TÉCNICO LISBOA

Phenotyping and Understanding Multimorbidity

Miguel Terras Froes

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Prof. Mário Jorge Costa Gaspar da Silva
Dr. Bernardo Alves Vieira Duque Neves

Examination Committee

Chairperson: Prof. João Miguel Raposo Sanches
Supervisor: Prof. Mário Jorge Costa Gaspar da Silva
Members of the Committee:
Dr. André Peralta-Santos
Prof. Susana de Almeida Mendes Vinga Martins

December 2020

Declaration

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

Preface

The work presented in this thesis was performed at Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (Lisbon, Portugal), during the period February-December 2020, under the supervision of Prof. Mário Gaspar da Silva and Prof. Bruno Martins. The thesis was co-supervised at Hospital da Luz by Dr. Bernardo Neves.

Acknowledgments

I would like to thank everyone that supported me during the adventure of doing my Masters and in the development of this dissertation. A special acknowledgement to those who never let me go through this process by myself.

Firstly, I would like to express my sincere gratitude to Prof. Mário Silva, Prof. Bruno Martins, and Dr. Bernardo Neves for the countless hours of meetings and e-mails exchanged, as well as for their mentorship throughout the entire process.

I am very much thankful to INESC-ID for providing me and my research colleagues a room where we could gather and work together on our projects.

I would also like to express my deep gratitude to my girlfriend. This work would not have been possible without your support and motivation. Thank you for your patience and for helping not to doubt myself.

To my family for supporting me in every possible way and for encouraging me to never give up and to achieve the best results in life. Thank you for being the example of effort and devotion, always granting me the best opportunities.

Finally, I would like to thank all my friends for helping me through every stage of this work. To their companionship and support, who are always there in the good and bad times, day after day.

Resumo

Esta dissertação propõe um pipeline de processamento de informação para a extração de dados fenotípicos e análise de multimorbidade. O pipeline consiste num processo de Extract, Transform, and Load (ETL) que é aplicado a dados de Registos Clínicos Eletrónicos (RCE), compilando os mesmos num Clinical Data Repository (CDR). O CDR organiza as informações de maneira estruturada e unificada, permitindo uma análise de multimorbidade. A multimorbidade, definida como a coocorrência de duas ou mais doenças crónicas, tem sérias implicações nos indivíduos e nos sistemas de saúde, e está previsto o aumento da sua prevalência nas gerações futuras. Porém, poucos recursos são investidos para identificar (ou seja, fenotipar) e caracterizar pacientes com multimorbidade. Os RCE podem desempenhar um papel importante na melhor compreensão da multimorbidade. Com este pipeline, três estudos foram realizados: (i) Desenvolvimento e avaliação de um modelo de Processamento de Linguagem Natural (PLN) para processar os resumos de alta da base de dados MIMIC-III, por forma a identificar doenças crónicas. O modelo foi avaliado usando dados rotulados de acordo com sistema de codificação CID-9 e atribuídos por especialistas após revisão manual, tendo alcançado F1-scores de 0.93 e 0.97, respetivamente; (ii) Avaliação do impacto e aumento dos riscos associados à multimorbidade na população infetada com COVID-19 em Portugal. Os resultados mostraram que a multimorbidade está significativamente associada a desfechos adversos; (iii) Estudo dos padrões e evolução temporal da multimorbidade em pacientes da base de dados Enroll-HD. Foram detetadas relações evidentes entre condições crónicas, nomeadamente hipertensão, dislipidemia e diabetes. No entanto, estes resultados devem ser lidos com um certo grau de reserva devido ao dataset utilizado.

Palavras-chave: Multimorbilidade, Registos Clínicos Eletrónicos, Fenotipagem de Registos Clínicos Eletrónicos, Processamento de Linguagem Natural

Abstract

This dissertation proposes an information processing pipeline for phenotype data extraction and multimorbidity analysis. The pipeline consists of an Extract, Transform, and Load (ETL) process that is applied to Electronic Health Record (EHR) data, collecting it in an Observable Clinical Data Repository (CDR). The CDR organizes information, in a unified structured manner, and supports a subsequent multimorbidity analysis. Multimorbidity, as the co-occurrence of two or more chronic conditions, has serious implications on individuals and healthcare systems, and its prevalence is expected to increase in future generations. However, few resources are invested in tools to identify (i.e., phenotype) and characterize patients with multimorbidity. EHRs could play an important role in better understanding multimorbidity. With this pipeline, three studies were developed: (i) Development and evaluation of a Natural Language Processing (NLP) model to process full-text contents of MIMIC-III discharge summaries, for identifying chronic conditions. The model was evaluated using human-assigned ICD-9 diagnostic codes and manually reviewed labels, having achieved averaged F1-scores of 0.93 and 0.97, respectively; (ii) Assessment of the impact and increased risks associated with multimorbidity in the COVID-19 infected population on the Portuguese SINAVE database. Findings showed that multimorbidity is significantly associated with poor outcomes in this population; (iii) Study on the patterns and temporal evolution of multimorbidity in clinical patient timelines on the Enroll-HD dataset. Clear relationships between chronic conditions, namely hypertension, dyslipidemia, and diabetes were detected. However, these should be seen with some degree of reservation because of the dataset used.

Keywords: Multimorbidity, Electronic Health Records, Electronic Phenotyping, Natural Language Processing

Contents

List of Tables

List of Figures

Chapter 1

Introduction

We live in a society where each generation is expected to outlive their ancestors. According to the World Health Organization, WHO (2011), the advances being made in medicine and the changes in lifestyle and diet lead to an increase in life expectancy, as well as rise in the prevalence of chronic conditions. With that being said, the number of people with multiple health conditions is set to rise. According to Van den Akker et al. (1998), multimorbidity is defined as the presence of two or more co-occurring chronic conditions. Navickas et al. (2016) estimates that out of the primary care population over 65 years old up to 95% are affected by it. Also, multimorbidity greatly affects the individual's well being and quality of life.

It is, therefore, of great importance to correctly characterise patients according to their single, or co-occurrent, chronic conditions. Identifying a patient's specific conditions or outcomes is known as phenotyping. A correct recognition of a patient's phenotype, and correspondent analysis, can bring several advantages to all steps of the healthcare process, such as identifying treatment pathways optimized for a specific subset of patients affected by a specific combination of chronic diseases. Clinical trials on new treatments or medication is another area where identifying patient cohorts, having in mind co-occurrent conditions, is of great value. In this case, knowing beforehand the different diseases that affect the population – besides the target disease of the study – can help understand unexpected outcomes, possible adverse or unwanted secondary effects, or interactions with medications prescribed for the treatment of other comorbid diseases.

In modern medicine, the Electronic Health Record (EHR) is the standard for managing patient information, containing both structured and unstructured data. Structured data includes demographics, diagnosis codes, procedure codes, lab values, and medication exposures, whereas unstructured data includes progress notes, discharge summaries, and imaging or pathology reports. The EHR is the cornerstone for conducting a phenotyping process. However, according to Banda et al. (2018), due to the diverse nature of EHR data, accurately characterizing patients according to their chronic conditions still remains a challenge.

As reminded by Shivade et al. (2014), identifying patient cohorts using structured data, namely using

diagnosis codes, laboratory results, and medications, can be useful and an easier challenge, but is not sufficient and cannot supplant the added clinical value offered by unstructured text data (e.g., radiology reports, discharge summaries, progress notes) offers. Most of the structured information that results from a patient-doctor interaction is focused on the disease that caused the visit and has administrative purposes. The majority of crucial information for EHR-based phenotyping is, on the other hand, stored in the form of clinical notes. It is, therefore, of the highest importance to study how this information can be extracted and treated so that clinical records can be truly utilised, patients correctly characterised, and treatments precisely customised and applied.

1.1 Objectives and Results

In my M.Sc. research project, I present an information processing pipeline, represented in Figure 1.1, for extraction of phenotype data for multimorbidity analysis. Within this pipeline, three studies were developed.

The first stage of the pipeline uses both structured and unstructured data from the EHR. An Extract, Transform, Load (ETL) process is applied to handle the different types of data in the EHR.

The structured data is selected based on diagnostic and procedures codes, following the International Classification of Diseases (ICD) system, lab results, and medication prescribed. This selection is focused on detecting certain chronic diseases and uses previous developed algorithms presented by Tonelli et al. (2015) and Hvidberg et al. (2016). The extraction of structured data, especially ICD codes, is mainly focused on validating future results obtained from the treatment of unstructured data.

The value of unstructured text data in the EHR supplants the contribution from structured data. The main focus of this ETL process is the extraction of phenotypes from clinical notes using Natural Language Processing (NLP) techniques. First study: an NLP algorithm that processes full-text contents of discharge summaries, capable of identifying different chronic conditions while detecting cases of disease negation.

Data produced by the ETL process is collected in an Observable Clinical Data Repository (CDR). The CDR consolidates data obtained from the previous step and presents it in a unified structured manner, independently of its original source. The repository is a necessary bridge between the two processes presented in Figure 1.1. It organizes information and supports a subsequent analysis with respect to multimorbidity. Besides the information selected from the EHR structured data, this CDR also contains pertinent data, such as chronic disease's onsets, history of hospitalisations and Intensive Care Unit (ICU) admission, and date and cause of death.

With organised and uniform data, in the form of the CDR, it is possible to move to the last presented process of the pipeline. The last step corresponds to the analysis of the collected and treated data with particular focus on the topic of multimorbidity. This analysis focus on comprehending the impacts of multimorbidity in the quality of life and, ultimately, identifying specific patient cohorts and possible specified treatment pathways. Understanding multimorbidity is paramount when taken into account its

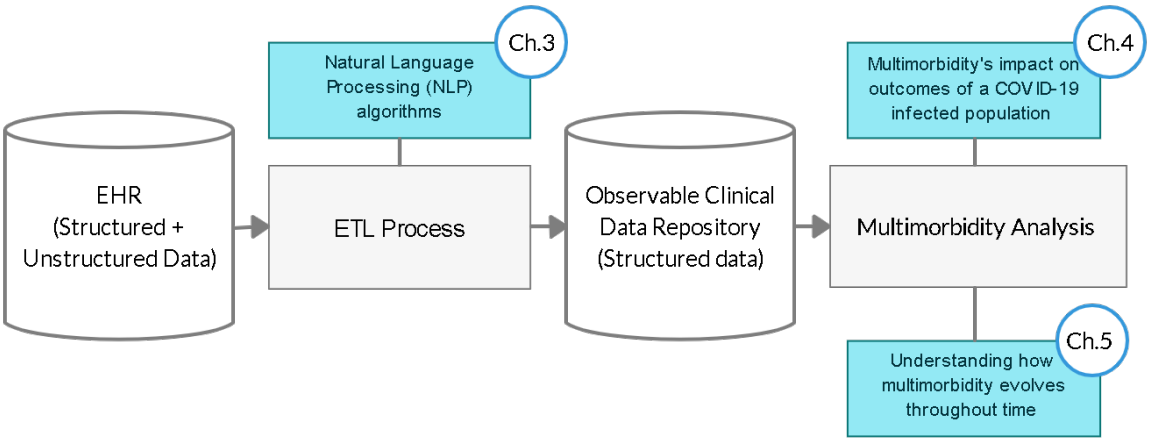


Figure 1.1: Proposed pipeline to analyse multimorbidity.

increased prevalence in older age groups in combination with the rise of life expectancy of our current generation.

The second study seeks to understand the impact and increased risks associated with multimorbidity on the different outcomes – Death, Hospitalization, and ICU admission – on the Coronavirus disease 2019 (COVID-19) infected Portuguese population. This enables to understand how COVID-19 interacts with chronic diseases and what added risks exist associated with the increased number of co-occurrent chronic conditions.

The third and last study focus on understanding the patterns and temporal evolution of multimorbidity in clinical patient timelines. Chronic diseases' onsets in combination with prescription history are used to find possible relations in the order of diagnosis of the conditions, and the time interval between onsets.

1.2 Methodology

In order to evaluate the proposed pipeline, a thorough, broad, and longitudinal dataset is necessary. Originally, Hospital da Luz (HL), as part of a continued partnership with Instituto Superior Técnico (IST) and Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID), planned to provide a dataset containing discharge summaries and other clinical information collected throughout a considerable period of time for this dissertation. Unfortunately, the current COVID-19 pandemic precluded the necessary treatment and anonymisation of the data because of the diminished access to the HL facilities and a homebound workforce, in time to be available for use in this thesis

To overcome this limitation, an alternative dataset that could be promptly accessed was searched for. Such dataset was not found. As a mean to guarantee the same level of coverage of the HL dataset, three datasets were used (see Table 1.1).

Table 1.1: Summary of the datasets used.

Dataset	Information Used	Main purpose of the dataset
MIMIC-III Critical Care Database	Discharge summaries ICD-9-CM (Diagnosis and Procedure Codes) Medication Lab results	*Dataset comprising information relating to patients admitted to critical care units that allows to create and test techniques to extract patient cohorts from unstructured data*
COVID-19 from SINAVE	Age Gender Medical diagnosis Hospitalization Mortality ICU Admission	*Study prevalence of certain chronic diseases in the Portuguese population. Understand the added risks of multimorbidity, in the midst of a pandemic, for mortality, hospitalizations and ICU admission*
Enroll-HD from CHDI Foundation	ICD-10-CM Medical diagnosis Diseases' onsets Medication	*Observational study, in the context of Huntington's disease, that allows for a better understanding on the temporal evolution, and relation, between chronic diseases' onsets*

It is also important to state that the original dataset would be in Portuguese and, therefore, the NLP phenotyping algorithms would be based on the Portuguese language. No dataset, of the same quality and dimensions, containing Portuguese discharge summaries could be found. Accordingly, the algorithms were based on the English language. However, the rules used were translated to Portuguese in order to facilitate any possible future application.

These alternative datasets fed the execution of the two stages of the proposed pipeline of Figure 1.1. The MIMIC-III Critical Care Database was used as a substitute for the EHR so that the ETL process, especially the text-mining algorithm, could be performed and studied. The datasets relative to the COVID-19 infected Portuguese population and Enroll-Huntington's disease (HD) observational study were both used to attest the last stage of the pipeline. For this purpose, both datasets only required a simple ETL process.

1.3 Thesis Outline

The organization of this dissertation is as follows:

- Chapter 2 gives context to all steps of the proposed pipeline. First, the focus is on the topic of multimorbidity, describing previous work focusing on multimorbidity analysis. Then, it presents the basis of EHR phenotyping, also discussing how these methods are evaluated. Finally, this chapter presents methodologies developed for EHR-based phenotyping, with a special focus on work concerning the use of NLP algorithms applied to clinical notes, as they were the main type of ETL process employed.
- Chapter 3 details the proposed approach considered for solving the problem of extracting information from clinical notes. It details the structure of the dataset used. Then, it presents the selection process of the dataset, along with statistical analysis, and the natural language processing (NLP) tool developed. Finally, it presents, explains, and discusses the results and performance of the proposed tool.
- Chapter 4 focuses on the proposed study related with understanding the impact of multimorbidity in the outcomes of a population affected by a life-threatening infection. It details the structure of the dataset used, including the pre-processing of the data and a statistical analysis of the dataset used in the experiment. Then, it presents and explains the experimental methodology used. Finally, this chapter presents the results obtained in the experiment discussing them.
- Chapter 5 presents the experimental evaluation of the proposed method to comprehend how multimorbidity evolves throughout a lifetime and how certain chronic conditions can impact the predisposition to the onset of other diseases. It details the structure of the dataset used. Then, it explains the selection process, as well as statistical analysis of the dataset, and the methodology developed for obtaining visual representations of the temporal relationships between chronic conditions. Finally, this chapter presents and discusses the representations of the relationships between chronic conditions' onsets.
- Finally, Chapter 6 summarises and outlines the main conclusions and limitations of this M.Sc. research project, also presenting suggestions for advancements in future work.

Chapter 2

Concepts and Related Work

This chapter describes fundamental concepts about the methodologies used in previous studies related to Electronic Health Record (EHR) phenotyping, also known as electronic phenotyping. In the medical sciences, phenotyping refers to the process of identifying observable characteristics or traits in patients who satisfy predefined criteria within a large population. The chapter also provides an overview on the topic of multimorbidity, describing previous work focusing on multimorbidity analysis. Section 2.1 briefly summarises previous work focusing on the characterisation of chronic diseases and multimorbidity. Section 2.2 presents a brief introduction to EHR phenotyping, also discussing how these methods are evaluated, as well as the most used coding systems in EHRs. Section 2.3 presents methodologies developed for EHR-based phenotyping, with a special focus on work concerning the use of NLP algorithms applied to clinical notes. Finally, Section 2.4 summarises the state of the art on EHR phenotyping and multimorbidity analysis.

2.1 Multimorbidity

Multimorbidity, defined by Van den Akker et al. (1998) as the co-occurrence of two or more chronic conditions, has serious direct and indirect implications on individuals and healthcare systems. Additionally, as pointed out by Knottnerus et al. (1992), multimorbidity also has a major impact on the well-being and lifestyle of the family and friends of those affected. According to the World Health Organization, WHO (2011), life expectancy is increasing, caused by better living conditions, and thus one can also expect an increase of chronic conditions prevalence in the next generation. Ultimately, this combination will result in a higher number of people afflicted by multimorbidity.

It is of the utmost importance to understand the risk factors and consequences of multimorbidity, at an individual level, to properly act on them. The most consistent risk factor is ageing, but the prevalence of multimorbidity does not exclusively affect the elderly. Van den Akker et al. (1998) identified cases of multimorbidity in all age groups in a general practice setting, although prevalence of multimorbidity increased with age, having started at 10% for all participants in the lowest age group (i.e., 0 – 19 years

old). In the Portuguese setting, Laires and Perelman (2019) performed a similar study and achieved similar results. This significant prevalence in low aged groups underlines the importance of identifying additional predisposing factors for multimorbidity. In the Finnish population, Wikström et al. (2015) identified smoking, physical inactivity, body mass index (BMI), hypertension, and low education as risk factors for a disease-free population.

The severity of the consequences of multimorbidity can vary. Fortin et al. (2004) showed an inverse relationship between all domains of quality of life (e.g., physical, psychological, and social) and multimorbidity. Additionally, multimorbidity is related with an increase of the number of interactions between a patient and healthcare providers. As pointed out by Navickas et al. (2016), patients with multimorbidity have more frequent and longer hospital admissions, while also having more interactions with different medical specialists. Ultimately, Menotti et al. (2001) associates people with multimorbidity to higher risks of premature death.

Beyond the individual level, healthcare systems are also greatly affected by multimorbidity. This impact is mainly in the form of excessive use of resources. Zulman et al. (2015) analysed the direct costs of multimorbidity in a large and integrated healthcare system, concluding that two-thirds of the top 5% highest-cost patients, which accounted for nearly half the total healthcare costs, had two or more chronic conditions. This is due to the fact that most healthcare systems are disease-oriented. Navickas et al. (2016) explains that in a disease-oriented system, more emphasis is given to managing each individual condition, while ignoring their interactions. This results in an inefficient, ineffective, and fragmented care for patients with multimorbidity. To solve this, as defended in Baker et al. (2018), the focus should be shifted from the disease to the patient (i.e., achieving the outcomes that matter for patients).

A patient-centred model should integrate patient cohort identification (i.e., phenotyping) tools, to accurately identify high risk multimorbidity patient groups, and a wider understanding of interactions between chronic diseases. Several studies have been focused on phenotyping techniques, but researches have usually focused on specific cohorts of patients. Hvidberg et al. (2016) and Tonelli et al. (2015) developed rule-based phenotyping methods to catalog chronic diseases. In particular, Hvidberg et al. (2016) payed special attention to the level of chronicity when creating the catalog of diseases. The diseases were divided into four categories according to their inclusion time (i.e., how long they will be present in the patient). Hassaine et al. (2019) developed an approach to account for multimorbidity patterns over time. The proposed method identifies the disease clusters and their temporal trajectories, and is capable of generating hypotheses of multimorbidity patterns over time.

2.2 Electronic Phenotyping

The Electronic Health Record (EHR) is a key health Information Technology (IT) component in modern healthcare. Besides allowing the recording of a patient's medical history, diagnoses, medications, and laboratory/test results, it can also be integrated with evidence-based tools and used on the decision-making process. EHRs contain both structured (e.g., diagnosis codes, laboratory results, medications)

and unstructured (e.g., radiology reports, discharge summaries, progress notes) data.

One of the major steps in utilising these EHRs, and the most significant to this thesis, is the process of phenotyping patients. There is no standard tool for electronic phenotyping that is easily available for use across institutions, and there are several barriers to the adoption of one such tool. Shivade et al. (2014) pointed out administrative roadblocks, collaboration running costs, and the sensitive nature of patient data as the primary reasons for the lack of cooperation between institutions to create a standard phenotyping technique, which would allow for faster and easily comparable phenotypes. This results in most institutions ending up creating their own systems tailored to their needs. However, some advances are still being made regarding phenotypic data. Phenopackets (2019), developed by the Monarch Initiative[1] and endorsed by the Global Alliance for Genomics and Health (GA4GH)[2], is a standard file format for the exchange of phenotypic data. The proposed file format comprises, anonymously, relevant demographic data, detailed descriptions of a patient's phenotypes (e.g., clinical diagnosis, age of onset, disease severity, lab tests results), any potential diagnoses, and genetic information. By removing some of the barriers associated with phenotypic data, namely the lack of uniformity across organisations and usage of different data formats, Phenopackets aspires to facilitate communication between organisations with the overall goal of improving the ability to understand, diagnose, and treat diseases, both rare and common.

The broader notion of phenotype is normally associated with genotype (i.e, genetic constitution of an individual organism). A phenotype is used to refer to the set of observable characteristics of an individual that result from the interaction of its genotype with the environment. Most electronic phenotyping methods associate phenotypes to the diseases/conditions that afflict a certain population; however, phenotypes can also be representative of exposure (i.e., medications prescribed, smoking status, BMI) and outcome criteria (i.e., death, hospitalization). There are several possible applications for electronic phenotyping. Banda et al. (2018) reviewed the uses and applications of EHRs regarding phenotyping. They concluded that EHR-based phenotyping was being used in a wide range of (i) cross-sectional studies (e.g., epidemiological research, quality measurement), (ii) association (case-control/cohort) studies (e.g., genome-wide association studies, pharmacovigilance, identifying clinical risk factors and protective factors), and (iii) experimental studies (e.g., clinical trial recruitment).

Due to the heterogeneity and incompleteness of EHR data, correctly identifying phenotypes in EHRs is a time consuming and challenging task. Electronic phenotyping methods can be divided into two families according to the type of data used:

- The first family uses structured, and mainly administrative, data from EHRs. One of the commonly used coding systems for EHR structured data is the International Classification of Diseases (ICD) System;
- The second kind of methods relies on the extraction of information from clinical documents, such as notes from the patient-doctor interaction. These methods are able to extract more accurate

[1] https://monarchinitiative.org
[2] https://www.ga4gh.org

and representative information, but also face the inherent challenges of correctly parsing complex narratives, which can be repleted with misspellings, ungrammatical text, and abbreviations.

2.2.1 The International Classification of Diseases System

In my M.Sc. research project, I have used ICD diagnostic and procedures codes as means to identify patient cohorts and as a gold standard for validating a phenotyping algorithm. Developed by the World Health Organization (WHO), the ICD is, according to WHO (2018), "the standard diagnostic tool for epidemiology, health management, and clinical purposes" including "the analysis of the general health situation of population groups". This classification system characterises the universe of all health related conditions and arranges it into a hierarchical structure.

As an example of how the ICD-9 system lists every code, we take "Acute coronary occlusion without myocardial infarction". This diagnosis is associated to code 411.8.1, which is inserted into block 411, respectively named "Other acute and subacute forms of ischemic heart disease". The fourth digit of the full-code states that the etiology is "Other" and the fifth indicates the anatomic site "coronary artery". This block is located into the "Ischemic Heart Disease" (410 – 414) section of chapter 7, "Diseases Of Circulatory System" (390 – 459).

ICD enables:

- Easy storage, retrieval and analysis of health information for evidenced-based decision-making;
- Sharing and comparing health information between hospitals, regions, settings and countries;
- Data comparisons in the same location across different time periods.

The ICD system is regularly updated and new revisions are presented from time to time, in order to keep up with the constant health developments and medical science advances. Currently, the most recent edition of ICD is the 11^{th}, which was launch on the 18^{th} of June 2018. It is expected that member-states start reporting in ICD-11 on January 1^{st} 2022. In Table 2.4 one can observe the main differences, pointed out by the American Medical Association (2015), between the 9^{th} and 10^{th} revisions, regarding diagnostic codes, of the ICD classification system.

