Computer-based Identification of Relationships Between Medical Concepts and Cluster Analysis in Clinical Notes

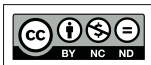
by

Ruth María REÁTEGUI ROJAS

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, APRIL 4, 2019

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC



BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mrs. Sylvie Ratté, Thesis Supervisor
Département de génie logiciel et des technologies de l'information, École de technologie supérieure

M. Maarouf Saad, President of the Board of Examiners
Département de génie électrique, École de technologie supérieure

M. Luc Duong, Member of the jury
Département de génie logiciel et des technologies de l'information, École de technologie supérieure

M. Jean-Guy Meunier, External Examiner
Département de philosophie, Université du Québec à Montréal

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON MARCH 27, 2019

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Prof. Sylvie Ratté, for her generosity, enthusiasm and help throughout this project. My thanks to Professors Luc Duong, Maarouf Saad and Jean-Guy Meunier for accepting to be part of the jury. I would also like to thank Drs. María Estefanía Baustista, Juan Francisco Beltrán, and Victor Duque for their valuable contribution to this work.

I am equally grateful to my lab mates from LINCS and LIVE for their kindness, advice and the coffee and lunches, which gave me the opportunity to know a little more about their culture and discover the marvelous people they are.

I would also like to thank the Universidad Técnica Particular de Loja and DCCE for the opportunity to start my studies. Also, I am grateful to SENESCYT for their financial support for this project.

Finally, I would like to express my thanks to my friends and family who encourage me and make me feel loved and accompanied wherever I go.

Identification informatique des relations entre les concepts médicaux et l'analyse par grappes dans les notes cliniques

Ruth María REÁTEGUI ROJAS

RÉSUMÉ

Les textes cliniques contiennent des informations variées qui mettent en relief des concepts médicaux ou des entités; on y trouve des formes de surface et des codes qui correspondent entre autres à des maladies, des traitements et des médicaments. Ces dernières –les entités– donnent au clinicien une impression générale et exhaustive de la santé du patient. L'analyse automatique de cette information riche est pertinente pour les experts et les chercheurs de la santé afin d'identifier des associations parmi les entités médicales. Cependant, l'extraction automatique d'information à partir des textes cliniques constitue un défi à cause de leur format narratif et leur structure libre.

Cette recherche décrit un processus pour extraire de manière automatique des entités médicales afin d'identifier des grappes de patients ainsi que les relations entre les maladies et les traitements. L'ensemble de données i2b2 2008 Obesity a été utilisé. Cet ensemble de données est composé de 1237 résumés sur le surpoids et les patients diabétiques, donc ce travail fixe son regard sur les maladies liées à l'obésité.

Pour l'extraction automatique des entités médicales, les outils MetaMap et cTAKES ont été utilisés pour comparer leur capacité d'extraction automatique. Les modules du Unified Medical Language System ont été mis à contribution pour ajouter des informations à propos des entités extraites. Pour l'identification des grappes de patients, deux approches sont proposées. Premièrement, l'algorithme de groupement K-moyen disperses est appliqué sur une matrice patient-maladie comportant 14 comorbidités liées à l'obésité. Deuxièmement, pour visualiser et analyser d'autres maladies présentes sur les données cliniques, 86 maladies ont été utilisées pour former des grappes selon une approche fondée sur des graphes. Les graphiques bipartites obtenus ont permis d'explorer les relations maladie-traitement corrélées avec les principales grappes obtenues.

Le résultat des expérimentations a montré que cTAKES est préférable à MetaMap, mais que cette situation peut changer si l'on modifie les choix de configuration des outils – les listes d'abréviations par exemple. De surcroît, l'ajout de concepts (avec des types sémantiques similaires ou différents) s'avère une bonne stratégie pour améliorer l'acquisition automatique d'entités médicales à partir de textes cliniques.

L'algorithme K-moyen disperses a distingué trois types de grappes (élevée, moyenne et basse); ces groupes ont été identifiés en fonction du nombre de comorbidités et du pourcentage de patients affectés par elles. Ces résultats montrent que le diabète, l'hypercholestérolémie, la maladie cardiovasculaire, l'insuffisance cardiaque congestive, l'apnée obstructive pendant le sommeil, et la dépression sont les maladies les plus répandues.

La construction des graphes a permis de visualiser et d'analyser l'information des patients; elle a permis l'identification de trois sous-graphes: des patients obèses avec des problèmes de métabolisme, des patients obèses avec problèmes infectieux, et des patients obèses avec des problèmes mécaniques. Les graphes bipartites pour une relation maladie-traitement mettent ainsi en relief les traitements pour différents types de maladie, les patients obèses souffrant de multiples troubles de santé.

Cette thèse confirme que les textes narratifs cliniques en forme libre constituent une source d'information très riche qui peut être utilisée pour explorer, visualiser, et analyser l'information des patients grâce à une méthode automatisée. D'autres travaux sont nécessaires pour explorer la relation entre les différentes entités médicales des textes cliniques et les autres ensembles de données médicales. L'aspect temporel des données devrait également être considéré dans de futurs travaux afin de former un portrait personnalisé des grappes, des maladies et des patients.

