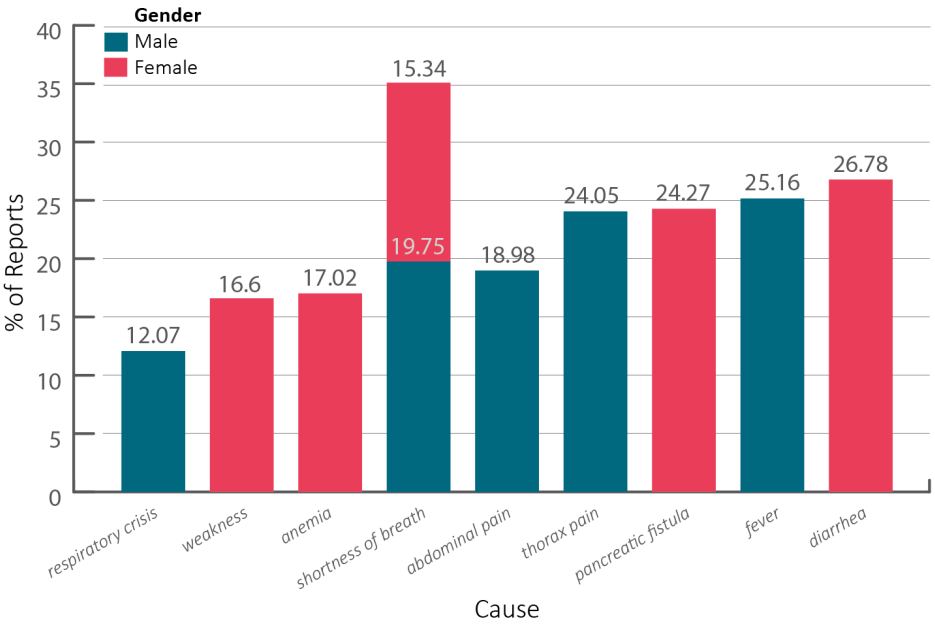

Figure 5.20: Top 5 causes of hospitalization by gender

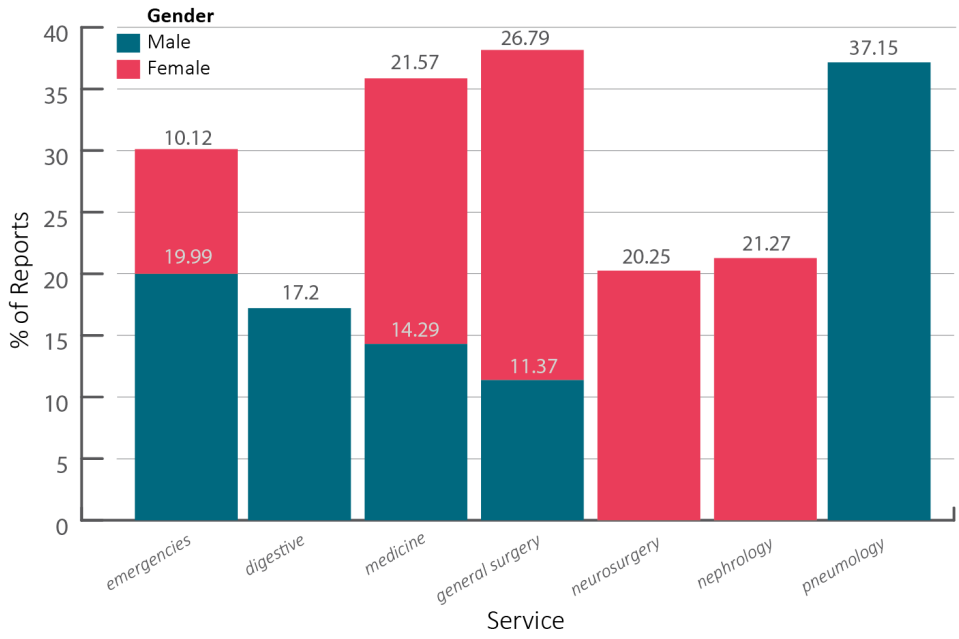

Figure 5.21: Top 5 consulted services by gender

Data Source: Services and Causes of Hospitalization Table.

Figure 5.20, shows the most common causes of hospitalization by gender of patients.