In my M.Sc. research project, both the 9^{th} and 10^{th} revision of the ICD classification system were used. The MIMIC-III dataset, proposed by Johnson et al. (2016) and presented in Chapter 3, is coded according to the ICD-9 system, whereas the Enroll dataset, sponsored by the CHDI Foundation (2012) and presented in Chapter 5, is coded with the ICD-10. The ICD systems are comprised of diagnostic and procedures codes. Diagnostic and procedure codes were used in Chapter 3, whereas only diagnostic codes where used in Chapter 5.

In the ICD-9-Clinical Modification (CM), the diagnostic codes are collected in Volume 1 and 2, presented in Table 2.1, while Volume 3 gathers the procedures codes, summarised in Table 2.2. In the ICD-10, the diagnostic and procedure codes are divided into ICD-10-CM and ICD-10-Procedure Coding

Table 2.1: ICD-9-CM chapters for diagnostic codes.

Chapter	Blocks	Title
I	001 – 139	*Infectious And Parasitic Diseases*
II	140 – 239	*Neoplasms*
III	240 – 279	*Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders*
IV	280 – 289	*Diseases Of The Blood And Blood-Forming Organs*
V	290 – 319	*Mental Disorders*
VI	320 – 389	*Diseases Of The Nervous System And Sense Organs*
VII	390 – 459	*Diseases Of The Circulatory System*
VIII	460 – 519	*Diseases Of The Respiratory System*
IX	520 – 579	*Diseases Of The Digestive System*
X	580 – 629	*Diseases Of The Genitourinary System*
XI	630 – 679	*Complications Of Pregnancy, Childbirth, And The Puerperium*
XII	680 – 709	*Diseases Of The Skin And Subcutaneous Tissue*
XIII	710 – 739	*Diseases Of The Musculoskeletal System And Connective Tissue*
XIV	740 – 759	*Congenital Anomalies*
XV	760 – 779	*Certain Conditions Originating In The Perinatal Period*
XVI	780 – 799	*Symptoms, Signs, And Ill-Defined Conditions*
XVII	800 – 999	*Injury And Poisoning*
	E800 – E999	*Supplementary Classification of External Causes of Injury and Poisoning*
	V01 – V89	*Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services*

System (PCS), respectively. Only the ICD-10-CM is used in this dissertation, and this can be seen in Table 2.3.

Overall, the ICD system is widely adopted by institutions and is the standard EHR diagnostic tool. However, the codes serve, primarily, a billing purpose, which leads to less prevalent or less diagnosed diseases being overshadowed by conditions associated with higher costs.

2.2.2 Phenotyping Evaluation

On what regards evaluation, the task of electronic phenotyping from any kind of data type, either structured or free-text, can be seen as a standard classification problem. Using the example of algorithms that use unstructured data, phenotypes are extracted from clinical narratives and evaluated against the gold standard, normally obtained by domain experts through manual revision of each clinical note.

Performance metrics are essential to assess the quality of detection/extraction methods, and phenotyping algorithms are no exception to this. A confusion matrix, presented in Table 2.5, is a common way to visualise the performance of an algorithm.

Several performance metrics have been used to evaluate the performance of electronic phenotyping algorithms. For example, Wei et al. (2016) used *Precision*, *Recall*, and *F1-score*, while Zeng et al. (2019)

Table 2.2: ICD-9-CM Volume 3 (procedures codes)

Section	Blocks	Title
nulla	00 – 00	*Procedures And Interventions , Not Elsewhere Classified*
I	01 – 05	*Operations On The Nervous System*
II	06 – 07	*Operations On The Endocrine System*
III	08 – 16	*Operations On The Eye*
IIIa	17 – 17	*Other Miscellaneous Diagnostic And Therapeutic Procedures*
IV	18 – 20	*Operations On The Ear*
V	21 – 29	*Operations On The Nose, Mouth, And Pharynx*
VI	30 – 34	*Operations On The Respiratory System*
VII	35 – 39	*Operations On The Cardiovascular System*
VIII	40 – 41	*Operations On The Hemic And Lymphatic System*
IX	42 – 54	*Operations On The Digestive System*
X	55 – 59	*Operations On The Urinary System*
XI	60 – 64	*Operations On The Male Genital Organs*
XII	65 – 71	*Operations On The Female Genital Organs*
XIII	72 – 75	*Obstetrical Procedures*
XIV	76 – 84	*Operations On The Musculoskeletal System*
XV	85 – 86	*Operations On The Integumentary System*
XVI	87 – 99	*Miscellaneous Diagnostic And Therapeutic Procedures*

described several techniques that used *Accuracy*, *Specificity*, and *Sensitivity* (i.e., the same as *Recall*). Regardless of the performance metrics being used, it is important to understand that there will always be a trade-off between pairs of performance metrics, where the increase of one results in the decrease of the other. The decision of which metric to favour depends on the nature of the problem. *Precision/Recall* and *Specificity/Sensitivity* are the most common trade-offs for classification problems.

Precision, also called Positive Predicted Value, represents the ratio of correctly predicted positive observations to the total predicted positive observations (see Equation 2.1). *Recall*, also known as True Positive Rate or Sensitivity, is the ratio of correctly predicted positive observations to the all observations in actual positive class (see Equation 2.2). *F1-score* is the weighted average of Precision and Recall (see Equation 2.3).

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2.1}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2.2}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.3}$$

Specificity/Sensitivity, widely used in medicine, are statistical measures that assess the performance of binary classification problems. *Specificity*, also called True Negative Rate, is the ratio of correctly

Table 2.3: ICD-10-CM chapters.

Chapter	Blocks	Title
I	A00–B99	*Certain infectious and parasitic diseases*
II	C00–D48	*Neoplasms*
III	D50–D89	*Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism*
IV	E00–E90	*Endocrine, nutritional and metabolic diseases*
V	F00–F99	*Mental and behavioural disorders*
VI	G00–G99	*Diseases of the nervous system*
VII	H00–H59	*Diseases of the eye and adnexa*
VIII	H60–H95	*Diseases of the ear and mastoid process*
IX	I00–I99	*Diseases of the circulatory system*
X	J00–J99	*Diseases of the respiratory system*
XI	K00–K93	*Diseases of the digestive system*
XII	L00–L99	*Diseases of the skin and subcutaneous tissue*
XIII	M00–M99	*Diseases of the musculoskeletal system and connective tissue*
XIV	N00–N99	*Diseases of the genitourinary system*
XV	O00–O99	*Pregnancy, childbirth and the puerperium*
XVI	P00–P96	*Certain conditions originating in the perinatal period*
XVII	Q00–Q99	*Congenital malformations, deformations and chromosomal abnormalities*
XVIII	R00–R99	*Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified*
XIX	S00–T98	*Injury, poisoning and certain other consequences of external causes*
XX	V01–Y98	*External causes of morbidity and mortality*
XXI	Z00–Z99	*Factors influencing health status and contact with health services*
XXII	U00–U99	*Codes for special purposes*

Table 2.4: Comparison between 9^{th} and 10^{th} revision of the ICD classification system, as provided by the American Medical Association (2015).

ICD-9-CM	ICD-10-CM
3-5 characters in length	3-7 characters in length
Approximately 13,000 codes	Approximately 68,000 available codes
First digit may be alpha (E or V) or numeric; digits 2-5 are numeric	First digit is alpha; digits 2 and 3 are numeric; digits 4-7 are alpha or numeric
Limited space for adding new codes	Flexible for adding new codes
Lacks detail	Very specific
Lacks laterality	Has laterality (i.e., codes identifying right vs. left)

predicted negative observations to all observations in the actual negative class (see Equation 2.4).

Table 2.5: Confusion matrix layout

		Actual class: Positive	Actual class: Negative
Predicted class	Positive	*TruePositive*	*FalsePositive*
	Negative	*FalseNegative*	*TrueNegative*

Sensitivity is already presented in Equation 2.2.

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{2.4}$$

Finally, we have that *Accuracy* represents the ratio of correctly predicted observation to the total observations. It is defined by Equation 2.5.

$$Accuracy = \frac{TrueNegatives + TruePositives}{AllSamples} \tag{2.5}$$

In multi-label classification (i.e., problem in which we can have multiple labels simultaneously associated with a single instance) it is important to evaluate the overall performance of the system. This can be achieved by dividing the original multi-label scenario into several binary classification problems for each of the different classes. Then, the overall performance metrics can be obtained either by computing the metric on individual class labels first and then averaged over all classes (i.e., macro-averaging), or by computing them globally over all instances and all class labels (i.e., micro-averaging). A macro-average computes the metric, independently for each class, and takes the average, treating all classes equally, while micro-average aggregates the contributions of all classes to compute the average metric, being preferable when dealing with imbalanced data.

2.3 Electronic Phenotyping Methods

A large variety of studies have been developed to tackle the challenge of identifying patient cohorts using all types of EHR data. The first phenotyping algorithms, corresponding to rule-based systems, consisted on simple rules applied to structured data. Recently, more complex systems have surfaced that allow for improved results by extracting valuable clinical data from clinical narratives. Banda et al. (2018) identified as primary systems for electronic phenotyping three different approaches: Rule-based, Natural Language Processing (NLP), and Machine Learning (ML). Rule-based and ML systems are considered to belong to the family of administrative phenotyping algorithms (i.e., algorithms that use structured data collected from statistics extracted from EHRs). On the other hand, NLP methods are considered to be all methods, either rule-based or using ML, that at some point extract information from clinical texts to obtain patients' phenotypes. These three types of methods for electronic phenotyping will be discussed

throughout this section.

2.3.1 Rule-Based Methods

Rule-based methods are the traditional approach to EHR-based phenotyping. They normally require clinicians to specify certain criteria for inclusion and exclusion. These methods have a widespread use and can achieve robust results. Shivade et al. (2014) pointed out that rule-based systems commonly used diagnosis codes and patient demographics as primary data sources. However, other structured data elements can be used in these methods, such as electronic prescriptions, lab measurements, and procedure codes.

Phenotypes with clear procedures and diagnosis codes are easily identified using rule-based systems. Jensen et al. (2012) examined publications, between 1997 and 2008, that used EHRs to identify patients with Atrial Fibrillation (AF). They concluded that the ICD-9 code 427.31 alone was sufficient to classify cases of AF with a relatively high performance – positive predicted values (precision) ranging from 70% to 96% and a median value of 89%. Tonelli et al. (2015) gathered and validated algorithms to recognise the presence or absence of 30 chronic conditions. These algorithms relied on mentions of ICD-9-CM/ICD-10-CM codes and evidences of hospitalisations. Algorithms with both positive predictive value and sensitivity greater than 70% were considered to be of "high validity". 16 conditions were identified for which the algorithm had "high validity". Hvidberg et al. (2016) conducted a similar study and created a catalog of 199 chronic conditions. This catalog used ICD-10-CM as the primary data source for disease categorisation; however, registers of prescribed drugs and use of practitioners' services were used in 35 conditions.

The use of multimodal search queries (i.e., queries comprising information from different data sources) usually shows better performance when comparing to single code queries. Wei et al. (2016) showed that queries using only one diagnostic code, applied to 10 diseases (e.g., Alzheimer, Parkinson, or multiple sclerosis) had significant variances in F1-scores, with an average value of 0.17. However, when using two or more diagnosis codes, the average F1-score significantly increased to 0.60. The combination of diagnostic codes with medication references or even keyword mentions from unstructured data sources, like clinical notes, also improved query performance – to an average score of 0.70.

Several studies have been dedicated to the development of rule-based algorithms to characterise specific diseases. Highly prevalent and chronic diseases are usually the main focus of these studies as they have a greater impact on health. Martucci et al. (2013) built a rule-based classifier for Chronic Obstructive Pulmonary Disease (COPD) identification that required the presence of three or more ICD codes. This algorithm had a sensitivity of 97.6%, a specificity of 76.0%, and a positive predicted value of 57.1%. Another algorithm, combining ICD codes and a mention of oxygen use in clinical notes, was also developed and achieved higher values of sensitivity and positive predicted value, respectively of 95.0% and 86.5%. Both Franchini et al. (2018) and Tison et al. (2017) developed algorithms for identifying Heart Failure (HF). In the first case, Franchini et al. (2018) proposed the CARPEDIEM algorithm which used ICD-9 codes and drug prescriptions as markers of HF. Regarding Tison et al. (2017), their algorithm,

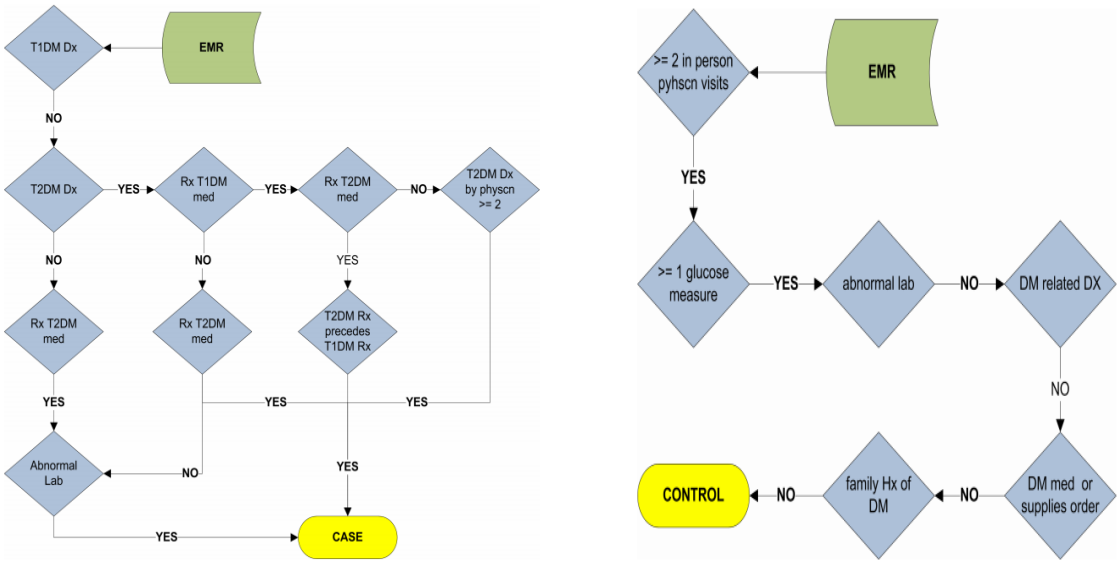


Figure 2.1: The algorithm proposed by Pacheco and Thompson (2012) for identifying both T2DM cases and controls.

very similar to the one developed by Franchini et al. (2018), considered elevated N-terminal pro B-type Natriuretic Peptide (NT-proBNP) lab results as an additional marker of HF, having achieved results agreeing with those of the CARPEDIEM method.

Overall, rule-based systems are fairly easy and fast to implement, especially considering limited datasets. However, most of these systems are never properly validated, as they are only used in a specific dataset and never shared throughout different healthcare settings (i.e., tested on other datasets apart from the one which they were created on). Also, rule-based phenotyping methods can be limited by the complexity of the phenotypes under analysis, and by the level of standardisation of the datasets used. With the objective of allowing transportability (i.e., shareability between different institutions) of electronic phenotype algorithms, Kirby et al. (2016) proposed PheKB (Phenotype Knowledgebase) to work as a repository of phenotypes. As of 2018, according to Banda et al. (2018), the majority of publicly available phenotypes on PheKB were rule-based. As an example of a rule-base algorithm from PheKB, Pacheco and Thompson (2012), from the Northwestern University, developed an algorithm for extracting both Type 2 Diabetes Mellitus (T2DM) cases and T2DM controls from medical records (see Figure 2.1). According to this algorithm, to be included in the control group, a patient's record must include two or more interactions with a physician, including at least one glucose measure, while having no abnormal glucose results, no personal or family history of diabetes, and no prescription of medication associated with diabetes.

PheKB is still expanding, and it currently already encompasses NLP methods for phenotyping. However, Halpern et al. (2016) pointed out that PheKB does not focus on algorithms that activate clinical decision support in real time and that, by requiring a rigorous developing process and clinician consensus, is very time consuming. These authors have developed an alternative phenotype library with 42 publicly available interpretable and fast to build phenotypes. Instead of using rule-based systems, this library

used established ML techniques as a resource to create statistical methods to estimate the phenotypes.

2.3.2 Machine Learning Methods

Machine Learning (ML) has been embraced by the field of biomedical informatics for a variety of tasks. These methods were recently adopted for computational phenotyping due to their high accuracy and scalability. ML approaches represent each patient as a vector of features, and they can be divided in three major categories (i.e., supervised, semi-supervised, and unsupervised). All machine learning methods require training in order to achieve results. The training data is said to be labeled when it has the correct answers attached to it. Additionally, classical statistical machine learning methods, mainly supervised ones, are commonly used in phenotyping due to their capacity to provide confidence estimates on the obtained classification.

Supervised learning algorithms require labelling of each sample in the training set. According to Zeng et al. (2019), logistic regression, Bayesian networks, and Support Vector Machine (SVM) classifiers are among the most popular supervised statistical machine learning methods used in electronic phenotyping. Shao et al. (2019) used a logistic regression model to detect probable dementia cases in patients without a dementia-related diagnosis. The model was developed using features from structured and unstructured data (i.e., ICD-9 codes, medications, Current Procedural Terminology (CPT) codes, or mentions in clinical notes). The logistic regression model was compared to manually reviewed records and it obtained a high level of agreement with them. Figueroa and Flores (2016) presented a method for automatic identification of obesity and categorisation of obesity status (i.e., super obesity, morbid obesity, severe obesity, or moderate obesity). They used and compared Naïve Bayes and SVM models to evaluate the performance of each approach, with SVM obtaining the best overall performance.

Unsupervised learning, in contrast with supervised learning, is able to automatically predict labels from unlabeled samples by clustering samples with similar patterns into groups. This eliminates the need for the time-consuming and labor-intensive task of labeling clinical data. As explained by Zeng et al. (2019), unsupervised learning can detect easily missed data patterns and, consequently, discover novel phenotypes. The discovery of new phenotypes has the potential to offer useful clinical information, but is also associated with increased challenges regarding their interpretations. One example of unsupervised learning applied to computational phenotyping is the work of Roque et al. (2011) which represented patients' profiles as vectors of ICD-10 codes. The cosine similarity (i.e., a measure of similarity between two non-zero vectors of an inner product space) scores between pairs of vectors was used as distance metric, and hierarchical clustering (i.e., grouping of similar objects into clusters) allowed for the identification of 26 clusters within $2,584$ patients.

2.3.3 Natural Language Processing Methods

Clinical narratives present the main source of information for a correct phenotyping process, as well as the greatest challenge. NLP, commonly referred to, in the medical field, as text-mining, evolved out of

the field of linguistics with the rise of computers. It allows one to extract knowledge from unstructured text in a high-throughput way. The earliest methods consisted on pattern-matching against standard vocabularies. More recently, most NLP techniques focus on analysing the semantic relationships within text. NLP-based algorithms have become crucial for electronic phenotyping. These methods can either consist of rule-based or ML approaches, supervised and unsupervised.

When integrating rule-based systems with NLP techniques, keyword search and term extraction are the least complex and easily implemented algorithms for computational phenotyping. More complex NLP systems use semantics to identify the context of certain detected concepts. A textual instance, such as a clinical note, can be seen as a hierarchical structure. This reflects the idea that a text is composed of sentences that in turn are composed of words. Semantics studies the meaning or relationship between words or set of words. The use of NLP systems that consider semantics allows for detecting uncertainty, negation, and parsing temporal relationships. Detecting negations and uncertainties of concepts in clinical text can significantly improve the precision and recall of the phenotyping algorithm.

With keyword search systems, algorithms use keywords, derivations of keywords, or a combination of keywords to identify phenotypes. Keywords can be related to, for example, prescribed medications, diagnostics, procedures, family history, or demographic data. Nath et al. (2016) created an NLP based approach, named EchoInfer, to analyse echocardiography reports. EchoInfer used regular expressions and specific keywords, such as “aortic prosthesis” and “aortic regurgitation”, to extract information regarding valvular heart disease. This tool achieved a precision of 94.06%, a recall of 92.21%, and an F1-score of 93.12%. Ware et al. (2009) developed an NLP framework, focused on obesity and its co-morbidities, that used keywords to extract clinical findings and diagnoses. The framework considered several medical concepts such as medications (e.g., salbutamol, albuterol, and proventil), numerical features (e.g., BMI, weight, and height), and observed diseases. The results presented, overall, a good level of agreement with the physician annotated gold standard.

Regarding term extraction, most studies use tools that map textual elements and obtain the corresponding Unified Medical Language System (UMLS) concepts (Bodenreider (2004)). Carroll et al. (2011) used the KnowledgeMap Concept Identifier (KMCI), from Denny et al. (2003), to process clinical notes and obtain UMLS Concept Unique Identifiers (CUI) together with additional qualifiers, such as negation status (i.e., whether the identified concept is negated in the text or not). The extracted concepts were then filtered and used to identify a cohort of patients with rheumatoid arthritis. Nguyen et al. (2010) developed a classification system able to identify lung cancer stages using textual information. To achieve this, they used a medical text extraction system, named MEDical Text EXtraction (MEDTEX), that mapped Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) concepts, from SNOMED International (2020), from free-text, while also identifying negation and possibility phrases. Figure 2.2 presents the MEDTEX pipeline, which first obtains UMLS concepts using the UMLS annotator. SNOMED-CT concepts are obtained using a UMLS to SNOMED-CT mapper. The stage of the cancer was predicted from the SNOMED-CT concepts using rules based on lung cancer staging guidance. The performance of this system was compared to that of SVM-based text classification systems.

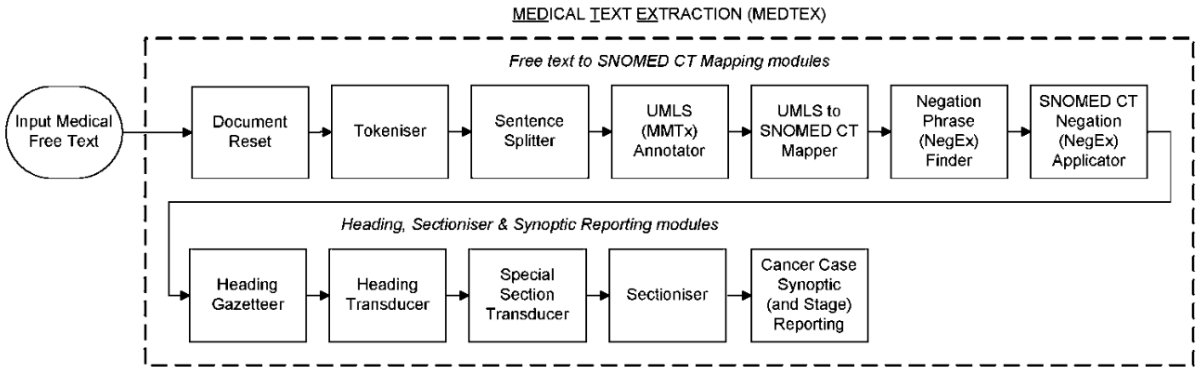


Figure 2.2: Medical text extraction (MEDTEX) pipeline application used to classify cancer stages, as originally proposed by Nguyen et al. (2010).

Working with semantic relationships, Byrd et al. (2017) developed an NLP pipeline for early detection of Heart Failure (HF). This pipeline involves two distinct phases, starting by the detection of mentions for HF that respect the Framingham criteria, from King et al. (2012). Then, a system analyses semantic relationships and labels each encounter as asserted, denied, or unknown. Both the mention extraction and labeling processes were able to achieve good results.

Several studies that check the presence or absence of a finding or disease mentioned in text use NegEx as a complement to their own algorithm (i.e., a system developed by Chapman et al. (2002), which uses regular expressions to search negation terms on the vicinity of the findings or disease mentions). This algorithm, despite being extremely simple, performs reasonably well. It achieves a specificity of 94.5%, a precision of 84.5%, and a recall of 77.8%. Peng et al. (2018) developed NegBio, that improves on NegEx by detecting uncertainties and increasing the searched word distance from the finding. Comparing to NegEx, NegBio had an average 9.5% improvement in precision, and 5.1% in F1-score.

More recently, with the integration of ML methods and specifically neural network models, several NLP techniques have been developed with very promising results. In the clinical domain, Wu et al. (2019a) identify word embeddings and Recurrent Neural Networks (RNN) as the state-of-the-art models used for natural language processing. Kwak and Hui (2019) define word embedding as the collective name for a set of techniques that maps words to vectors of real numbers.

