Instituto Universitário de Lisboa

Department of Information Science and Technology

Structured and Unstructured Data Integration with Electronic Medical Records

Diogo Veiga Amorim Santos Baptista

A Thesis presented in partial fulfilment of the Requirements for the Degree of

Master in Decision Support Systems

Supervisor:
Prof. Dr. João Carlos Amaro Ferreira, Assistant Professor
ISCTE-IUL

Co-Supervisor:
Prof. Dr. Rúben Filipe de Sousa Pereira, Assistant Professor
ISCTE-IUL

October 2019

Acknowledgments

I would like to thank my supervisor Dr. João Ferreira and co-supervisor Dr. Rúben Pereira for the constant support, counselling and guidance given throughout this work, since without them this work wouldn’t have been possible.

I would also like to thank three of my closest friends Carina Abreu, Nuno Mendonça and João Silva for the support and patience to help me both technically as also motivationally.

Finally, I would like to thank my family and friends for all their love and support throughout this year, without you this wouldn’t have been the journey that it was, and I am very thankful for that.

Resumo

Nos últimos anos tem-se assistido a uma grande evolução populacional e tecnológica por todo o mundo. Paralelamente, mais áreas para além da tecnologia e informática têm-se também desenvolvido, nomeadamente a área da medicina, o que tem permitido um aumento na esperança média de vida que por sua vez leva a uma maior necessidade de cuidados de saúde.

Com o intuito de fornecer os melhores serviços de saúde possíveis, nos dias que hoje os hospitais guardam nos seus sistemas informáticos grandes quantidades de dados relativamente aos pacientes e doenças (sobre a forma de registos médicos eletrónicos) ou relativos à logística de alguns departamentos dos hospitais, etc. Por conseguinte, a estes dados têm vindo a ser utilizadas técnicas da área das ciências da computação como o *data mining* e o processamento da língua natural para extrair conhecimento e valor dessas fontes ricas em informação com o intuito não só de desenvolver, por exemplo, novos modelos de predição de doenças, como também de melhorar processos já existentes em centros de saúde e hospitais. Este armazenamento de dados pode ser feito em uma de três formas: de forma estruturada, não estruturada ou semi-estruturada.

Neste trabalho o autor testou a integração de dados estruturados e não estruturados de dois departamentos diferentes do mesmo hospital português, com o intuito de extrair conhecimento e melhorar os processos do hospital. Com o intuito de reduzir a perda do armazenamento de dados que não são utilizados.

Palavras-Chave: Dados Estruturados, Dados Não Estruturados, Processamento Natural da Língua, Integração de Dados, Registos Médicos Eletrónicos.

Abstract

In recent years there has been a great population and technological evolution all over the world. At the same time, more areas beyond technology and information technology have also developed, namely medicine, which has led to an increase in average life expectancy which in turn, leads to a greater need for healthcare.

In order to provide the best possible treatments and healthcare services, nowadays the hospitals store large amounts of data regarding patients and diseases (in the form of electronic medical records) or the logistics of some departments in their storage systems. Therefore, computer science techniques such as data mining and natural language processing have been used to extract knowledge and value from these information-rich sources in order not only to develop, for example, new models for disease prediction, as well as improving existing processes in healthcare centres and hospitals. This data storage can be done in one of three ways: structured, unstructured or semi-structured.

In this paper, the author tested the integration of structured and unstructured data from two different departments of the same Portuguese hospital, in order to extract knowledge and improve hospital processes. Aiming to reduce the value loss of loading data that is not used in the healthcare providers systems.

Keywords: Structured Data, Unstructured Data, Natural Language Processing, Data Integration, Electronic Medical Records.

Contents

Index of Tables

Index of Figures

List of Abbreviations

AD – Appointments Department

DM – Data Mining

ED – Emergency Department

EMR – Electronic Medical Records

ICD – International Code of Diseases

NLP – Natural Language Processing

TF-IDF – Term Frequency-Inverse Document Frequency

TM – Text Mining

WHO – World Health Organization

Chapter 1 – Introduction

1.1. Context

With the evolution of the medical sector, life expectancy has been increasing. On the one hand, this evolution leads to positive events, such as the end of some diseases (smallpox, plague, etc.) but, on the other hand, new problems arise like the manifestation of new ones such as dementia, cancer, etc. But, as medicine grew and developed, so many other areas like computer science matured and can be used to help to fight these new challenges (Baptista, Ferreira, Pereira, & Baptista, 2019).

Nowadays, healthcare providers store loads of data, medical, and non-medical (Baptista et al., 2019). This data can be about multiple things, such as drug prescription, treatment records, general check-up information, physician's notes, surgical information or financial and administrative and is stored in either legacy systems or electronic medical records (EMR) (Luo et al., 2016). The EMR is computerized medical information systems that collect, store and display patient information (McLane, 2005).

The usage of EMR accommodates multiple advantages. According with (Yamamoto & Khan, 2006) those advantages could be summarized as "optimizing the documentation of patient encounters, improving communication of information to physicians, improving access to patient medical information, reduction of errors, optimizing billing and improving reimbursement for services, forming a data repository for research and quality improvement, and reduction of paper".

The format in which this data is recorded is very important since its format is directly related to the way it can be used to extract further insight.

1.2. Biomedical Data Types

There are three different types of data: structured, structured, unstructured, and semi-structured (Yadav, Steinbach, Kumar, & Simon, 2017). Considering the project in hand, the latter is not considered. Therefore, focusing only on the structured and unstructured data types.

Associated with these different types of data structures, there are different data manipulation techniques, such as data mining (DM) and text mining (TM) (also known as Text Data Mining (Sun et al., 2018)).

1.3.Data, Information, and Knowledge

Data, information, and knowledge, are three commonly misunderstood concept words that are, sometimes, used as equals which may be the cause of some misunderstandings (Lamy, 2018).

The three terms are related, in a pyramid/chain like relationship (Cooper, 2017), yet they do not represent the same. The complexity and understanding increase from data to information and finally to knowledge.

Data is a value, like a clinical measurement, such as heart rate = 50 beats per minute (bpm) (Cooper, 2017), or simply raw data, such as the clinical narrative (Lamy, 2018). And of the three, it is the least informative.

Information, the next level of the hierarchy, "is accumulated, assembled, or processed data through processes such as referential, type, purpose, relevance, and interpretation" (Allen, 2004), in other words, is put into context, acquiring some meaning. Continuing the example, in the context of a small child, a heart rate of 50 bpm gives some information to a doctor about the child, yet that same information could have different meanings if an adult presents the same values (Cooper, 2017).

The third level of the scale, knowledge, is the most informative of the three presented. When information is structured and organized as a result of cognitive processing (Cooper, 2017) to offer understanding, experience, and accumulated learning (Allen, 2004) and validation become knowledge.

So, to recapitulate, while data are raw values, information is when those values are put into context, and finally, knowledge is information that is structured, organized and processed and may be used to improve procedures or other processes.

1.4. Motivation

The increasing amount of people in the world brings with it an increasing amount of people in the hospitals, generating huge amounts of stored data, under multiple formats. This can have several uses depending on its structure typology. While structured data can be used in prediction models for heart disease risk (Amin, Agarwal, & Beg, 2013), diabetes (Simon et al., 2015) or re-hospitalization (Basu Roy et al., 2015) through DM techniques, with the unstructured data, through the usage of TM, it is possible to assess coronary artery diseases risk (Jonnagaddala et al., 2015) and the identification certain

case of injuries (Luther et al., 2015). But the usage of these data types together is seldom seen. This research aims to assess the validity of the usage of both data types together in order to get further insight into the stored data. Aiming to reduce the value loss of loading data that is not used in the healthcare providers systems.

1.5.Objectives/Goals

This work intends to develop a proof of concept able to integrate both structured and unstructured medical data, from different data sources, and enable the extraction of medical information and knowledge.

The biomedical datasets used to test and evaluate the system are both from the same Portuguese hospital, and while one contains structured data regarding the emergency department, the second contains data from medical appointments, where important data is written in the form of free text.

The extraction of both the information and knowledge is supported by a combined application of data mining (DM) and text mining (TM) techniques, such as natural language processing (NLP).

1.6.Research Questions

With this work, it is intended that the following questions, present in Table I, are answered by its end.

Table I – Research Questions

Question ID	***Question***
RQ1	Is it possible to extract structured information from unstructured clinical data contained in EMRs?
RQ2	Is it possible to successfully extract clinical knowledge from the combined information of two departments from the same hospital?

1.7. Dissertation's Structure of Organization

The remaining part of this dissertation is comprised of six chapters: (1) Literature Review, (2) System Architecture, (3) Unstructured Data, (4) Structured and Mixed Data, (5) Results and Discussion and finally (6) Conclusion.

The second chapter, Literature Review, gives an overview of the related work developed throughout the last years regarding the use of structured, unstructured, and mixed data in the medical field.

The third chapter, System Architecture, presents an overview of all process regarding the dissertation. It enlightens all the steps of the process, although from a more high-level point of view.

The fourth chapter, Unstructured Data, contains the presentation of the unstructured dataset and the explanation of its' processing, in detail, since the data's raw form until its' final stage.

The fifth chapter, Structured and Mixed Data, contemplates the same explanation of the structured dataset and its processing, as it happens in the previous chapter with the unstructured dataset. Nevertheless, this chapter also displays how both structured and unstructured data are combined.

The sixth chapter, Results, and Discussion present the final results of the dissertation, a reflection of its significance, and also a comparison with the planned objectives.

Finally, the last chapter, Conclusion, summarizes all the work done in this dissertation, the attained objectives, answers the research questions, reflects on the limitations of this work and also contemplates on the future work that can be done to continue with the developed work and how to improve it.

Chapter 2 – Literature Review

This chapter enlightens the evolution of what has been developed in the previous years related to information extraction from electronic medical records (EMR's) data, structured, unstructured, and mixed. Few use cases are shown and analyzed for structured and unstructured data, where techniques and frameworks are compared and evaluated. This chapter is composed of three main chapters: (1) Structured Data, (2) Unstructured Data and (3) Mixed Data Approaches.

On the one hand, regarding the structured data front, the case studies presented approach the data mining techniques most commonly used, its utility and results, and the data attributes themselves most commonly used.

On the other hand, about the unstructured data, several approaches have been made to further analyze and use this type of data. Whereas through the classification of text or simply through the extraction of clinical entities from medical records. Because of it, different natural language processing (NLP) systems in the biomedical field are approached and analyzed as the attained results. This overview also explains why the chosen NLP system for the project in hand.

Lastly, regarding the mixed data approaches, different cases studies are an overview and analyzed regarding the usage of structured and unstructured data together and what was achieved with these methodologies.

2.1.Structured Data

Structured data is the name given to data that is stored in a fixed schema database and out of the three existing data types, this one is the easiest of being managed. The fact that the data is stored in a structured schema makes it off easier and faster access.

2.1.1. Biomedical Structured Data

The most common data in this scenario is demographic information (e.g. race, ethnicity, birth date), admission and discharge dates, diagnosis codes (historic and current), procedure codes, laboratory results, medications, allergies, social information (e.g., tobacco usage), and some vital signs (blood pressure, pulse, weight, height) (Garets & Davis, 2006).

The way to handle this type of data, transform it into information, and enable knowledge extraction is through Data Mining as aforementioned.

2.1.2. Data Mining in Healthcare

DM has many different definitions, for example, according to (Koh & Tan, 2005), DM is described as "(…) the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns." But in this scenario, the DM definition to be considered is Gartner's, "the process of discovering meaningful correlations, patterns, and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques" (Gartner, 2019a).

Associated with DM, there are a lot of different techniques, benefits, and areas that are used throughout the world, considering the situation.

In the next chapter, some case studies about DM in healthcare are analyzed and reviewed to understand the diversity of techniques available, in which areas it is used and the impact of each one.

2.1.3. Data Mining Case Studies

To facilitate the understanding of how structured data is managed and used, through the use of DM in the healthcare sector, some cases are presented in Table II.

As it can be viewed, the DM analysis utilized in the respective papers can be divided into five categories (Sun, Cai, Liu, Fang, & Wang, 2017): Classification, Clustering, Time Series, Regression Algorithms and Hybrid-Model.

And for all these DM types, many methods were used, showing the wide variety of techniques, applications, and possibilities of this technology.

Considering the overall cases studies (Baba et al., 2015; Basu Roy et al., 2015; Ravindranath, 2015; Somanchi, Adhikari, Lin, Eneva, & Ghani, 2015), it is possible to access that there are a lot of cases, in this scenario, 9 out of 12, that use statistical methods as base for classification models.

Table II – Overview of case studies of DM techniques in the Healthcare

DM Type	*Methods*	*Application*	*Speciality*	*Reference*
Hybrid Model	Neural Networks and Genetic Algorithms	Heart disease risk prediction	Cardiology	(Amin et al., 2013)
Time Series Mining	Time-series mining	Risk prediction of heart disease	Cardiology	(Chia & Syed, 2014)
Hybrid Model	K-means and clustering technologies	EMR data processing	Operational	(Sumana & Santhanam, 2015)
Association Rule Mining	Association rule mining	Diabetes risk prediction	Endocrinology	(Simon et al., 2015)
Classification Technology	Dynamic classification and hierarchical model	Re-hospitalization risk prediction	Operational	(Basu Roy et al., 2015)
Classification Technology	Multi-classifier method	Noncommunicable disease pre-diction	Epidemiology	(Baba et al., 2015)
Classification Technology	Support Vector Machine	ICU risk prediction	Cardiology	(Somanchi et al., 2015)
Classification Technology	Decision Tree	Decision Support Systems	Operational	(Ravindranath, 2015)
Clustering Technology	Dynamic feature selection	Quality of medical service	Quality of Service	(Rabbi, Mamun, & Islam, 2015)
Regression Algorithm	Ordinal regression framework	Suicide risk prediction	Psychiatry	(Tran, Phung, Luo, & Venkatesh, 2015)
Time Series Mining	Time-series mining	Risk prediction of colorectal cancer	Oncology	(Kop, Hoogendoorn, Moons, Numans, & ten Teije, 2015)
Regression Algorithm	Multivariate logistic regression	Analysis of Cardiac Surgical Bed Demand	Operational	(Toerper et al., 2016)

On the one hand, the majority of these case studies, use DM techniques to train models and later apply them to predictions in many sectors, some more technical, as cardiology

(Amin et al., 2013; Chia & Syed, 2014; Somanchi et al., 2015), oncology (Kop et al., 2015), psychiatry (Tran et al., 2015), endocrinology (Simon et al., 2015), other more operational, as quality of service (Rabbi et al., 2015), risk of patients re-hospitalization (Basu Roy et al., 2015) and fore-casting the daily bed needs (Toerper et al., 2016).

On the other hand, as shown in (Ravindranath, 2015), predictions are not the only use of these structured data. K. R. Ravindranath not only explains that there are many ways to diagnose heart diseases, proposing a new Decision Support System for effect based on decision trees.

2.2. Unstructured Data

While in the previous chapters, it was discussed what is structured data, how it can be used to boost the health sector capabilities and were shown some practical examples, during this subchapter both unstructured data and TM techniques were reviewed. Amongst the three aforementioned data types, unstructured data is the one that offers the maximal flexibility (Yadav et al., 2017).

2.2.1. Clinical Diaries

The unstructured data of an EMR is present in clinical notes, surgical records, discharge records, radiology reports, and pathology reports (Sun et al., 2018).

Clinical notes are documents written, in free text, (Yadav et al., 2017) by the doctors, nurses and staff providing care to a patient, and offer increased detail beyond what may be inferred from a patient's diagnosis codes (Feldman, Hazekamp, & Chawla, 2016).

The information contained in clinical notes maybe about a patient's medical history (diseases, interventions, etc.), familiar history of diseases, environmental exposures, and lifestyle data (Yadav et al., 2017). And according to (Savova et al., 2010), the knowledge of said notes are retrieved "by employing domain experts to manually curate such narratives". Hence the usage of an automatic method of interpretation of these clinical notes and records is of the utmost importance.

As explained, DM techniques that could be applied to structured data, cannot be applied to this type of data without some previous structuring (pre-processing).

Unfortunately, because this type of data is represented as free text, there is no common framework, there may be improper grammatical use, spelling errors, local dialects (Sun et al., 2018), short phrases and/or abbreviations (Yadav et al., 2017).

Due to such difficulties, data processing and analysis becomes harder. A great deal of this difficulty is precisely the pre-processing of said free text. Natural Language Processing (NLP) tools and techniques are used and have proven very useful when comes to extract knowledge (Yadav et al., 2017).

Therefore, with the help of methods like NLP, it becomes possible to structure the free text and apply DM techniques. Some examples of the types of operations that are done in pre-processing are, removal of digits, anonymization, punctuation removal, etc.