It can be noticed that in men, causes of hospitalization are more related to lung cancer symptoms than in women.

Figure 5.21, shows the most used services by gender of patients.

It is interesting to notice that only men patients used Pneumology in this population.

5.3.8 Services and Causes of Hospitalization by Stage

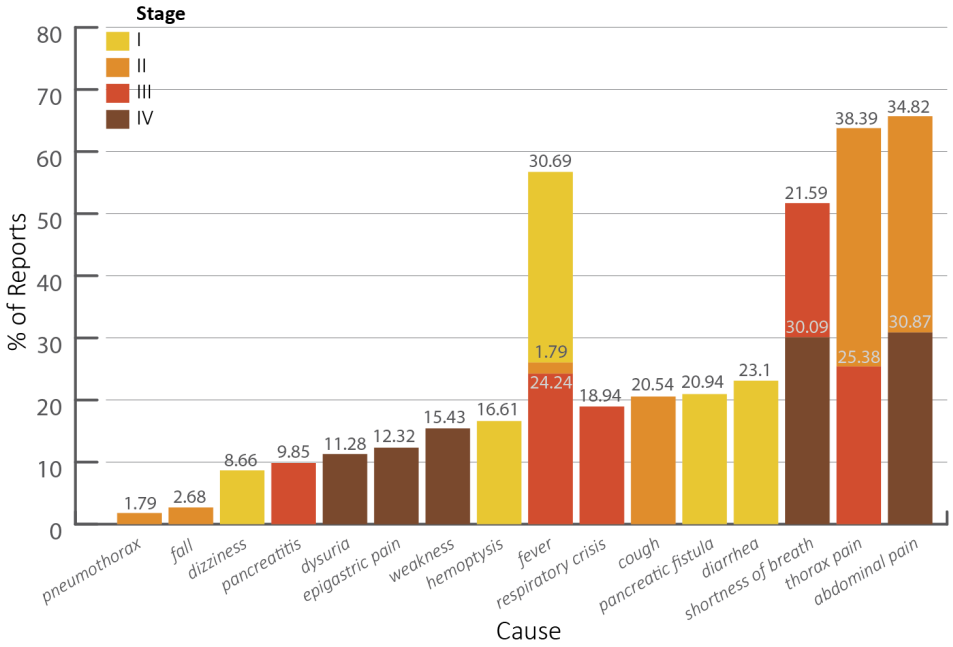

Figure 5.22: Top 5 causes of hospitalization by stage

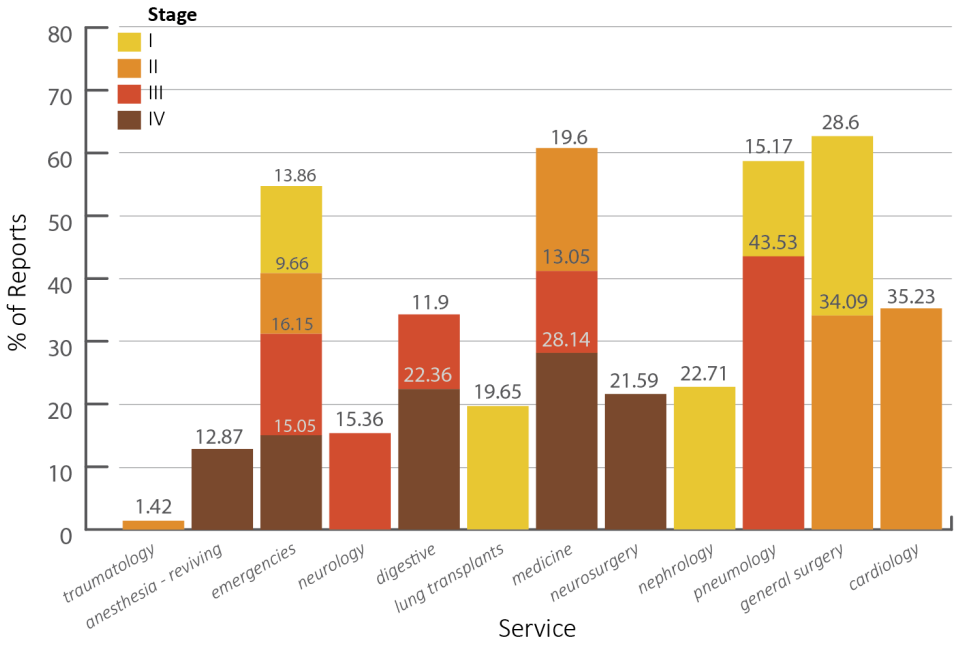

Figure 5.23: Top 5 consulted services by stage

Data Source: Services and Causes of Hospitalization Table.