Word embeddings are a representation of a document vocabulary, capable of capturing several textual attributes e.g., word context, as well as semantic and syntactic similarity. These embeddings are used as input for neural networks models. Word2vec, developed by Mikolov et al. (2013) at Google, is the most popular group of models to produce word embeddings. In word2vec, word vectors are located in the vector space such that words sharing common contexts, in the corpus of text, are positioned in close proximity to one another. Mikolov et al. (2013) introduced word2vec with two model architectures for distributed representation of words: continuous bag-of-words (CBOW) and skip-gram. The CBOW architecture is used to predict the current word from the surrounding context words, while the skip-gram architecture uses the current word to predict a surrounding window of context words. Several word embedding techniques have also been developed to be used with medical data, some tailored to the problem of

phenotyping. Patel et al. (2017) presented a method to add task-specific information to pre-trained word embeddings. In order to do so, they adapted the CBOW model from the word2vec package. Information from medical coding data was added to different pre-trained word embeddings, as well as the first level from the hierarchy of the ICD-10 medical code set. Their goal was to deal with medical synonyms and abbreviations. The original embeddings were consistently outperformed by the modified ones, for all five pre-trained embeddings used. Bai et al. (2018) proposes the JointSkip-gram model, based on the skip-gram framework of word2vec, which embeds both diagnostic codes and words from clinical notes, of the same medical encounter, in the same vector space. In this model, each ICD-9 code is used to predict all other codes and words, and each word is used to predict all diagnostic codes and neighbouring words. The joint representations were obtained using the MIMIC-III dataset. These representations proved to be effective at extracting phenotypes of different conditions and useful in predicting the reason for the following visit. Glicksberg et al. (2018) presented an unusual approach and used word2vec to create unsupervised embeddings of the phenotype space within an EHR system. The embeddings were learned using medical concepts from structured EHR data. These embeddings are used to summarise a patient's history over time windows and, ultimately, represent each patient as a vector of medical concepts. A query containing related medical concepts is created for each disease of interest. The disease cohort is obtained by selecting the patient vectors whose cosine distance to the query vector is below a certain threshold. This method was applied to research-grade cohorts for five diseases, and compared against the established PheKB electronic phenotyping algorithms for each disease. Although the performance at the disease level varied, the overall evaluation metrics show promising results with average F1-score of 0.57 and AUC of 0.98.

Recurrent Neural Networks (RNNs), as the name indicates, are neural networks that repeat themselves over time. These are a class of artificial neural networks that, contrary to the traditional neural networks, consider all inputs and outputs as dependent of each other. They do so by sequentially updating a hidden state based on the activation of the current input and the prior hidden state. By sharing parameters across different stages, the total number of parameters that an RNN has to learn is reduced. As noted by Kwak and Hui (2019), RNNs are specialized for time-series data and natural language. The self-loop connections and shared parameters enable this class of neural networks to, according to Pham et al. (2017), memorise previous inputs and capture longer dependencies than those obtained with alternative sequential models used in healthcare, such as hidden Markov models. However, as pointed out by Bengio et al. (1994), standard RNN models have limitations regarding long term dependencies in long input sequences. To overcome these limitations, RNN variants were developed, with Long Short-Term Memory (LSTM), proposed by Hochreiter and Schmidhuber (1997), and Gated Recurrent Units (GRU), proposed by Cho (2006), as the most popular.

The aforementioned RNN variants have been successfully applied to medical data. Pham et al. (2017) created DeepCare, built on LSTM, to deal with the episodic nature and time irregularities in medical records. This model analyses healthcare observations, registers previous disease history, deduces the present illness states, and, ultimately, predicts future medical outcomes. Baumel et al. (2017) compared four models on multi-label ICD classification of discharge summaries. The approaches presented by the

authors consisted on an SVM-based one-vs-all classifier, a continuous bag-of-words (CBOW) model, a Convolutional Neural Network (CNN), and a Hierarchical Attention-bidirectional Gated Recurrent Unit (HA-GRU) model. The HA-GRU method, created by the authors, is a hierarchical model, with two levels of bidirectional GRU encoding, to assign labels to the document by identifying the sentences relevant for each label. Out of the proposed approaches, the HA-GRU model performed the best, achieving state-of-the-art results. Dubois et al. (2017) used an RNN, with a bag-of-concepts representation of patient notes as time steps, to predict disease categories of patients. Rumeng et al. (2017) proposes a hybrid neural network model (HNN) for clinical text mining, which integrates an RNN. This approach sought to predict the presence (i.e., present, absent, possible, conditionally present, hypothetically present, and not associated with patients) and period (i.e., current, history, future, and unknown) assertions values associated with medical events. The proposed method was compared to standard SVM and LSTM models and obtained competitive results on both the period and presence tasks.

Similarly to rule-based phenotyping algorithms, the problem of creating interoperable pipelines for NLP techniques, that can be used across disciplines and test sites, also exists. Most authors end up developing their own NLP tool tailored to a specific task and rarely make them freely available. As pointed out by Wu et al. (2019b), reusing these models in new settings remains a heavy task that requires retraining and validation on new data. To overcome this, some frameworks have been created. Friedman et al. (1995) developed the medical language extraction and encoding (MedLEE) system to extract coded concepts from radiology reports. This tool is only available through licensing and lacks the capability to extract relations among identified entities. Aronson (2001) presented MetaMap as a framework to identify and map concepts, exclusively, from clinical narratives to the UMLS Metathesaurus, which is seen as a limitation. More recently, Savova et al. (2010), developed the clinical Text Analysis and Knowledge Extraction System (cTAKES), an open-source, large-scale, modular NLP system that combines rule-based and machine learning techniques to extract and process semantically viable information (e.g., diagnoses, symptoms, medication exposures) from clinical text. Reátegui and Ratté (2018) compared MetaMAP and cTAKES and concluded that cTAKES slightly outperformed MetaMAP.

In summary, NLP techniques add a great value to the task of electronic phenotyping by taking advantage of information stored in unstructured data, which has traditionally been neglected. Combining structured data with NLP yields significant benefits to both rule-based and ML phenotyping algorithms. The ability of being used to directly recognize phenotypes or to derive features, for ML approaches, strengthens the position of NLP as a cornerstone to the current and future electronic phenotyping toolkit.

2.4 Overview

Multimorbidity prevalence is set to rise. Understanding how chronic conditions occur together, their shared risks, and consequences for individuals should be the main health goals going forward. Some studies have already been conducted, where multimorbidity is the main focus, although disease-oriented studies are still the norm. This thesis proposes two studies of multimorbidity. A study on the consequences

Table 2.6: Primary methods for electronic phenotyping, with respective advantages and implementation challenges.

METHODS		ADVANTAGES	CHALLENGES
Rule-based		Straightforward construction; Easy implementation; High accuracy.	Rule developing is time-consuming and laborious; Requires domain expert; Rigid rules can lead to low recall.
Natural Language Processing		Takes advantage of unstructured data; Easily integrated with rule-based or machine learning methods.	Heterogeneity of clinical expression; Understanding causality and temporally between concepts; Detecting negations and uncertainties.
Machine Learning	Supervised Learning	High accuracy; High scalability.	Labelling process is time-consuming and laborious.
	Unsupervised Learning	High scalability; Low dependency on experts; Novel phenotype discovery.	Lower performance when compared to supervised learning; Novel phenotype interpretation requires domain experts.

(i.e., hospital admissions, ICU stays, mortality) of multimorbidity on an individual level, and an analysis of multimorbidity patterns over time.

Computational phenotyping is a field of great importance and interest in biomedical informatics. As seen above, several methods and approaches have been developed throughout the years with phenotyping in mind. Rasmussen et al. (2014), Mo et al. (2015), and Shang et al. (2019) provide helpful guides on optimising the implementation of a phenotyping algorithm. Table 2.6 summarises the advantages and challenges of each major family of phenotyping algorithms (i.e., rule-based, NLP, and ML).

Being able to correctly identify patients affected by multimorbidity is the first step to understand it. From what it was possible to infer from the related work revision, state-of-the-art phenotyping algorithms rely on NLP methods due to their ability to use clinical narratives. This dissertation intends to develop an NLP approach, using both structured and unstructured data from EHR, to identify patient cohorts for 12 predefined chronic conditions and their co-occurrence. The method, based on NegEx and NegBio, is designed to detect negations of disease mentions, and is assessed in the MIMIC-III database.

Chapter 3

Multimorbidity Information Extraction

This chapter presents an electronic phenotyping study developed for extracting information from clinical notes. This study was originally planned for processing a dataset from Hospital da Luz. Due to the current COVID-19 pandemic, the necessary treatment and anonymisation of the data was not made available. I have used, as an alternative, the Medical Information Mart for Intensive Care (MIMIC)-III Critical Care Database, from Johnson et al. (2016), to develop and test the method. Section 3.1 details the structure of the MIMIC-III dataset. Section 3.2 explains the selection process of the MIMIC-III dataset, along with statistical analysis, and the natural language processing (NLP) tool developed for Multimorbidity Information Extraction (MIE). Section 3.3 details the performance of the proposed MIE tool when applied to the MIMIC-III dataset. These results are finally discussed in Section 3.4.

3.1 MIMIC-III

MIMIC-III is a relational database containing 26 different data tables regarding patients who stayed within the Intensive Care Unit (ICU) at Beth Israel Deaconess Medical Center from June 2001 to October 2012 (see Figure 3.1). The database is de-identified according with the Health Insurance Portability and Accountability Act (HIPAA) standard. Fields such as patient name, telephone number, and address are removed, while dates are shifted into the future by a random offset for each individual patient in a consistent manner to preserve intervals. After the required institutional and ethical approvals, a *.csv* file was obtained for each data table. A table is a data storage structure which is similar to a spreadsheet: each column contains consistent information (e.g., patient identifiers), and each row contains an instantiation of that information. For this work, only 8 tables were needed to test and evaluate the created MIE tool (see Figure 3.2).

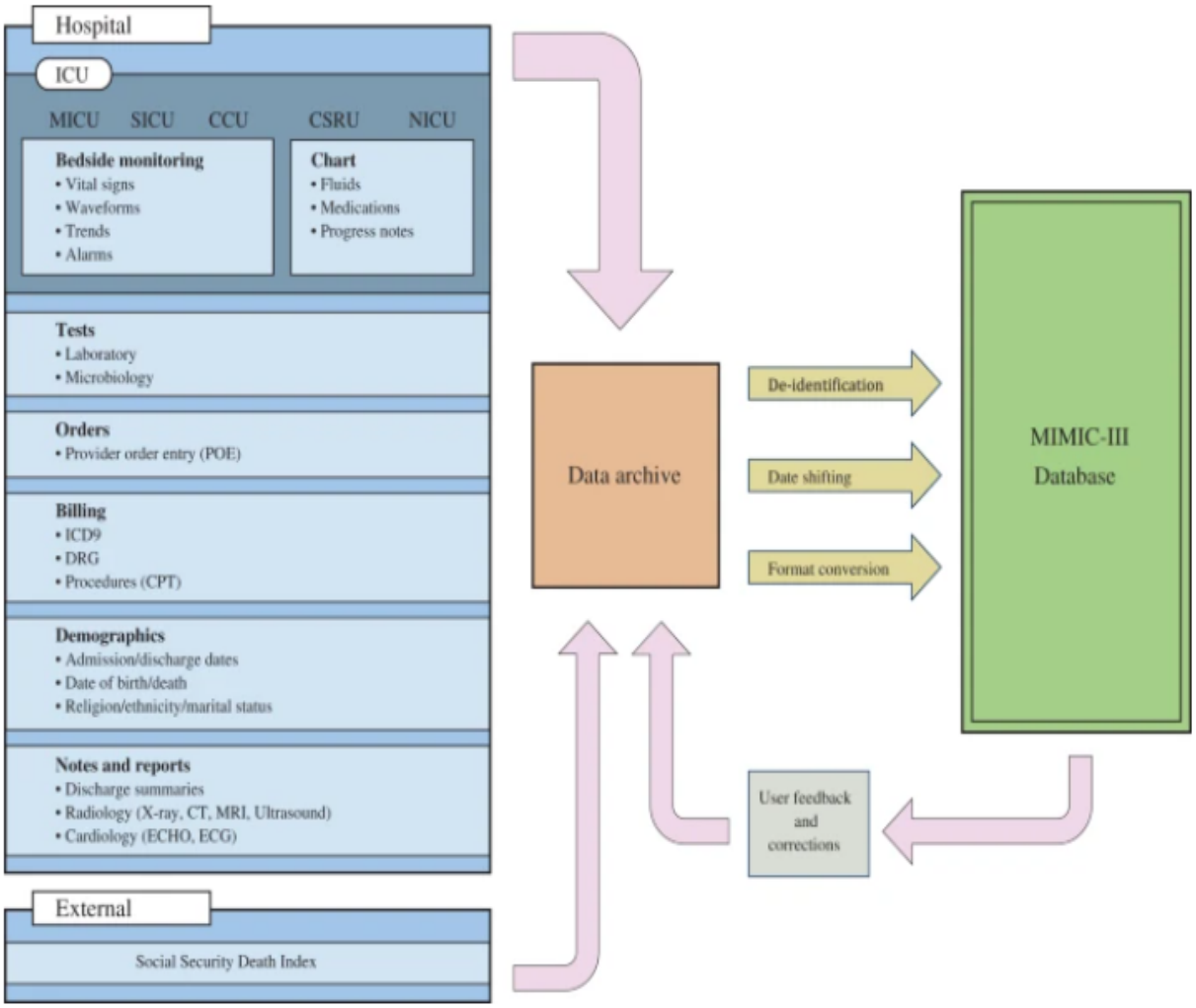


Figure 3.1: MIMIC-III critical care database overview from Johnson et al. (2016). The MIMIC-III database was populated with data of archives from critical care information systems, hospital electronic health record databases, and the Social Security Administration Death Master File. The data imported into the MIMIC-III database is first subjected to pre-processing operations: de-identification, date-shifting, and format conversion. Sensitive information was removed from free-text fields using a de-identification system based on extensive dictionary look-ups and pattern-matching with regular expressions. Each table is stored as a *.csv* file. User feedback is used to correct potential flaws in the pre-processing operations.

3.2 Electronic Phenotyping Methodology

This study focuses, primarily, on identifying patients with 12 chronic conditions. The latter selected following clinical experts' opinion on their clinical utility and practicality for the presented study. The pipeline of the proposed method for obtaining phenotypes from the MIMIC-III database is described in Figure 3.3. This section describes the selection and extraction processes used in the proposed method as well as a statistical analysis of the pre- and post-selection populations.

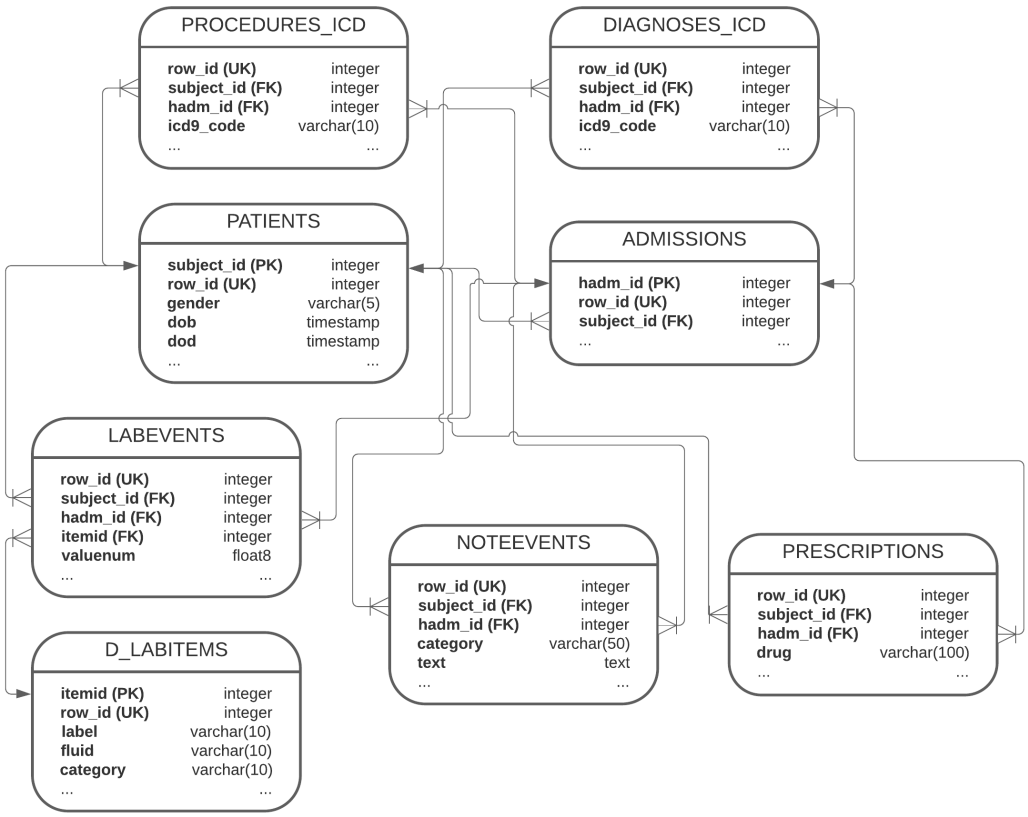


Figure 3.2: MIMIC-III database schema for used tables. **PATIENTS**: Every unique patient in the database. **ADMISSIONS**: Every unique hospitalisation for each patient in the database. **NOTEEVENTS**: De-identified clinical notes. **DIAGNOSES_ICD**: Hospital assigned diagnoses, coded using the ICD-9 system. **PROCEDURES_ICD**: Patient procedures, coded using the ICD-9 system. **LABEVENTS**: Laboratory measurements for patients, both within the hospital and in out patient clinic. **D_LABITEMS**: Dictionary of items in the laboratory database that relate to laboratory tests. **PRESCRIPTIONS**: Medications ordered for each patient.

Figure 3.3: Pipeline for the obtainment of phenotypes using the MIMIC-III database.

3.2.1 Data Selection and Analysis

Using all the tables, excluding *NOTEEVENTS*, described in Figure 3.2, I have extracted all the relevant information to characterise the dataset. This included information regarding a patients' age, gender, mortality, and number of admissions as well as previous diagnoses. Table 3.1 presents the statistical profile of the dataset before and after the selection process. I have defined rules, inspired by those from Hvidberg et al. (2016) and Tonelli et al. (2015), to identify the chronic conditions of interest using structured data (i.e., diagnostic and procedure codes, medications, lab results). The rules are described

Table 3.1: Statistical characterisation of the original MIMIC-III dataset and after pre-processing.

	Original		Selected	
	Total	Diseased	Total	Diseased
Number of patients	46 520	33 116	41 314	32 407
Number of male patients	26 121	18 803	23 306	18 408
Number of female patients	20 399	14 313	18 008	13 999
Number of admissions	58 976	44 848	53 691	44 132
Atrial Fibrillation prevalence	22.68%	31.86%	25.08%	31.97%
Chronic Kidney Disease prevalence	26.42%	37.12%	29.19%	37.21%
Chronic Obstructive Pulmonary Disease prevalence	13.98%	19.64%	15.46%	19.71%
Deafness/Hearing Loss prevalence	0.45%	0.63%	0.50%	0.64%
Dementia prevalence	4.01%	5.63%	4.40%	5.61%
Diabetes prevalence	22.41%	31.49%	24.76%	31.56%
Dyslipidemia prevalence	36.88%	51.81%	40.90%	52.14%
Heart Failure prevalence	24.09%	33.84%	26.74%	34.09%
Hypertension prevalence	47.08%	66.14%	52.00%	66.29%
Ischemic Cardiomyopathy prevalence	29.42%	41.33%	32.60%	41.56%
Obesity prevalence	4.87%	6.85%	5.44%	6.93%
Osteoarthritis prevalence	2.74%	3.85%	3.05%	3.88%
Percentage of diseased patients ($\geqslant 1$ morbidity)	71.19%	100%	78.44%	100%
Percentage of patients with multimorbidity ($\geqslant 2$ morbidity)	58.65%	82.38%	64.75%	82.55%

in Table 3.2. Figure 3.4 represents the 25 most common, single and co-occurring, combinations of the chronic health conditions in the original population.

Prior to handing the MIMIC-III data to the MIE tool, a selection process took place. The *NOTEEVENTS* table has a total of $2,083,180$ instances distributed over 15 different categories of clinical narratives (e.g., nursing, physician notes, radiology, discharge summaries, nutrition, social work). I have considered that only three categories (i.e., nursing, physician notes, discharge summaries) were enough to gather all relevant information for phenotyping, while reducing the total number of instances analysed. These categories result from direct contact between patient and care provider and summarised information from different sources (e.g., radiology reports, pharmacy reports). This selection process reduced the number of clinical narratives to $391,031$, after also removing duplicates.

The study population changed after the selection of clinical narratives. Compared to the numbers represented in the second and third columns of Table 3.1, the number of patients decreased to $41,314$, with a male predominance ($23,306$ males and $18,008$ females), which accounted for $53,691$ admissions. 78.44% of the population had been diagnosed, according to the criteria of Table 3.2, with at least one of the analysed diseases, and 64.75% of the individuals were a case of multimorbidity. The prevalence of the studied chronic conditions in the total and diseased pre-processed populations is displayed in the last two columns of Table 3.1.

Table 3.2: Rules applied to structured data from MIMIC-III to detect chronic diseases.

	Diagnostic Code	Procedures	Medication	Lab Results
Atrial Fibrillation	427.3X	N/A	N/A	N/A
Chronic Kidney Disease	583.XX; 584.XX; 585.XX; 586.XX; 592.XX; 593.9X	Dialysis	N/A	N/A
Chronic Obstructive Pulmonary Disease	416.8X; 416.9X; 490.XX; 491.XX; 492.XX; 494.XX; 495.XX; 496.XX; 497.XX; 498.XX; 499.XX; 500.XX; 501.XX; 502.XX; 503.XX; 504.XX; 505.XX; 506.4X; 508.XX	N/A	N/A	N/A
Deafness	389.9X	N/A	N/A	N/A
Dementia	290.XX; 941.XX; 294.11; 331.XX	N/A	Donepezil; Galantamine; Rivastigmine; Memantine	N/A
Diabetes	250.XX	N/A	N/A	Glucose levels >200 mg/dL
Dyslipidemia	272.XX	N/A	N/A	Total Cholesterol >5.2 mmol/L HDL-C <1.0 mmol/L LDL-C >3.4 mmol/L Triglycerides >1.7 mmol/L
Heart Failure	428.XX	N/A	N/A	NT-proBNP >450 pg/mL
Hypertension	401.XX; 402.XX; 403.XX; 404.XX; 405.XX	N/A	N/A	N/A
Ischemic Cardiomyopathy	410.XX; 411.XX; 412.XX; 413.XX; 414.XX	N/A	N/A	N/A
Obesity	278.XX	N/A	N/A	N/A
Osteoarthritis	715.XX	N/A	N/A	N/A

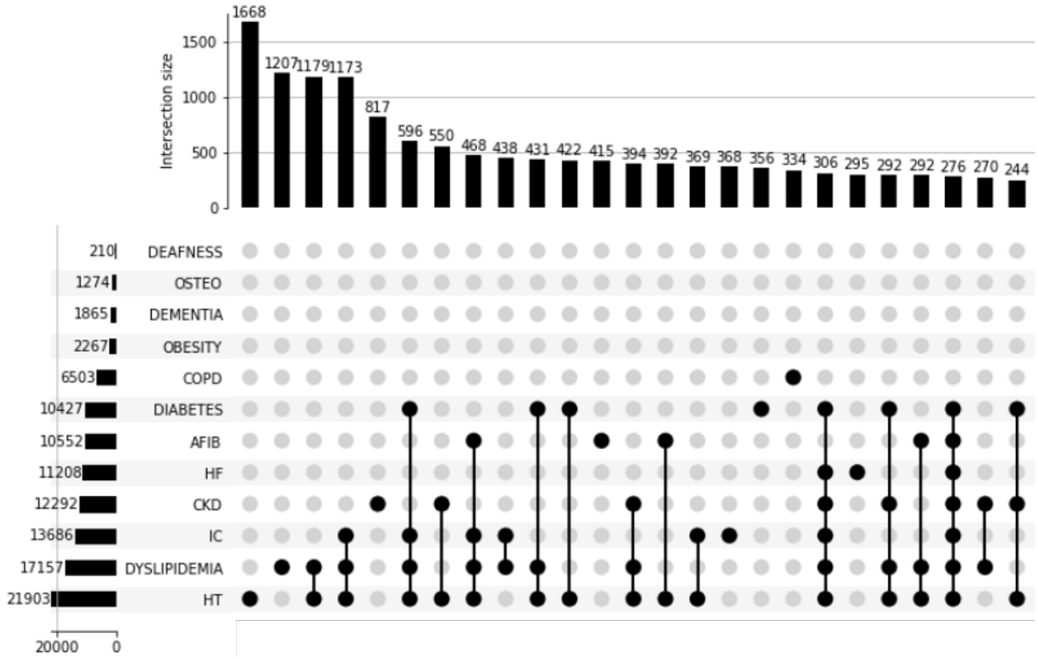


Figure 3.4: UpSet plot for the 25 most common, single and co-occurring, chronic health conditions in the original MIMIC-III dataset. Atrial Fibrillation: AFIB; Chronic Kidney Disease: CKD; Chronic Obstructive Pulmonary Disorder: COPD; Heart Failure: HF, Hypertension: HT; Ischemic Cardiomyopathy: IC; Osteoarthritis: OSTEO.

