Phenotyping and Understanding Multimorbidity

Miguel Froes

Abstract—This paper 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.

Index Terms—Multimorbidity, Electronic Health Records, Electronic Phenotyping, Natural Language Processing

1 Introduction

Multimorbidity is defined by Van den Akker et al. (1998) as the presence of two or more co-occurring chronic conditions, and has serious implications on individuals and healthcare systems. Due to an increase in life expectancy, prevalence of chronic conditions, and consequently multimorbidity, is set to rise. Correctly characterising patients according to their single, or co-occurrent, chronic conditions is the first step on understanding how to tackle multimorbidity. 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 optimised for a specific subset of patients affected by a specific combination of chronic diseases.

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. 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.

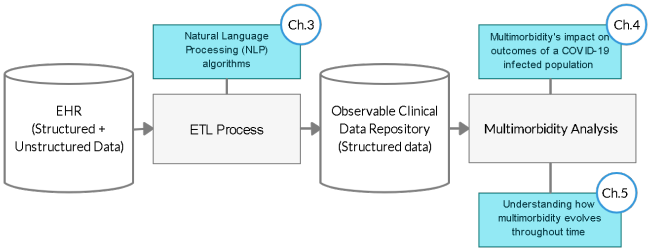

Fig. 1: Proposed pipeline to analyse multimorbidity.

Several methodologies have been developed and applied to identify and characterise patients with chronic conditions, but very few consider the presence and interactions of different conditions. I propose the usage of an information processing pipeline, represented in Figure 1, for phenotype data extraction and multimorbidity analysis.

The first stage of the pipeline uses both structured and unstructured data from the EHR. An Extract, Transform, and 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 clin-

ical 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. It organizes information and supports a subsequent analysis whit 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 hospitalizations 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 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 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 to find possible relations in the order of diagnosis of the conditions, and the time interval between onsets.

The rest of this article is organised as follows. Section 2 surveys important concepts on multimorbidity analysis, and previous related work on EHR-based phenotyping, specifically, on natural language processing methodologies applied on clinical notes. Section 3 details the proposed approach considered for solving the problem of extracting information from clinical notes. Section 4 focus on the proposed study related with understanding the impact of multimorbidity in the outcomes of a population affected by a life-threatening infection. Section 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. At last, Section 6 outlines the main conclusions and possible developments for future work.

2 Concepts and Related Work

This section provides an overview on the topic of multimorbidity, presenting previous work focusing on multimorbidity analysis, and describes fundamental concepts on EHR-based phenotyping and a related work revision.

2.1 Multimorbidity

Understanding the risk factors and consequences of multimorbidity, at both an individual and healthcare system level, is essential 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. 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. Ultimately, Menotti et al. (2001) associates people with multimorbidity to higher risks of premature death.

A patient-centred healthcare 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.

2.2 Electronic Phenotyping

The 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 utilizing these EHRs, and the most significant to this article, 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 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.

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).

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. 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.

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.

Several studies have been dedicated to the development of rule-based algorithms to characterize 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. 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 considered elevated 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 health care 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 standardization of the datasets used.

2.3.2 Machine Learning Methods

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. Classical statistical machine learning methods, manly 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, developed to detect probable dementia cases in patients without a dementia-related diagnosis. Figueroa and Flores (2016) presented a method for automatic identification of obesity and categorization 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.

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. 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 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. 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 to extracted information regarding valvular heart disease.

Regarding term extraction, most studies use tools that map textual elements and obtain the corresponding Unified Medical Language System (UMLS) concepts. 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 MEDTEX, that mapped Systematised Nomenclature of Medicine Clinical Terms (SNOMED-CT) concepts from free-text, while also identifying negation and possibility phrases.

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.

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. (2019) identify word embeddings and Recurrent Neural Network (RNNs) as the state-of-the-art models used for natural language processing. 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. RNNs, as the name indicates, are neural networks that repeat themselves over time. These are a class of artificial neural networks that consider all inputs and outputs as dependent of each other. As noted by Kwak and Hui (2019), RNNs are specialised for time-series data and natural language, due to their ability to memorize previous inputs and capture longer dependencies than those obtained with alternative sequential models, such as hidden Markov models.