Figure 5.22, presents the most common causes of hospitalization by cancer stage of patients.

For stage IV abdominal pain and shortness of breath are the most common causes, together with patients having stage II and stage III. For stage I it is interesting to notice a high number of patients being hospitalized for fever.

Figure 5.23, shows instead the most used services by cancer stage of patients.

It is interesting to notice a high number of patients at stage III visiting Pneumology and an high number of patients having stage II visiting cardiology. No interesting insights can be observed for stages I and IV.

5.3.9 Services and Causes of Hospitalization by Smoking Habit

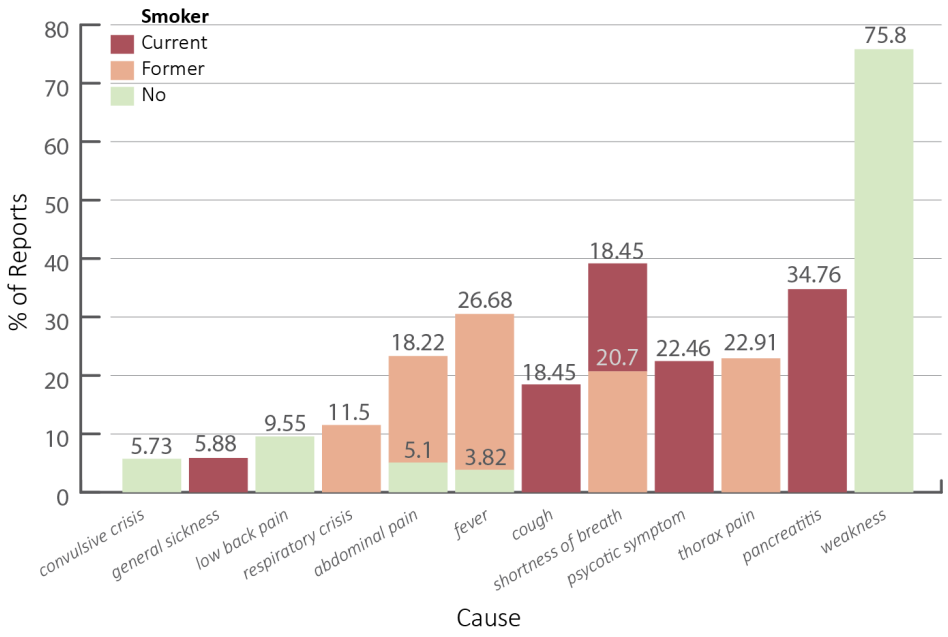

Figure 5.24: Top 5 causes of hospitalization by smoking habit

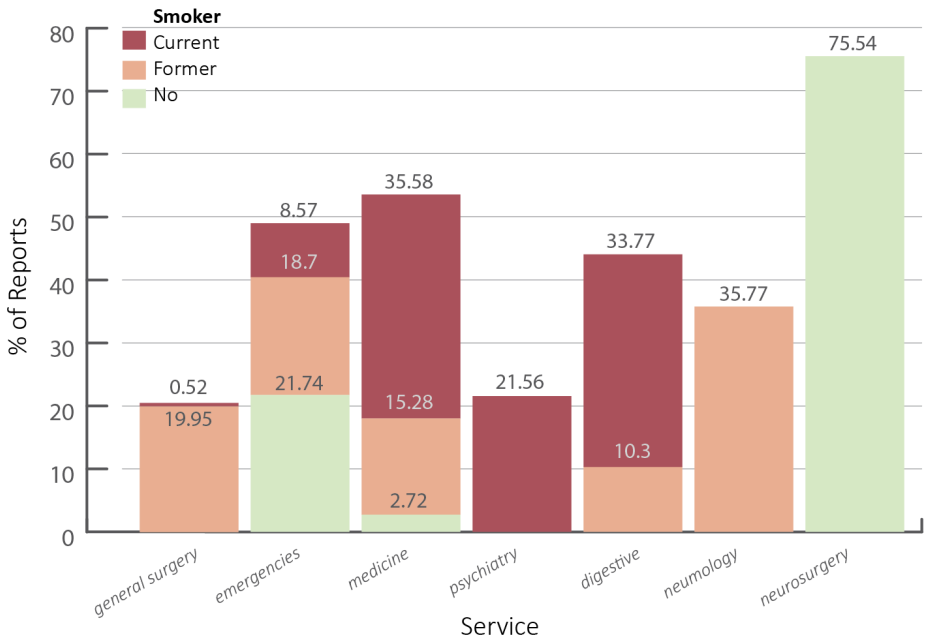

Figure 5.25: Top 5 consulted services by smoking habit

Data Source: Services and Causes of Hospitalization Table.

Figure 5.24, shows the most common causes of hospitalization by smoking habit of patients. Smoking habits have been divided in 3 categories: non-smoker (no), current smoker (current) and used to be a smoker (former).

It can be noticed that former smokers are often hospitalized with abdominal pain, thorax pain and shortness of breath. Current smokers, instead, are often hospitalized with shortness of breath. Non-smokers have no causes related to lung cancer symptoms, apart from those few samples having abdominal pain.

Figure 5.25, instead, shows the most used services by smoking habit.

It is interesting to notice that former smokers are often consulted in the Pneumology service. While no other relevant facts are being extracted from this point of view.

6 Conclusions and Perspectives

This master thesis work makes possible to transform raw data into a structured form that can be used to extract knowledge that can help to improve Key Performance Indicators. Clinical documents were processed and enriched with additional data extracted from texts: hospitalization dates, discharge dates, services names which produced the report and causes of hospitalization. Hospitalization and discharge dates, were used to extract semantically related clinical documents into processes. To achieve this task different approaches have been tested, such as relying only on existing date ranges inside texts or associating documents created in consecutive dates to belong to the same process. Processes are then classified into a defined set of categories and ready to be used for further analysis. The processes obtained, allow us to help clinicians in establishing policies for a better care of patients and to early diagnose lung cancer tumors.