3.2.2 Information Extraction

The MIE tool proposed in this thesis takes as input the selected clinical reports and outputs labels for the presence or absence of the 12 chosen phenotypes. The tool, coded in the *Python 3.6* programming language, incorporates methods for identifying negated findings using regular expressions, taking inspiration from previous work on *NegEx* by Chapman et al. (2002). The proposed NLP pipeline has the following steps (see Figure 3.5a):

1) Newline control characters from the clinical notes are removed;

2) Reports are split into sentences according to the presence of full stops (i.e., "."");

3) Each sentence is matched for keywords associated with each of the phenotypes. Table 3.3 presents the list of keywords used to detect the presence or absence of each disease. The keywords were chosen based on previous studies, presented in Chapter 2, which used keyword mentions to identify patients afflicted with the chosen conditions. The lists also include, for each disease, the most popular abbreviations and synonyms for the main medical terms;

4) Matched sentences are cleaned of unnecessary characters (i.e., punctuation, symbols);

5) Matched sentences are divided into two separate segments. The pre-match and post-match, each including all the words occurring before and after the matched keyword in the original sentence, respectively. Figure 3.5b shows how the negation finding part of the algorithm works on a sentence;

6) Inspired by Chapman et al. (2002), the pre- and post-match sentences are searched, within a 6 word window, for expressions used to negate the mentioned keyword. Table 3.4 shows the negation phrases used to assert the negation of a keyword mention, depending on their position relative to the matched keyword;

7) For each identified disease, a corresponding label is assigned to the reports.

3.3 Evaluation

To evaluate the proposed NLP method for inferring phenotypes from clinical notes, the true ICD-9 diagnostic codes assigned in the MIMIC-III dataset (see Table 3.2) were compared to the algorithm's assertion regarding the presence of a corresponding disease. Several codes were assigned, for each patient stay, according to their priority – highest priority is associated with primary cause of visit. A true positive case was considered when the disease identified by the algorithm had an associated ICD-9 code throughout the patient's history. Additionally, for a small sample of instances associated to each chronic condition, the algorithm's results were compared to a gold standard obtained by medical doctors through manual revision. I have obtained measurements of *Precision*, *Recall*, and *F1-score*, using *Python*'s

Table 3.3: Keyword used to detect the presence or absence of each disease in clinical narrative text. We have a match if any of the positive keywords is found, for a certain condition, and none if any of the negative keywords is found.

	Positive keywords	Negative keywords
Atrial Fibrillation	Afib, AF, A-Fib	N/A
CKD	CKD, Chronic kidney disease, Chronic renal disease, CRD, Chronic renal failure, Chronic renal disorder, Chronic kidney failure, CRF, Chronic renal insufficiency	N/A
COPD	COPD, Chronic obstructive pulmonary disease, Chronic obstructive lung disease, Chronic obstructive airway disease	N/A
Deafness	Deaf, Deafness, Hypoacusis, Hearing loss, Hearing impairment, Hypoacuses	N/A
Dementia	Dementia, Memory loss, Cognitive impairment, mmse, Alzheimer, Senility	N/A
Diabetes	Diabetes, Diabetic, DM	N/A
Dyslipidemia	Dyslipidemia, Lipaemia, Lipemia, Lipidemia, Hyperlipidemia	N/A
Heart Failure	Ankle edema, Night cough, Heart failure, Cardiac failure, CF;	N/A
Hypertension	Hypertension, Hyper tension, HTN, High blood pressure	Pulm HTN, PHTN, Pulmonary hypertension, Pulmonary artery systolic hypertension
Ischemic Cardiomyopathy	Coronary artery bypass, Stent-LAD, Ischemic cardiomyopathy, Ischemic CM, Heart attack, Myocardial infarction, MI	N/A
Obesity	Obese, Obesity, BMI>30, Gastric bypass, Fatness, Corpulence, Overfatness, Overweight, Over-weight	N/A
Osteoarthritis	Osteoarthritis, Degenerative arthritis, Degenerative joint disease, Hypertrophic arthritis, Osteoarthrosis	N/A

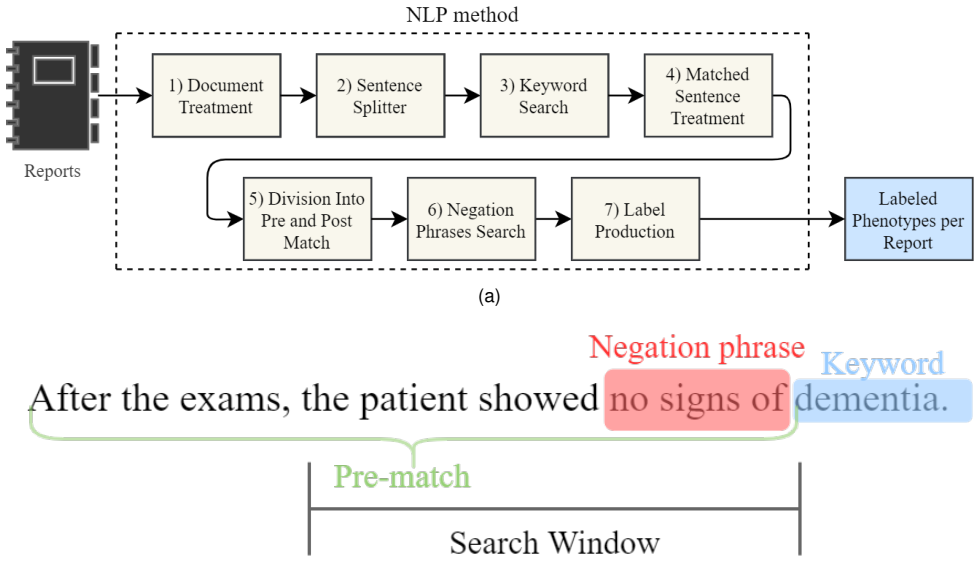


Figure 3.5: The MIE tool. (a) Pipeline for the extraction of phenotypes from clinical notes. (b) Negation finding process on example sentence.

Table 3.4: Negation phrases used in the negation finding part of the proposed NLP method.

	Pre-match	**Post-match**
Negation phrase	*no; not; absence of; declined; denies; denying; did not exhibit; no sign of; no signs of; not demonstrated; patient was not; rules out; ruled out; doubt; negative for; no cause of; no complaints of; no evidence of; without; without indication of; without sign of; no further; without any further; without further*	*was declined; unlikely; ruled out; was denied; was absent; not present*

Table 3.5: Performance metrics, with respective micro- and macro-averages, number of analysed instances and negations detected, for each chronic condition, for the NLP method against ICD-9 diagnostic codes presented in Table 3.2.

			Performance		
	Instances	Negations	Precision	Recall	F1-score
Atrial Fibrillation	173,562	2,015	95.62	98.92	97.24
CKD	38,743	211	97.85	99.46	98.65
COPD	76,986	1,172	85.88	98.92	91.94
Deafness	2,628	60	27.92	98.76	43.53
Dementia	24,973	229	45.78	99.37	62.68
Diabetes	132,499	3,369	91.07	98.86	94.81
Dyslipidemia	45,178	242	82.46	99.69	90.26
Heart Failure	138,216	2,020	92.47	97.88	95.10
Hypertension	248,901	1,992	89.73	99.37	94.31
Ischemic Cardiomyopathy	26,406	2,634	91.27	97.23	94.16
Obesity	53,512	5,861	53.39	93.05	67.85
Osteoarthritis	7,227	39	43.09	99.61	60.15
Micro-average	-	-	87.19	98.71	92.60
Macro-average	-	-	74.71	98.43	82.56
Total	968,831	19,844	-	-	-

SciKit-learn[1] package, for each disease. Additionally, by micro- and macro-averaging the performance metrics for each disease, I have obtained values of *Precision*, *Recall*, and *F1-score* for the overall performance of the method (micro-averaging is insensitive to the imbalance on the number of clinical notes per chronic condition).

Table 3.5 shows the performance of the proposed electronic phenotyping method in detecting the ICD-9 diagnostic codes from the clinical notes marked by the algorithm for each disease.

For each disease, a subset of 100 instances was created and sent to a medical doctor for manual revision and labeling. Additionally, in cases of disagreement between the algorithm and the doctor, a

[1] https://scikit-learn.org/stable/

Table 3.6: Performance metrics, with respective macro-averages, and number of negations detected, for each chronic condition, for the NLP method against labels obtained via expert manual revision of the 1200 clinical notes (100 from each class of ICD codes assigned on MIMIC-III).

		Performance		
	Negations	Precision	Recall	F1-score
Atrial Fibrillation	2	100	98.00	98.99
CKD	3	98.97	96.97	97.96
COPD	2	96.94	98.96	97.94
Deafness	2	97.96	97.96	97.96
Dementia	1	90.91	98.90	94.74
Diabetes	6	90.43	94.44	92.39
Dyslipidemia	3	100	98.98	99.49
Heart Failure	4	90.62	100	95.08
Hypertension	1	95.96	100	97.94
Ischemic Cardiomyopathy	1	100	99.00	99.50
Obesity	19	100	82.65	90.50
Osteoarthritis	1	100	100	100
Average	-	96.82	97.16	96.84
Total	45	-	-	-

small explanation was provided, justifying the reasoning behind the expert's label. Table 3.6 shows the performance of the proposed NLP method against the labels obtained by the expert's manual revision. In this case, by always using 100 instances for each disease, the data is balanced and standard average was used to obtain metrics for the overall performance of the algorithm.

3.4 Discussion

The proposed phenotyping method is capable of achieving good results, for most of the diseases under analysis, when using MIMIC-III assigned ICD-9 diagnostic codes as the gold standard. Independently of the condition studied, the values of *Recall* are always above 90%. This is due to the fact that the algorithm predicts mostly positive cases of keyword mentions, which increases the number of true positives. Regarding *Precision*, some diseases show significantly lower values than others.

To evaluate the performance of the NLP method used on each disease we can also look at results obtained in similar studies. I have searched for studies that used EHR data, preferably clinical notes, to identify patients with one or more of the chronic conditions studied. Unfortunately, none of the studies considered evaluated the MIMIC-III dataset, comprised of patients admitted to ICU, hence results are not directly comparable. Table 3.7 presents the performance, for each chronic condition in analysis, of the method developed in this thesis and methods developed in similar studies.

Deafness and osteoarthritis were the only conditions for which we found no previous study dedicated

Table 3.7: Performance metrics and number of analysed instances, for each disease, of methods developed in this thesis (first row of each group of rows) and in related work. Deafness and osteoarthritis not included due to no term of comparison. *Study not using NLP methods and clinical narratives.

		Performance		
	Instances	Precision	Recall	F1-score
Atrial Fibrillation	173,562	**95.62**	**98.92**	**97.24**
Wei et al. (2016)	1,732	72.00	3.00	7.00
CKD	38,743	**97.85**	**99.46**	**98.65**
Winkelmayer et al. (2005)*	1,852	91.60	20.7	33.77
COPD	76,986	85.88	**98.92**	**91.94**
Martucci et al. (2013)	200	**86.50**	97.00	91.00
Dementia	24,973	**45.78**	**99.37**	**62.68**
Shao et al. (2019)	1,861	N/A	82.50	N/A
Diabetes	132,499	**91.07**	**98.86**	**94.81**
Wei et al. (2016)	T1DM: 18,380 T2DM: 29,171	T1DM: 12.00 T2DM: 68.00	T1DM: 12.00 T2DM: 21.00	T1DM: 12.00 T2DM: 32.00
Dyslipidemia	45,178	82.46	**99.69**	90.26
Oake et al. (2017)*	4,400	**100**	94.00	**96.91**
Heart Failure	138,216	92.47	**97.88**	**95.10**
Byrd et al. (2017)	1,492	**92.52**	89.68	91.08
Hypertension	248,901	89.73	**99.37**	**94.31**
Teixeira et al. (2017)	631	**95.20**	90.2	92.63
Ischemic Cardiomyopathy	26,406	91.27	**97.23**	**94.16**
Ivers et al. (2011)	969	**91.30**	72.40	80.76
Obesity	53,512	53.39	93.05	67.85
Figueroa and Flores (2016)	3,015	UNK	UNK	**78.30**

to their phenotyping. This is representative of the level of importance given to these conditions, easily seen in Table 3.1 by the low prevalence in the studied population.

For some of chosen chronic conditions no studies were found that made use of clinical narratives, and employed NLP methods, to phenotype them. Despite this, to obtain some validation data for our method, I am still presenting the performance results of studies that only used structured data for phenotyping. Chronic kidney disease and dyslipidemia are the two conditions for which no NLP phenotyping method was found. Winkelmayer et al. (2005) used diagnostic codes and laboratory results to identify patients with Chronic kidney disease. On the other hand, Oake et al. (2017) identified patients with dyslipidemia using ICD codes and reports of abnormal lipid level in laboratory data. This method outperformed the proposed NLP electronic phenotyping algorithm. These high results reinforce the advantages of using rules to identify dyslipidemia from structured data in Table 3.2, and illuminate the benefits, in certain cases, of combining both structured and unstructured data in a phenotyping algorithm.

Wei et al. (2016) used different types of EHR data (i.e., billing codes, clinical notes, prescription history), both separately and combined, to obtained phenotypes for several diseases, including atrial

fibrillation and diabetes (type 1 and type 2 separately). The results were evaluated against manual reviewed labels. For comparison, we considered the values of precision and recall obtained using only the clinical narratives, as these are more representative of the results obtained in this thesis. All the values obtained are significantly lower than those obtained by the proposed algorithm. However, it is important to note that, there is no information regarding the keywords used to identify the diseases in the Wei et al. (2016) study. Additionally, my evaluation does not distinguish between types of diabetes, which would probably change the results shown in Table 3.5.

Regarding the low prevalence diseases, Shao et al. (2019) and Figueroa and Flores (2016) developed methodologies to identify patients with dementia and obesity, respectively. Shao et al. (2019) used a logistic regression model to detect probable dementia cases in patients without a dementia-related diagnosis. The model was developed using features from structured and unstructured data (i.e., ICD-9 codes, medications, CPT codes, clinical notes). Shao et al. (2019) obtained values for sensitivity (recall) and specificity of 82.5% and 83.2%, respectively. The proposed algorithm was not evaluated using specificity, but following Equation 2.4, and the obtained dementia's confusion matrix (i.e., TP=$11,480$, FP=$13,599$, FN=73, TN=156), the specificity associated with dementia for our method is 1.13%. Our method is clearly outperformed by the method developed by Shao et al. (2019), with respect to sensitivity; however, this may likely result from the underdiagnosis of this condition. Also, Shao et al. (2019) used labels from experts manual revision, while this thesis uses billing codes to obtain the gold standard.

Figueroa and Flores (2016) presented a method for automatic identification of obesity and categorization of obesity status (e.g., moderate obesity, morbid obesity) from clinical narratives using a Bag of Words (BOW) approach. They used and compared Naïve Bayes and SVM models to evaluate the performance of this approach, with SVM model obtaining the best overall performance for the task of identifying obesity. Once again, this method was evaluated against annotated records which are more representative of the true prevalence of a disease than billing codes.

Martucci et al. (2013) identified patients with Chronic Obstructive Pulmonary Disease using a combination of ICD codes and mentions of oxygen use in text. Byrd et al. (2017) developed an NLP pipeline for early detection of patients afflicted with heart failure.

Teixeira et al. (2017) combined diagnostic codes, medications, vital signs, and Unified Medical Language System (UMLS) concepts extracted using NLP to identify hypertensive patients.

Lastly, Ivers et al. (2011) identified patients with ischemic cardiomyopathy using the free-text within the medical history fields of the EHRs.

The results obtained, from the presented studies, for the four previous diseases (i.e., COPD, Heart failure, Hypertension, Ischemic cardiomyopathy) are all similar to those obtained in this thesis (see Table 3.7). Having in mind that these studies used labels resulting from manual revision, the similarity between results in the literature and this dissertation – besides proving the efficacy of the proposed method – can be justified by the fact that the diseases in question are well represented and, therefore, reported in the MIMIC-III dataset.

One major characteristic of the proposed NLP method is its ability to identify negated findings. Therefore, it is important to evaluate its overall results (see Table 3.5) against a similar algorithm. NegEx, developed by Chapman et al. (2002), was the chosen algorithm for this purpose, having been used as inspiration for the created pipeline. In Chapman et al. (2002), NegEx achieved a precision of 84.5% and a recall of 77.8% on the task of identifying whether a finding or disease mentioned within a clinical narrative is present or absent. Despite showing better results than NegEx, it is not reasonable to conclude that this thesis' method is superior to that of Chapman et al. (2002). It is crucial to state that NegEx does not narrow its search to 12 chronic conditions, but instead to all UMLS terms identified in the text. Additionally, NegEx is evaluated against annotated records and tested in a dataset where half of the matched sentences contain negation phrases. This is not the case of the MIMIC-III dataset, where the prevalence of instances containing negation phrases is significantly lower than 50%. Having said that, the method reported on this dissertation is able to identify negated findings, but has not been properly evaluated on its ability to do so. It would be interesting to evaluate this method on a dataset similar to that used by NegEx.

Similarly to most of the studies discussed above, the results from the proposed method were also compared to labels obtained from manual revision (see Table 3.6). However, manual revision was only obtained for 100 randomly selected instances per disease. Overall, the algorithm presented a high performance, regardless of the disease, which confirms that billing codes are not effective at representing the true prevalence of certain diseases. This is especially visible when looking at chronic conditions that are rarely coded using the ICD system (i.e., deafness, dementia, obesity, osteoarthritis). The combination of mostly positive predicted labels with fewer positive true labels results in a higher number of false positives and, consequently, lower values of *Precision*. Nevertheless, the main advantage of manual revision is the ability to analyse, one-by-one, the matched phrases and to see what are the algorithm's primary causes of error. The main reasons for disagreement between the algorithm evaluation and the medical doctor's revision were:

1. The keyword detected was used to describe family history of the patient and, therefore, was not directly connected to the patient's diagnosis. This resulted in an increase of false positives;

2. The doctor identified uncertainties and, therefore, the manual revision could not determine the true diagnostic;

3. The matched keyword was negated outside the word window considered by the algorithm – of size 6. This was typically seen when the clinical narrative included enumeration of diseases or symptoms. This resulted in an increase of false negatives;

4. Some clinical notes were structured in a way that the algorithm was not able to detect different sentences (i.e., instead of "." the doctor used "-", "#", capitalization, or simply just a space as sentence delimiters). The full stop was the general rule used to split clinical notes into sentences due to a low prevalence in the usage of the described alternative sentence delimiters. This resulted in some keywords being negated due to the word window including a previous phrase containing a

negation term. This results in an increase of false negatives;

5. References to medications for a certain disease were negated instead of the disease itself. For example: “Patient is currently not on diabetic medication”. This resulted in an increase of false negatives.

3.5 Overview

Clinical narratives play an important, and traditionally insufficient, role in the task of phenotyping from EHR data. This chapter describes all steps of development, implementation, and evaluation of an NLP-based electronic phenotyping tool, for the detection chronic conditions, using clinical narratives. Additionally, this algorithm is able to filter out disease mentions that appear to be negated. The algorithm was applied to the MIMIC-III Critical Care Database, which required data pre-processing to allow for model evaluation.

The proposed method was evaluated using true ICD-9 diagnostic codes and manual reviewed labels. The performance on the true ICD-9 diagnostic codes labels was compared against similar studies, for each disease. Overall, the method outperformed most of the analysed studies. NegEx, from Chapman et al. (2002), was compared with proposed method. NegEx was tested in a dataset containing a high prevalence of negation phrases and was not restricted to 12 conditions as in my evaluation. Therefore, no conclusion can be taken regarding superiority of the proposed method in comparison with NegEx, until both algorithms are tested in similar conditions.

The evaluation against labels obtained via expert manual revision of the clinical notes showed a high performance of the proposed method, for all tested conditions. Additionally, it provided crucial information for the future improvement of the method and on the ineffectiveness of using ICD-9 diagnostic codes to phenotype certain diseases.

Chapter 4

Comparison of Multimorbidity in COVID-19 Infected and General Population in Portugal

This study was developed in the special context of the COVID-19 pandemic and was published by Froes et al. (2020), made available in MedRxiv. The study was first published in June 2020 and later updated in August 2020. Since its release, several studies, focusing on the impact of COVID-19, were developed that might invalidate some of the statements made by Froes et al. (2020).

COVID-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that recently challenged health systems. According to WHO (2020), $21,200,000$ cases and $761,000$ deaths have been reported globally by August 16, 2020 [1]. Early evidence from the pandemic suggested that older patients with chronic conditions were over-represented and may herald a poor clinical course. However, as stated in Chen et al. (2020); Grasselli et al. (2020), data supporting these analyses is still largely limited to China and Italy. Given that multimorbidity increases with age, the specific risk of different chronic conditions and their combination in terms of poor health outcomes needs to be adjusted for age, which, according to Mason et al. (2020), is not routinely done. Understanding which groups are at higher risk is important to better inform public health policies and resource allocation, and to advance knowledge about this novel condition.

Directionate-General of Health (DGS), the Portuguese public health authority, developed and operates National Epidemiological Surveillance System (SINAVE) (i.e., the national public health surveillance system). SINAVE, DGS (2018), is used to collect, update, analyse, and disclose data related to infectious diseases with mandatory reporting, and other public health hazards. Symptoms, previous medical history, disease course, and laboratory results are introduced by physicians in charge of COVID-19 patients. Although SINAVE was not specifically designed for this outbreak, it has been used since the beginning of

[1] $82,835,563$ cases and $1,807,638$ deaths reported globally by December 31, 2020

the pandemic as the main official source of data about COVID-19 cases in Portugal.

This study evaluates the prevalence of multimorbidity and age-adjusted risk of hospitalisation, ICU admission, and death, in the Portuguese population from official data, based on a dataset[2] extracted from SINAVE containing all confirmed cases of COVID-19 infection, in Portugal, by June 30, 2020.

4.1 Dataset Description

This analysis used data from the DGS/SINAVE dataset, after the required institutional and ethical approvals, spanning the full period (i.e., the June version of the dataset, which is the most recent at the time of this report), that updates an initial (April) version of the dataset based on cases reported until April 28, 2020, adding two more months and providing more data about the initial cases.

The sample population consists of all the Portuguese population with SARS-CoV-2 confirmed infection, as notified by clinicians by June 30, 2020. A broad range of clinical and demographic variables are present in this dataset. In this study, variables corresponding to age, gender, hospital admission, admission in intensive care unit, mortality, and patient's underlying conditions were used.

Chronic conditions were originally provided as categorical variables on the presence, absence, or unknown status of the following conditions:

- Asthma
- Malignancy
- Chronic hematological disorder
- Diabetes
- HIV/other immune deficiency
- Renal disease
- Liver disease
- Chronic lung disease
- Neuromuscular/Neurological disorder

A field containing *raw* textual input from doctors was also taken into account, to better complement the cases where the chronic conditions were left as unknown. This alternative information was very useful, particularly on what regards cardiovascular disorders (including hypertension and other cardiovascular diseases), which were not included in the dataset as a categorical variable and could, therefore, not be detected if not for the *raw* input.

Addressed outcomes were hospitalisation, admission to ICU unit, and reported death. A composite outcome of any of these events was also analysed.