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 been traditionally 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.

3 Multimorbidity Information Extraction

This section 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 MIMIC-III Critical Care Database, from Johnson et al. (2016), to develop and test the method.

MIMIC-III is a relational database containing 26 different data tables regarding patients who stayed within the ICU at Beth Israel Deaconess Medical Center from June 2001 to October 2012. For this work, only 8 tables were needed to test and evaluate the created Multimorbidity Information Extraction (MIE) tool.

3.1 Electronic Phenotyping Methodology

Using tables that gather structured data, I have extracted all the relevant MIMIC-III information to characterise the dataset. This included information regarding a patients' age, gender, mortality, and number of admissions, as well as previous diagnoses. Table 1 presents the statistical profile of the dataset before and after the selection process. The chronic conditions were detected using rules, inspired by those from Hvidberg et al. (2016) and Tonelli et al. (2015), on structured data (i.e., diagnostic and procedure codes, medications, lab results).

The selection process consisted of filtering the population according to the category of clinical narratives. There were 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 $391,031$.

To perform the extraction of information from the MIMIC-III dataset, I developed an information extraction pipeline. The MIE tool takes as input the selected clinical reports and outputs labels for the presence or absence of the 12 chosen phenotypes. The tool incorporates methods for identifying negated findings using regular expressions, taking inspiration from previous work on *NegEx* by Chapman et al. (2002). The NLP pipeline has the following steps:

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. The keywords were chosen based on previous studies, 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;

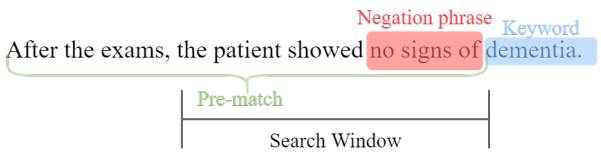

Fig. 2: Negation finding process on example sentence.

TABLE 1: Statistical characterization 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%

TABLE 2: Negation phrases used in the negation finding part of the proposed NLP method.

	Negation phrase
Pre-match	*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*
Post-match	*was declined; unlikely; ruled out; was denied; was absent; not present*

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 2 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 2 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.

3.2 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 were compared to the algorithm's assertion regarding the presence of a corresponding disease. A true positive case was considered when the disease identified by the algorithm had an associated ICD-9 code throughout the patient's history. I have obtained measurements of *Precision*, *Recall*, and *F1-score* for each disease. Table 3 presents the performance, for each chronic condition in analysis, of the method developed in this project and methods developed in similar studies.

3.3 Discussion

The proposed phenotyping method is capable of achieving good results, for most of the diseases under analysis. 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, hence results are not directly comparable. Table 3 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 study dedicated to its phenotyping. This is representative of the level of importance given to this condition, easily seen in Table 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, I am still presenting the performance results of studies that only used structured data for phenotyping.

One major characteristic of the proposed NLP method is its ability to identify negated findings. Therefore, it is important to evaluate its overall results 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 pipeline that was created. 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 findings or disease mentioned within a clinical narrative is present

TABLE 3: 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 not include 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**

or absent. By micro-averaging of the performance metrics, the overall precision and recall of the proposed method were 87.2% and 98.7%, respectively. Despite showing better results than NegEx, it is not reasonable to conclude that this project's method is superior to that of Chapman et al. (2002). It is very important 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 percentage of instances containing negation phrases is way lower than 50%. Having said that, the method reported on this article 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.

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 in MedRxiv (see Froes et al. (2020)). Since its release, a lot more studies, focusing on the impacts of COVID-19, were developed that might invalidate some of the statements made.

4.1 Dataset Description and Methodology

This study evaluates the prevalence of multimorbidity and age-adjusted risk of hospitalization, ICU admission, and death, in the Portuguese population from official data, based on a dataset [1] extracted from National Epidemiological Surveillance System (SINAVE) containing all confirmed cases of COVID-19 infection, in Portugal, by June 30, 2020.