2.2.2. Biomedical Natural Language Processing

Natural Language Processing (NLP), according to (Gartner, 2019b), "technology involves the ability to turn text or audio speech into encoded, structured information, based on an appropriate ontology". This technology has many utilities from text classification, to sentiment analysis, to automatic summarization. And one of the many sectors to benefit from this technology is the clinical domain, where many successes have been achieved with NLP (Savova et al., 2010).

Considering the free text used from the EMR are clinical notes as exposed above, the processing of the written natural language may vary when using NLP algorithms. Nonetheless, the initials steps usually are the same. Initially, the texts are broken down as separated sentences and then tokenized. After that usually a process of Part-of-Speech Tagging (POS-Tagging) comes into play where each word of the sentence is classified according to its morphological class (verb, noun, adverb, etc.).

In the next chapter, some case studies about NLP and TM techniques and developed tools for the healthcare sector are analyzed and overviewed.

2.2.3. Text Mining Case Studies

Regarding the NLP and TM application to process free text in the healthcare sector, several case studies are presented in order to perceive how, and how much the use of such techniques can better the healthcare sector.

Table III – Overview of case studies of TM and NLP techniques in the Healthcare

NLP Methods	***Methods***	***Application***	***Reference***
Stemming, stop words, numbers and dates removed	Logistic Regression	Detection of Adverse Childhood Experiences	(Araneo & Celozzi, 2015)
Stop words, lowercase, IDF	Support Vector Machine	Identification of fall-related injuries	(Luther et al., 2015)
Self-Developed Algorithm	-	Coronary artery disease risk assessment	(Jonnagaddala et al., 2015)
Tokenization, TF-IDF, Removal of Stop words	Forward Neural Network with Back Propagation	Automatic diagnosis prediction	(Pulmano & Estuar, 2016)

Depicted in Table III are four case studies where the authors used NLP and/or TM to extract information from medical records. This information would be later used in DM models for prediction (Pulmano & Estuar, 2016) and disease identification (Araneo & Celozzi, 2015; Jonnagaddala et al., 2015; Luther et al., 2015).

As previously explained, usually it is necessary some pre-processing of the free text from EMR before it is in conditions to be used from information extraction. There are some techniques that are quite common such as, the removal of stop words, the lowercase treatment. Other techniques are not that common as Term Frequency-Inverse Document Frequency (TF-IDF) and Stemming. The pre-processing depends on the data in hands and what wants to be done with it.

Although the use of NLP methodologies and tools was different in all scenarios, as the DM methodologies used after, the results were very satisfactory. Another great aspect of these studies is that the unstructured data of the EMR that would not be used in some cases is indeed used and proven useful.

Yet, throughout the years, some tools were developed in order to facilitate the detection and extraction of medical knowledge from EMR's. Some of those developed programs are presented in Table IV.

Table IV – Clinical NLP Programs Case Studies

NLP Software	***Date***	***Reference***
MedLEE	1994	(Friedman, Liu, Shagina, Johnson, & Hripcsak, 2001; Jain & Friedman, 1997; Sevenster, Van Ommering, & Qian, 2012)
GATE	1996	(Liu, Mitchell, Chapman, & Crowley, 2005; Wu et al., 2013)
cTAKES	2010	(Afzal et al., 2016; Ananthakrishnan et al., 2013; Kidwai et al., 2011; Savova et al., 2010; Sohn, Kocher, Chute, & Savova, 2011)

The MedLEE (Medical Language Extraction and Encoding System) is "an MLP [Medical Language Processing] system that extracts, structures, and encodes relevant clinical information that occurs inpatient reports" (Friedman et al., 2001). It has been used throughout the years in many ways, such as for the detection of finding suspicious of breast cancer from mammography reports (Jain & Friedman, 1997), it was also used in "a system that extracts clinical findings and body locations from radiology reports and correlates them" (Sevenster et al., 2012).

The GATE (General Architecture for Text Engineering) is an open-source (Liu et al., 2005) NLP software designed to extract information from open-text fields (Wu et al., 2013). This program has been used in systems to detect smoking status information from open-text fields (Wu et al., 2013) and has also been incorporated in a system that, in an automated way, allowed the extraction of tissue annotation from surgical pathology reports (Liu et al., 2005).

And finally, cTAKES (the clinical Text Analysis and Knowledge Extraction System) is "a comprehensive clinical NLP system based on the Unstructured Information Management Architecture (UIMA)" (Kidwai et al., 2011). It is composed of six components that are interconnected and work together to achieve this system's purpose. The cTAKES enables the detection and extraction of medical information from clinical narratives (Savova et al., 2010). This NLP system has been used to identify peripheral arterial diseases from clinical notes (Afzal et al., 2016), extraction of side effects of some drugs from clinical narratives regarding the medical specialties of psychiatry and psychology (Sohn et al., 2011) and case definition of crohn's disease and ulcerative colitis processing electronic medical records (Ananthakrishnan et al., 2013).

Based on these three clinical NLP programs, the one that was selected to be used in this work was cTAKES. According to (Lamy, 2018), which made a comparison between,

not only these but also some other NLP systems, the cTAKES system not only presented good results in the case studies where it was used, as it is open-source software, which allows for versatility of usage. Allied to the fact that it is also the most recent, the cTAKES was the selected NLP software used in this work.

In the next chapter, the case studies that are presented go over the use of structured and unstructured data together and how the healthcare sector can benefit from the mixed-used of these two types of data.

2.3. Mixed Data Approach

2.3.1. A Mixed Technique

Until now, both DM and NLP case studies have been shown and analyzed, yet, in all of them, there is not one where the structured and unstructured data are used together. While in some, only the structured data is used, others only use the structured data extracted from the unstructured free text from de NLP processing used.

The author considers another way of using all this data. Instead of only using the already structured data or only the structured data retrieved from the free text with TM. Which may end-up with the loss of potential of good information, either way.

2.3.2. Mixed Approach Case Studies

For this effect, a few case studies that used a mix of information from both structured and unstructured data are presented and analyzed.

The case studies are presented in Table V and are organized by application, reference and NLP methods used.

Table V – Overview of Mixed Data Case Studies

Methods	***Application***	***Speciality***	***Reference***
Topic Modelling (LDA)	Identify Related Patient Safety Events	Security	(Fong, Hettinger, & Ratwani, 2015)
Naïve Bayes, Bayesian Belief Networks and DT	Diagnosing Early Stages of Dementia	Psychiatry	(Moreira & Namen, 2018)
Logistic Regression	Predict Hospital Re-admissions	Cardiology	(Sundararaman, Valady Ramanathan, & Thati, 2018)
Clustering	Geriatric Syndrome Detection	Geriatric	(Kharrazi et al., 2018)

As it can be perceived, only four items are present in the Table V. This is because not many case studies make of both structured and unstructured data at the same time, yet the majority is from 2018 which could show a possible tendency.

In these case studies, not only NLP techniques were applied, as DM techniques were used.

In (Moreira & Namen, 2018), not only both techniques were used as its results were compared. In other words, to detect early stages of dementia, the authors used two different techniques to assist specialists in the diagnosis of patients with clinical suspicion of dementia. And concluded that the model that used the structured data and the clustering of the texts written in free format by the physicians integrated, improved the accuracy of predictive models in all pathologies (Moreira & Namen, 2018).

According to (Sundararaman et al., 2018), similar results were obtained when the authors applied the logistic regression to five different iterations. The first one only using structured data, the second one using only unstructured data, the third used feature selection, and the last two used mixed data. And after the five iterations, according to the observed results, the authors concluded "it is recommended that iteration 5 be chosen for such research and applications" (Sundararaman et al., 2018).

Finally, the third case had the objective of comparing the number of geriatric syndrome cases identified using structured claims and structured and unstructured EHR data in order to understand the added value of the latter (Kharrazi et al., 2018). Of the obtained results,

once again, the conclusions were similar to those already explained, that the results improved when combining both models. This type of facts led the authors to encourage "incorporating NLP methods to increase the sensitivity of identification of individuals with geriatric syndromes" (Kharrazi et al., 2018).

Chapter 3 – System Architecture

This chapter aims to give a high-level view of the pipeline developed. Its' objective is to integrate structured and unstructured medical data from two different departments, (the Emergency Department (ED) and the Appointments Department (AD), respectively, in a unique database to allow the extraction of knowledge.

During this process the datasets are passed on through a series of steps, (1) Pre-processing, (2) Attribute Mapping, (3) Translation, (4) NLP Systems and (5) Integration in a common database. Presented in Figure 1 is the overall process with the steps of the process required for each dataset from its "raw form" to its "processed form" ready to be explored.

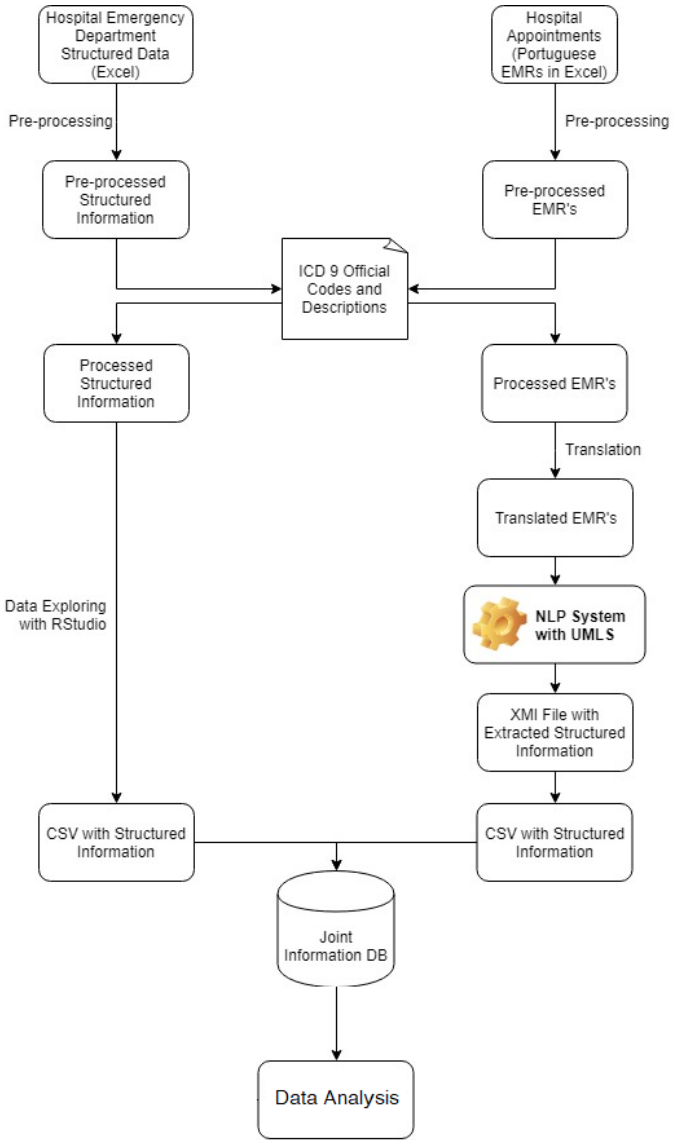


Figure 1 – Overview of the process of integration of both datasets, from the excel files to the final conjoint database

3.1. Pre-Processing

The pre-processing is a fundamental step in every project where datasets are used. The importance of this phase is due to the data cleaning that is done. Data cleaning is the process of preparing the data from its initial state (how it comes out from the original data sources) to a state ready to be used. It's during this phase that problems such as duplicates, null or "NA", wrong formatting problems are solved.

3.2. Attribute Mapping

The attribute mapping phase aims to establish that variables that are present in both datasets are represented the same way. This phase is required because of the finality of

this work. Considering that the final objective is to successfully integrate data from two different datasets, as it can be perceived in Figure 1, it is looked-for linking attributes that allow said linkage.

3.3. Translation

On the one hand, in the previous chapters was mentioned that in this work, an NLP system was used conjointly with a clinical-based of knowledge. On the other hand, the datasets used in this work are from a Portuguese hospital and, as such, the medical diaries, as a few other attributes of the datasets, are in Portuguese. Hence, the need and importance of this step.

3.4. NLP System

The NLP system, working together with the clinical base of knowledge, represents yet another important phase of the pipeline regarding the unstructured dataset. It is in this phase that the medical diaries go through the biggest transition, from unstructured to structured.

3.5. Integration

The final process before the data exploration rests in the integration of the processed data from both datasets, structured, and unstructured. Good integration is important because it may impact in a direct way the results to be obtained in the data exploration.

3.6. Data Analysis

The final phase of the process emphasizes on the exploration of the integrated data. It is at this stage when the data is analyzed through tables and graphs in order to extract clinical knowledge and make conclusions out of it.

In the next chapter, the full pipeline is explained in greater detail. The flow that the data from each dataset had to go through and the results from each stage of the process.

Chapter 4 – Unstructured Data

This chapter aims to explain in detail the constitution of the dataset of unstructured data and its' flow. This chapter is organized in two subchapters: (1) Unstructured Dataset Description and (2) Unstructured Data Processing.

4.1. Unstructured Dataset Description

The unstructured dataset consists of a set of 11.137 EMRs from a Portuguese hospital in an excel file regarding the appointment's department. In other words, each line from the excel is related to an individual appointment.

Since this dataset is related to the appointment's department, as stated previously, it is treated as Appointment Department (AD) dataset. This dataset contains data relative to the year 2017. Also, the dataset contains information relative to several medical specialities, such as oncology, rheumatology, gastroenterology, paediatrics, paediatric haematology and urology, amongst others.

Next, present in Table VI, it is represented the structure of an entry of the AD dataset, across its 10 attributes and some examples of what those attributes can have.

Table VI – AD Dataset Attributes

Attribute #	***Attribute Name***	***Data Type***	***Example***
1	HDI Episode	Integer	16008525
2	Sequential Number	Integer	608268
3	Date Init Treatment	Date	17.01.11
4	Speciality Code	Integer	40695
5	Speciality Description	String	Rheumatology
6	Diagnosis Code	Integer	6954
7	Diagnosis Description	String	Lupus erythematosus
8	Date	Date	17.01.11
9	Module	String	HDI
10	Diary	Text	Complaints of coughing dragged. Initially with yellowish expectoration, currently without expectoration. Apyretic. Pulmonary auscultation: Vesicular murmur maintained and symmetrical. No adventitious noise. X-ray thorax without changes. Put off therapeutics 1 week.

The first two attributes, “HDI Episode” and “Sequential Number”, are identifiers of the EMR’s.

The “Date Init Treatment” is represented in a date format (yy.mm.dd) and shows the date at which the treatment of the patient began.

The “Speciality Description” attribute, as the name suggests, infers the medical speciality of the appointment, whereas the “Speciality Code” represents the respective code of the medical speciality.

The “Diagnosis Code” contains the code of the diagnosis given by the doctor and the “Diagnosis Description” represents the respective description of the disease, both variables are in agreement with the International Code of Diseases 9 (ICD 9). The ICD 9 which is regulated by the World Health Organization (WHO) comes to standardize disease description and “(…) promote international comparability in the collection, processing, classification, and presentation of mortality statistics.”(“ICD - ICD-9 -

International Classification of Diseases, Ninth Revision," 2019). The ICD 9 comes The ICD 9 CM is composed of 19 chapters based on the subject of the ICD codes each chapter contains. Each chapter is identified by a number and a description ("ICD-9-CM Chapters List," 2019). In Table VII is an excerpt of the chapters' numbers and descriptions[1]. Each set of ICD codes from each chapter is specified by a range showing the first three digits of the code range included. In other words, the first three digits of an Infectious and Parasitic Disease is comprised between 001 and 139.

Table VII – ICD Chapters Excerpt

Chapter #	***Code Range***	***Description***
1	001-139	Infectious and Parasitic Diseases
2	140-239	Neoplasms
3	240-279	Endocrine, Nutritional and Metabolic Diseases, And Immunity Disorders
4	280-289	Diseases of The Blood and Blood-Forming Organs

The "Date" is represented the same way as the variable "Date Init Treatment" but represents the date of the appointment.

Finally, the "Diary", maybe the most valuable variable of this dataset, contains a descriptive narrative of the appointment, written by the doctors themselves. This free text can contain a broad of different elements that can give valuable knowledge of what has been done, the drugs prescribed for each disease, its dosages, etc.

The attributes "Speciality Description", "Diagnosis Description" and "Diary" are written in Portuguese, but to show an example, as performed with the AD attributes, a translated example was shown.

These attributes could be a bridge between datasets and enable the extraction of valuable knowledge needs to be rectified, so both attributes have the same configuration, both integers, facilitating possible cross-referencing.

The following subchapter explains in detail the processing through which the unstructured dataset was passed.