In particular, this data-driven approach makes possible to extract length, causes of hospitalization, most frequent processes depending on patient stage, performance status or treatment and analyze processes that happen prior to death. All these parameters are of paramount importance to establishing polices for better patient care.

In particular, we have focused on the analysis of the services and causes of hospitalization that patients have visited prior to the tumor diagnose in order to be able to find if there is any indicator that can help doctors to anticipate the tumor diagnose and consequently send them to an oncology consultation as soon as possible. We have done this firstly by visualizing the most common causes of hospitalization and most used services, then by focusing the attention on three main services, often related with lung cancer patients: Cardiology and Pneumology. Then including also the Emergencies service.

Despite the promising results, this works only represents preliminary results, and further work is required to find among others, the relation of the objectives analyzed with demographics, habits, and comorbidities has been done as a first work. More work is also expected on associating this information with bibliographic information.

Furthermore, processes, to which the patient has been subjected to, can be used as features describing patients. An interesting use case of these additional features would be to use them to predict, with a supervised learning algorithm, whether the patient has a short or long survival.

Bibliography

[1] "Data science concepts you need to know! - available at https://towardsdatascience.com/introduction-to-statistics-e9d72d818745," 2018.

[2] W. Vorhies, "Crisp-dm – a standard methodology to ensure a good outcome - available at https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome," 2016.

[3] E. S. Pearson, "The test of significance for the correlation coefficient," *Journal of the American Statistical Association*, vol. 26, no. 174, pp. 128–134, 1931.

[4] "The digitization of the worldfrom edge to core - available at https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf," 2018.

[5] "Ibm - big data industry insight - available at https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/," 2013.

[6] "Bigmedilytics eu - available at https://www.bigmedilytics.eu/pilot/lung-cancer/,"

[7] "Airc - cancer numbers - available at https://www.airc.it/cancro/informazioni-tumori/cose-il-cancro/numeri-del-cancro," 2018.