The sample population consists of all the Portuguese population with SARS-CoV-2 confirmed infection, as notified by clinician. 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: (1) Asthma; (2) Malignancy; (3) Chronic hematological disorder; (4) Diabetes; (5) HIV/other immune deficiency; (6) Renal disease; (7) Liver disease; (8) Chronic lung disease; (9) 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.

A text mining script, using keywords associated with all the previously mentioned conditions, was created to better capture the prevalence of the diseases and to detect cases of cardiovascular disorders. The keywords were chosen, based on an empirical analysis of the textual field, in order to cover different cases, considering misspellings or abbreviations. The following keywords (in Portuguese) were used, in con-

1. https://covid19.min-saude.pt/disponibilizacao-de-dados/

nection to each of the diseases that was 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.

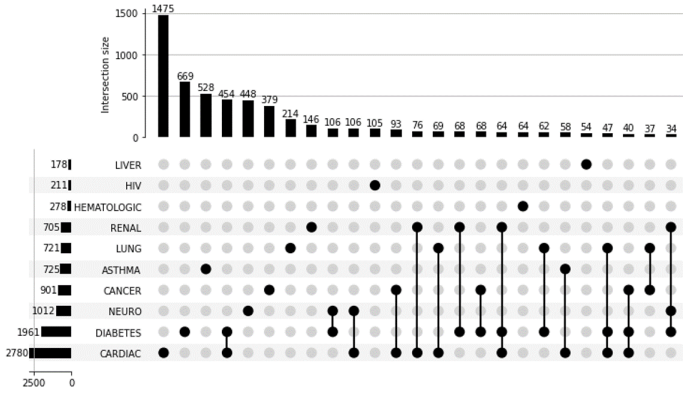

Fig. 4: UpSet plot of the 25 most common, single and co-occurring, chronic health conditions in the COVID-19 infected population.

4.2 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 disorder was the most commonly reported condition, present in 43.33% of the patients with any morbidity. Table 4 shows the reported prevalence of different chronic conditions in the studied population. Multimorbidity, as previously defined, was present in 6.77% of the cases. Figure 3 plots the prevalence of multimorbidity by age group for the COVID-19 infected hospitalised population. To analyze 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. Additionally, Figure 4 presents the 25 most common, single and co-occurring, chronic health conditions.

Data regarding hospitalization and ICU admission was available for only $32,945$ patients (90.90% of the overall study population). Within this population, hospitalization 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%. All chronic conditions, except for asthma, were associated with increased risk of mortality and hospitalization. Age, diabetes, renal disease, lung disease, and neuromuscular disorders, were all associated with increased risk of ICU admission. Additionally, every additional chronic condition increases the risk for the patients of the composite outcome of death, hospitalization, or ICU admission, by 123.3% (OR 2.22; CI 95%: $2.13 - 2.32$).

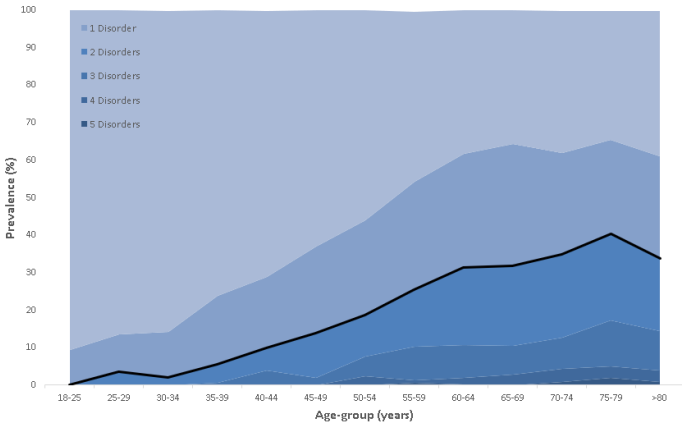

Fig. 3: Prevalence of multimorbidity by age group for the COVID-19 infected Portuguese hospitalised population. The lighter shade of blue is representative of the absence of conditions and the black line represents the prevalence of multimorbidity.