[1] The complete table is in the Annex A

4.2.AD Processing

The AD dataset required a complex process due to the unstructured nature of attribute "Diary". Since it is written in free text, more operations were required in order to prepare the data. This chapter is segmented in four subchapters: (1) Pre-processing, (2) ICD 9 Mapping, (3) Translation, (4) cTAKES and UMLS.

4.2.1. AD Pre-processing

As previously stated, the pre-processing is a fundamental stage of every project that uses datasets because it is at this stage that the data is prepared.

The first stage consists of the removal of any entry that contains any empty fields. Cleaning these rows eliminates the change of getting future "Nulls" and/or "N/A" in the database to be created, which improves its quality.

Secondly, the author searched for duplicated rows in order to remove repeated data (between rows) which can create problems in the future if not dealt with.

Following the elimination of the duplicated rows, some of the AD attributes were eliminated from the dataset since some of them represented duplicated information (between columns), shown always the same value or showed other problems. The removed columns and the respective reason for removal are presented in Table VIII.

Table VIII – AD Removed Attributes and Justification

Attribute #	***Attribute Name***	***Data Type***	***Reason***
1	HDI Episode	Integer	Value Repetition
2	Sequential Number	Integer	Value Repetition
3	Date Init Treatment	Date	Column Repetition
9	Module	String	One Value

The first two attributes were removed since the only information it transmitted was the internal identifiers of the hospital system. Being identifiers, these attributes should have unique values, which was not the case. For this reason, and because said attributes did not add great value to the dataset, both columns got removed.

The third column was removed since it presented always the same value as the attribute "Date". Finally, the ninth attribute, "Module", was also removed because it only showed one value, not adding any value to the dataset.

The next step in the pre-processing stage was focused on the most complex and important attribute of the dataset, "Diary". Considering the column is composed of free-text, a specific treatment is required in order to improve its' quality. Through the excel, the free text was analyzed and corrected, in other words, the unfold of abbreviations, acronyms and the correction of misspelling was done. Another important aspect of this phase was the elimination of every identification, such as names, phone numbers and addresses of both patients as practitioners and/or nurses. This was done as a privacy measure. This step is considered an important one since the text is in Portuguese and it still needs to be translated to English, the higher the quality of the text in Portuguese the better the results to be obtained after the translation. Hence, ensuring the quality of the Portuguese text is of the utmost importance.

After the stage of the pre-processing is done, the stage of mapping is done to ensure the common variables between dataset are in conformance.

4.2.2. ICD 9 Mapping

The stage of the ICD 9 mapping is important since the variables which contain the ICD 9 codes and designations, "Diagnosis Code" and "Diagnosis Description", respectively, as explained previously, are present in both datasets but show different values and/or formatting. Standardizing these attributes creates a useful link between datasets.

In order to map these attributes, an official list of the ICD 9 codes and descriptions, version 31 was retrieved from ("ICD9 Provider Diagnostic Codes," 2014). The official listing is composed of three columns: Diagnosis Code, Long Description and Short Description. The mapping process is represented in Figure 2.

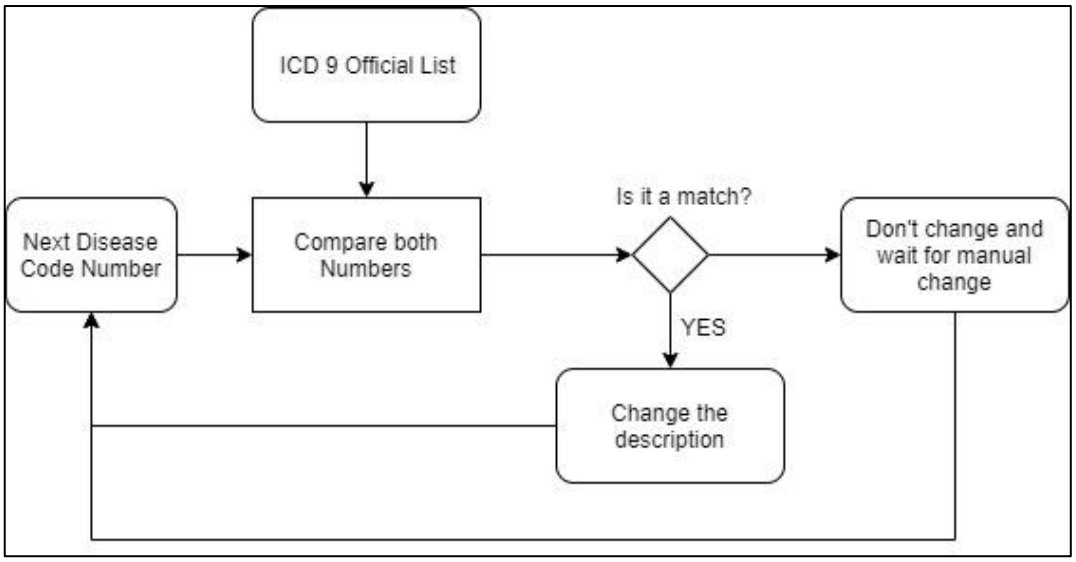


Figure 2 – Mapping Process

This process works entry by entry of the dataset. For each row, it is made a lookup for the "Diagnosis Code" in the official listing. In case there is a match, the field "Diagnosis Description" was overwritten with the respective long description, contained in the official list. In case there is not a match, the author searched manually ("ICD-9-CM Chapters List," 2019) [2] for the code's description and updated the official list.

After the dataset has had its attributes correctly mapped, the dataset is ready to be translated in the next phase of the process. Hence the excel file was imported into a data frame structure using python and Jupyter Notebook.

4.2.3. AD Translation

As the third phase of the AD, dataset processing comes the translation of the Portuguese elements. There are three main reasons for this need for translation of the data.

Firstly, both datasets present attributes in both Portuguese and English. Therefore, to ensure that both datasets are aligned and coherent, the language must be the same. The chosen language was English.

Secondly, to extract the medical entities and information from the free text present in the EMR's an NLP software was used, and this uses the English language as default.

[2] https://icd.codes/icd9cm

The translator used to perform the translation of the attributes was Google Translate. The author developed a python script which used a Google Translate API. The attributes translated using the developed script were “Diary” and “Speciality Description”.

Presented in Figure 3 is an excerpt of the data frame containing the EMRs and, as it is observable, both columns, “Des Especialidade Eng” and “Diario Eng”, contain its respective values translated.

	HDI_EPISODIO	NUM_SEQUENCIAL	DATA	COD_ESPECIALIDADE	DES_ESPECIALIDADE_ENG	COD_DIAGNOSTICO	DES_DIAGNOSTICO	COD_MODULO	DIARIO_ENG
0	17002848	949978	17.04.18	40691	Infecciology	136	Other and unspecified infectious and parasiti...	HDI	Administered Ivermectin 3mg comp - 15mg oral (...
1	17009615	868667	17.11.07	40730	Immunohemotherapy	1028	Latent yaws	HDI	HDI\nPunctured peripheral venous access. Made ...
2	17009092	189370	17.11.07	40695	Rheumatology	69510	Erythema multiforme, unspecified	HDI	Patient Dermatology, with Psoriasis, to perfor...

Figure 3 – AD Dataset After the Translation

After the translation of the aforementioned attributes, the dataset is finally in one language only. In the next phase, the medical diaries are fed to the NLP software. Since said software only processes text files (TXT files), the author developed a python script that created a txt file for each row of the data frame. Shown in Figure 4 is an example of what one of the TXT files would contain before being fed into de NLP software.

```
Systemic lupus erythematosus under Belimumab.
No complaints of lupus activity, referring to fatigue.
Note.
Hemodynamically stable.
No arthritis.
Cardiopulmonary auscultation without changes.
Made analysis today.
```

Figure 4 – Example of the content of a TXT file

4.2.4. cTAKES and UMLS

The cTAKES (clinical Text Analysis and Knowledge Extraction System) consists of a “(…) modular system of pipelined components combining rule-based and machine learning techniques aiming at information extraction from the clinical narrative”(Savova et al., 2010). In other words, cTAKES it is a pipeline of six components that aims to process clinical texts/narratives and enables the extraction of medical information from them. The clinical texts fed to the cTAKES can be in the form of plain texts or clinical document architecture-compliant XML documents (Savova et al., 2010).

The six components that make the cTAKES pipeline are:

- Sentence Bound Detector;
- Tokenizer;
- Normalizer;
- Part-of-Speech (POS) tagger;
- Shallow Parser;
- Named entity recognition (NER) annotator (which includes status and negation annotators).

The Sentence Bound Detector is the first component of the cTAKES pipeline, and its function consists of detecting the ending of sentences, in other words, this element segments a text into sentences. To help further understand how each stage of the cTAKES pipeline works, the sentence presented in Figure 5 shows an example of the various stages of the pipeline. This example is was retrieved from (Savova et al., 2010)[3].

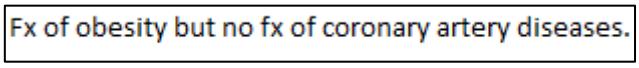


Figure 5 – Initial State of the Sentence to be Processed

The Tokenizer is the element that follows the Sentence Bound Detector and has the goal of splitting the sentences previously detected into tokens (words and punctuations). An example of a tokenized sentence is presented in Figure 6.

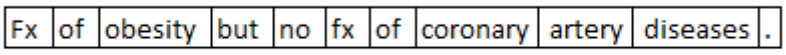


Figure 6 – Tokenized Sentence

After the sentence is split into tokens, the Normalizer is the next element in the cTAKES pipeline. The Normalizer applies a process named stemming. Stemming is the process of "(…) reducing word's inflectional and derivational forms to a common basic form, by performing morphological analysis in texts (…)" (Moreira & Namen, 2018). Presented in Figure 7 is an example of a sentence after the Normalizer. Highlighted, in

[3] "Fx" stands for "Family History".

blue, is the cell that contains the word “disease”, which shows how the word “diseases” was transformed into its basic form.

Fx	of	obesity	but	no	fx	of	coronary	artery	disease	.

Figure 7 – Normalized Sentence

The next step in the pipeline is the POS Tagger. Part-of-Speech tagging is the capability of a computer to understand the meaning of a word according to its context, in other words, it classifies the words of a sentence as a noun, a verb, an adverb, etc. Meaning it could distinguish the meaning of, for example, “flies” as a verb (to fly) or as a noun (plural of fly, the animal). Presented in Figure 8 is an example of the POS Tagging of a sentence. In this example the POS Tagger identified all elements of the sentence according to its morphological class, e.g. in this scenario “obesity”, “fx” and “artery” were classified as Nouns (NN), while “coronary” was classified as an Adjective (JJ), and so on.

Fx	of	obesity	but	no	fx	of	coronary	artery	diseases	.
NN	IN	NN	CC	DT	NN	IN	JJ	NN	NNS	.

Figure 8 – Output of the POS Tagger Component

The Shallow Parser has the responsibility of analyzing the tokens classified by the POS Tagger and, according to the context of the sentence, associate them into logical groups. Shown in Figure 9 is an example of how the shallow parser would analyze a sentence. In this scenario, out of the logical groups detected, those that contain more than one component are the Noun Phrases (NP) with “no fx” (no family history) and “coronary heart diseases”.

Fx	of	obesity	but	no	fx	of	coronary	artery	diseases	.
NP	PP	NP		NP		PP	NP			.

Figure 9 – Shallow Parser Output

Finally, the last element of the cTAKES pipeline is the Name Entity Recognition (NER) Annotator.

This element works with the help of a dictionary and allows for detection of words and clinical information, through its dictionary lookup algorithm. The dictionary used by

cTAKES can be configured according to its clinical terms and relationships. The algorithm gathers all the detected entities by the previous elements of the pipeline and uses the dictionary to map a concept to the detected entity. The detected clinical terms can be categorized into five distinct groups:

- Disorders/diseases;
- Signs/symptoms;
- Procedures;
- Anatomical Sites;
- Drugs/medications.

Furthermore, this component can also detect if the identified entities are negated in the sentences it processes. Presented in Figures 10 and 11 are examples of the output of the Name Entity Recognition Annotator regarding the identification of the clinical entities, as of its negation status.

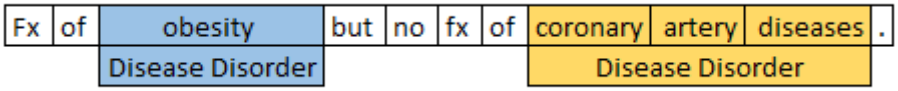


Figure 10 – NER Output with Detected Medical Entities

In this case were identified five medical entities, two Diseases/Disorders. As for the negation analysis, the Figure 11 shows that in both cases was correctly identified, family history of obesity was classified and "not negated", while the family history of coronary artery diseases was classified is "negated".

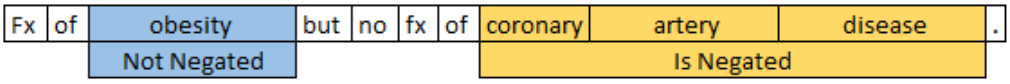


Figure 11 – NER Output with Negation Status

Since the cTAKES generic form of operating has been explained, follows a more specific explanation of its two modes available. The cTAKES software, as aforementioned, presents two different modes in which it operates, the CAS Visual Debugger (CVD) and the Collection Processing Engine (CPE). The difference between these two modes rests in the number of documents it can analyze at a time. While the first

only processes a document at a time focusing on its display, the latter was developed to process collections of documents.

Presented in Figure 12 is the print of the CVD menu. This mode is more indicated to visualize the analysis performed by the cTAKES software. The menu is composed of two different screens: the text area and the NLP analysis result area split into two, where said analysis is presented as a tree. Looking at the example presented, it is visible that this mode is good to have a more visual image of the process.

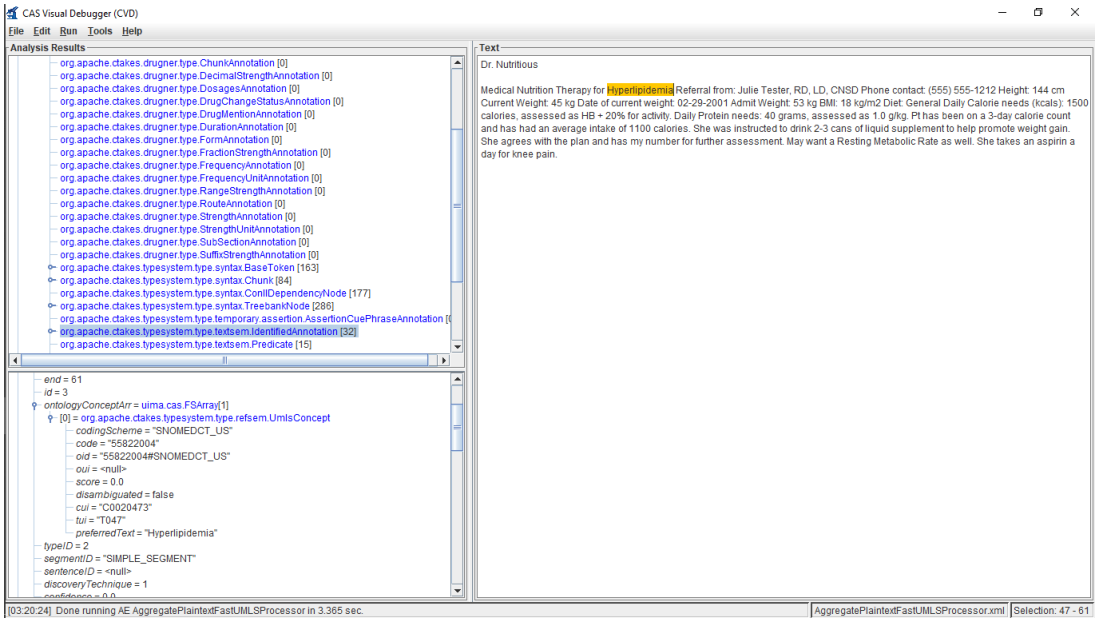


Figure 12 – cTAKES CVD Menu

In the left screens shows that cTAKES detected a disease mention which is highlighted in the right screen with the value “Hyperlipidemia” which is indeed a disease mention. Nevertheless, for only one text document, it took 3.365 seconds to process and analyze. Since there are 11.134 text files to be processed, the CVE would take a larger amount of time to process everything. Hence, the CPE is the model used in this work.

The CPE model of the cTAKES software, contrarily to the CVD, does not allow for visual analysis of the document, as it can be seen in Figure 13. This mode’s menus do not allow for more visual analysis of the NLP processing since it aims to process collections of documents instead of a document at a time.