[2] https://covid19.min-saude.pt/disponibilizacao-de-dados/

4.2 Methodology

A text-mining script, using keywords associated with all the previously mentioned conditions, was used in order to better capture the prevalence of the diseases (i.e., in many cases, it was noticed that although the categorical variables corresponding to chronic conditions were left with an unknown status, the textual inputs contained relevant mentions to chronic conditions) and to detect cases of cardiovascular disorders. The following keywords, in Portuguese, were used, in connection to each of the diseases that were considered in the study:

- **Asthma:** asma;
- **Malignancy:** neo, cancro, carcinoma, linfoma;
- **Cardiovascular disorders (including hypertension and other cardiovascular diseases):** cardio, cárdio, miocar, cardia, cardía, hta, auricular, arterial, venosa;
- **Chronic hematological disorder:** hematológica;
- **Diabetes:** diabetes, DM;
- **HIV/other immune deficiency:** hiv, vih;
- **Renal disease:** renal;
- **Liver disease:** hepatomegalia;
- **Chronic lung disease:** dpoc, pulmonar;
- **Neuromuscular/Neurological disorder:** alz, parkinson, epilepsia.

The keywords were chosen, based on an empirical analysis of the textual field, in order to cover different cases, considering misspellings or abbreviations. For instance, *neoplasia* was commonly abbreviated to *neo* and the latter was therefore preferred (and also matched as a prefix). Another example is the use of the prefix *alz* to detect Alzheimer's disease, that was often misspelled as *Alzaimer*.

The contribution of using the *raw* textual input, when preparing the data, was also analysed. All the reported statistics were also measured in a version of the data that only considered the structured information. The main difference, as stated above, was the complete absence of any information on cardiovascular disorders, when not considering the textual information. There were also some differences in the prevalence of the other chronic conditions, although only reaching up to 0.16%.

Python 3.6 was the programming language considered due to its abundance of data analysis methods and support resources available. Several libraries were used in this project to analyse and visualise the data such as NumPy[3], SciPy[4], Pandas[5], Seaborn[6], and UpSetPlot[7]. Additionally, *IBM SPSS Statistics*[8]

[3]https://numpy.org
[4]https://www.scipy.org
[5]https://pandas.pydata.org
[6]https://seaborn.pydata.org
[7]https://pypi.org/project/UpSetPlot/
[8]https://www.ibm.com/products/spss-statistics

was used to obtained the age-adjusted risk of hospitalisation, ICU admission, and death.

Categorical variables were presented to analysis as counts and percentages with 95% confidence intervals, and comparisons were made using using the χ^2 test. Univariate regression analysis of each individual chronic condition was performed adjusting only for age. A multivariate logistic regression was performed adjusting for age and every other chronic condition with significant statistical association on univariate analysis. Results were considered statistically significant when $P < 0,05$.

4.3 Results

The overall sample contained $36,244$ adult patient cases, with women being more prevalent (56.66%). Among the cases, 18.79% had at least one chronic condition. Cardiovascular disorders were the most commonly reported condition, present in 43.33% of the patients with any morbidity. Table 4.1 shows the reported prevalence of different chronic conditions in the studied population.

Multimorbidity, as previously defined, was present in 6.77% of the cases (this number would be reduced to 4.01% if ignoring the textual input). Figures 4.1 and 4.2 plot the prevalence of multimorbidity by age group, respectively for (a) the COVID-19 infected general population, and for (b) the hospitalised population.

To analyse the Odd Ratio and prevalence of co-occurring pairs of chronic diseases, people with unknown disease prevalence were excluded, which resulted in a population of $33,283$ adult patients. The prevalence of co-occurring pairs of chronic health conditions, plotted in Figure 4.4, shows Cardiovascular disorders and Diabetes as the most common dyad of chronic diseases. Additionally, Figure 4.5 presents the 25 most common, single and co-occurring, chronic health conditions.

Data regarding hospitalisation and ICU admission was available for only $32,945$ patients (90.90% of the overall study population). Within this population, hospitalisation occurred in 12.89% of the patients, with a male predominance (50.66%), and ICU admission was required for 4.11% of the patients, with a female predominance (51.73%). Observed mortality was 3.19%. Tables 4.2 to 4.5 are representative of the univariate regression analysis, and Tables 4.6 to 4.10 of the multivariate regression analysis. All chronic conditions, except for asthma, were associated with increased risk of mortality and hospitalisation (Tables 4.6 and 4.7, respectively). Age, diabetes, renal disease, lung disease, and neuromuscular disorders, were all associated with increased risk of ICU admission (Table 4.8). Additionally, every additional chronic condition increases the risk for the patients of the composite outcome of death, hospitalisation, or ICU admission, by 123.3% (OR 2.22; CI 95%: $2.13 - 2.32$).

When comparing the two versions of the dataset that were released by the DGS, it is possible to state that the June version is more complete and representative than the April one. This is easily verified by the prevalence of cardiovascular disorders in the total population, with a value of 0.28% in the April version that rose to 8.14% in the most recent one. Concerning chronic diseases and multimorbidity prevalence, the first version displayed values of 16.39% and 4.49%, respectively. With respect to Odd Ratios, the June

Table 4.1: Percentage of COVID-19 infected total Portuguese population affected by each comorbidity.

	Asthma	Cancer	Cardiovascular Disorders	Diabetes	Hematological Disorder	HIV	Renal Disease	Liver Disease	Lung Disease	Neuromuscular Disorder
Population (%)	2.10	2.68	8.14	5.92	0.89	0.60	2.09	0.57	2.18	3.09

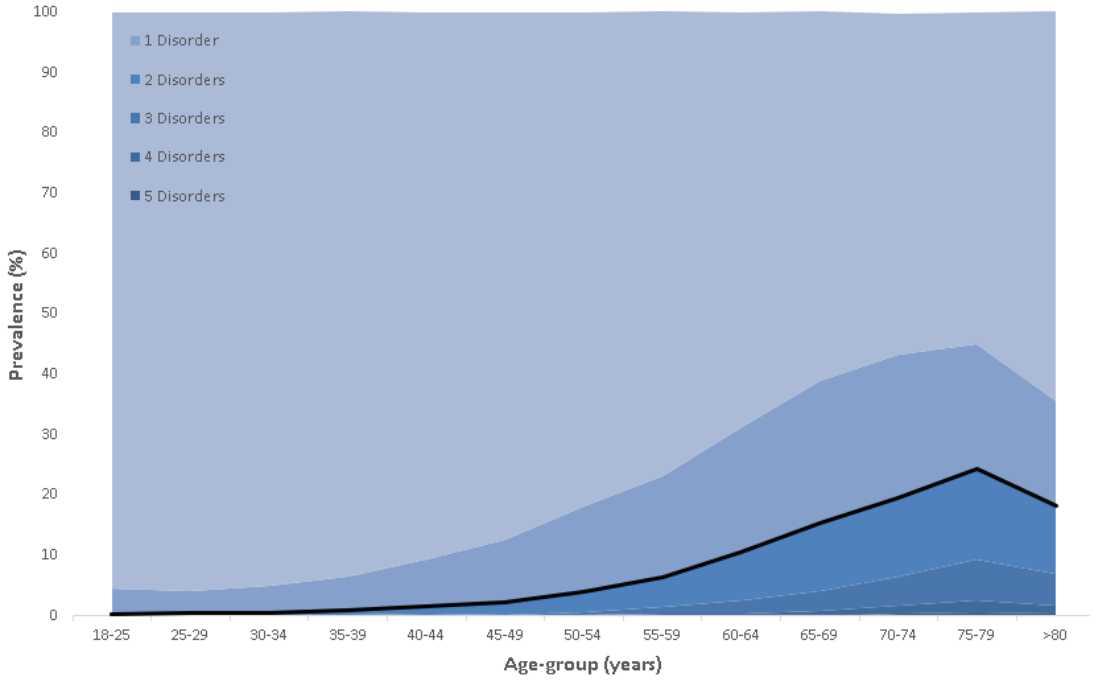


Figure 4.1: Prevalence of multimorbidity by age group for the COVID-19 infected Portuguese population. The lighter shade of blue is representative of the absence of conditions and the black line represents the prevalence of multimorbidity.

version mainly showed differences in higher ratios for cardiovascular disorders. In the April version, due to the small sample size of patients with cardiovascular disorders, some Odd Ratios got as high as 91.72 (CI 95%: $28.61 - 294.06$), as was the case for the composite outcome of death, hospitalisation, or ICU admission (see Table 4.5).

The way by which the structured information on chronic diseases was encoded changed amidst the two versions of the dataset. Besides the difference in terms of encoding the data as a single versus multiple categorical variables, the most recent version makes a distinction between *present*, *not present*, and *unknown*, while the April version listed the conditions that were *present* or *not present*. This difference can influence the veracity of the results, and it is believed that the June version can better depict the population affected by the different diseases and outcomes.

4.4 Discussion

This study shows that multimorbidity is significantly associated with adverse outcomes for COVID-19 infection in the Portuguese population, independently from age. All chronic conditions, except for asthma, lead to increased risk of hospitalisation. However, only diabetes, chronic kidney disease, chronic respiratory diseases, and neuromuscular disorders, are associated with more severe cases requiring ICU

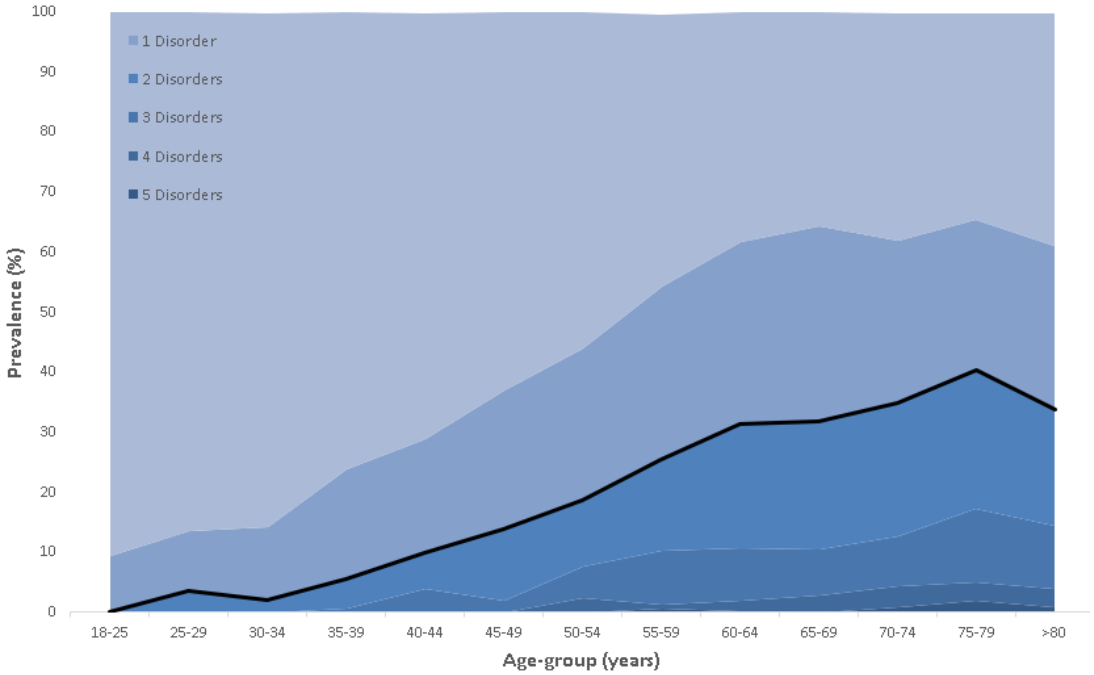


Figure 4.2: Prevalence of multimorbidity by age group for the COVID-19 infected Portuguese hospitalised population. The lighter shade of blue is representative of the absence of conditions and the black line represents the prevalence of multimorbidity.

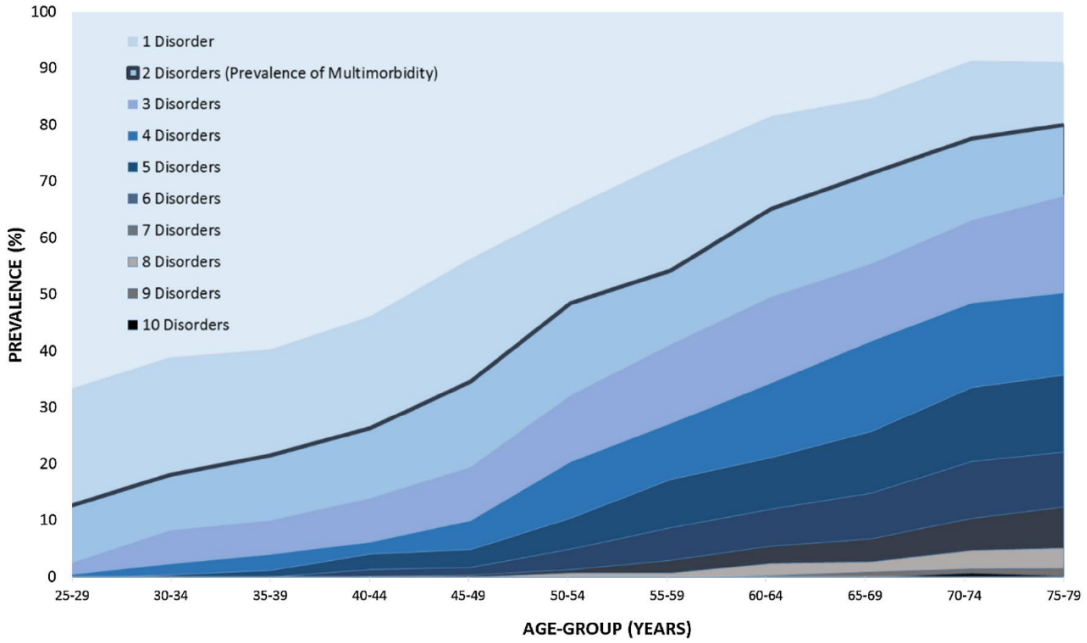


Figure 4.3: Prevalence of multimorbidity by age group using data from the Portuguese Fifth National Health Interview Survey (Laires et al. 2019). The lighter shade of blue is representative of the absence of conditions and the black line represents the prevalence of multimorbidity.

admission. These results are in line with previous reports by Grasselli et al. (2020); Petrilli et al. (2020) where chronic diseases are associated with poorer outcomes. Although the strength of association differs between diseases, every additional morbidity leads to an increased risk of the composite outcome of hospitalisation, ICU admission, and mortality.

Disease	Percentage (%)
CARDIAC DISORDER	43.86
DIABETES	30.94
NEUROMUSCULAR DISORDER	15.97
CANCER	14.22
ASTHMA	11.44
LUNG DISEASE	11.38
RENAL DISEASE	11.12
HEMATOLOGIC DISEASE	4.39
HIV	3.33
LIVER DISEASE	2.81

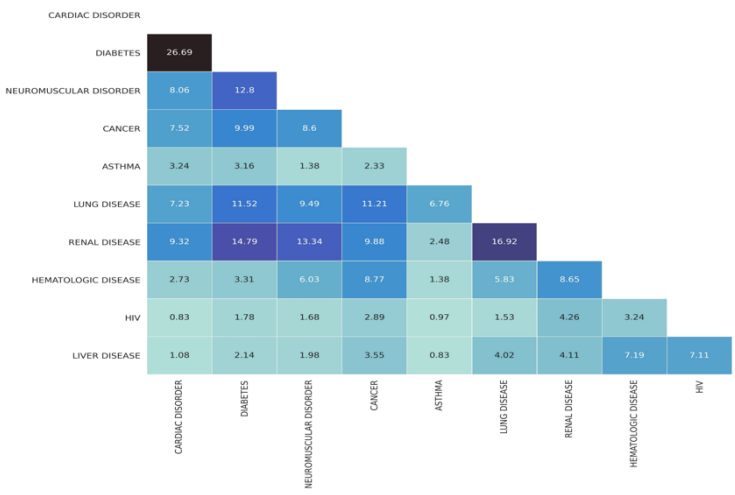


Figure 4.4: Prevalence of single (left) and co-occurring pairs (right) of chronic health conditions. Left: Prevalence of the disease in the Portuguese population with at least one disease. Right: Prevalence of the disease, rows, in the Portuguese population affected by another disease, columns. All values are presented in percentage.

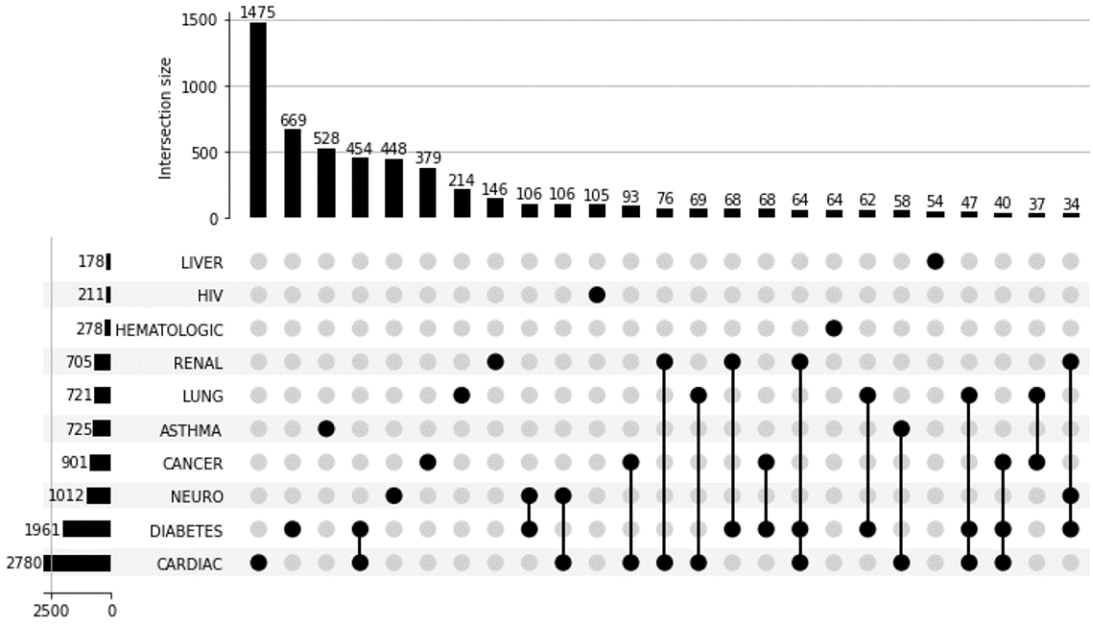


Figure 4.5: UpSet plot of the 25 most common, single and co-occurring, chronic health conditions in the COVID-19 infected population.

Multimorbidity was previously studied by Laires and Perelman (2019) for the general Portuguese population in 2014, using data from the fifth National Health Interview Survey (Inquérito Nacional de Saúde, INS). Although only individuals aged $25 - 79$ were included in the study from 2014, the choice of individuals constitutes a robust sample for the study of morbidity prevalence in Portugal. It is possible to observe, in Figures 4.1 to 4.3, a rise in chronic health conditions with increasing age, as expected. However, multimorbidity is much less prevalent in the COVID-19 study population (6.77% vs 43.9%).

Since the beginning of the COVID-19 pandemic there has been significant public awareness regarding the higher risk of older people with comorbidities. This may have induced efforts to protect and isolate

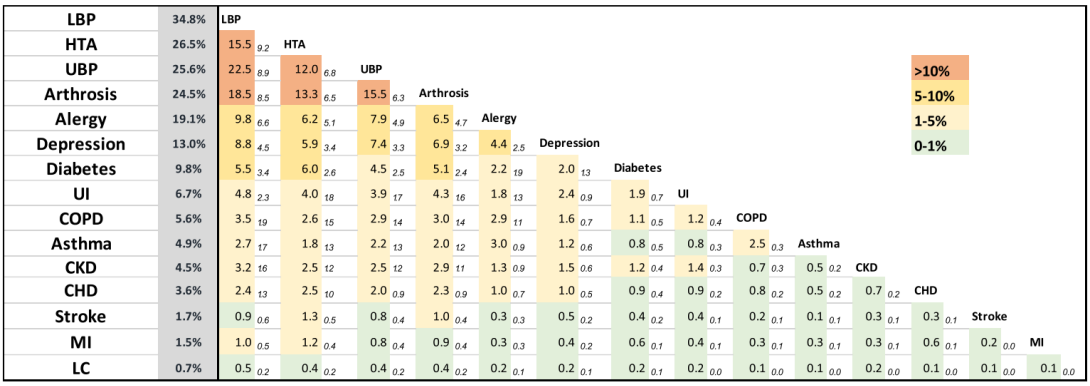


Figure 4.6: Percentage of observed and expected prevalence of co-occurring pairs of chronic health conditions (data from the Portuguese Fifth National Health Interview Survey) Laires and Perelman (2019). The shaded bar depicts the prevalence of each chronic health condition. In the matrix, the first value for each pair is the observed frequency, while the second (italic) is the expected one after multiplying the respective prevalence of each disorder. χ^2 tests were used to determine whether observed frequencies were significantly different from expected frequencies. All p-values were inferior to 5%.) LBP low back pain, HTA hypertension, UBP upper back pain, UI urinary incontinence, COPD chronic obstructive pulmonary disease, CKD chronic kidney disease, CHD coronary heart disease, MI previous myocardial infarction, and LC liver cirrhosis

Table 4.2: Odds Ratio for the outcome (Death) for the Age variable and analysed comorbidities. The real value of Age was used instead of an age group. Being a continuous variable, an Odds Ratio of 1.097 implies that, as the Age variable increases by 1, the probability of the patient having the outcome increases by 9.7%. **B**, **S.E**, and **Wald** are the unstandardised regression weight, how much the unstandardised regression weight can vary by, and test statistic for the individual predictor variable, respectively.

							95% C.I	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Age	.092	.002	1446.655	1	<0.001	1.097	1.092	1.102
Asthma	.188	.262	.515	1	.473	1.207	.722	2.017
Cancer	1.021	.113	82.084	1	<0.001	2.775	2.225	3.461
Cardiovascular Disorders	.520	.081	41.384	1	<0.001	1.682	1.435	1.970
Diabetes	.723	.087	69.373	1	<0.001	2.061	1.739	2.444
Hematologic Disease	.997	.175	32.438	1	<0.001	2.711	1.924	3.822
HIV	1.330	.306	18.837	1	<0.001	3.780	2.074	6.892
Renal Disease	1.177	.108	118.925	1	<0.001	3.244	2.626	4.008
Liver Disease	1.437	.253	32.137	1	<0.001	4.207	2.560	6.914
Lung Disease	.897	.117	58.276	1	<0.001	2.452	1.948	3.087
Neuromuscular Disease	1.092	.091	143.005	1	<0.001	2.981	2.492	3.565

this population, with findings suggesting a positive effect of such measures, given the younger and healthier population in the COVID-19 dataset. The discrepancy may, however, be related to different reporting methods. For instance, the maximum number of significant reported simultaneous morbidities in the COVID-19 infected population was 5 disorders (there were a total of 5 people with simultaneous morbidities ranging from 6 to 8), which is half of the maximum number of morbidities found in the study population of Laires and Perelman (2019). Since the total number of conditions considered in both datasets is not so different (COVID-19: 10 diseases; INS: 13 diseases), a possible explanation for the higher number of co-occurring conditions in the INS population can be the combination of self-diagnoses with the presence of more *subjective* disorders, such as lower and upper back pain, allergies, depression, and urinary incontinence.

Table 4.3: Odds Ratio for the outcome (hospitalisation) for the Age variable and analysed comorbidities. The real value of Age was used instead of an age group. Being a continuous variable, an Odds Ratio of 1.056 implies that, as the Age variable increases by 1, the probability of the patient having the outcome increases by 5.6%. **B**, **S.E**, and **Wald** are the unstandardised regression weight, how much the unstandardised regression weight can vary by, and test statistic for the individual predictor variable, respectively.

							95% C.I	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Age	.054	.001	3180.583	1	<0.001	1.056	1.054	1.058
Asthma	.085	.130	.429	1	.513	1.088	.844	1.403
Cancer	1.251	.076	271.427	1	<0.001	3.493	3.010	4.054
Cardiovascular Disorders	.815	.049	277.596	1	<0.001	2.258	2.052	2.486
Diabetes	1.205	.054	494.291	1	<0.001	3.335	2.999	3.709
Hematologic Disease	1.772	.144	151.367	1	<0.001	5.882	4.435	7.801
HIV	1.538	.158	94.268	1	<0.001	4.654	3.412	6.348
Renal Disease	2.013	.091	478.288	1	<0.001	7.387	6.175	8.836
Liver Disease	1.885	.166	128.229	1	<0.001	6.584	4.751	9.123
Lung Disease	1.307	.086	232.600	1	<0.001	3.694	3.123	4.369
Neuromuscular Disease	1.536	.076	409.383	1	<0.001	4.645	4.003	5.390

Table 4.4: Odds Ratio for the outcome (ICU stay) for the Age variable and analysed comorbidities. The real value of Age was used instead of an age group. Being a continuous variable, an Odds Ratio of 1.052 implies that, as the Age variable increases by 1, the probability of the patient having the outcome increases by 5.2. **B**, **S.E**, and **Wald** are the unstandardised regression weight, how much the unstandardised regression weight can vary by, and test statistic for the individual predictor variable, respectively.