4.3 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 asthma, lead to increased risk of hospitalization. However, only diabetes, chronic kidney disease, chronic respiratory diseases, and neuromuscular disorders, are associated with more severe cases requiring ICU admission. Although the strength of association differs between diseases, every additional morbidity leads to an increased risk of the composite outcome of hospitalization, ICU admission, and mortality.

Multimorbidity was previously studied by Laires and Perelman (2019) for the general Portuguese population. 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. In both studies it is possible to observe a rise in chronic health conditions with increasing age. However, multimorbidity is much less prevalent in the COVID-19 study population (6.77% vs 43.9%). The studies also differ in maximum number of co-occurring conditions (i.e., 5 co-occurring disorders in the COVID-19 infected population vs 10 in the general population). 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: 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

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. Therefore, measured morbidities may herald heterogeneous groups of diseases with different degrees of severity, which may influence outcomes. Another important concern is related to the risk of under-reporting, which becomes obvious by analyzing reported cardiovascular disorders. According to Polonia et al. (2014), cardiovascular diseases, particularly hypertension, are very 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.

Although it is acknowledge that the DGS/SINAVE dataset was not primarily generated for research, but rather for public health proceedings and government information, it is believe that a better user interface design and a more rational set of chronic conditions could effortlessly improve the quality of the recorded data.

5 Analysis on the Temporal Evolution of Chronic Conditions and their Onsets

This section 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 [2] dataset, a clinical research platform and longitudinal observational study for Huntington's disease (HD) families intended to accelerate progress towards therapeutics.

5.1 Data Selection and Analysis

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 Section 3. These chronic conditions were identified using rules on the presence of associated ICD-10 diagnostic codes.

The study population is considered to be any participant identified has currently having at least one of the selected chronic conditions determined. Table 5 presents the statistical profile of the original and study populations.

2. https://www.enroll-hd.org/acknowledgments/

TABLE 5: Statistical characterization 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%

5.2 Temporal Evolution Analysis

For all participants in the study population, I have created a timeline of their chronic conditions' onsets. To allow comparison between participants, each timeline was offset so that time-zero corresponds to the onset of the first condition identified. Figure 5 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. Each node represents a chronic condition and the amount of participants having it, and each edge displays the average

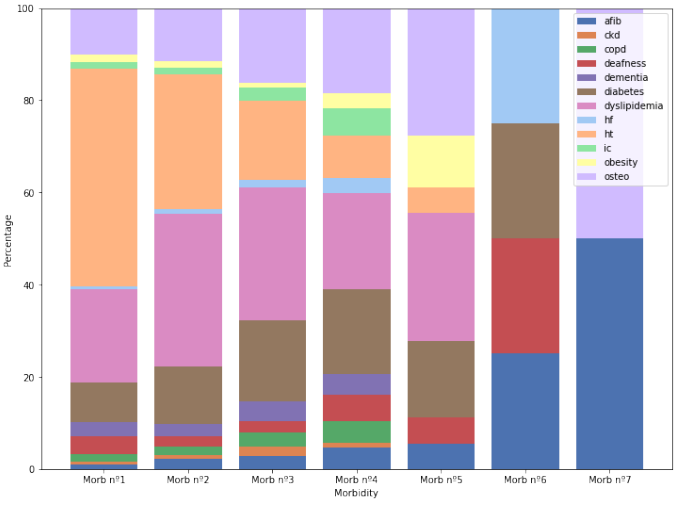

Fig. 5: Prevalence of the different chronic conditions according to their order of diagnosis. **Morb n°1**: $4,097$ participants. **Morb n°2**: $1,353$ participants. **Morb n°3**: 404 participants. **Morb n°4**: 87 participants. **Morb n°5**: 18 participants. **Morb n°6**: 4 participants. **Morb n°7**: 2 participants.