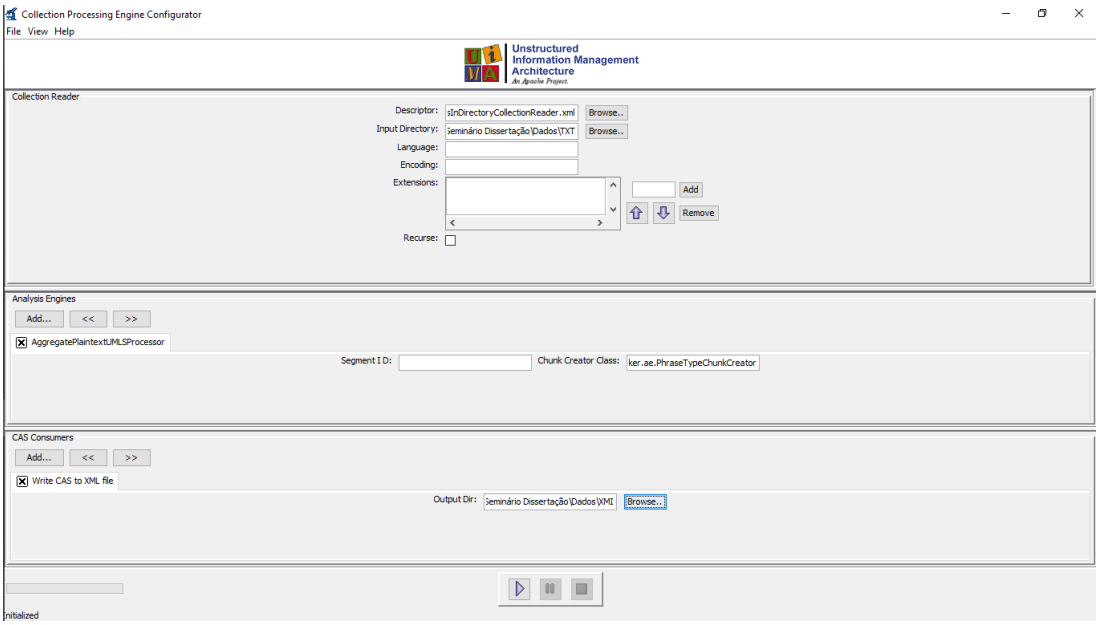


Figure 13 – cTAKES CPE Menu

This mode's menu is composed of three different screens: the collection of the reader, the analysis engine, and the CAS consumer. The first element allows the user to choose the element to be processed, it can from a directory to a database.

The second component, the analysis engine, is where the user selects the type of engine to be used according to its needs. It can go from a simple NLP analysis with normalizations and part-of-speech tagging to a full detection of medical entities in the text. For the latter scenario, cTAKES uses as a clinical base of knowledge of the Unified Medical Language System (UMLS).

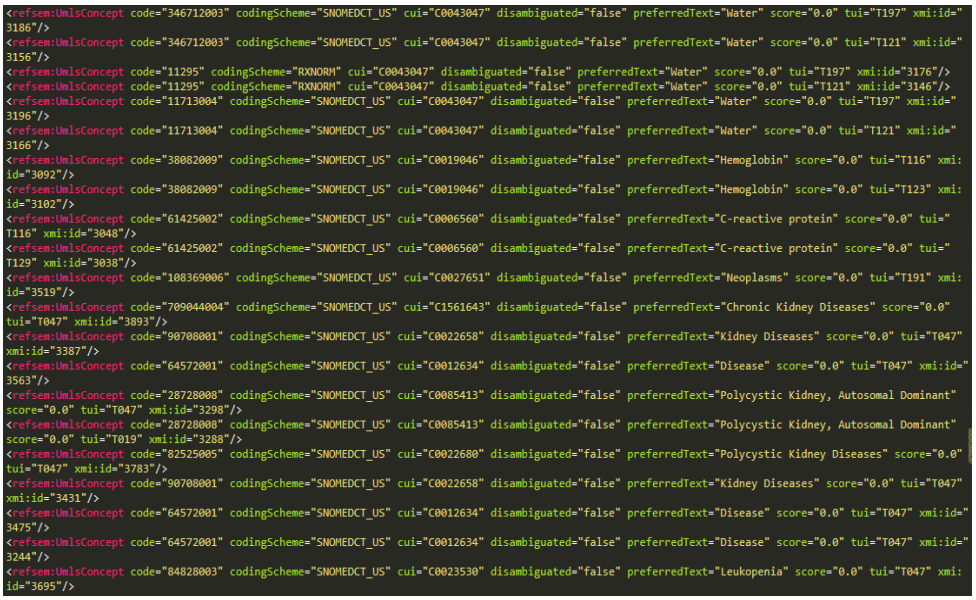
The UMLS is "a repository of biomedical vocabularies developed by the US National Library of Medicine. The UMLS integrates over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts" (Bodenreider, 2004) which serves as a support to the cTAKES software. The UMLS contains three different knowledge sources:

- Metathesaurus – contains terms and codes form many vocabularies, hierarchies, definitions and other relationships and attributes, and is the main module. Some of the vocabularies contained in this knowledge source are MedSH (Medical Subject Headings), RxNorm and SNOMED CT;
- Semantic Network – has broad categories (semantic types) and their relationships (semantic relations);

- SPECIALIST Lexicon and Lexical Tools – is a large syntactic lexicon of, not only, biomedical but also general English and some tools.

In order to manage the identification and extraction of the identified entities in the medical diaries, a dictionary is required. Said dictionary, by default, is a subsection of the Metathesaurus which used two of its vocabularies, the RxNorm, and the SNOMED CT.

Lastly, the third component is where the data type of the output is select. There are three data types the cTAKES can have as output, XML, XMI (XML Metadata Interchange) or HTML. The chosen output format was XMI because it is relatively easy to manipulate using Python scripts. Presented in Figure 14 is an excerpt of one of the XMI file outputted from the cTAKES.



```
<refsem:UmlsConcept code="346712003" codingScheme="SNOMEDCT_US" cui="C0043047" disambiguated="false" preferredText="Water" score="0.0" tui="T197" xmi:id="3186"/>
<refsem:UmlsConcept code="346712003" codingScheme="SNOMEDCT_US" cui="C0043047" disambiguated="false" preferredText="Water" score="0.0" tui="T121" xmi:id="3156"/>
<refsem:UmlsConcept code="11295" codingScheme="RXNORM" cui="C0043047" disambiguated="false" preferredText="Water" score="0.0" tui="T197" xmi:id="3176"/>
<refsem:UmlsConcept code="11295" codingScheme="RXNORM" cui="C0043047" disambiguated="false" preferredText="Water" score="0.0" tui="T121" xmi:id="3146"/>
<refsem:UmlsConcept code="11713004" codingScheme="SNOMEDCT_US" cui="C0043047" disambiguated="false" preferredText="Water" score="0.0" tui="T197" xmi:id="3196"/>
<refsem:UmlsConcept code="11713004" codingScheme="SNOMEDCT_US" cui="C0043047" disambiguated="false" preferredText="Water" score="0.0" tui="T121" xmi:id="3166"/>
<refsem:UmlsConcept code="38082009" codingScheme="SNOMEDCT_US" cui="C0019046" disambiguated="false" preferredText="Hemoglobin" score="0.0" tui="T116" xmi:id="3092"/>
<refsem:UmlsConcept code="38082009" codingScheme="SNOMEDCT_US" cui="C0019046" disambiguated="false" preferredText="Hemoglobin" score="0.0" tui="T123" xmi:id="3102"/>
<refsem:UmlsConcept code="61425002" codingScheme="SNOMEDCT_US" cui="C0006560" disambiguated="false" preferredText="C-reactive protein" score="0.0" tui="T116" xmi:id="3048"/>
<refsem:UmlsConcept code="61425002" codingScheme="SNOMEDCT_US" cui="C0006560" disambiguated="false" preferredText="C-reactive protein" score="0.0" tui="T129" xmi:id="3038"/>
<refsem:UmlsConcept code="108369006" codingScheme="SNOMEDCT_US" cui="C0027651" disambiguated="false" preferredText="Neoplasms" score="0.0" tui="T191" xmi:id="3519"/>
<refsem:UmlsConcept code="709044004" codingScheme="SNOMEDCT_US" cui="C1561643" disambiguated="false" preferredText="Chronic Kidney Diseases" score="0.0" tui="T047" xmi:id="3893"/>
<refsem:UmlsConcept code="90708001" codingScheme="SNOMEDCT_US" cui="C0022658" disambiguated="false" preferredText="Kidney Diseases" score="0.0" tui="T047" xmi:id="3387"/>
<refsem:UmlsConcept code="64572001" codingScheme="SNOMEDCT_US" cui="C0012634" disambiguated="false" preferredText="Disease" score="0.0" tui="T047" xmi:id="3563"/>
<refsem:UmlsConcept code="28728008" codingScheme="SNOMEDCT_US" cui="C0085413" disambiguated="false" preferredText="Polycystic Kidney, Autosomal Dominant" score="0.0" tui="T047" xmi:id="3298"/>
<refsem:UmlsConcept code="28728008" codingScheme="SNOMEDCT_US" cui="C0085413" disambiguated="false" preferredText="Polycystic Kidney, Autosomal Dominant" score="0.0" tui="T019" xmi:id="3288"/>
<refsem:UmlsConcept code="82525005" codingScheme="SNOMEDCT_US" cui="C0022680" disambiguated="false" preferredText="Polycystic Kidney Diseases" score="0.0" tui="T047" xmi:id="3783"/>
<refsem:UmlsConcept code="90708001" codingScheme="SNOMEDCT_US" cui="C0022658" disambiguated="false" preferredText="Kidney Diseases" score="0.0" tui="T047" xmi:id="3431"/>
<refsem:UmlsConcept code="64572001" codingScheme="SNOMEDCT_US" cui="C0012634" disambiguated="false" preferredText="Disease" score="0.0" tui="T047" xmi:id="3475"/>
<refsem:UmlsConcept code="64572001" codingScheme="SNOMEDCT_US" cui="C0012634" disambiguated="false" preferredText="Disease" score="0.0" tui="T047" xmi:id="3244"/>
<refsem:UmlsConcept code="84828003" codingScheme="SNOMEDCT_US" cui="C0023530" disambiguated="false" preferredText="Leukopenia" score="0.0" tui="T047" xmi:id="3695"/>
```


Figure 14 – cTAKES XMI Output Excerpt

Through its analysis, it is possible to observe all the entities that were detected in one document. Each detected entity is characterized with several attributes: (1) code, (2) coding scheme, (3) CUI, (4) disambiguated, (5) preferred Text, (6) score, (7) TUI and (8) XMI id. Focusing on the most relevant, the coding scheme identifies the vocabulary from where the term concept was select (RxNorm and SNOMED CT). The CUI (Concept Unique Identifier), as the name implies, shows the concept unique number. Preferred Text contains the concept detected in the text. From the cTAKES and UMLS pipeline comes out one XMI file for each TXT file that entered. In other words, the outcome of the pipeline is a collection of XMI files.

At the moment the unstructured dataset is divided between the XMI file collection where the medical entities related to the clinical diaries and the excel dataset with the remaining attributes such as medical specialities, date, diagnostics, among others.

In order to structure everything into a single database table, a python script was developed by the author. This script resulted in a CSV (Comma Separated Values) file with 81.103 rows.

Chapter 5 – Structured and Mixed Data

This chapter follows a structure similar to the previous one, but in this case, instead of explaining the unstructured data, it is the structured and mixed data that are approached. This chapter is organized in two subchapters: (1) Structured Data and (2) Mixed Data.

5.1 Structured Data

5.1.1 Structured Dataset Description

The structured dataset used in this work contains information regarding the Emergency Department (ED) of a Portuguese hospital. This dataset contains data regarding said department between the dates of January 1st of 2015 and December 31st of 2017.The aforementioned data has been extracted from the hospital's database in the form of an excel file, and it contains 108.295 records and 19 attributes. The attributes, its data types and some examples are represented in Table IX.

Table IX – ED Dataset Attributes

Attribute #	***Attribute Name***	***Data Type***	***Example***
1	Episode Alert	Integer	44172901
2	Cod Epis Type Ext	String	URG
3	Patient	String	D84
4	Department	String	SU_Geral
5	Admission Date	Timestamp	16/02/2017 09:42:32
6	Triage Colour	String	(4) GREEN
7	Episode Sonho	Integer	17020483
8	Triage Date	Timestamp	16/02/2017 09:55:17
9	Nurse	String	E5
10	Date First Medical Observation	Timestamp	16/02/2017 10:45:21
11	Doctor (1st Observation)	String	M34
12	Medical Discharge Date	Timestamp	16/02/2017 19:29:09
13	Discharge Doctor	String	M34
14	Administrative Discharge Date	Timestamp	16/02/2017 20:21:08
15	Discharge State	Char (1)	A
16	Destination	String	Family Doctor - Unspecified Health Centre
17	Readmission	Integer	6312104
18	ICD Code	Double	873.5
19	ICD Description	String	Complicated Open Wound of Face

Each entry of the file represents an interaction of a patient with the ED. Some entries represent the same interaction but with different diagnostic codes and descriptions due to the identification of multiple diagnostics. The ED flow can be resumed in five simple steps, admission, triage, observation, discharge, and administrative discharge.

It all begins when a patient enters the ED and is admitted by the service and waits for further evaluation (admission). The next step, the triage, consists of an evaluation of the patient, performed by a nurse, and results in the patient's classification according to the Manchester Triage System (MTS). This triage system "(…) enables nurses to assign a clinical priority to patients, based on presenting signs and symptoms, without making any assumption about the underlying diagnosis. The MTS allocates patients to one out of five urgency categories, which determine the maximum time to first contact with a physician", being one of the most common triage systems used in Europe (Zachariasse et al., 2017). Presented in Table X are the possible classifications of this system and respective colour and time.

Table X – Manchester Triage System Classification and Times

Priority #	***Urgency Level***	***Code/Colour***	***Target Time to See the Patient (minutes)***
1	Immediate	Red	0
2	Very Urgent	Orange	10
3	Urgent	Yellow	60
4	Standard	Green	120
5	Non-Urgent	Blue	240

The third step of this process is the observation that corresponds to the first medical observation from a doctor when a diagnose is assessed.

The fourth step is the discharge, whether to home, to another department of the hospital or even to another facility, this step removes the patient from the ED.

Finally, the fifth step corresponds to the evaluation and acceptance of the paperwork related to step four. As aforementioned, the structured dataset contains 19 variables, all presented in Table IX.

The “Episode Alert” and “Episode Sonho” consist of simple identifiers of the specific encounter. The “Patient”, “Nurse”, “Doctor (1st Observation)” and “Discharge Doctor” are identifiers of the patients, nurses, and doctors involved. This identification is necessary as a privacy measure since the data in hand is relative to the health of people.

The “Department” represents the local where the event happened and is written in Portuguese. The “Cod Epis Type Ext” also represents the location of where the event happened.

The “Admission Date”, “Triage Date”, “Date First Medical Observation”, “Medical Discharge Date” and “Administrative Discharge Date” are, as the name implies, dates recorded as timestamps (dd/mm/yyyy hh:mm:ss) and represent, respectively, the date and time at which the patient was admitted, when it went through triage, when the first observation was made to the patient from a doctor, the date at which the patient got the discharge clearance and the date at which the medical discharge papers were accepted.

The “Triage Colour” represents the category in which the nurse considered the patient according to the MTS, explained above, and is also written in Portuguese.

The “Discharge State” is an attribute and character that works as a flag, which shows the value “A” when the patient has been discharged. The “Destination” shows the destination to where the patient was discharged, it also is written in Portuguese.

The “Readmission” is a numeric attribute that represents the readmission number of the encounter. Otherwise, the attribute is empty.

Finally, the disease codes are represented as numeric codes the, “ICD Code”, and its description “ICD Description”. All these descriptions and codes are from International Code of Diseases 9 (ICD 9), which were explained in the previous chapter.

The attributes which are written in Portuguese in the original dataset were translated to English, so the example of data presented in Table IX could be presented.

5.1.2 Structured Data Processing

The processing of the structured dataset is simpler when comparing to the unstructured dataset since the ED data is already structured since it requires less processing until its ready to be used. Nevertheless, some processing is required to clean the data and standardize some attributes.

5.1.2.1 ED Pre-Processing

The pre-processing phase of the ED dataset shows a similarity with the pre-processing phase of the AD dataset. Hence all the steps have already been explained previously.

In the first phase, all rows that had any attribute fields empty got removed to avoid the existence of "Nulls" and "N/A" further ahead in the process. Secondly, the duplicated rows were removed to avoid the duplication of information, row-wise.

The next stage, just like it was done with the AD dataset, the removal of unnecessary attributes was done. Presented in Table XI are the attributes that got eliminated, its number, and the reason behind its removal.

Table XI – ED Removed Attribute and Justification

Attribute #	***Attribute Name***	***Data Type***	***Reason***
2	Cod Epis Type Ext	String	One Value
4	Department	String	One Value
7	Episode Sonho	Integer	Identifier Column
15	Discharge State	Char (1)	One Value

The attributes "Cod Epis Type Ext", "Department" and "Discharge State" were both removed since each one presented only one value along all rows, not adding any insight or quality to the dataset. The variable "Episode Sonho" was removed since it was another identifier of each episode and the variable "Episode Alert" was already being used as an encounter identifier.