[8] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo, "Analysis of length of hospital stay using electronic health records: A statistical and data mining approach," *PloS one*, vol. 13, no. 4, p. e0195901, 2018.

[9] R. Houthooft, J. Ruyssinck, J. van der Herten, S. Stijven, I. Couckuyt, B. Gadeyne, F. Ongenae, K. Colpaert, J. Decruyenaere, T. Dhaene, *et al.*, "Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores," *Artificial intelligence in medicine*, vol. 63, no. 3, pp. 191–207, 2015.

[10] J. Zuckerman, M. Ades, L. Mullie, A. Trnkus, J.-F. Morin, Y. Langlois, F. Ma, M. Levental, J. A. Morais, and J. Afilalo, "Psoas muscle area and length of stay in older adults undergoing cardiac operations," *The Annals of thoracic surgery*, vol. 103, no. 5, pp. 1498–1504, 2017.

[11] A. Almashrafi, H. Alsabti, M. Mukaddirov, B. Balan, and P. Aylin, "Factors associated with prolonged length of stay following cardiac surgery in a major referral hospital in oman: a retrospective observational study," *BMJ open*, vol. 6, no. 6, p. e010764, 2016.

[12] M. C. Wong, X. Q. Lao, K.-F. Ho, W. B. Goggins, and L. Shelly, "Incidence and mortality of lung cancer: global trends and association with socioeconomic status," *Scientific reports*, vol. 7, no. 1, p. 14300, 2017.

[13] S. B. Edge and C. C. Compton, "The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm," *Annals of surgical oncology*, vol. 17, no. 6, pp. 1471–1474, 2010.

[14] P. M. Ellis and R. Vandermeer, "Delays in the diagnosis of lung cancer," *Journal of thoracic disease*, vol. 3, no. 3, p. 183, 2011.

[15] A. Noone, N. Howlader, M. Krapcho, *et al.*, "Seer cancer statistics review (csr) 1975-2015 national cancer institute web site," 2019.

[16] "Lung cancer risk factors - cancer.org - available at https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/risk-factors.html,"

[17] M. H. Yarmohammadian, H. Ebrahimipour, and F. Doosty, "Improvement of hospital processes through business process management in qaem teaching hospital: A work in progress," *Journal of education and health promotion*, vol. 3, 2014.

[18] "Health at a glance 2017 - available at https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2017/consultations-with-doctors_health_glance-2017-60-enl,"

[19] "Kpi basics - available at https://kpi.org/KPI-Basics,"

[20] "Kpi characteristics - available at https://entrinsik.com/5-characteristics-effective-kpis/,"

[21] "Kpi for healthcare - available at https://www.datapine.com/kpi-examples-and-templates/healthcare,"

[22] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, pp. 37–37, 1996.

[23] "A day in data - available at http://res.cloudinary.com/yumyoshojin/image/upload/v1/pdf/future-data-2019.pdf," 2019.

[24] "How much data do we create every day? the mind-blowing stats everyone should read - available at https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#76826aeb60ba," 2018.

[25] "Crisp-dm - available at http://crisp-dm.eu/,"

[26] "Data cleansing - available at https://en.wikipedia.org/wiki/Data_cleansing,"

[27] "A short introduction to model selection - towards data science - available at https://towardsdatascience.com/a-short-introduction-to-model-selection-bb1bb9c73376,"

[28] Y. Li, S. Lipsky Gorman, and N. Elhadad, "Section classification in clinical notes using supervised hidden markov model," in *Proceedings of the 1st ACM International Health Informatics Symposium*, pp. 744–750, ACM, 2010.

[29] "Nih u.s. national library of medicine - umls - available at https://www.nlm.nih.gov/research/umls/index.html,"

[30] "ggplot2 package - tidyverse - available at https://ggplot2.tidyverse.org/,"

[31] "Stringr package - available at https://cran.r-project.org/web/packages/stringr/vignettes/stringr.html,"

[32] "dplyr package - tidyverse - available at https://dplyr.tidyverse.org/,"

[33] "Rmysql: Database interface and 'mysql' driver for r - available at https://cran.r-project.org/web/packages/RMySQL/index.html,"

[34] J. H. S. K. C. C. Tye Rattenbury, Joseph M. Hellerstein, *Principles of Data Wrangling*. 2017.