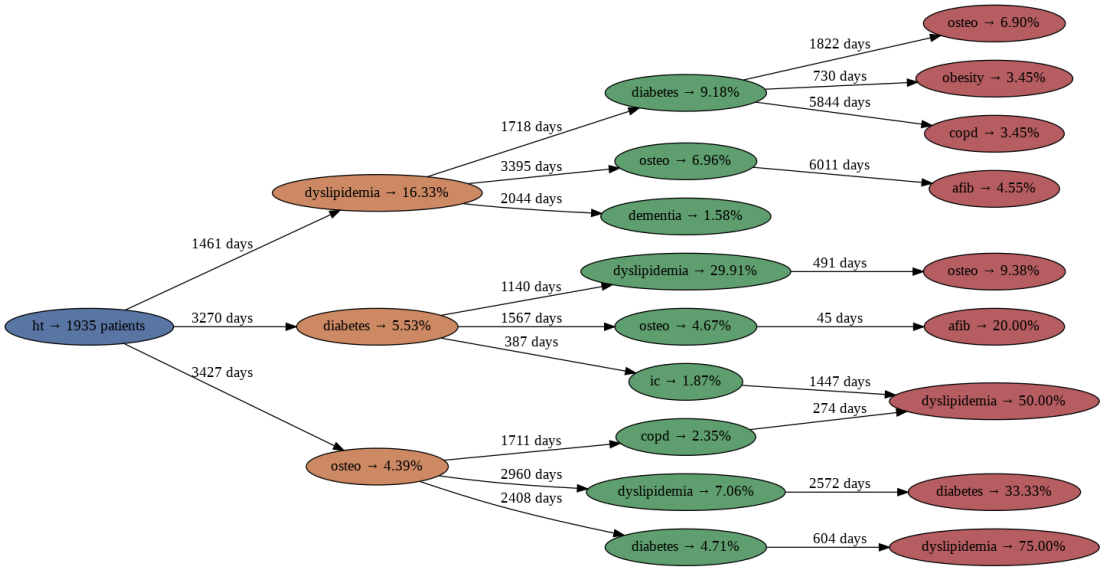

Fig. 6: Directed graph for subset of participants with hypertension as their first identified chronic condition.

number of days between parent and child nodes. To allow for a proper visualization 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 6 presents the directed graph for patients with hypertension as their first diagnosed condition.

5.3 Discussion

The directed graph shown in Figure 6 offers valuable information, but also has 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, 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 **??** and Chapter **??**, 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 inline with that of the general population, or if it is related to under-reporting caused by the EHR data used to phenotype conditions in the Enroll-HD dataset (i.e., ICD-10 diagnostic codes and medication). 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.

Secondly, the graph should not be generalised as a predictor 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. 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.

6 Conclusions and Future Work

The developed phenotyping tool can be easily adapted to different datasets containing clinical narratives, enforcing only minor alterations. The achieved experimental results outperformed, in most cases, the literature methods found for the same chronic conditions. Although 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 (i.e., deafness and osteoarthritis), which further motivates the study of NLP methods for certain conditions. 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. Considering disease severity and stage, instead of only disease mentions, and more vast input information, besides just clinical narratives, are some future work considerations to be held.

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. Future work should 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 increased since the last available version (i.e., all confirmed cases of COVID-19 as of June 30, 2020) and will continue to increase.

The study of the temporal evolution of multimorbidity showed interesting results, but these should be look 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. Future work would dwell on using a bigger and more general population, and complementing the patients' timeline with additional information, besides just chronic conditions and the days between their onsets. This could be done by integrating additional information such as, for example, gender, age group, ethnicity, smoking status, and dietary habits. Ultimately, the resulting graphs could be used to train a model to predict a range of possible outcomes.

Overall, future work should focus on applying the presented studies to a single complete and longitudinal dataset, which would allow 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.

REFERENCES

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.

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.

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. pages 462–479, 2016.

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.

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.

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.

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.

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.

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.

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.

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.

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, 2019.

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

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, 2016.

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.

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.

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.

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.

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.

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.