							95% C.I	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Age	.050	.002	1100.303	1	<0.001	1.052	1.048	1.055
Asthma	-.332	.251	1.744	1	.187	.718	.439	1.174
Cancer	.011	.136	.007	1	.935	1.011	.775	1.319
Cardiovascular Disorders	.026	.083	.101	1	.751	1.027	.873	1.208
Diabetes	.268	.088	9.255	1	.002	1.307	1.100	1.554
Hematologic Disease	.113	.215	.276	1	.600	1.119	.735	1.705
HIV	-.252	.419	.362	1	.547	.777	.342	1.767
Renal Disease	.369	.125	8.800	1	.003	1.447	1.134	1.847
Liver Disease	.392	.287	1.867	1	.172	1.479	.844	2.594
Lung Disease	.366	.128	8.152	1	.004	1.442	1.122	1.855
Neuromuscular Disease	.295	.107	7.605	1	.006	1.343	1.089	1.655

This study has several important limitations. First of all, the cross-sectional nature of the COVID-19 dataset makes it impossible to account for incomplete outcomes, since several patients could ultimately be hospitalised or die after the end of observation. Reported data on outcomes may, therefore, be underestimated, so careful interpretation is advised until more data is available. More importantly, despite the fact that no standard set of conditions is established to define multimorbidity, chronic conditions were given on broad groups and there is no specific information on individual conditions. For example, diabetes is given as one group and no distinction is made between type 1 and type 2 diabetes. Therefore, measured morbidities may herald heterogeneous groups of diseases with different degrees of severity, which may influence outcomes. Future datasets should ideally include more accurate information on chronic conditions.

Table 4.5: Odds Ratio for the outcome (Death + hospitalisation + ICU stay) for the Age variable and analysed comorbidities. The real value of Age was used instead of an age group. Being a continuous variable, an Odds Ratio of 1.061 implies that, as the Age variable increases by 1, the probability of the patient having the outcome increases by 6.1%. **B**, **S.E**, and **Wald** are the unstandardised regression weight, how much the unstandardised regression weight can vary by, and test statistic for the individual predictor variable, respectively.

							95% C.I	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Age	.059	.001	3654.861	1	<0.001	1.061	1.059	1.063
Asthma	.032	.130	.060	1	.806	1.033	.800	1.333
Cancer	1.216	.076	253.311	1	<0.001	3.373	2.904	3.918
Cardiovascular Disorders	.768	.049	248.499	1	<0.001	2.156	1.959	2.372
Diabetes	1.170	.054	464.558	1	<0.001	3.221	2.896	3.583
Hematologic Disease	1.747	.147	140.286	1	<0.001	5.736	4.296	7.658
HIV	1.537	.159	93.498	1	<0.001	4.649	3.405	6.348
Renal Disease	2.032	.094	463.142	1	<0.001	7.627	6.338	9.177
Liver Disease	1.811	.168	116.114	1	<0.001	6.115	4.399	8.501
Lung Disease	1.320	.087	231.704	1	<0.001	3.743	3.158	4.437
Neuromuscular Disease	1.586	.078	414.126	1	<0.001	4.886	4.194	5.693

Table 4.6: Odds Ratio for the association between the age, categories of comorbidity and outcome (Death) in patients with COVID-19. The continuous value of Age was used instead of an age group. Being a continuous variable, an Odds Ratio of 1.093 implies that, as the Age variable increases by 1, the probability of the patient having the outcome increases by 9.3%. **B**, **S.E**, and **Wald** are the unstandardised regression weight, how much the unstandardised regression weight can vary by, and test statistic for the individual predictor variable, respectively.

							95% C.I	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Age	.089	.003	1157.213	1	<0.001	1.093	1.088	1.099
Cancer	.711	.119	35.639	1	<0.001	2.037	1.613	2.573
Cardiovascular Disorders	.257	.086	8.903	1	.003	1.293	1.092	1.530
Diabetes	.253	.096	6.887	1	.009	1.288	1.066	1.556
Hematologic Disease	.253	.186	1.864	1	.172	1.288	.896	1.854
HIV	.890	.320	7.750	1	.005	2.436	1.301	4.558
Renal Disease	.698	.118	35.218	1	<0.001	2.009	1.596	2.530
Liver Disease	.851	.270	9.917	1	.002	2.342	1.379	3.977
Lung Disease	.471	.125	14.146	1	<0.001	1.602	1.253	2.048
Neuromuscular Disease	.886	.095	87.857	1	<0.001	2.425	2.015	2.919

Another important concern is related to the risk of under-reporting, which becomes obvious by analyzing reported cardiovascular diseases. According to Polonia et al. (2014), cardiovascular diseases, particularly hypertension, are vastly prevalent in the Portuguese population. The observed prevalence of 8.14% in the COVID-19 study population highly suggests that under-reporting may have occurred. In addition, the prevalence of reported cardiovascular diseases significantly increased from the April version of the DGS dataset, which showed a much lower prevalence in the general population of 0.28%. Surprisingly, cardiovascular diseases are absent from the available list of previous conditions in SINAVE's reporting page, which could have contributed to a lower notification of this comorbidity (see Figure 4.7b). One explanation to the considerable increase in cardiovascular disorder prevalence is the fact that the *raw* textual input from doctors was included in the June version of the DGS/SINAVE dataset and used in the analysis through the consideration of keywords, instead of relying on the categorical variable for the

Table 4.7: Odds Ratio for the association between the age, categories of comorbidity and outcome (hospitalisation) in patients with COVID-19. The continuous value of Age was used instead of an age group. Being a continuous variable, an Odds Ratio of 1.045 implies that, as the Age variable increases by 1, the probability of the patient having the outcome increases by 4.5%. **B**, **S.E**, and **Wald** are the unstandardised regression weight, how much the unstandardised regression weight can vary by, and test statistic for the individual predictor variable, respectively.

							95% C.I	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Age	.044	.001	1721.450	1	<0.001	1.045	1.043	1.047
Cancer	.916	.083	120.287	1	<0.001	2.498	2.121	2.942
Cardiovascular Disorders	.526	.054	93.696	1	<0.001	1.692	1.521	1.881
Diabetes	.719	.061	137.713	1	<0.001	2.053	1.821	2.315
Hematologic Disease	1.002	.162	38.278	1	<0.001	2.724	1.983	3.741
HIV	1.079	.176	37.594	1	<0.001	2.941	2.083	4.152
Renal Disease	1.396	.099	198.500	1	<0.001	4.041	3.327	4.907
Liver Disease	1.309	.184	50.593	1	<0.001	3.704	2.582	5.314
Lung Disease	.794	.095	69.540	1	<0.001	2.213	1.836	2.667
Neuromuscular Disease	1.323	.080	275.328	1	<0.001	3.756	3.212	4.391

Table 4.8: Odds Ratio for the association between the age, categories of comorbidity and outcome (ICU stay) in patients with COVID-19. The continuous value of Age was used instead of an age group. Being a continuous variable, an Odds Ratio of 1.050 implies that, as the Age variable increases by 1, the probability of the patient having the outcome increases by 5.0%. **B**, **S.E**, and **Wald** are the unstandardised regression weight, how much the unstandardised regression weight can vary by, and test statistic for the individual predictor variable, respectively.

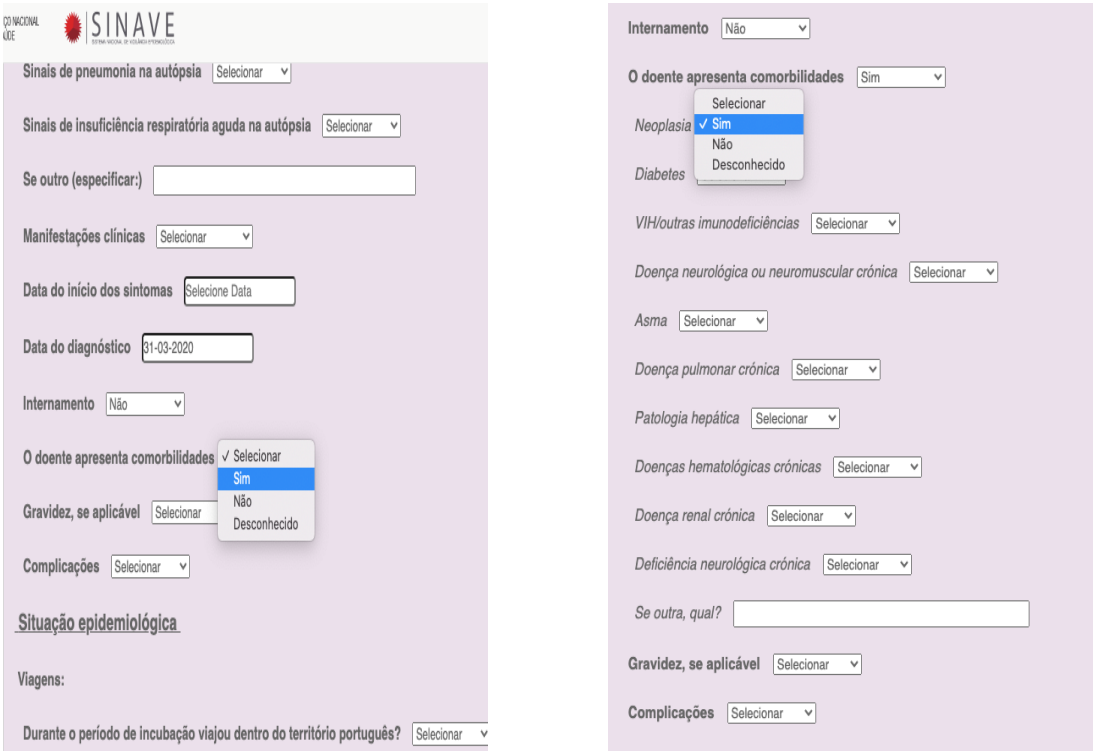
							95% C.I	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Age	.048	.002	943.562	1	<0.001	1.050	1.046	1.053
Diabetes	.177	.093	3.652	1	.056	1.194	.995	1.431
Renal Disease	.238	.130	3.335	1	.068	1.269	.983	1.638
Lung Disease	.270	.132	4.187	1	.041	1.309	1.011	1.695
Neuromuscular Disease	.231	.108	4.553	1	.033	1.260	1.019	1.558

Table 4.9: Odds Ratio for the association between the age, categories of comorbidity and outcome (Death + hospitalisation + ICU stay) in patients with COVID-19. The continuous value of Age was used instead of an age group. Being a continuous variable, an Odds Ratio of 1.050 implies that, as the Age variable increases by 1, the probability of the patient having the outcome increases by 5.0%. **B**, **S.E**, and **Wald** are the unstandardised regression weight, how much the unstandardised regression weight can vary by, and test statistic for the individual predictor variable, respectively.

							95% C.I	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Age	.049	.001	2160.553	1	<0.001	1.050	1.048	1.052
Cancer	.885	.084	111.355	1	<0.001	2.423	2.055	2.855
Cardiovascular Disorders	.480	.054	78.761	1	<0.001	1.616	1.454	1.797
Diabetes	.694	.061	127.705	1	<0.001	2.002	1.775	2.258
Hematologic Disease	.983	.166	35.153	1	<0.001	2.671	1.930	3.697
HIV	1.083	.176	37.806	1	<0.001	2.955	2.092	4.174
Renal Disease	1.446	.102	201.261	1	<0.001	4.246	3.477	5.184
Liver Disease	1.227	.186	43.534	1	<0.001	3.409	2.368	4.908
Lung Disease	.830	.096	74.757	1	<0.001	2.293	1.900	2.767
Neuromuscular Disease	1.375	.081	285.182	1	<0.001	3.957	3.373	4.642

Table 4.10: Odds Ratio for the association between the age, number of comorbidities and outcome (Death + hospitalisation + ICU stay) in patients with COVID-19. Both Age and Nº of comorbidities are continuous variables, i.e, an Odds Ratio of 2.223 implies that, as the Nº of comorbidities increases by 1, the probability of the patient having the outcome increases by 122.3%. **B**, **S.E**, and **Wald** are the unstandardised regression weight, how much the unstandardised regression weight can vary by, and test statistic for the individual predictor variable, respectively.

							95% C.I	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Age	.050	.001	2321.879	1	<0.001	1.051	1.049	1.053
Nºof comorbidities	.799	.023	1257.322	1	<0.001	2.223	2.127	2.323



(a) (b)

Figure 4.7: Screenshots of SINAVE's interface regarding the report of known comorbidities: (a) Reporting the presence of comorbidities; (b) Reporting the known comorbidities

presence or absence of this disease, as in the case of the April version.

Although it is acknowledged that the DGS/SINAVE dataset was not primarily generated for research, but rather for public health proceedings and government information, it is believed that a better user interface design and a more rational set of chronic conditions could effortlessly improve the quality of the recorded data. One valuable lesson learned from this pandemic is the important contribution that quickly gathered relevant clinical data through effective health information systems, such as SINAVE, can have in this setting.

SINAVE is suspected to be overly susceptible to under-reporting, especially regarding multimorbidity. The user interface is somewhat confusing and non-practical, allowing doctors to skip filling some important fields in the entry form. Some suggestions to improve future studies regarding multimorbidity, using SINAVE's data, include:

- Redesign the reporting form, making data entry more effective and faster. The current user interface could be simplified, while encouraging the input of relevant comorbidity information;
- Integrate SINAVE data with the patient's health record or with data from the new *Trace COVID-19* system, to provide richer data relevant to COVID-19. This could be a way of implementing the redesigned user interface suggested above;
- Make it mandatory to input if comorbidities are absent; if present, the filling of the entry form inputs related to comorbidities and the specific chronic conditions should be mandatory. As seen in Figure 4.7, the list of comorbidities only becomes visible if a previous parameter is filled;
- Emphasise to healthcare professionals the importance of entering all known chronic conditions in the system;
- Add cardiovascular diseases to the list of comorbidities.

4.5 Overview

It is crucial to understand what impacts multimorbidity can have in a patient's life. This chapter described the development steps and implementation of a descriptive analysis for the impact of multimorbidity in the population with COVID-19 infection. This analysis used data from the DGS/SINAVE dataset, after the required institutional and ethical approvals.

Multimorbidity was present in 6.77% of the $36,244$ infected patients. These patients showed an increased risk of hospitalisation, ICU admission, and mortality with OR 2.22 (CI 95%: 2.13-2.32) for every additional morbidity. Further studies should confirm these findings and special attention should be made on data collection, to ensure proper recording of patient's comorbidities.

Chapter 5

Analysis on the Temporal Evolution of Chronic Conditions and their Onsets

This chapter presents a study developed for understanding the temporal evolution of patients with multimorbidity and possible relationships between chronic conditions' onsets. To achieve this, I have used the Enroll-HD [1] dataset. Section 5.1 details the structure of the Enroll-HD dataset. Section 5.2 explains the selection process, as well as statistical analysis of the dataset, and the methodology developed and results regarding the representations of the temporal relationships between chronic conditions. These results are discussed in Section 5.3.

5.1 Enroll-HD

Enroll-HD is a clinical research platform and longitudinal observational study for Huntington's disease (HD) families intended to accelerate progress towards therapeutics. It integrates two former HD registries — REGISTRY in Europe, and COHORT in North America and Australasia — while also expanding to include sites in Latin America. It is sponsored by Cure Huntington's Disease Initiative (CHDI) Foundation, a nonprofit biomedical research organisation exclusively dedicated to collaboratively developing therapeutics for HD.

Enroll-HD is a relational database containing 11 different data files, divided into participant, study, and visit-based (see Figure 5.1). The database is de-identified and all date values are represented as relative number of days since the date of Enroll–HD's baseline visit (i.e., 11/01/2014). After all the required institutional and ethical approvals, a *.sql* file was obtained for each data file and loaded into *SQLite*, a relational database management system. Only 3 data tables (i.e., profile, comorbid, pharmacotx) were used to study the temporal evolution of chronic conditions' onsets in multimorbidity afflicted patients.

[1] https://www.enroll-hd.org/acknowledgments/

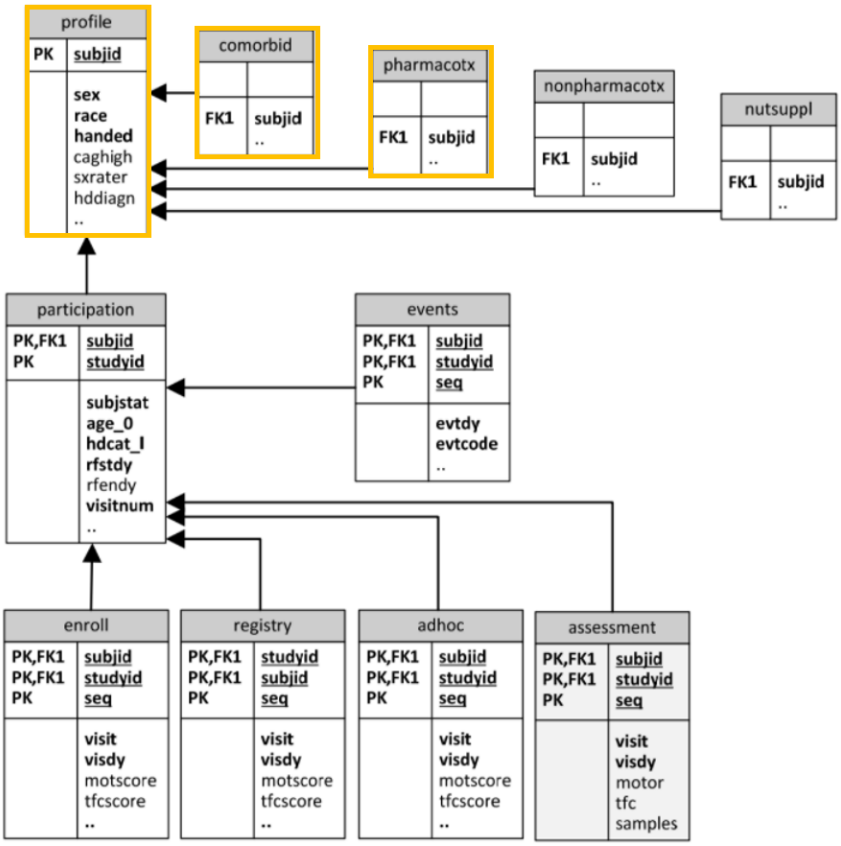


Figure 5.1: Enroll-HD Entity Relationship Diagram from CHDI Foundation (2012), used tables highlighted. **PROFILE**: General and annually updated information about each participant. **COMORBID**: Information about comorbid conditions. **PHARMACOTX**: Information about pharmacologic therapies. **NONPHARMACOTX**: Information about non-pharmacologic therapies. **NUTSUPPL**: Information about nutritional supplements. **PARTICIPATION**: Study-specific information about participant taken part in the study. **EVENTS**: Information about reportable events happened within Enroll-HD. **ENROLL**: Enroll data contains forms from Enroll-HD study. **REGISTRY**: REGISTRY visits contains combined forms from REGISTRY3 and REGISTRY2. **ADHOC**: Ad Hoc data containing forms: Variable, Motor, Function, TFC, MMSE, Cognitive. **ASSESSMENT**: Visit-specific information about which assessments were done at existing regular visits performed in any study.

5.2 Methodology and Results

The presented study focuses on identifying relationships between chronic conditions' onsets. This section describes the selection and temporal analysis processes used, as well as a statistical analysis of the population. *Python 3.6* was the programming language considered for both processes described in this section. NumPy, Pandas, Seaborn, Matplotlib [2], and UpSetPlot were used to analyse and visualise the data, while Graphviz [3] was used to visualise the temporal relationships between chronic conditions' onsets.

[2] https://matplotlib.org
[3] https://www.graphviz.org

5.2.1 Data Selection and Analysis

For the 3 used tables, I have extracted the following information:

- *PROFILE*: Unique identifier, Participant's gender, Participant's status (i.e., alive or dead);
- *COMORBID*: Unique identifier, ICD-10 diagnostic codes, Condition's onset, Status of the condition (i.e., ongoing or treated), Condition's end date;
- *PHARMACOTX*: Unique identifier, Condition treated, Status of the prescription (i.e., ongoing or finished), Start date of the prescription, End date of the prescription;

The Enroll-HD database gathers information about $15,300$ participants. Participants with no information regarding the onset of a condition were excluded. This reduced the number of participants to $12,759$, henceforward considered as the original database.

There are $4,492$ distinct conditions identified in the database. To simplify, I have only considered the chronic conditions used in Chapter 3. These chronic conditions were identified using rules, inspired by Hvidberg et al. (2016) and Tonelli et al. (2015), on the presence of associated ICD-10 diagnostic codes (see Table 5.1). Additionally, *PHARMACOTX* was used to identify patients on medication related to any of the studied chronic conditions, for the study population. Using this table, which directly identifies the conditions that caused the prescriptions, eliminates the need for application of specific rules.

The study population is considered to be any participant identified has currently having at least one of the selected chronic conditions determined, either using the ICD-10 diagnostic codes or the prescription history. For example, participants who were considered "cured" of diabetes have not been considered. Table 5.2 presents the statistical profile of the original and study populations. Figure 5.2 represents the 25 most common, single and co-occurring, combinations of the chronic health conditions in the study population.

5.2.2 Temporal Evolution Analysis

For all participants in the study population, I have created a timeline of their chronic conditions' onsets. To allow for comparison between participants, each timeline was offset so that time-zero corresponds to the onset of the first condition identified. A boxplot of the time intervals between chronic conditions can be seen in Figure 5.3. It is important to point out that several chronic conditions were identified at the same time, for a significant portion of the participants. This results in a relatively high prevalence of intervals between diseases of zero days. Figure 5.4 shows the prevalence of each disease according to their order of diagnosis.

To study the temporal evolution of chronic conditions, I have used directed graphs to represent the "route" of diseases throughout the Enroll-HD's participant lives. In this approach, each node represents a chronic condition and the amount of participants having it, and each edge displays the average number of days between parent and child nodes. The resulting directed graph has a funnel-like structure where

Table 5.1: Rules applied to Enroll-HD's data to detect chronic diseases.

	ICD-10 diagnostic code
Atrial Fibrillation	I48
Chronic Kidney Disease	N02; N03; N04; N05; N06; N07; N08; N09; N10; N11; N17; N18; N19
Chronic Obstructive Pulmonary Disease	J44
Deafness	H90; H91
Dementia	F00; F01; F02; F03; G30; G31
Diabetes	E10; E11; E12; E13; E14
Dyslipidemia	E78
Heart Failure	I11.0; I50
Hypertension	I10; I11; I13; I15
Ischemic Cardiomyopathy	I25.5; I25.6; I25.9; I42
Obesity	E66
Osteoarthritis	M15; M16; M17; M18; M19

Table 5.2: Statistical characterisation of the original Enroll-HD population and study population.

	Original		Study
	Total	Diseased	Total
Number of participants	12 759	3 768	4 097
Number of male participants	5 553	1 808	1 947
Number of female participants	7 206	1 960	2 127
Atrial Fibrillation prevalence	0.87%	2.95%	2.22%
Chronic Kidney Disease prevalence	0.43%	1.46%	0.95%
Chronic Obstructive Pulmonary Disease prevalence	0.85%	2.89%	2.81%
Deafness/Hearing Loss prevalence	1.69%	5.71%	5.00%
Dementia prevalence	0.50%	1.70%	4.32%
Diabetes prevalence	4.73%	16.03%	15.11%
Dyslipidemia prevalence	11.87%	40.21%	34.61%
Heart Failure prevalence	0.30%	1.01%	1.17%
Hypertension prevalence	16.76%	56.74%	58.80%
Ischemic Cardiomyopathy prevalence	0.49%	1.67%	2.12%
Obesity prevalence	0.81%	2.73%	2.37%
Osteoarthritis prevalence	2.83%	9.58%	16.11%
Percentage of diseased participants ($\geqslant 1$ morbidity)	29.53%	100%	100%
Percentage of participants with multimorbidity ($\geqslant 2$ morbidity)	9.45%	32.01%	33.02%

the ancestor nodes contain a higher number of participants than their descendants. To allow for a proper visualisation of the graphs, only the top three most common child nodes were represented after each parent node, and the hierarchical level of the graph was limited to three (excluding the root node).