After cleaning the dataset from problematic rows and unusable columns, the attribute "Destination" was translated manually. The reason behind the manual translation resides on the fact that most of the values refer to locations, and some of them contained geographic locations that would be wrongly translated.

After the translation phase, the pre-processing of the ED dataset is concluded, and the ICD mapping was implemented.

5.1.2.2 ICD 9 Mapping

The way the mapping of the ICD 9 codes and descriptions was made is the exact same process as it was done with the AD dataset, pictured in Figure 2, as shown in chapter 4.2.2. This way, both datasets are in conformance in these two attributes.

5.1.2.3 Data Exploring

Continuing the processing of the structured dataset, the author explored the dataset, analyzing it, and seeing what could be done to improve its quality and manageability. For this exploitation, the author used the R language through the RStudio software.

Contemplating the variables in hand, there were several actions that could be made to improve the dataset, such as, creation of new variables based on the ones already present, changing the nature of some attributes, etc.

Looking at the dataset, it is immediately noticeable the presence of timestamp variables. Considering the structured data is about an emergency department of a hospital, knowing the amount of time that the patients must wait during the overall process of the emergency room is important. Hence, using the variables "Admission Date", "Triage Date", "Date First Observation", "Medical Discharge Date" and "Administrative Discharge Date" five new variables were created: "Time Until Triage", "Time Until First Obs", "Time Until Medical Discharge", "Time Until Admin Discharge" and "Days In Hospital". Presented in Table XII are shown the new variables the calculating operations and data types.

Table XII – Created Time Variables

Attribute #	***Attribute Name***	***Data Type / Unit***	***Operation (Attribute Numbers)***[4]	***Example***
20	Time Until Triage	Integer/Minutes	(8) – (5)	22
21	Time Until First Obs	Integer/Minutes	(10) – (8)	223
22	Time Until Medical Discharge	Integer/Minutes	(12) – (10)	18
23	Time Until Admin Discharge	Integer/Minutes	(14) – (12)	21
24	Days in Hospital	Integer/Days	(14) – (5)	0

[4] The attribute numbers used refer to the attribute numbers of table IX

Each attribute, from 20 to 23, account for the time each step of the triaging process took, while the “Days in Hospital” account for the time the overall event took. With the creation of these new variables, it was possible to detect more rows that suffered from inconsistencies. This problem accrues from the negative time values that some of these new attributes were presenting. Since timespans cannot show negative values, those rows were eliminated.

After creating new variables, two other attributes were worked on, “Triage Colour” and “Readmission”.

So far, the “Triage Colour” was a string attribute which contained a number and a text. The number was related to the colour code with which the patient had been tagged and the text as the colour in written in Portuguese, e.g. “(4) VERDE”, which stand for “(4) GREEN”. To simplify the variable, the text part of this attribute was removed, keeping it only the numeric chunk. Hence, the data type of this variable changed from string to integer.

Regarding the attribute “Readmission”, as explained previously, if an entry in the emergency department is readmission of a patient, then said entry would be represented with an integer identifier. To increase the quality of this attribute, the author changed its nature, shifting the type of the variable, from an integer to a flag. The flag variables only present two values, 0 and/or 1. In this scenario, the rows which were readmissions changed its numeric value to 1, the remains rows kept the 0.

After all the changes to the structured dataset had been made, using RStudio, the dataset was formatted into a CSV file to facilitate its integration into the database. In the next chapter, how the structured and unstructured data were imported into the database, and its structure is explained in greater detail.

5.2 Mixed Data

After both datasets were fully processed, the final results were two CSV files, one for each. In order to facilitate the integration process, the CSV files were inserted into tables in a database. For this work, the selected database tool was SQLite Studio software. The SQLite Studio allows the creation and managing of SQLite databases. SQLite implements a small, fast, high-reliability, full-featured SQL database engine. The author opted for the SQLite studio since this tool allows the creation of a simple database that can be fast and

efficient and because that dataset used are not very large. SQLite3 also allows for the importing data through python. Which is what was done.

The developed database is comprised of two different tables, AD_SQLite and ED_SQLite, for the AD and ED datasets, correspondingly. The ED_SQLite contains the 20 attributes that have been explained throughout the chapter 5.1.

The AD_SQLite, on the other hand, has eight attributes: “Entity Value”, “Entity Type”, “Medical Speciality Code”, “Medical Speciality Description”, “Diagnosis Description”, “Diagnosis Code “, “Date” and “File ID”.

The first two attributes contain the entities found by the cTAKES and UMLS pipeline and its type (medication, anatomical site, etc.). The “File ID” presents the name of the text file from where the entities were identified. The remaining variables were previously explained in subchapter 4.1. Presented in Figures 15 and 16 are the excerpts of the ED_SQLite and AD_SQLite tables, accordingly.

	Episode Alert	Patient ID	Admission Date	Triage Color	Triage Date	Nurse ID	Date First Obs	Doctor First Observation ID	Medical Discharge Date	Doctor Discharge ID
1	45160894	D15087	2017-04-07 23:56:53	3	2017-04-08 00:19:11	E30	2017-04-08 04:02:07	M371	2017-04-08 04:19:50	M371
2	45296887	D3524	2017-04-14 20:23:39	3	2017-04-14 20:25:30	E49	2017-04-14 20:46:49	M371	2017-04-14 23:57:08	M371
3	46438424	D26681	2017-06-28 21:21:21	3	2017-06-28 21:28:38	E20	2017-06-28 22:25:32	M4	2017-06-29 04:55:17	M371
4	46500886	D30519	2017-07-04 02:06:17	4	2017-07-04 02:13:57	E20	2017-07-04 05:09:47	M371	2017-07-04 06:50:19	M371
5	48901903	D44245	2017-10-30 16:57:09	4	2017-10-30 17:11:16	E41	2017-10-30 23:42:28	M371	2017-10-31 03:56:32	M371
6	48615387	D45025	2017-10-10 03:26:23	3	2017-10-10 03:28:56	E30	2017-10-10 04:43:01	M82	2017-10-10 06:00:51	M82
7	49118406	D49134	2017-11-17 21:39:22	3	2017-11-17 21:43:46	E48	2017-11-17 22:42:22	M525	2017-11-18 02:05:20	M525
8	43972887	D5908	2017-02-04 13:37:48	3	2017-02-04 13:39:28	E26	2017-02-04 14:03:15	M79	2017-02-04 16:25:52	M79
9	45225888	D18355	2017-04-12 02:35:32	3	2017-04-12 02:38:00	E39	2017-04-12 02:49:29	M87	2017-04-12 06:15:21	M232
10	45225889	D18356	2017-04-12 02:36:29	3	2017-04-12 02:40:26	E39	2017-04-12 02:58:19	M87	2017-04-12 06:13:22	M232
11	46049904	D14615	2017-05-30 07:52:39	2	2017-05-30 07:54:19	E11	2017-05-30 07:57:51	M317	2017-05-30 12:13:18	M164
12	46392889	D30865	2017-06-25 07:40:38	4	2017-06-25 07:43:50	E23	2017-06-25 10:07:53	M195	2017-06-25 15:12:46	M195
13	47527889	D31677	2017-08-27 16:01:20	4	2017-08-27 16:04:54	E12	2017-08-27 18:54:28	M195	2017-08-27 22:43:30	M195
14	46817997	D32766	2017-07-23 16:09:31	4	2017-07-23 16:22:09	E3	2017-07-23 19:49:01	M195	2017-07-23 20:12:37	M195
15	47467905	D39203	2017-08-25 17:23:01	3	2017-08-25 17:26:41	E22	2017-08-25 17:55:17	M11	2017-08-25 18:20:17	M11
16	48425889	D41964	2017-10-02 09:33:40	4	2017-10-02 09:37:31	E20	2017-10-02 14:56:08	M31	2017-10-02 15:17:38	M31
17	48670104	D14512	2017-10-14 23:30:15	3	2017-10-14 23:35:22	E20	2017-10-15 00:09:16	M57	2017-10-15 07:33:57	M112
18	48670104	D14512	2017-10-14 23:30:15	2	2017-10-14 23:35:22	E20	2017-10-15 00:09:16	M57	2017-10-15 07:33:57	M112
19	48685386	D46528	2017-10-16 05:17:08	4	2017-10-16 05:20:41	E20	2017-10-16 09:24:51	M31	2017-10-16 09:48:13	M31
20	49593955	D51645	2017-12-14 10:30:03	3	2017-12-14 10:34:12	E66	2017-12-14 10:46:49	M11	2017-12-14 11:44:02	M11
21	48892930	D3450	2017-10-30 09:53:14	5	2017-10-30 10:02:37	E11	2017-10-30 14:08:54	M31	2017-10-30 14:41:32	M31
22	46074002	D5007	2017-06-01 17:06:21	3	2017-06-01 17:13:42	E16	2017-06-01 18:49:06	M39	2017-06-02 15:27:28	M153
23	46074002	D5007	2017-06-01 17:06:21	3	2017-06-01 17:13:42	E12	2017-06-01 18:49:06	M39	2017-06-02 15:27:28	M153
24	46631394	D30374	2017-07-12 22:26:13	3	2017-07-12 22:29:39	E41	2017-07-12 22:48:07	M69	2017-07-15 00:36:16	M432

Figure 15 – Excerpt ED_SQLite Table

	Entity Type	Entity Value	Medical Specialty Code	Medical Specialty Description	Diagnosis Description	Diagnosis	Dat
1	Medication	Ivermectin	40691	Infecciology	Infectious And Parasitic Diseases, Nop Or Not Specified	136	201
2	Medication	Oral Dosage Form	40691	Infecciology	Infectious And Parasitic Diseases, Nop Or Not Specified	136	201
3	AnatomicalSite	Oral cavity	40691	Infecciology	Infectious And Parasitic Diseases, Nop Or Not Specified	136	201
4	AnatomicalSite	Veins	40730	Immunohemotherapy	Latent Yaws	1028	201
5	Procedure	Interventional procedure	40730	Immunohemotherapy	Latent Yaws	1028	201
6	SignSymptom	Services	40730	Immunohemotherapy	Latent Yaws	1028	201
7	SignSymptom	Chief complaint (finding)	40730	Immunohemotherapy	Latent Yaws	1028	201
8	Medication	rituximab	40695	Rheumatology	Lupus Erythematosus	6954	201
9	Medication	prednisolone	40695	Rheumatology	Lupus Erythematosus	6954	201
10	DiseaseDisorder	Lupus Erythematosus	40695	Rheumatology	Lupus Erythematosus	6954	201
11	Medication	Today	40694	Nephrology	Chronic Renal Insufficiency	585	201
12	Medication	Hemoglobin	40694	Nephrology	Chronic Renal Insufficiency	585	201
13	Medication	Urea	40694	Nephrology	Chronic Renal Insufficiency	585	201
14	Medication	Creatinine	40694	Nephrology	Chronic Renal Insufficiency	585	201
15	AnatomicalSite	Kidney	40694	Nephrology	Chronic Renal Insufficiency	585	201
16	AnatomicalSite	Lymphatic vessel	40694	Nephrology	Chronic Renal Insufficiency	585	201
17	AnatomicalSite	Veins	40694	Nephrology	Chronic Renal Insufficiency	585	201
18	AnatomicalSite	Internal jugular vein structure	40694	Nephrology	Chronic Renal Insufficiency	585	201
19	AnatomicalSite	Structure of jugular vein	40694	Nephrology	Chronic Renal Insufficiency	585	201
20	AnatomicalSite	Blood	40694	Nephrology	Chronic Renal Insufficiency	585	201
21	AnatomicalSite	Leukocytes	40694	Nephrology	Chronic Renal Insufficiency	585	201
22	AnatomicalSite	neutrophil	40694	Nephrology	Chronic Renal Insufficiency	585	201
23	Procedure	Dialysis procedure	40694	Nephrology	Chronic Renal Insufficiency	585	201
24	Procedure	Therapeutic procedure	40694	Nephrology	Chronic Renal Insufficiency	585	201

Figure 16 – Excerpt AD_SQLite Table

Chapter 6 – Results and Discussion

This chapter aims to demonstrate the attained results during this work. With this intent, the author used the Power BI software to illustrate through tables and graphs the results achieved.

Power BI is a business analytics solution that allows the viewing of data through dynamic dashboards and reports. Considering that the data is stored in an SQLite database, this tool is also very useful because it allows the migration of the data from the database into Power BI through an Open Database Connection (ODBC) which is a safe way to transfer information.

This chapter is organized in three separate subchapters, each regarding the results of a specific dataset: (1) Structured Data Results, (2) Unstructured Data Results, and (3) Mixed Data Results. The initial two chapters' purpose is to show that both the AD dataset and ED dataset allow the extraction of information on its own, while the third chapter shows what kind of results are obtained from the mixed usage of these datasets.

6.1 Structured Data Results

The structured data results consist of a set of graphs and tables that result from statistical analysis of the ED dataset. To better understand the affluence of the emergency department of this Portuguese hospital through the year 2017, the graph presented in Figure 17 was made.

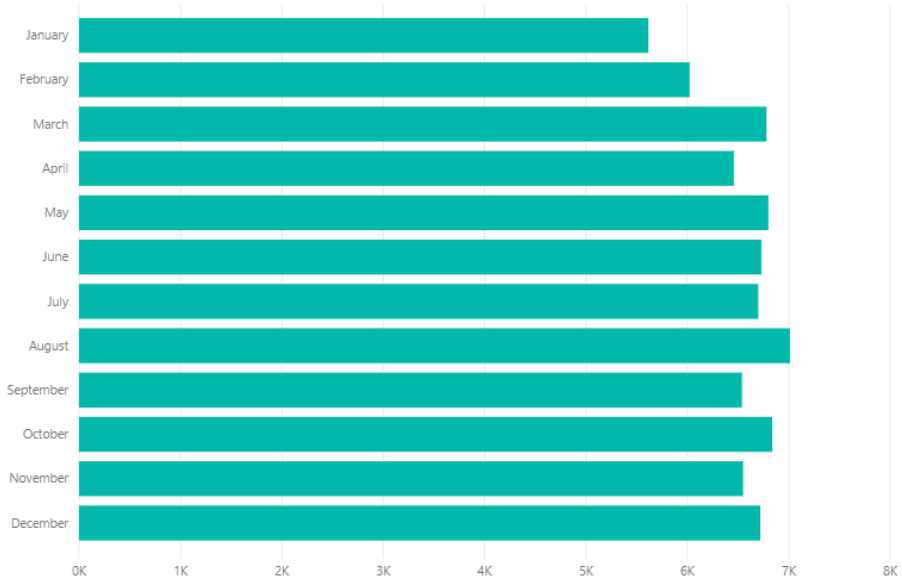


Figure 17 – Monthly Patient Affluence Distribution

Through its analysis, it is clear that the Top 3 months with the most patients to be admitted to the emergency department are August, October and March. In contrast, the months with the least number of patients are January, February and April.

To deepen this analysis, it was verified how the hospital's emergency department admittances are distributed amongst the various classifications of the screening system used.

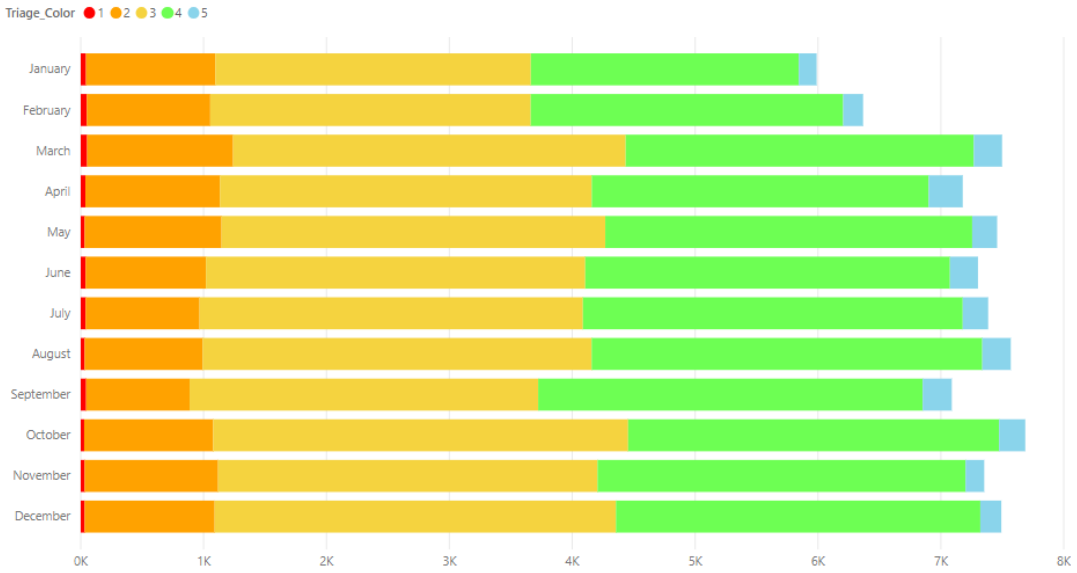


Figure 18 – Distribution of Diseases according to Triage Classification

Analyzing Figure 18, one fact is immediately verified, the amount of entries each month contains is superior to the amount shown in Figure 17. This is because, as previously explained in chapter 5.1, there are records that, despite being regarding the same patient, i.e. having the same identifier (Episode Alert attribute), show different diagnoses. In other words, for some admissions, multiple diagnoses were recorded. Hence, the difference between the number of entries in both graphs. Progressing with the analysis, it is perceived that although the month that received more patients was August, the most illnesses were registered in October. Followed by August and March. Throughout the year it should be noted that the vast majority of the cases that appeared in the emergency room of this hospital are categorized as urgent (yellow) or standard (green). The month in which most entries were appearing as very urgent (red) is March.

An important set of an emergency department is the waiting time since it can mean the difference between losing or saving a patient’s life. Following this path, presented in Figure 19 is a graph that evaluates the average waiting time (this measure was calculated

with the help of Power BI, and it is the sum of the attributes Time Until Triage and Time Until First Observation) throughout the year.

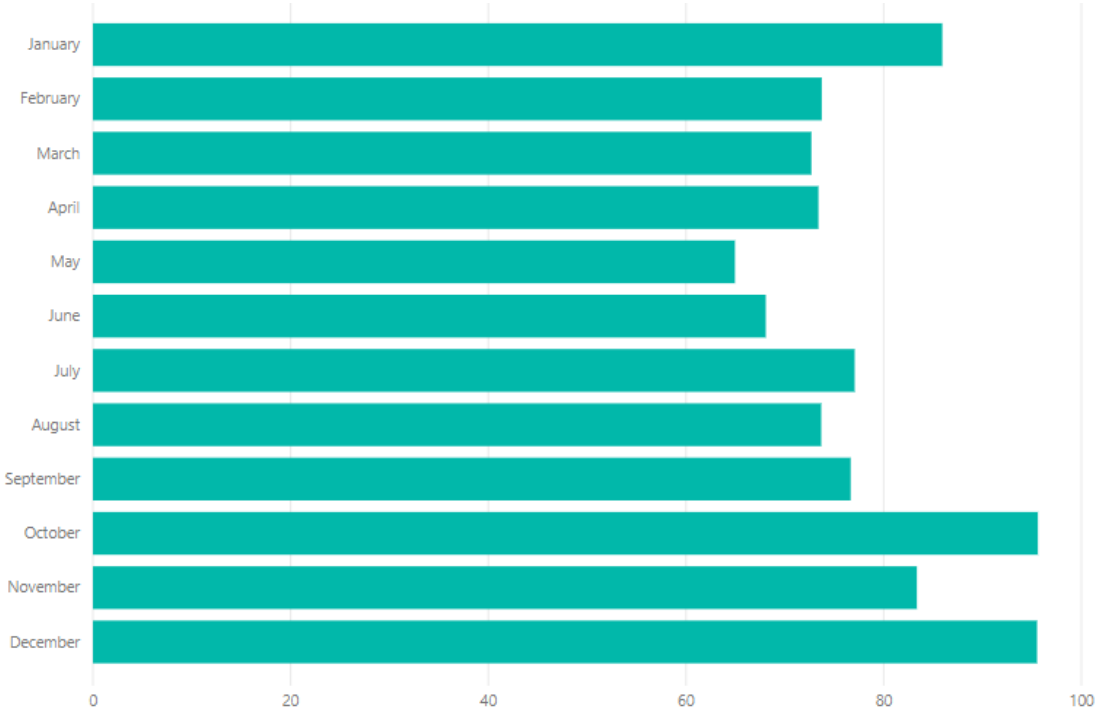


Figure 19 – Average Waiting Time per Month

Interestingly enough, one of the highest Waiting Time average month is January, even though it is the month with the least amount of patient records. Whilst the other two months with higher average waiting times are October and December.

As for the type of diseases that appear in this department, Figure 20 shows the top 5 most common diseases during the year of 2017.

ICD_9_Description	ICD_9_Code	Count of ICD_9_Code
Lumbago	7242	3769
Urinary tract infection, site not specified	5990	3077
Abdominal pain	7890	2979
Chest pain	7865	1811
Head injury, unspecified	95901	1575
Total		**13211**

Figure 20 – Top 5 Diseases in the Emergency Department

Because of this analysis, it can be concluded that diagnostics most commonly presented in this hospital's ED are "Lumbago", "Urinary Tract Infections", "Abdominal Pain", "Chest Pain" and "Head Injuries". As it can be checked, all the diagnosis codes contain at least four digits which are good since it means that some specificity is given to each case that appears. Also, by the description, such as "Urinary tract infection, site not

specified" shows that that site of the infection is not specified, whilst if the site was known, another code and description would be given.

The analysis presented in this chapter had the finality of showing that the structured data contained in the ED dataset allowed for an exploratory analysis across a broad number of angles. Having this dataset prepared for information extraction was an important step in order to facilitate the extraction of the mixed data further ahead. In the next subchapter, a similar analysis is performed, yet it is done in the structured information extracted from EMR's.

6.2 Unstructured Data Results

This chapter focusses on the presentation of the obtained results using only the data present in the AD _SQLite table. In other words, only the structured information extracted from the unstructured data of the EMR is analyzed.

For a first analysis, it has been developed the following table, Figure 21, where all detected entities are counted. As it can be analyzed, from the five categories that the cTAKES classify an entity, the most detected category is "SignSymptom", followed by "Medication" and "Procedure".

Entity_Type	Count of Entity_Type
SignSymptom	24345
Medication	18209
Procedure	18145
AnatomicalSite	14945
DiseaseDisorder	5459
Total	**81103**

Figure 21 – Count of All Entities Found Grouped by Type

Before understanding how these entity types are distributed through the medical specialities, it was ascertained how many files existed initially of each speciality. Presented in Figure 22 is the table with that analysis.

Medical_Specialty_Description	Contagem de File_ID
Medical Oncology	2885
Pain unit	1966
Hematology	1608
Infecciology	1363
Immunohemotherapy	618
Rheumatology	616
Pneumology	495
Urology / Oncology	244
Neurology	193
Pediatrics Hematology	143
Pediatrics	129
Gastroenterology	83
Pediatrics Infectious Diseases	59
Nephrology	44
Total	**10446**

Figure 22 – Medical Speciality Count

On the one hand, this way is possible to perceive that of the 10.446 EMR's, the Top 3 most common medical specialities, in this dataset, is "Medical Oncology"; "Pain Unit" and "Hematology". As for the 3 least frequent specialities, those are "Nephrology", "Paediatrics Infectious Diseases" and "Gastroenterology".

On the other hand, through a similar but completely different analysis, presented in Figure 23, can be perceived which specialities had the most entities detected in its clinical diaries. Looking at the Top 3 specialities with the most detected entities, the first two remain the same "Medical Oncology" and "Pain Unit" but as for the third, instead of being "Hematology", is "Rheumatology".

Medical_Specialty_Description	Contagem de Entity_Value
Medical Oncology	16989
Pain unit	13438
Rheumatology	11990
Hematology	11265
Immunohemotherapy	6250
Infecciology	5999
Pneumology	3282
Neurology	2893
Pediatrics Hematology	2792
Pediatrics	2011
Pediatrics Infectious Diseases	1300
Gastroenterology	1141
Urology / Oncology	965
Nephrology	788
Total	**81103**

Figure 23 – Medical Specialties by Number of Entities Detected

As for the top 3 specialities with the least detected entities, those are “Nephrology”, “Urology/Oncology” and “Gastroenterology”. Through this analysis, it can be concluded that the richest specialities, in terms of quantity of detected entities, are not, mandatorily, the most frequent ones.

Regarding the top 5 entities detected in the richest speciality, further analysis was done. Presented in Figures 24 and 25 are represented the most common entities in five categories: Medication, Sign/Symptom, Anatomical Site, Procedure and Disease/Disorder, respectively, of the “Medical Oncology” speciality.

Entity_Value	Count of Entity_Value
bevacizumab	179
capecitabine	153
Fluorouracil	188
Solution Dosage Form	378
Valine	179
Total	**1077**

Entity_Value	Count of Entity_Value
Administration occupational activities	388
Blood Pressure	322
Complication	392
Illness (finding)	310
Pressure (finding)	326
Total	**1738**

Figure 24 – Top 5 Medications and Symptoms (Medical Oncology)

As it can be perceived, for the speciality of “Medical Oncology”, the medications most commonly used, of the detected ones, are “bevacizumab”, “capecitabine” and “fluorouracil”. As for the detected symptoms, the most common is “Administration Occupational Activities”, “Blood Pressure”, “Complication”, “Illness (finding)” and “Pressure (finding)”.

Whilst analyzing Figure 25 and looking at the bottom table, it is possible to note the existence of more than five lines. This happens since the last three lines, “Pad Mass”, “Peripheral Arterial Diseases” and “Traumatic Injury” were identified the same number of times. This type of evaluation can be done for all the medical specialities present in the dataset.

Entity_Value	Count of Entity_Value
Blood	879
Skin	187
Oral cavity	139
Heart	115
Breast	108
Total	**1428**

Entity_Value	Count of Entity_Value
Therapeutic procedure	685
Administration procedure	388
Chemotherapy	384
Analysis of substances	186
Infusion procedures	141
Total	**1784**

Entity_Value	Count of Entity_Value
Tension	104
Injury wounds	62
Adenohypophyseal Diseases	56
Syndrome	49
Pad Mass	34
Peripheral Arterial Diseases	34
Traumatic injury	34
Total	**373**

Figure 25 – Top 5 Anatomical Sites, Procedures and Diseases (Medical Oncology)

Regarding the diagnosis attributes present in the unstructured dataset, are presented in Figure 26 are the top 5 diagnosis descriptions, codes and respective counts of the overall specialities.

Diagnosis_Description	Diagnosis_Code	Count of Diagnosis_Code
Neoplasms	14	21527
Diseases Of The Blood And Blood-Forming Organs	28	17147
Diseases Of The Musculoskeletal System And Connective Tissue	71	13228
Rheumatoid arthritis	7140	8175
Infectious And Parasitic Diseases	10	6104
Total		**66181**

Figure 26 – Top 5 Diagnostics (Overall)

Represented in Figure 27 are represented the most frequent diagnostics of the "Medical Oncology" speciality.

Diagnosis_Description	Diagnosis_Code	Count of Diagnosis_Code
Neoplasms	14	16354
Endocrine, Nutritional And Metabolic Diseases, And I mmunity Disorders	24	582
Diseases Of The Blood And Blood-Forming Organs	28	46
Diseases Of The Genitourinary System	58	5
Acute Glomerunonefritis	580	1
Diseases Of The Musculoskeletal System And Connective Tissue	71	1
Total		**16989**

Figure 27 – Top 5 Diagnostics (Medical Oncology)

As it can be compared between Figures 26 and 27, the most common ICD 9 in this hospital is the same, Neoplasms.

An important aspect to consider is that of the most common diagnosis descriptions present in the appointments department (Overall), four out of five are generic descriptions of the diagnosis itself. In other words, the descriptions are the names of the chapters in which those diagnostics are contained. Being the respective code the first entry number of those chapters, e.g. Neoplasms is the name of the second chapter of the ICD 9 codes[5], and 14 is respective to 014 which is the first number of the codes respective to that chapter. The same happens with "Diseases of the Blood and Blood-Forming Organs" with code 28, which is chapter 4, and so on. The main implication of this fact is that, although the emergency department has its codes very specific about the majority of its diseases, the appointment department is very generic, which influences the number of matching pairs by ICD 9 codes in the combined dataset analysis.

This subchapter, as explained at the beginning of chapter 6, was created to show that the outcome of the unstructured dataset's processing was fruitful and allowed for an analysis of information retrieved from the medical diaries. In the next subchapter, the two tables, so far analyzed individually, are linked, and a conjoint examination is done.

6.3 Mixed Data Results

In the previous chapters, 6.1 and 6.2, were presented examples of the knowledge that could be extracted from both the emergency and the appointments' departments alone, respectively. In order to extract information and knowledge from these two tables together, it was necessary to link the tables together through a common attribute. The chapter is divided into two subchapters, (1) Simple Key Linkage and (2) Compound Key Linkage, regarding the two different attributes used to connect the tables: ICD 9 Code and a compound key, correspondingly.

During these analyses, it was assumed that the treatments that the patients got during the appointments, such as medications, symptomatology, the procedures they incur, etc., are the same in case such cases appear in the emergency department.

[5] Present in Annex A

6.3.1. Simple Key Linkage

For the first approach to link both tables, it was used the attributes “ICD 9 Code” and “Diagnosis Code” from the ED and AD tables, correspondingly. The reason behind the selection of this variable dwell in the fact that these are some of the few common attributes between tables. The remaining attributes that exist mutually are the “Date” of the AD_SQLite table and the rest of the time attributes of the ED_SQLite table, “Admission Date”, “Date First Observation”, etc...

Presented in Figure 28 is shown the model of both tables with the linking attributes highlighted.

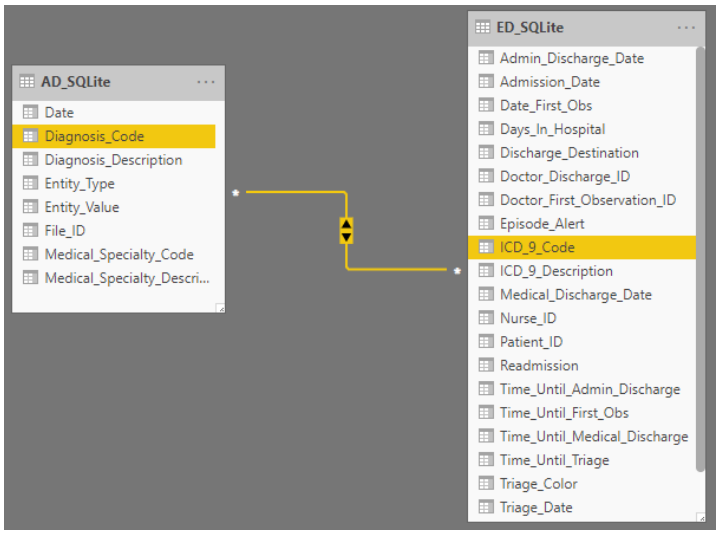


Figure 28 – Unique Attribute Connection

Because none of these tables contains a unique identifier, as it was explained in both chapters 4 and 5, a “many to many” relation is the only way to link the two tables. The resultant table of this merge has more entries than the sum of the number of rows of both tables that share the linking attribute since every “Diagnosis Code” tries to match with every “ICD 9 Code”. Due to this limitation, the analysis is slightly compromised when it comes to which viewpoints can be used to analyze the data.

To check which diseases were shared by both hospital departments, Figure 29 was created. Unfortunately, while the AD and the ED table contained, 26 and 2.835 different diagnostics, correspondently, after the linking both tables, only six diagnostics remained.

ICD_9_Description	Diagnosis_Code
Rheumatoid arthritis	7140
Ankylosing spondylitis	7200
Chronic kidney disease (ckd)	585
Regional enteritis	555
Lupus erythematosus	6954
Erythema Multiforme	6951

Figure 29 – Count of Common ICD 9 Codes and Descriptions (Simple Key Linkage)

From its analysis, it is also possible to observe that the six diagnosis descriptions that are common in both departments, are "Rheumatoid arthritis", "Ankylosing spondylitis", Chronic kidney diseases (ckd)", "Regional enteritis", "Lupus erythematosus" and "Erythema Multiforme". To better understand how these diagnostics are related to the hospital's medical specialities, the table in Figure 30 was created.

Medical_Specialty_Description	ICD_9_Description
Gastroenterology	Regional enteritis
Immunohemotherapy	Chronic kidney disease (ckd)
Immunohemotherapy	Regional enteritis
Immunohemotherapy	Rheumatoid arthritis
Nephrology	Chronic kidney disease (ckd)
Rheumatology	Ankylosing spondylitis
Rheumatology	Erythema Multiforme
Rheumatology	Lupus erythematosus
Rheumatology	Rheumatoid arthritis

Figure 30 – Medical Specialties per Diagnosis (Simple Key Linkage)

In this scenario, it is visible that the medical specialities that are coexistent in both departments are "Gastroenterology", "Immunohemotherapy", "Nephrology" and "Rheumatology". Despite "Nephrology" and "Gastroenterology" only having one diagnostic associated with each one, both "Rheumatology" and "Immunohemotherapy" present multiple diseases accompanying. Being the "Rheumatology" the most overarching speciality with four different diagnosis descriptions.