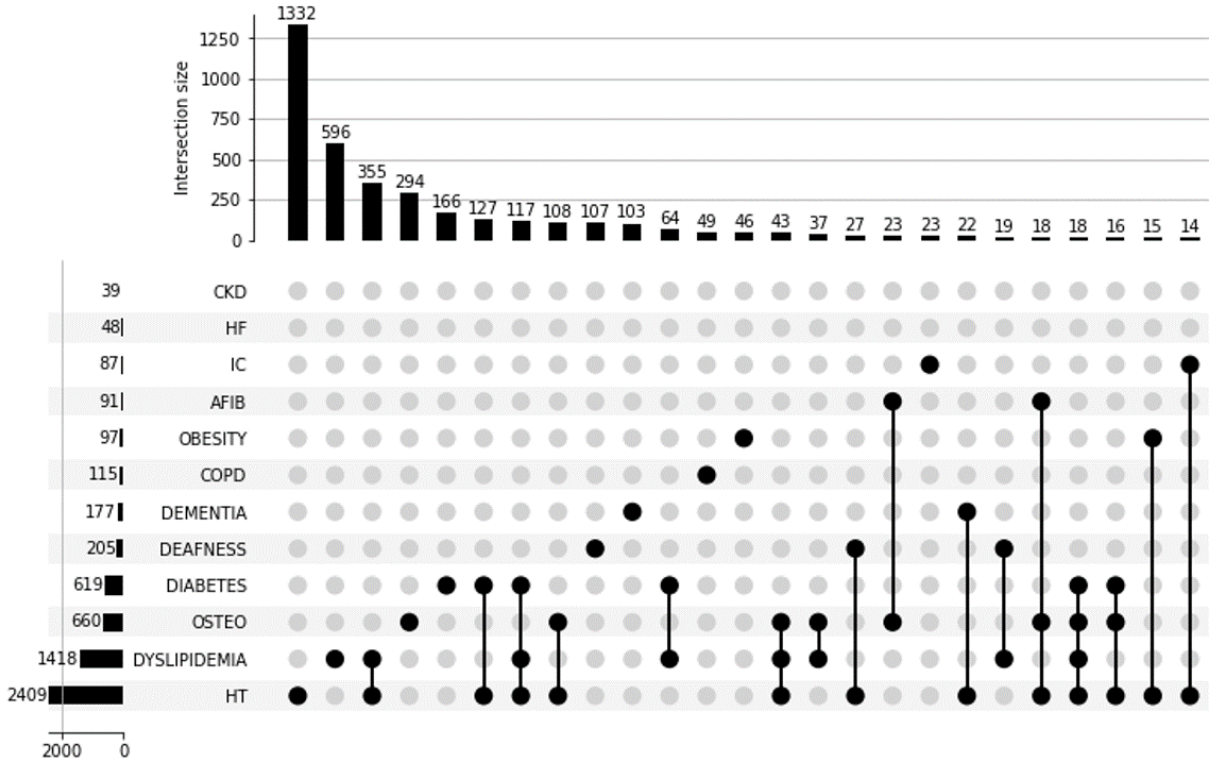


Figure 5.2: UpSet plot of the 25 most common, single and co-occurring, chronic health conditions in the selected Enroll-HD population. Atrial Fibrillation: AFIB; Chronic Kidney Disease: CKD; Chronic Obstructive Pulmonary Disorder: COPD; Heart Failure: HF, Hypertension: HT; Ischemic Cardiomyopathy: IC; Osteoarthritis: OSTEO.

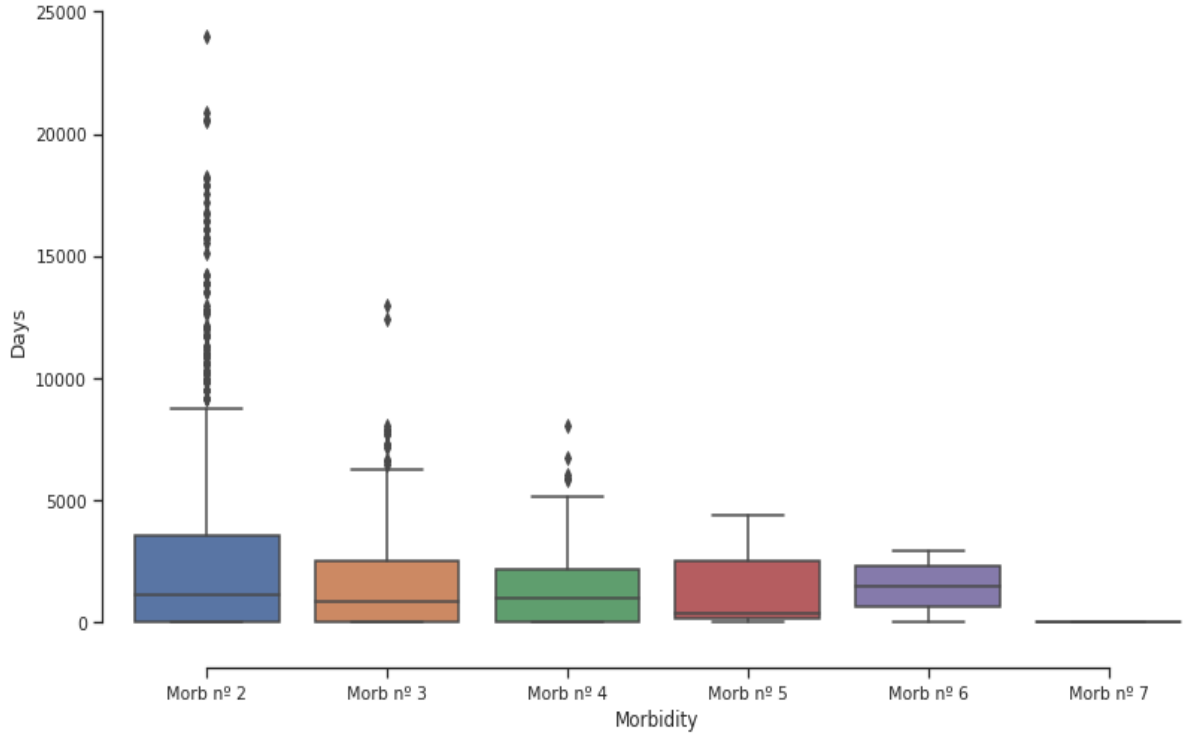


Figure 5.3: Distribution of days between onsets of different chronic conditions. The intervals are always relative to the previous disease's onset. For example, **Morb nº2** is relative to the interval between the first and second identified diseases and **Morb nº3** is relative to the interval between the second and third.

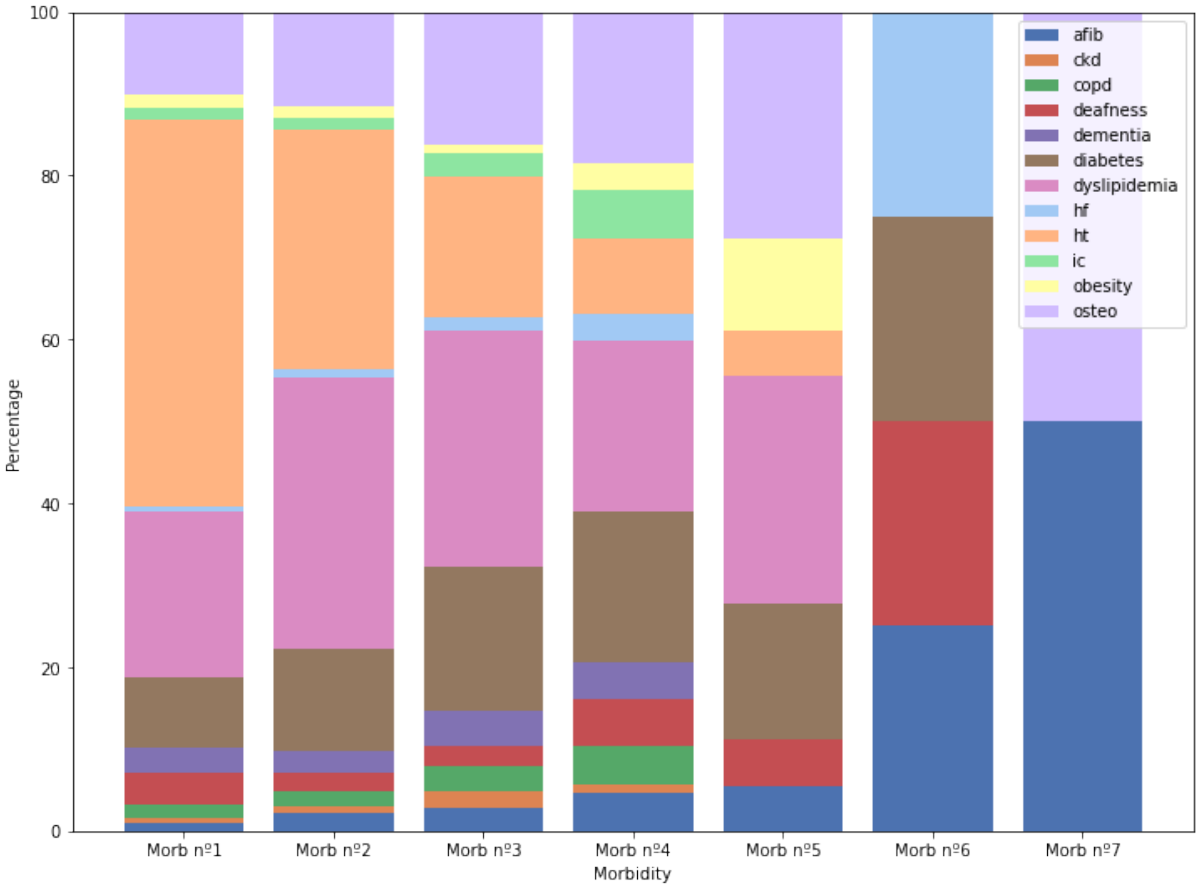


Figure 5.4: Prevalence of the different chronic conditions according to their order of diagnosis. **Morb nº1**: $4,097$ participants. **Morb nº2**: $1,353$ participants. **Morb nº3**: 404 participants. **Morb nº4**: 87 participants. **Morb nº5**: 18 participants. **Morb nº6**: 4 participants. **Morb nº7**: 2 participants.

As seen in Figure 5.4, all chronic conditions – besides hypertension, dyslipidemia, osteoarthritis, and diabetes – have a low prevalence as the first identified disease. Having this in mind, and in order to have a reasonable sized population as the root node, a directed graph was obtained for the top four most common first diagnosed conditions. These graphs can be seen from Figure 5.5 to Figure 5.8.

5.3 Discussion

When characterising the study population, considering prescription history, besides only ICD-10 diagnostic codes, proved beneficial to phenotyping certain chronic conditions. Comparing to the original diseased population, presented in the third column of Table 5.2, the study population contains 329 additional participants. However, this increase in the number of participants was not followed by a proportional increase in prevalence of the different chronic conditions. In reality, only 5 out of the 12 chronic conditions had an increase in prevalence in the study population, namely dementia, heart failure, hypertension, ischemic cardiomyopathy, and osteoarthritis. Dementia and osteoarthritis were the conditions with more significant increases (i.e., 1.70% to 4.32%, and 9.58% to 16.11%, respectively). Having said that, prescription history is essential to correctly identify participants with certain chronic

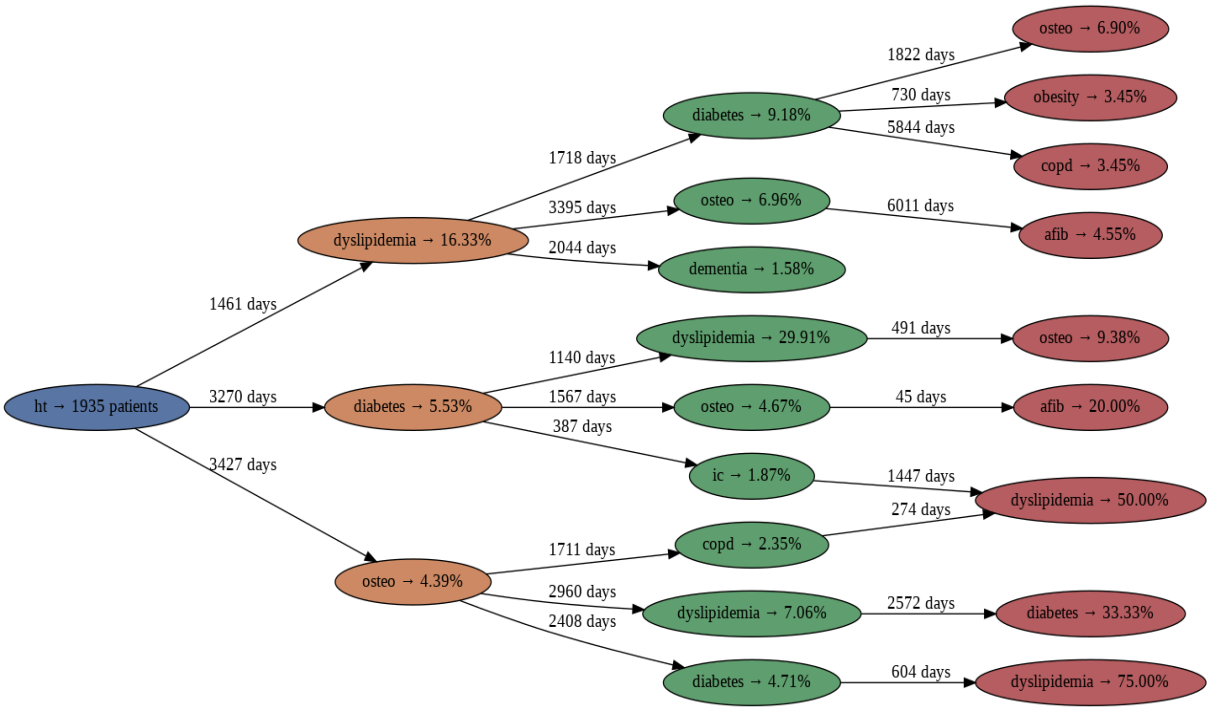


Figure 5.5: Directed graph for subset of participants with hypertension as their first identified chronic condition. Each node, besides the root node, presents the percentage of participants, with respect to the size of its parent node, with a certain condition. Each edge displays the average number of days between two consecutive diseases, for all the participants with the same connecting nodes.

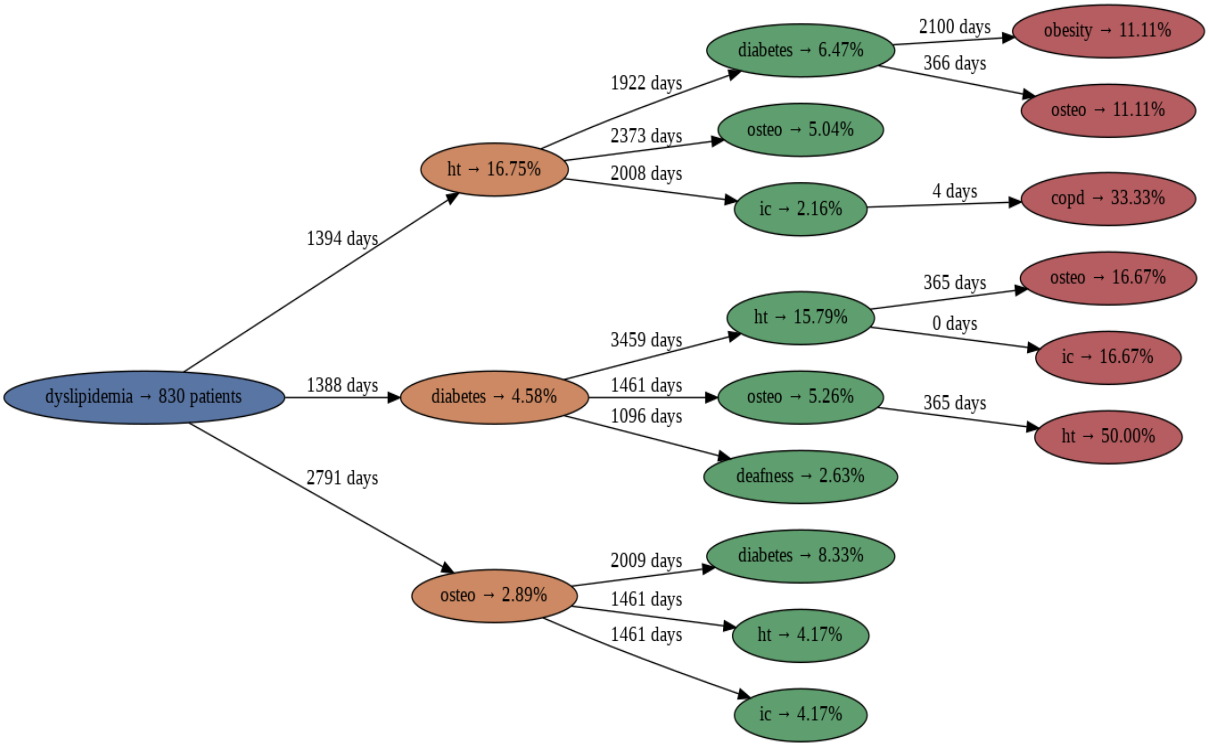


Figure 5.6: Directed graph for subset of participants with dyslipidemia as their first identified chronic condition. Each node, besides the root node, presents the percentage of participants, with respect to the size of its parent node, with a certain condition. Each edge displays the average number of days between two consecutive diseases, for all the participants with the same connecting nodes.

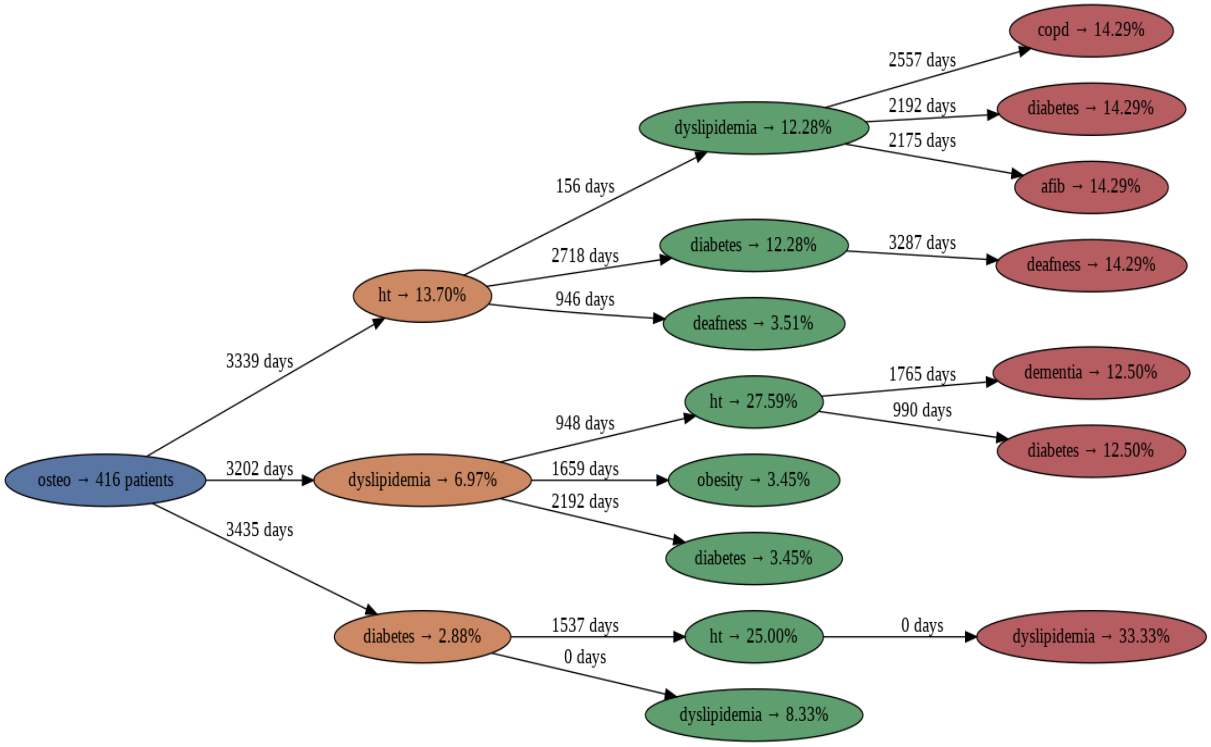


Figure 5.7: Directed graph for subset of participants with osteoarthritis as their first identified chronic condition. Each node, besides the root node, presents the percentage of participants, with respect to the size of its parent node, with a certain condition. Each edge displays the average number of days between two consecutive diseases, for all the participants with the same connecting nodes.

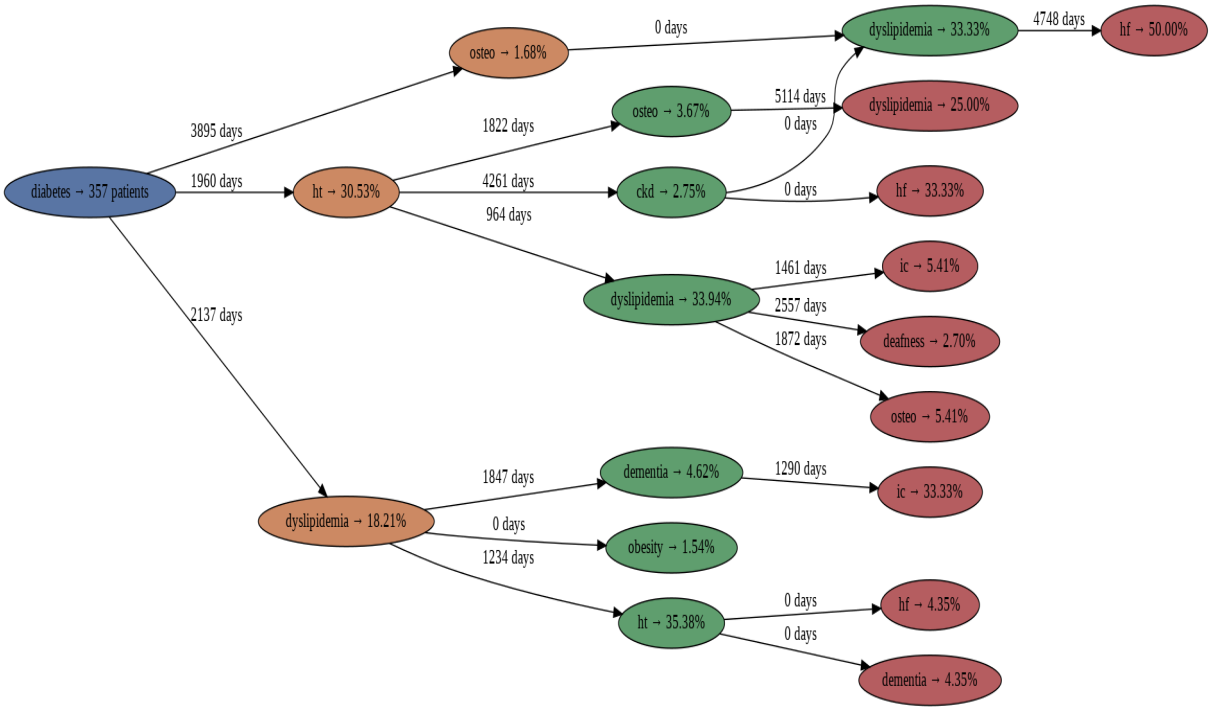


Figure 5.8: Directed graph for subset of participants with diabetes as their first identified chronic condition. Each node, besides the root node, presents the percentage of participants, with respect to the size of its parent node, with a certain condition. Each edge displays the average number of days between two consecutive diseases, for all the participants with the same connecting nodes.

conditions, proving that the use of ICD-10 diagnostic codes alone is insufficient to correctly characterise the Enroll-HD dataset, supporting the findings of Wei et al. (2016).

The directed graphs shown in Figures 5.5 to 5.8 offer valuable information, but also have some limitations. First of all, it is clear that there is an explicit over-representation of the top-4 first diagnosed diseases. This was already visible in Table 5.2, but becomes more evident when almost all paths of the presented graphs contain three, if not all, of the most prevalent chronic conditions (i.e., out of the 43 paths represented only 9 have less than three instances of either hypertension, dyslipidemia, osteoarthritis, or diabetes). However, this is not exclusive to the Enroll-HD dataset. In Chapter 3 and Chapter 4, hypertension, dyslipidemia, and diabetes have also been three of the most common individual and co-occurring chronic conditions. However, this is not unexpected, as these conditions share the same risk factors and are themselves risk factors of each other. Having said that, the discussion can be redirected to understanding if the prevalence of the remaining conditions is in line with that of the general population, or if it is related to under-reporting caused by: (i) EHR data used to phenotype conditions in the Enroll-HD dataset (i.e., ICD-10 diagnostic codes and medication); (ii) relationship with Huntington's disease. It is important to point out that the Enroll-HD database was not created with the intent of correctly phenotyping chronic conditions, but to accelerate progress towards therapeutics for HD. The tables used to gather all the information for this study scratch the surface of the level of complexity of the knowledge stored in this dataset. In other words, there is information (e.g., ethnicity, non pharmacological therapies, nutritional supplements, stages of HD progression) in the unused Enroll-HD's tables that could enrich the analysis.