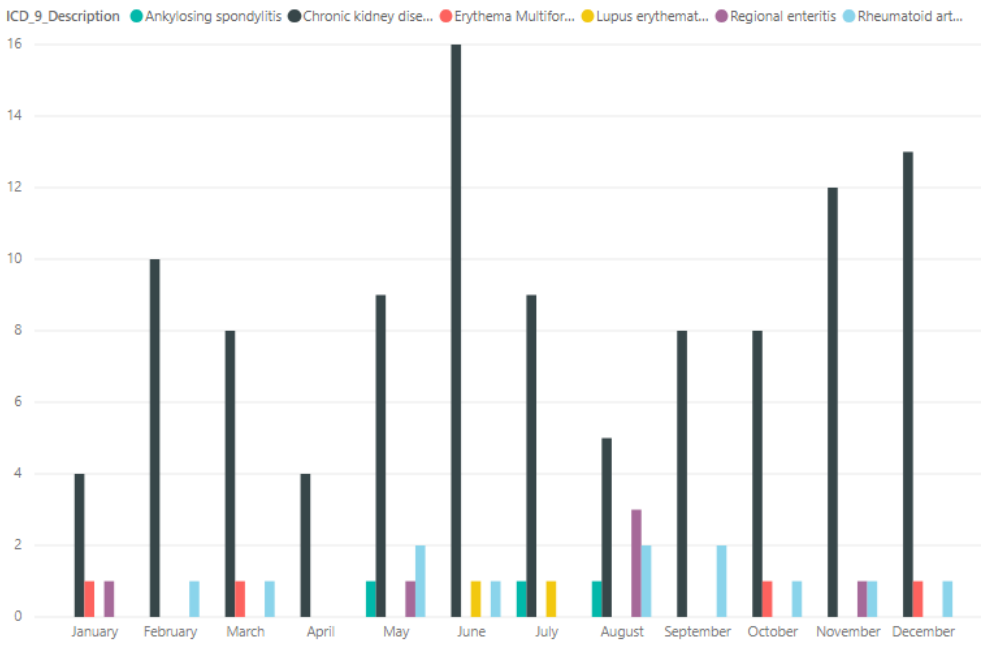


Figure 31 – Monthly Diagnosis Distribution (Simple Key Linkage)

In order to understand the affluence of such diagnosis to the hospital's emergency, it was created the chart in Figure 31. Throughout the year the most common diagnoses, by a great amount, out the of the six contemplated in this analysis, is "Chronic kidney disease (ckd)", from now on handled as CKD. The second most frequent diagnose is "Regional Enteritis" even though it is not present over the whole year, it has a considerable presence during the month of August. Since the CKD is the most prominent diagnose out of the six, to assess the number of medications, symptoms, body parts, diseases and procedures are associated with this disease, and Figure 32 was analyzed.

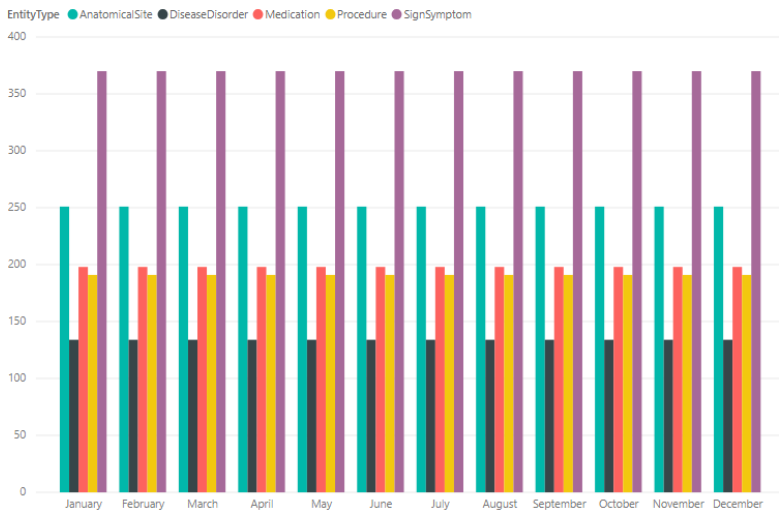


Figure 32 – CKD's Identified Clinical Entities (Simple Key Linkage)

Comparing the graphs Figures 31 and 32 it is immediately visible that the connection "many to many" between tables are starting to have repercussions. This happens because, for each CKD code present, every entity detected is associated with it. Because of this, the monthly quantities of detected entities are always the same.

Following this path invalidates a possible evaluation of the evolution the medications, symptomology and other entity types of the ED throughout the year. Nevertheless, to show the availability of information of this scenario, presented in Figure 33, are the most common detected entities for the diagnosis CKD.

EntityType	EntityValue	Count of EntityValue
Medication	Today	15
Medication	Bleach	10
Medication	Hemoglobin	10
Total		**35**

EntityType	EntityValue	Count of EntityValue
Procedure	Administration procedure	18
Procedure	Blood Transfusion	14
Procedure	Transfusion (procedure)	14
Total		**46**

EntityType	EntityValue	Count of EntityValue
AnatomicalSite	Blood	27
AnatomicalSite	Heart	21
AnatomicalSite	Veins	21
Total		**69**

EntityType	EntityValue	Count of EntityValue
DiseaseDisorder	Tension	13
DiseaseDisorder	Disease	11
DiseaseDisorder	Kidney Diseases	8
Total		**32**

EntityType	EntityValue	Count of EntityValue
SignSymptom	Blood Pressure	23
SignSymptom	Chief complaint (finding)	23
SignSymptom	Pressure (finding)	23
Total		**69**

Figure 33 – Most Common Entities of CKD Diagnosis (Simple Key Linkage)

Even though through this type of linkage is not possible to understand the evolution of the necessities of the ED facing this diagnosis, it is possible to assess that, according to the available information, the patients that went to the ED and suffered from CKD had symptomology of "Blood Pressure", "Chief Complaint (finding)" and "Pressure (finding)".Regarding the procedures administrated to the patients, the most common was "Administration Procedure", "Blood Transfusion" and "Transfusion (procedure)". As for the detected diseases, logically, the most common is "Tension"; "Disease" and Kidney Disease". Strangely, for the anatomical sites most common for this diagnosis are "Blood"; "Heart" and "Veins". For entity value, the author expected the "Kidney" to appear as one of the most common body parts, but it only appears in sixth place with 13 occurrences. As for the most common medications related to CKD, the top 3 is made of "Today", "Bleach" and "Hemoglobin". This is indeed and strange result since the first two results

are not, at first sight, related to the medication. The entity “Today” is associated with the day when the patient went to the hospital. The “Bleach” entity value is one example of the limitations of this project. The word “bleach” was translated from the word “descorado” which, also stand for “pale”, meaning that this word could have been detected as a symptom and was classified and medication due to translation errors. Nevertheless, ignoring these two entities, the top of medications for the CKD would be composed “Hemoglobin”, “Endoglin” and “Furosemide”.

In order to attempt as more real-time analysis and try to get around the problem of the “many to many” connection, another type of linkage was done. In the following chapter, its results are shown and discussed.

6.3.2. Compound Key Linkage

This second approach was created to try to mitigate the problem of lacking unique identifiers in both tables, which limits the analysis and the “many to many” connection. In order to reduce the number of entries of the resulting table, a compound key was created, using the diagnostic codes and the dates, as presented in Table XIII.

Table XIII - Compound Key Attributes of Each Table

Table Name	Compound Key Attributes	Example
AD_SQLite	“Date” and “Diagnosis Code”	9/20/2017-7140
ED_SQLite	“Admission Date” and “ICD 9 Code”	

In the example presented, the compound key is related to entry, from both tables from the 20th of September of 2017, where the diagnosis code was 7140 (Rheumatoid arthritis). Unfortunately, this key was not sufficiently specific to serve as a unique identifier for both tables but reduced the number of matches, increasing the credibility of the tables and graphs shown. In order to perceive the difference between the two utilized linking methods, the same angles in which the dataset was analyzed in the subchapter 6.3.1, are shown again.

It can be seen from Figures 34 and 35 that due to the change in the linking attribute, the number of diagnosis matches has sharply decreased, from six to two.

ICD_9_Description	Diagnosis_Code
Chronic kidney disease (ckd)	585
Rheumatoid arthritis	7140

Figure 34 – Common ICD 9 Codes and Descriptions (Compound Key Linkage)

Naturally, the number of diseases per medical speciality also decreased. Nevertheless, of all specialities previously represented, only "Gastroenterology" disappeared. Remaining for this analysis, the clinical diagnosis of "Chronic kidney disease (ckd)" and "Rheumatic arthritis", according to Figure 35. The latter from now on is treated as RA.

Medical_Specialty_Description	ICD_9_Description
Immunohemotherapy	Chronic kidney disease (ckd)
Immunohemotherapy	Rheumatoid arthritis
Nephrology	Chronic kidney disease (ckd)
Rheumatology	Rheumatoid arthritis

Figure 35 – Medical Specialties per Diagnosis (Compound Key Linkage)

To understand the impact of the new linkage methodology in the distribution of the diagnosis throughout the year it was necessary for the analysis of the graph of Figure 36.

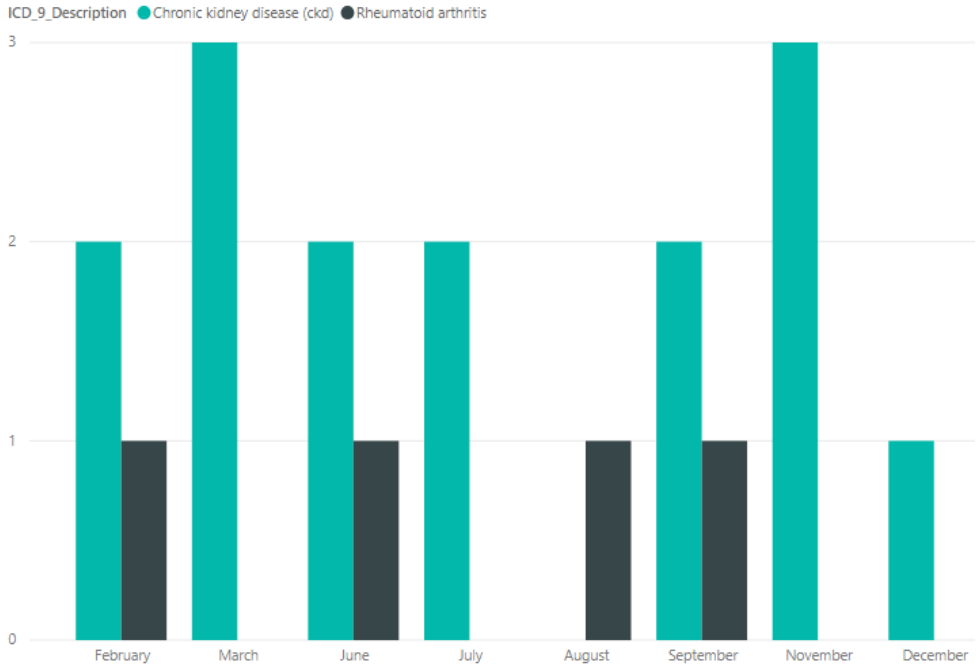


Figure 36 – Monthly Diagnosis Distribution (Compound Key Linkage)

Looking at the graph, the first thing important to notice is that not all months are represented. The other distinct feature of this chart is the number of entries in the ED which have the diagnosis of CKD or RA. This happens because a record of being present in this chart has to have a respective record in the appointments department with both the

same date and diagnosis code. Hence, the decrease in both records and months. From an overall analysis, it can be stated that the most common months of CKD patients are November and March whilst the RA continues to not show a prevalent month.

Depicted in Figure 37 is the distribution of detected entities for CKD diagnostics. Contrarily to what was shown in Figure 32, it is now more prominent the existence of months when the entities are more present than the others.

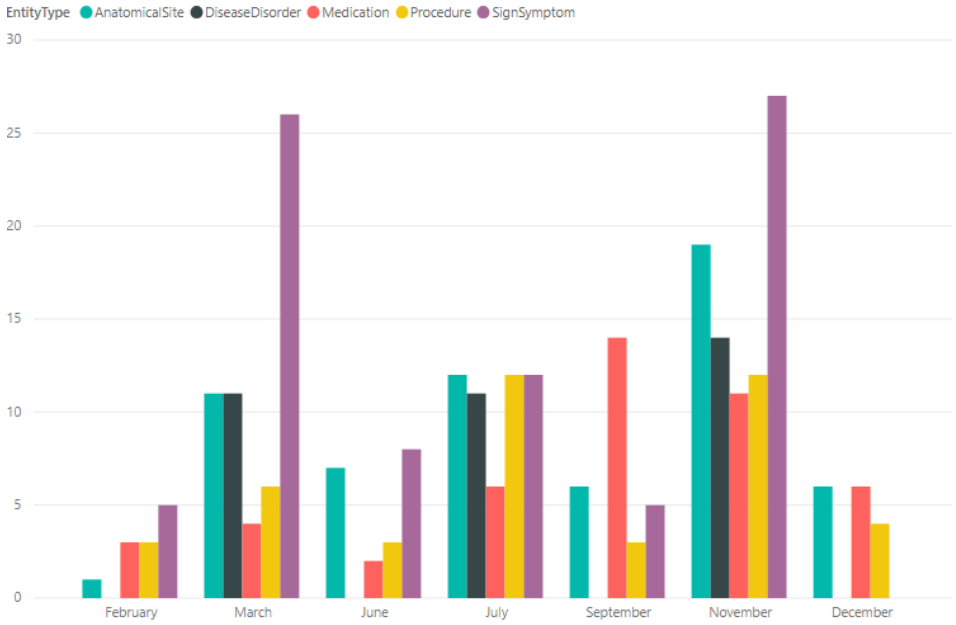


Figure 37 – CKD's Identified Clinical Entities (Compound Key Linkage)

On the one hand, according to the graph, is possible to assess the existence of months when the symptomatology is more registered by the doctors probably to the fact of it being more prominent. Said months are March and November.

On the other hand, the months when most medications are more presented is during the months of September and November. Just like it was done previously, it was analyzed the most common entity types for the diagnosis CKD, Figure 38. Yet, this time it was only regarding the month of November since the majority of its entity types counts are above average.

EntityType	EntityValue	Count of EntityValue
AnatomicalSite	Arteries	2
AnatomicalSite	Blood	2
AnatomicalSite	Heart	2

EntityType	EntityValue	Count of EntityValue
Procedure	Administration procedure	2
Procedure	Analysis of substances	1
Procedure	Blood Transfusion	1
Procedure	Oxygen saturation measurement	1
Procedure	Plain x-ray	1

EntityType	EntityValue	Count of EntityValue
Medication	Today	2
Medication	Antibodies	1
Medication	Antineutrophil Cytoplasmic Antibodies	1
Medication	Bleach	1
Medication	Cryoglobulins	1

EntityType	EntityValue	Count of EntityValue
DiseaseDisorder	Tension	2
DiseaseDisorder	Acute congestive heart failure	1
DiseaseDisorder	Atrial Fibrillation	1
DiseaseDisorder	Cardiac Arrest	1
DiseaseDisorder	Chronic Kidney Diseases	1

EntityType	EntityValue	Count of EntityValue
SignSymptom	Administration occupational activities	2
SignSymptom	Blood Pressure	2
SignSymptom	Pressure (finding)	2
SignSymptom	Systemic arterial pressure	2

Figure 38 – November's Most Common Entities of CKD Diagnosis (Compound Key Linkage)

Through its analysis, some things can be concluded regarding the most common entity value for the five entity types. Concerning the most common body parts of November for the diagnosis of CKD, stays the same comparing to the analysis with the previous linkage key. Regarding the most common procedures done, comparing with the results of the previous linkage, new procedures arose such as, "Analysis of Substances", "Oxygen Saturation Measurement" and "Plain X-ray". Moving on to the Medications detected, also new elements were detected, "Antibodies", "Antineutrophil Cytoplasmic Antibodies" and "Cryoglobulins". Also, the entity type of diseases detected new values "Cardiac arrest", "Atrial Fibrillation" and "Acute congestive heart failure". Lastly, two new symptoms were detected in comparison to the previous linkage, "Administration Occupational Activities" and "Systemic Arterial Pressure".

Unfortunately, all these results have counts between one and two occurrences. This is obviously due to the fact that there are few records in the month of November, nevertheless, counts of one and two are not enough to conclude anything with certainty. In other words, it can be concluded that through the usage of the more specific key linkage, the records that remained were only sufficient to assess that based on the dataset sets given, the most common months with CKD are March and November, being the latter the month when more symptomatology and anatomical sites is described. Considering that these conclusions were taken under the assumption that the medications, symptomatology, procedures and the remaining entity types that are taken into account during the appointments in the AD are also taken into account in the ED.

Chapter 7 – Conclusion

With the development of both medical and computer science domains in the last years, a lot of progressions have been made in each field alone but also combining both universes. The amount of data that is stored, in all its forms (structured, unstructured and semi-structured) nowadays in any institution is vast but unfortunately not all of it used or analyzed, the same happens in hospitals.

The purpose of this work was to analyses a proof of concept regarding the possibility of knowledge extraction from the combination of structured and unstructured data using EMR's from two different departments of the same Portuguese hospital, Emergency and Appointment's departments, correspondingly. Through an NLP system, denominated cTAKES (with the help of the Unified Medical Language System), it was possible to extract structured clinal information from the clinical diaries in the EMR's. With the use of Power BI, it was possible to link the data form the two departments in order to extract knowledge. Said linkage was done in two ways, one using the international code of diseases, the other using a compound key (the date and the international code of diseases).

This way, it can be concluded that the objectives were attained since not only data integration was possible, as well as its analysis and exploration and knowledge extraction, enabling a new perspective about the potential needs of the emergency department during the year. As the final output of this work, it was possible to assess how the CKD adherence to the hospital varied throughout the years and perceive the month when it shows more described symptomatology and anatomical sites. This work already created a conference proceeding (Baptista et al., 2019).

7.1 Research Questions Answers

Concerning the first research questions presented at the beginning of this investigation, it can be concluded that it was indeed possible to extract structured information from the unstructured data of the appointments' clinical diaries, through the use of the cTAKES software.

Regarding the second research question, it was possible to extract knowledge from the integrated data, since it was possible to perceive the month when one type of diagnostic shows more documentation in the ED through the use of both departments' structured information. It could also be what that documentation contained regarding symptomatology, anatomical site, medication, diseases and procedures.

7.2 Limitations

Nevertheless, this work found some obstacles during its development. Firstly, the difference between departments and language of the datasets even though they are from the same hospital. With the difference between the department, it can be difficult to have common attributes, and since the translation is not a perfect process, some value was lost during the translations of the clinical diaries, as it was detected with the word “Bleach”.

The second limitation is related to the fact that while in the structured dataset, the anonymization of the all the actors was done, for the unstructured dataset that process was not done. Since it was the author that had to do the anonymization, all chances of relating the actors from either side with the other to extract knowledge were lost.

The third limitation of this work falls over the fact that despite the two datasets are from the same Portuguese hospital, while on the emergency dataset each entry had a specific diagnostic code, the appointment’s dataset was too generic, which ended up for limit the number of diagnostics that could be analyzed from both sides.

7.3 Future Work

As to continue and improve this work it was thought to be useful to outspread this work to the hospital to get a uniform anonymization process between datasets, gaining more linking attributes, such as doctors, patients and nurses. Another improvement that could be made would be the aggregation of some diagnostics codes, i.e. instead of considering “Regional Enteritis of Small Intestine” and Regional Enteritis of Large Intestine” as two different diagnostics, consider it only as “Regional Enteritis”. This way, it would be possible to get more results for each aggregated diagnostic.

Bibliography

Afzal, N., Sohn, S., Abram, S., Liu, H., Kullo, I. J., & Arruda-Olson, A. M. (2016). Identifying peripheral arterial disease cases using natural language processing of clinical notes. *3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016*, 126–131. https://doi.org/10.1109/BHI.2016.7455851

Allen, G. D. (2004). Hierarchy of Knowledge – from Data to Wisdom. *International Journal of Current Research in Multidisciplinary (IJCRM)*, *2*(1), 15–23.

Amin, S. U., Agarwal, K., & Beg, R. (2013). Genetic neural network based data mining in prediction of heart disease using risk factors. *2013 IEEE Conference on Information and Communication Technologies, ICT 2013*, (Ict), 1227–1231. https://doi.org/10.1109/CICT.2013.6558288

Ananthakrishnan, A. N., Cai, T., Savova, G., Cheng, S. C., Chen, P., Perez, R. G., … Liao, K. P. (2013). Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: A novel informatics approach. *Inflammatory Bowel Diseases*, *19*(7), 1411–1420. https://doi.org/10.1097/MIB.0b013e31828133fd

Araneo, R., & Celozzi, S. (2015). The Feasibility of Using Large-Scale Text Mining to Detect Adverse Childhood Experiences in a VA-Treated Population. *Applied Computational Electromagnetics Society Journal*, *28*, 505–514. https://doi.org/10.1002/jts.

Baba, Y., Hiramatsu, T., Kimura, M., Shimizu, S., Kobayashi, K., Tsuda, K., … Inoue, S. (2015). Predictive Approaches for Low-Cost Preventive Medicine Program in Developing Countries. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 1681–1690. https://doi.org/10.1145/2783258.2788587

Baptista, D., Ferreira, J. C., Pereira, R., & Baptista, M. (2019). Structured and Unstructured Data Integration with Electronic Medical Records. In *Proceedings of the World Congress on Engineering 2019 WCE 2019, July 3-5, 2019, London, U.K.* London.

Basu Roy, S., Teredesai, A., Zolfaghar, K., Liu, R., Hazel, D., Newman, S., & Marinez, A. (2015). Dynamic Hierarchical Classification for Patient Risk-of-Readmission.

Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15, 1691–1700. https://doi.org/10.1145/2783258.2788585

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(DATABASE ISS.), 267–270. https://doi.org/10.1093/nar/gkh061

Chia, C.-C., & Syed, Z. (2014). Scalable noise mining in long-term electrocardiographic time-series to predict death following heart attacks. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14*, 125–134. https://doi.org/10.1145/2623330.2623702

Cooper, P. (2017). Data, information, knowledge and wisdom. *Anaesthesia and Intensive Care Medicine*, *18*(1), 55–56. https://doi.org/10.1016/j.mpaic.2016.10.006

Feldman, K., Hazekamp, N., & Chawla, N. V. (2016). Mining the Clinical Narrative: All Text are Not Equal. *Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016*, 271–280. https://doi.org/10.1109/ICHI.2016.37

Fong, A., Hettinger, A. Z., & Ratwani, R. M. (2015). Exploring methods for identifying related patient safety events using structured and unstructured data. *Journal of Biomedical Informatics*, *58*, 89–95. https://doi.org/10.1016/j.jbi.2015.09.011

Friedman, C., Liu, H., Shagina, L., Johnson, S., & Hripcsak, G. (2001). Evaluating the UMLS as a source of lexical knowledge for medical language processing. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 189–193.

Garets, D., & Davis, M. (2006). Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference A HIMSS Analytics TM White Paper Source: HIMSS Analytics Database (derived from the Dorenfest IHDS+ Database TM), 1–14. Retrieved from www.himssanalytics.org

Gartner. (2019a). Data Mining. Retrieved October 23, 2019, from https://www.gartner.com/en/information-technology/glossary/data-mining (accessed in 23-10-2019)

Gartner. (2019b). Natural-language Processing (nlp). Retrieved October 23, 2019, from https://www.gartner.com/en/information-technology/glossary/natural-language-processing-nlp (accessed in 23-10-2019)

ICD-9-CM Chapters List. (2019). Retrieved September 25, 2019, from https://icd.codes/icd9cm (accessed in 25-09-2019)

ICD - ICD-9 - International Classification of Diseases, Ninth Revision. (2019). Retrieved March 18, 2019, from https://www.cdc.gov/nchs/icd/icd9.htm (accessed in 18-03-2019)

ICD9 Provider Diagnostic Codes. (2014). Retrieved September 22, 2019, from https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes.html (accessed in 22-09-2019)

Jain, N. L., & Friedman, C. (1997). Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proceedings : A Conference of the American Medical Informatics Association. AMIA Fall Symposium*, 829–833. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9357741%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2233320

Jonnagaddala, J., Liaw, S. T., Ray, P., Kumar, M., Chang, N. W., & Dai, H. J. (2015). Coronary artery disease risk assessment from unstructured electronic health records using text mining. *Journal of Biomedical Informatics*, *58*, S203–S210. https://doi.org/10.1016/j.jbi.2015.08.003

Kharrazi, H., Anzaldi, L. J., Hernandez, L., Davison, A., Boyd, C. M., Leff, B., … Weiner, J. P. (2018). The Value of Unstructured Electronic Health Record Data in Geriatric Syndrome Case Identification. *Journal of the American Geriatrics Society*, *66*(8), 1499–1507. https://doi.org/10.1111/jgs.15411

Kidwai, F., Justice, A., Re, V. Lo, Brandt, C., Scotch, M., Garla, V., … Dorey-Stein, Z. (2011). The Yale cTAKES extensions for document classification: architecture and application. *Journal of the American Medical Informatics Association*, *18*(5), 614–620. https://doi.org/10.1136/amiajnl-2011-000093

Koh, H. C., & Tan, G. (2005). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, *19*(2), 64–72.

Kop, R., Hoogendoorn, M., Moons, L. M. G., Numans, M. E., & ten Teije, A. (2015). On the advantage of using dedicated data mining techniques to predict colorectal cancer. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial*

Intelligence and Lecture Notes in Bioinformatics), *9105*, 133–142. https://doi.org/10.1007/978-3-319-19551-3_16

Lamy, M. M. V. P. de M. (2018). *Extracting Clinical Knowledge from Electronic Medical Records*. Master thesis from METI at ISCTE-IUL. Retrieved from http://hdl.handle.net/10071/17591

Liu, K., Mitchell, K. J., Chapman, W. W., & Crowley, R. S. (2005). Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *11*(Figure 1), 460–464.

Luo, L., Li, L., Hu, J., Wang, X., Hou, B., Zhang, T., & Zhao, L. P. (2016). A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. *BMC Medical Informatics and Decision Making*, *16*(1), 1–15. https://doi.org/10.1186/s12911-016-0357-5

Luther, S. L., McCart, J. A., Berndt, D. J., Hahm, B., Finch, D., Jarman, J., … Powell-Cope, G. (2015). Improving identification of fall-related injuries in ambulatory care using statistical text mining. *American Journal of Public Health*, *105*(6), 1168–1173. https://doi.org/10.2105/AJPH.2014.302440

McLane, S. (2005). Designing an EMR planning process based on staff attitudes toward and opinions about computers in healthcare. *CIN - Computers Informatics Nursing*, *23*(2), 85–92. https://doi.org/10.1097/00024665-200503000-00008

Moreira, L. B., & Namen, A. A. (2018). A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. *Computer Methods and Programs in Biomedicine*, *165*, 139–149. https://doi.org/10.1016/j.cmpb.2018.08.016

Pulmano, C. E., & Estuar, M. R. J. E. (2016). Towards Developing an Intelligent Agent to Assist in Patient Diagnosis Using Neural Networks on Unstructured Patient Clinical Notes: Initial Analysis and Models. *Procedia Computer Science*, *100*, 263–270. https://doi.org/10.1016/j.procs.2016.09.153

Rabbi, K., Mamun, Q., & Islam, M. D. R. (2015). Dynamic feature selection (DFS) based Data clustering technique on sensory data streaming in eHealth record system. *Proceedings of the 2015 10th IEEE Conference on Industrial Electronics and Applications, ICIEA 2015*, 661–665. https://doi.org/10.1109/ICIEA.2015.7334192

Ravindranath, K. R. (2015). Clinical Decision Support System for heart diseases using Extended sub tree. *2015 International Conference on Pervasive Computing: Advance Communication Technology and Application for Society, ICPC 2015*, *00*(c), 1–5. https://doi.org/10.1109/PERVASIVE.2015.7087026

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507–513. https://doi.org/10.1136/jamia.2009.001560

Sevenster, M., Van Ommering, R., & Qian, Y. (2012). Automatically correlating clinical findings and body locations in radiology reports using MedLEE. *Journal of Digital Imaging*, *25*(2), 240–249. https://doi.org/10.1007/s10278-011-9411-0

Simon, G. J., Caraballo, P. J., Therneau, T. M., Cha, S. S., Castro, M. R., & Li, P. W. (2015). Extending association rule summarization techniques to assess risk of diabetes mellitus. *IEEE Transactions on Knowledge and Data Engineering*, *27*(1), 130–141. https://doi.org/10.1109/TKDE.2013.76

Sohn, S., Kocher, J. P. A., Chute, C. G., & Savova, G. K. (2011). Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *Journal of the American Medical Informatics Association*, *18*(SUPPL. 1), 144–149. https://doi.org/10.1136/amiajnl-2011-000351

Somanchi, S., Adhikari, S., Lin, A., Eneva, E., & Ghani, R. (2015). Early Prediction of Cardiac Arrest (Code Blue) using Electronic Medical Records. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 2119–2126. https://doi.org/10.1145/2783258.2788588

Sumana, B. V., & Santhanam, T. (2015). Prediction of diseases by cascading clustering and classification. *2014 International Conference on Advances in Electronics, Computers and Communications, ICAECC 2014*, 1–8. https://doi.org/10.1109/ICAECC.2014.7002426

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: A review. *Journal of Healthcare Engineering*, *2018*. https://doi.org/10.1155/2018/4302425

Sun, W., Cai, Z., Liu, F., Fang, S., & Wang, G. (2017). A survey of data mining technology on electronic medical records. *2017 IEEE 19th International Conference on E-Health Networking, Applications and Services (Healthcom), e-Health Networking, Applications and Services (Healthcom), 2017 IEEE 19th International Conference On*. https://doi.org/10.1109/HealthCom.2017.8210774

Sundararaman, A., Valady Ramanathan, S., & Thati, R. (2018). Novel Approach to Predict Hospital Readmissions Using Feature Selection from Unstructured Data with Class Imbalance. *Big Data Research*, *13*, 65–75. https://doi.org/10.1016/j.bdr.2018.05.004

Toerper, M. F., Flanagan, E., Siddiqui, S., Appelbaum, J., Kasper, E. K., & Levin, S. (2016). Cardiac catheterization laboratory inpatient forecast tool: A prospective evaluation. *Journal of the American Medical Informatics Association*, *23*(e1), e49–e57. https://doi.org/10.1093/jamia/ocv124

Tran, T., Phung, D., Luo, W., & Venkatesh, S. (2015). Stabilized sparse ordinal regression for medical risk stratification. *Knowledge and Information Systems*, *43*(3), 555–582. https://doi.org/10.1007/s10115-014-0740-4

Wu, C. Y., Chang, C. K., Robson, D., Jackson, R., Chen, S. J., Hayes, R. D., & Stewart, R. (2013). Evaluation of Smoking Status Identification Using Electronic Health Records and Open-Text Information in a Large Mental Health Case Register. *PLoS ONE*, *8*(9), 1–8. https://doi.org/10.1371/journal.pone.0074262

Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2017). Mining Electronic Health Records: A Survey, *50*(6), 1–41. https://doi.org/1539-9087/2016/04-ART1

Yamamoto, L. G., & Khan, A. N. G. A. (2006). Challenges of Electronic Medical Record Implementation in the Emergency Department. *Pediatric Emergency Care*, *22*(3). Retrieved from https://journals.lww.com/pec-online/Fulltext/2006/03000/Challenges_of_Electronic_Medical_Record.12.aspx

Zachariasse, J. M., Seiger, N., Rood, P. P. M., Alves, C. F., Freitas, P., Smit, F. J., … Moll, H. A. (2017). Validity of the Manchester triage system in emergency care: A prospective observational study. *PLoS ONE*, *12*(2), 1–14. https://doi.org/10.1371/journal.pone.0170811

Annex and Appendix

Annex A

Chapter #	Code Range	Description
1	001-139	Infectious and Parasitic Diseases
2	140-239	Neoplasms
3	240-279	Endocrine, Nutritional and Metabolic Diseases, And Immunity Disorders
4	280-289	Diseases of The Blood and Blood-Forming Organs
5	290-319	Mental Disorders
6	320-389	Diseases of The Nervous System and Sense Organs
7	390-459	Diseases of The Circulatory System
8	460-519	Diseases of The Respiratory System
9	520-579	Diseases of The Digestive System
10	580-629	Diseases of The Genitourinary System
11	630-679	Complications of Pregnancy, Childbirth, And the Puerperium
12	680-709	Diseases of The Skin and Subcutaneous Tissue
13	710-739	Diseases of The Musculoskeletal System and Connective Tissue
14	740-759	Congenital Anomalies
15	760-779	Certain Conditions Originating in The Perinatal Period
16	780-799	Symptoms, Signs, And Ill-Defined Conditions
17	800-999	Injury and Poisoning
18	V01-V91	Supplementary Classification of Factors Influencing Health Status and Contact with Health Services
19	E000-E999	Supplementary Classification of External Causes of Injury and Poisoning

Annex B

Appendix A

Appendix B