Secondly, the graphs should not be generalised as predictors for a person's timeline of chronic conditions' onsets. Rather, they should be seen as an attempt to understand if there is a visible pathway for the onset of certain diseases. The number of days between the onset of different conditions is also a topic that requires special attention, given that some participants have several conditions identified at the same moment in time. When creating the timeline, this results in a high number of intervals between diseases being equal to zero days (see Figure 5.3). I have decided to keep "0 days" intervals due to need of a specific order of chronic conditions in order for the proposed method to work. These intervals obviously influence the calculation of the average number of days between onsets of chronic conditions, which becomes more visible at deeper levels of the graph. Figures 5.6, 5.7, and 5.8 have at least one instance of a "0 days" edge connecting two nodes. Instead of being looked at as parent and child nodes, they should be seen as siblings, where the size of one of the siblings is presented with respect to the other. Additionally, the fact that the Enroll-HD dataset portrays patients with HD and their families makes it difficult to derive any finding to the general population.

5.4 Overview

After identifying patients with multimorbidity, it is crucial to understand how multimorbidity evolves throughout a lifetime and how certain diseases can impact the predisposition to the onset of other conditions. This chapter described the development steps and implementation of a visualisation tool,

using directed graphs, for the timeline of onsets of certain chronic conditions. The Enroll-HD database was used has the data source, but this method could be applied, with few alterations, to similar datasets with multimorbidity information.

Due to the nature of the used dataset (i.e., Huntington's disease patients and their families), the resulting graphs cannot be used to hypothesise about the general population. The way by which diseases are identified in the dataset also undermines the interpretation of the obtained results. However, some clear relationships are visible when observing the graphs. Namely, the relation between the hypertension, dyslipidemia, and diabetes triad is more than evident.

Chapter 6

Conclusions and Future Work

In my M.Sc. research project, I presented an NLP tool for phenotyping patients with multimorbidity from clinical notes, applied to a real-world dataset of patients admitted to an ICU, a study of the impact of multimorbidity on the COVID-19 infected Portuguese population, and a study on the relationships between chronic conditions' onsets in patients with multimorbidity. Besides the different perspectives on multimorbidity studied, this dissertation also used inherently distinct datasets namely EHR data from ICU admitted patients, operational database of the COVID-19 infection in Portugal, and observational study for Huntington's disease. This adds values to this project, as it allows to discuss the problems associated with each dataset for multimorbidity analysis. This chapter summarises the main conclusions and limitations of this work and outlines possible directions for future work.

6.1 Conclusions and Limitations

This dissertation was motivated by the increased prevalence and impact of multimorbidity in today's society. Even though multimorbidity has serious, direct and indirect, implications on individuals and healthcare systems, little to no efforts are made to better identify and understand it. This is a result of most healthcare systems being disease-oriented (i.e., primarily focusing on managing each individual condition while ignoring their interactions). This can also be said regarding the way most medical datasets are built. Statistical analysis of the different datasets explored in this thesis showed clear cases of under-reporting for some of the analysed chronic conditions. This comes from the fact that medical datasets mainly rely on structured data to store information. However, most of this information results from a patient-doctor interaction focused on the disease that caused the visit for administrative purposes.

Concerning the problem of identifying patients with multimorbidity, the scientific literature revealed that using structured data, namely diagnosis codes, laboratory results, and medications, can be useful and less challenging, but is not sufficient and cannot supplant the added clinical value offered by unstructured text data (e.g., radiology reports, discharge summaries, progress notes). Following the objectives set for this work, this thesis developed an NLP tool that solely relied on clinical narratives to identify patients

afflicted with a wide range of chronic conditions.

The developed phenotyping tool can be easily adapted to different datasets containing clinical narratives, enforcing only minor alterations. Its application has the potential to ease up and enhance manual phenotyping performance, reducing the dimension of the team of trained physicians and time required to complete the assignment, which is intimately associated to a reduction of costs.

The achieved experimental results are promising, but showed that there is still some work to be done to achieve a phenotyping tool that uses clinical narratives to extract relevant information for multimorbidity analysis. Free-text inputs laden with alternative spellings, misspellings and unstructured information, and enormous quantities of data, which is not consistently relevant, make this task a difficult challenge to handle. Additionally, the tool relies on structured data for validation of the results. Despite that, experimental results of this dissertation outperformed, in most cases, the literature methods found for the same chronic conditions, and were in line with the classification obtained via expert manual revision of a portion of the clinical notes. However, some optimisation is still needed especially regarding the implemented algorithm for detecting negated findings, which are prevalent in clinical narratives.

Despite the literature reports and increased use of NLP methods for electronic phenotyping, we could not find any study for some of the selected chronic conditions, which further motivates the study of NLP methods for certain conditions (i.e., deafness and osteoarthritis). In these cases, the experimental results were compared to those of methods that used structured data for phenotyping, revealing the true usefulness of NLP phenotyping methods. There are clear advantages of using NLP methods when the structured data is lacking. If the structured data is complete and well-reported, rule-based methods are extremely effective at phenotyping patients, while also being easier to implement.

Concerning the problems of better understanding the impact and temporal evolution of multimorbidity, both developed studies shed a light on their respective questions. However, it is important to state that these studies cannot be easily generalised due to fact that both study populations were comprised of people affected by a specific disease (i.e., COVID-19 and Huntington's disease).

Findings in the study of the impact of multimorbidity showed that multimorbidity is significantly associated with poor outcomes in COVID-19 infection. Further data is needed to inform about the strength of this association and about the significance of observed differences in multimorbidity prevalence between infected patients and the general population of Portugal. It is believed that data collection problems may have occurred and influenced outcome measurement. Despite having been conducted in a highly specific setting (i.e., COVID-19 infection), the difficulties faced in this analysis are replicable to other settings, which reinforces the importance of this study for the topic of multimorbidity's impact. This study also provides recommendations for improving the data collection user interface, that could ultimately improve quality of health information about the COVID-19 infected population, while increasing confidence in the SINAVE data.

Findings in the study of the temporal evolution of multimorbidity showed interesting results, but these should be looked at with special attention. Out of the 12 studied chronic conditions, 4 were clearly present in most patient timelines. Namely, hypertension, dyslipidemia, and diabetes proved to be constantly

associated with each other. This, however, could be the result of lower prevalence of the remaining conditions. It is important to understand if their prevalence is in line with that of the general population (i.e., not only HD patients and their families), or if is related to limitations of the used dataset.

Overall, constant communication and feedback acquisition with both medical and engineering parts were key to allow the comprehension of the task and promoted a full involvement of a multidisciplinary team to perform a project as challenging and demanding as this was.

6.2 Future Work

Data quality is essential for the success of any study, especially one as comprehensive as this thesis. This work had to be developed without the availability of a single complete and longitudinal dataset where the three dimensions of multimorbidity analysis could be made. A single dataset would have allowed for an integration between studies and, consequently, a higher clinical significance of the results. This single dataset should be available in the near future, due to the Intelligent Care project. Intelligent Care is a project that gathers Hospital da Luz Learning Health, Instituto de Sistemas de Robótica (ISR), INESC-ID, Carnegie Mellon University (CMU), Priberam, and Hospital da Luz Lisboa to find artificial intelligence (AI) solutions, that will allow to establish better treatment methodologies for patients with multimorbidity.

With respect to the electronic phenotyping tool, even though valuable results were achieved, there are still several possibilities for future work. The simplest improvement to be made is regarding code optimisation and addition of new features. Currently, the tool is limited to detecting disease mentions and not considering disease severity and stage, which would result in a more impactful analysis. Also, the tool only distinguishes between negated and non-negated diseases, failing to identify cases of uncertainty and mentions of family or medical history. Considering more vast input information, besides just clinical narratives, such as structured data (e.g., diagnosis codes, laboratory results, and medications), would allow for a more robust method. Another appealing approach would be to explore alternative methods like word embedding for the same task. Lastly, it would be interesting to adapt the tool to the Portuguese language, as it was initially planned.

Still related to electronic phenotyping, an engaging future study would consist of developing techniques to validate NLP methods using only unstructured data, which would eliminate the dependency on structure data for validation.

Regarding the study of the impact of multimorbidity, future work should also include validation of the obtained results in a larger population. This could be performed with a more recent version of the SINAVE dataset, if it were to be made available, since the COVID-19 infected Portuguese population had a near tenfold increase since the last available version (i.e., all confirmed cases of COVID-19 as of June 30, 2020) and will continue to increase.

Concerning the study of the temporal evolution of multimorbidity, future work would also dwell on using a larger and more general (i.e., not related to any prior condition) population, preferably, including

more information besides ICD-10 diagnostic codes, such as clinical narratives. Additionally, it would be interesting to complement the patients' timeline with more information besides just chronic conditions and the days between their onsets. This could be done by integrating additional information such as gender, age group, ethnicity, smoking status, and dietary habits. Ultimately, the graphs that resulted from this study could be used to train a model to predict a range of possible outcomes, namely: future chronic conditions, hospitalizations, or surgeries.

Bibliography

WHO. Global health and aging. October 2011.

M. Van den Akker, F. Buntix, J. F. Metsemakers, S. Roos, and J. A. Knottnerus. Multimorbidity in general practice: Prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. *Journal of Clinical Epidemiology*, 51(5):367–375, 1998.

R. Navickas, V.-K. Petric, A. B. Feigl, and M. Seychell. Multimorbidity: What Do We Know? What Should We Do? *Journal of Comorbidity*, 6(1):4–11, 2016.

J. M. Banda, M. Seneviratne, T. Hernandez-Boussard, and N. H. Shah. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annual Review of Biomedical Data Science*, 1(1):53–68, 2018.

C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, 2014.

M. Tonelli, N. Wiebe, M. Fortin, B. Guthrie, B. R. Hemmelgarn, M. T. James, S. W. Klarenbach, R. Lewanczuk, B. J. Manns, P. Ronksley, P. Sargious, and S. Straus. Methods for identifying 30 chronic conditions : application to administrative data. 2015.

M. F. Hvidberg, S. P. Johnsen, C. Glümer, K. D. Petersen, A. V. Olesen, and L. Ehlers. Catalog of 199 register-based definitions of chronic conditions. (March):462–479, 2016.

J. Knottnerus, J. Metsemakers, P. Hoppener, and C. Limonard. Chronic illness in the community and the concept of social prevalence. *Family Practice*, 9(1):15–21, 1992.

P. A. Laires and J. Perelman. The current and projected burden of multimorbidity: a cross-sectional study in a Southern Europe population. *European Journal of Ageing*, 16(2):181–192, 2019.

K. Wikström, J. Lindström, K. Harald, M. Peltonen, and T. Laatikainen. Clinical and lifestyle-related risk factors for incident multimorbidity: 10-year follow-up of Finnish population-based cohorts 1982-2012. *European Journal of Internal Medicine*, 26(3): 211–216, 2015.

M. Fortin, L. Lapointe, C. Hudon, A. Vanasse, A. L. Ntetu, and D. Maltais. Multimorbidity and quality of life in primary care: A systematic review. *Health and Quality of Life Outcomes*, 2, 2004. ISSN 14777525. doi: 10.1186/1477-7525-2-51.

A. Menotti, I. Mulder, A. Nissinen, S. Giampaoli, E. J. Feskens, and D. Kromhout. Prevalence of morbidity and multimorbidity in elderly male populations and their impact on 10-year all-cause mortalitym: The FINE study (Finland, Italy, Netherlands, elderly). *Journal of Clinical Epidemiology*, 54(7):680–686, 2001.

D. M. Zulman, C. P. Chee, T. H. Wagner, J. Yoon, D. M. Cohen, T. H. Holmes, C. Ritchie, and S. M. Asch. Multimorbidity and healthcare utilisation among high-cost patients in the US Veterans Affairs Health Care System. *BMJ Open*, 5(4):1–10, 2015.

J. M. Baker, R. W. Grant, and A. Gopalan. A systematic review of care management interventions targeting multimorbidity and high care utilization. *BMC Health Services Research*, 18(1):1–9, 2018.

A. Hassaine, D. Canoy, J. R. A. Solares, Y. Zhu, S. Rao, Y. Li, M. Zottoli, K. Rahimi, and G. Salimi-Khorshidi. Learning Multimorbidity Patterns from Electronic Health Records Using Non-negative Matrix Factorisation. 2019.

Phenopackets. Standardizing and exchanging patient phenotypic data, 2019. URL https://www.ga4gh.org/news/phenopackets-standardizing-and-exchanging-patient-phenotypic-data/. Accessed: 2020-09-18.

WHO. International classification of diseases (icd), 2018. URL http://wwwhо.int/classifications/icd/en/. Accessed: 2020-08-15.

American Medical Association. The differences between icd-9 and icd-10, 2015. URL https://www.ama-assn.org/media/7546/download. Accessed: 2020-08-16.

A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

I. CHDI Foundation. Enroll -hd: A prospective registry study in a global huntington's disease cohort, 2012. URL https://clinicaltrials.gov/ct2/show/NCT01574053. Accessed: 2020-07-18.

W. Q. Wei, P. L. Teixeira, H. Mo, R. M. Cronin, J. L. Warner, and J. C. Denny. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association*, 23(e1):20–27, 2016.

Z. Zeng, Y. Deng, X. Li, T. Naumann, and Y. Luo. Natural Language Processing for EHR-Based Computational Phenotyping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(1):139–153, 2019.

P. N. Jensen, K. Johnson, J. Floyd, S. R. Heckbert, R. Carnahan, and S. Dublin. Identifying atrial fibrillation from electronic medical data: a systematic review. *Pharmacoepidemiology and drug safety*, 21(0 1):141–147, 2012.

V. L. Martucci, N. Liu, V. E. Kerchberger, T. J. Osterman, T. Eric, B. Richmond, and M. C. Aldrich. A Clinical Phenotyping Algorithm to Identify Cases of Chronic Obstructive Pulmonary Disease in Electronic Health Records. *Journal of Chemical Information and Modeling*, 53(9):1689–1699, 2013.

M. Franchini, S. Pieroni, C. Passino, M. Emdin, and S. Molinaro. The CARPEDIEM Algorithm: A Rule-Based System for Identifying Heart Failure Phenotype with a Precision Public Health Approach. *Frontiers in Public Health*, 6(January):1–10, 2018.

G. H. Tison, A. M. Chamberlain, M. J. Pletcher, S. M. Dunlay, S. A. Weston, J. M. Killian, J. E. Olgin, and V. L. Roger. Identifying Heart Failure using EMR-based algorithms. *Physiology and Behavior*, 176(10):139–148, 2017.

J. Pacheco and W. Thompson. Type 2 diabetes mellitus, 2012. URL https://phekb.org/phenotype/18. Accessed: 2020-11-03.

J. C. Kirby, P. Speltz, L. V. Rasmussen, M. Basford, O. Gottesman, P. L. Peissig, J. A. Pacheco, G. Tromp, J. Pathak, D. S. Carrell, S. B. Ellis, T. Lingren, W. K. Thompson, G. Savova, J. Haines, D. M. Roden, P. A. Harris, and J. C. Denny. PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association*, 23(6):1046–1052, 2016.

Y. Halpern, S. Horng, Y. Choi, and D. Sontag. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4):731–740, 2016.

Y. Shao, Q. T. Zeng, K. K. Chen, A. Shutes-David, S. M. Thielke, and D. W. Tsuang. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Medical Informatics and Decision Making*, 19(1):1–11, 2019.

R. L. Figueroa and C. A. Flores. Extracting Information from Electronic Medical Records to Identify the Obesity Status of a Patient Based on Comorbidities and Bodyweight Measures. *Journal of Medical Systems*, 40(8), 2016.

F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, L. J. Jensen, and S. Brunak. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Computational Biology*, 7(8), 2011.

C. Nath, M. S. Albaghdadi, and S. R. Jonnalagadda. A natural language processing tool for large-scale data extraction from echocardiography reports. *PLoS ONE*, 11(4):1–15, 2016.

H. Ware, C. J. Mullett, and V. Jagannathan. Natural Language Processing Framework to Assess Clinical Conditions. *Journal of the American Medical Informatics Association*, 16(4):585–589, 2009.

O. Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32 (DATABASE ISS.):267–270, 2004.

R. J. Carroll, A. E. Eyler, and J. C. Denny. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:189–196, 2011.

J. C. Denny, J. D. Smithers, R. A. Miller, and A. Spickard. "Understanding" medical school curriculum content using KnowledgeMap. *Journal of the American Medical Informatics Association*, 10(4):351–362, 2003.

A. N. Nguyen, M. J. Lawley, D. P. Hansen, R. V. Bowman, B. E. Clarke, E. E. Duhig, and S. Colquist. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association*, 17 (4):440–445, 2010.

SNOMED International. Snomed-ct, 2020. URL `http://www.snomed.org`. Accessed: 2020-11-03.

R. J. Byrd, S. R. Steinhubl, J. Sun, S. Ebadollahi, and W. F. Stewart. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *Physiology and Behavior*, 176(12):139–148, 2017.

M. King, J. Kingery, and B. Casey. Diagnosis and evaluation of heart failure. *American Family Physician*, 85(12):1161–1168, 2012.

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. 310(2001):301–310, 2002.

Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:188–196, 2018.

S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 0(0):1–14, 2019a.

G. H.-J. Kwak and P. Hui. DeepHealth: Deep Learning for Health Informatics. 2019.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pages 1–12, 2013.

K. Patel, D. Patel, M. Golakiya, P. Bhattacharyya, and N. Birari. Adapting Pre-trained Word Embeddings For Use In Medical Coding. pages 302–306, 2017.

T. Bai, A. K. Chanda, B. L. Egleston, and S. Vucetic. EHR phenotyping via jointly embedding medical concepts and words into a unified vector space. *BMC Medical Informatics and Decision Making*, 18(Suppl 4), 2018.

B. S. Glicksberg, R. Miotto, K. W. Johnson, K. Shameer, L. Li, R. Chen, and J. T. Dudley. Automated disease cohort selection using word embeddings from Electronic Health Records the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License. HHS Public Access. *Pac Symp Biocomput*, 23:145–156, 2018.

T. Pham, T. Tran, D. Phung, and S. Venkatesh. DeepCare: A deep dynamic memory model for predictive medicine. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9652 LNAI (i):30–41, 2017.

Y. Bengio, P. Simard, and P. Frasconi. Learning Long-Term Dependencies with Gradient Descent is Difficult, 1994.

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

K. Cho. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation Kyunghyun. *Journal of Biological Chemistry*, 281(49):37275–37281, 2006.

T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad. Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment. 2017.

S. Dubois, N. Romano, D. C. Kale, N. Shah, and K. Jung. Effective Representations of Clinical Notes. pages 1–20, 2017.

L. Rumeng, J. Abhyuday N, and Y. Hong. A hybrid Neural Network Model for Joint Prediction of Presence and Period Assertions of Medical Events in Clinical Notes. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2017:1149–1158, 2017.

H. Wu, K. Hodgson, S. Dyson, K. I. Morley, Z. M. Ibrahim, E. Iqbal, R. Stewart, R. J. Dobson, and C. Sudlow. Efficient reuse of natural language processing models for phenotype-mention identification in free-text electronic medical records: A phenotype embedding approach. *Journal of Medical Internet Research*, 21(12):1–14, 2019b.

C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson, and P. D. Clayton. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(1):83–108, 1995.

A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21, 2001.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 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, 2010.

R. Reátegui and S. Ratté. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making*, 18(Suppl 3), 2018.

L. V. Rasmussen, W. K. Thompson, J. A. Pacheco, A. N. Kho, D. S. Carrell, J. Pathak, P. L. Peissig, G. Tromp, J. C. Denny, and J. B. Starren. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. *Journal of Biomedical Informatics*, 51:280–286, 2014.

H. Mo, W. K. Thompson, L. V. Rasmussen, J. A. Pacheco, G. Jiang, R. Kiefer, Q. Zhu, J. Xu, E. Montague, D. S. Carrell, T. Lingren, F. D. Mentch, Y. Ni, F. H. Wehbe, P. L. Peissig, G. Tromp, E. B. Larson, C. G. Chute, J. Pathak, J. C. Denny, P. Speltz, A. N. Kho, G. P. Jarvik, C. A. Bejan, M. S. Williams, K. Borthwick, T. E. Kitchner, D. M. Roden, and P. A. Harris. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *Journal of the American Medical Informatics Association*, 22(6):1220–1230, 2015.

N. Shang, C. Liu, L. V. Rasmussen, C. N. Ta, R. J. Caroll, B. Benoit, T. Lingren, O. Dikilitas, F. D. Mentch, D. S. Carrell, W. Q. Wei, Y. Luo, V. S. Gainer, I. J. Kullo, J. A. Pacheco, H. Hakonarson, T. L. Walunas, J. C. Denny, K. Wiley, S. N. Murphy, G. Hripcsak, and C. Weng. Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network. *Journal of Biomedical Informatics*, 99(June):103293, 2019.

W. C. Winkelmayer, S. Schneeweiss, H. Mogun, A. R. Patrick, J. Avorn, and D. H. Solomon. Identification of individuals with CKD from medicare claims data: A validation study. *American Journal of Kidney Diseases*, 46(2):225–232, 2005.

J. Oake, E. Aref-Eshghi, M. Godwin, K. Collins, K. Aubrey-Bassler, P. Duke, M. Mahdavian, and S. Asghari. Using Electronic Medical Record to Identify Patients With Dyslipidemia in Primary Care Settings: International Classification of Disease Code Matters From One Region to a National Database. *Biomedical Informatics Insights*, 9, 2017.

P. L. Teixeira, W. Q. Wei, R. M. Cronin, H. Mo, J. P. VanHouten, R. J. Carroll, E. Larose, L. A. Bastarache, S. Trent Rosenbloom, T. L. Edwards, D. M. Roden, T. A. Lasko, R. A. Dart, A. M. Nikolai, P. L. Peissig, and J. C. Denny. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *Journal of the American Medical Informatics Association*, 24(1):162–171, 2017.

N. Ivers, B. Pylypenko, and K. Tu. Identifying Patients With Ischemic Heart Disease in an Electronic Medical Record. *Journal of Primary Care & Community Health*, 2(1):49–53, 2011.

M. Froes, B. Martins, B. Neves, and M. J. Silva. Comparison of Multimorbidity in Covid-19 Infected and General Population in Portugal. pages 1–20, 2020.

WHO. Novel coronavirus (2019-ncov) situation reports, 2020. URL www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/. Accessed: 2020-08-16.

N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, J. Xia, T. Yu, X. Zhang, and L. Zhang. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet (London, England)*, 395(10223):507–513, 2020.

G. Grasselli, A. Zangrillo, A. Zanella, M. Antonelli, L. Cabrini, A. Castelli, D. Cereda, A. Coluccello, G. Foti, R. Fumagalli, G. Iotti, N. Latronico, L. Lorini, S. Merler, G. Natalini, A. Piatti, M. V. Ranieri, A. M. Scandroglio, E. Storti, M. Cecconi, A. Pesenti, and C.-. L. I. Network. Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. *JAMA*, 323(16):1574, 2020.

K. E. Mason, P. McHale, A. Pennington, G. Maudsley, J. Day, and B. Barr. Age-adjusted associations between comorbidity and outcomes of COVID-19: a review of the evidence. *medRxiv*, 2020.

DGS. Sinave, 2018. URL https://www.dgs.pt/servicos-on-line1/sinave-sistema-nacional-de-vigilancia-epidemiologica.aspx. Accessed: 2020-06-09.

C. M. Petrilli, S. A. Jones, J. Yang, H. Rajagopalan, L. O'Donnell, Y. Chernyak, K. A. Tobin, R. J. Cerfolio, F. Francois, and L. I. Horwitz. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. *BMJ*, 369:m1966, 2020.

J. Polonia, L. Martins, F. Pinto, and J. Nazare. Prevalence, awareness, treatment and control of hypertension and salt intake in Portugal: changes over a decade. The PHYSA study. *Journal of Hypertension*, 32(6):1211–1221, 2014.