Mots-clés: analyse par grappes, approche basée sur les réseaux, k-moyennes disperses, données cliniques, obésité

Computer-based Identification of Relationships Between Medical Concepts and Cluster Analysis in Clinical Notes

Ruth María REÁTEGUI ROJAS

ABSTRACT

Clinical notes contain information about medical concepts or entities (such as diseases, treatments and drugs) that provide a comprehensive and overall impression of the patient's health. The automatic extraction of these entities is relevant for health experts and researchers as they identify associations between the latter. However, automatically extracting information from clinical notes is challenging, due to their narrative format.

This research describes a process to automatically extract and aggregate medical entities from clinical notes, as well as the process to identify clusters of patients and disease-treatment relationships. The i2b2 2008 Obesity dataset was used, and consists of 1237 discharge summaries of overweight and diabetic patients. Therefore, this thesis is focused on obesity diseases.

For the automatic extraction of medical entities, MetaMap and cTAKES were used, and the automatic extraction capacity of both tools compared. Also, UMLS enabled the aggregation of the extracted entities. Two approaches were applied for cluster analysis. Firstly, a sparse K-means algorithm was used over a patient-disease matrix with 14 comorbidities related to obesity. Secondly, to visualize and analyze other diseases present in the clinical notes, 86 diseases were used to identify clusters of patients with a network-based approach. Furthermore, bipartite graphs were used to explore disease-treatment relationships among some of the clusters obtained.

The result of the experiments we conducted show cTAKES slightly outperforming MetaMap, but this situation can change, considering other configuration options in the respective tools, including an abbreviation list. Moreover, concept aggregation (with similar and different semantic types) was shown to be a good strategy for improving medical entity extraction.

The sparse K-means enabled identification of three types of clusters (high, medium and low), based on the number of comorbidities and the percentage of patients suffering from them. These results show that diabetes, hypercholesterolemia, atherosclerotic cardiovascular diseases, congestive heart failure, obstructive sleep apnea, and depression were the most prevalent diseases.

With the network approach, it was possible to visualize and analyze patient information. In it, three sub-graphs or clusters were identified: obese patients with metabolic problems, obese patients with infection problems, and obese patients with a mechanical problem. Bipartite graphs for a disease-treatment relationship showed treatments for different types of diseases, which means that obese patients are suffering from multiple diseases.

This work shows that clinical notes are a rich source of information, and they can be used to explore, visualize, and analyze patient's information by applying different approaches. More work is needed to explore the relationship between the different medical entities from clinical notes and from different disease datasets. Also, considering that some medical documents express events in time, this characteristic should be considered in future works to form a personalized portrait of clusters, diseases and patients.

Keywords: cluster analysis, network-based approach, graph, sparse K-means, clinical notes, obesity

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABREVIATIONS

BMI	Body Mass Index
CAD	Atherosclerotic Cardiovascular Disease
CCY	Gallstones/Cholecystectomy
CHF	Congestive Heart Failure
CNN	Convolutional Neural Networks
CRF	Conditional Random Fields
CUI	Concept Unique Identifier
DW	Disease Weight
EHR	Electronic Health Records
EMR	Electronic Medical Records
FN	False Negatives
FP	False Positives
GERD	Gastroesophageal Reflux Disease
HC	High Comorbidity
HCL	Hypercholesterolemia
HMMs	Hidden Markov Models
ICD	International Classification of Diseases
LC	Low Comorbidity
MC	Medium Comorbidity

MEMMs	Maximum Entropy Markov Models
NER	Named Entity Recognition
NLP	Natural Language Processing
NLS	National Library of Medicine
OA	Osteoarthritis
OSA	Obstructive Sleep Apnea
POS	Part-of-speech
PVD	Peripheral Vascular Disease
RNN	Recurrent Neural Networks
SVM	Support Vector Machine
TP	True Positives
UMLS	Unified Medical Language System
VI	Venous Insufficiency
WSD	Word Sense Disambiguation

INTRODUCTION

Nowadays, most health institutions employ Electronic Medical Records (EMR) or Electronic Health Records (EHR) with a large quantity of clinical data in a structured format (e.g., ICD and SNOMED-CT codes) and unstructured or narrative format (e.g., discharge records, radiology reports). The narrative format of clinical notes, such as in discharge summaries, provides a comprehensive and overall impression of the patient's health (Lyalina *et al.*, 2013; Alnazzawi *et al.*, 2015). As an example, the symptoms expressed by patients in the course of their illness, and recorded by health professionals, present more complete descriptions of a disease than the diagnosis expressed by a code (Jackson *et al.*, 2017).

Clinical notes contain medical concepts or entities, including diseases, treatments and drugs. Health professionals and researchers are interested in extracting medical entities (Chiaramello *et al.*, 2016; Pradhan *et al.*, 2015; Becker & Bockmann, 2016) from clinical notes as a means to understand a patient's characterization. Furthermore, medical entities form the basis for other analytical tasks, such as cluster analysis and disease-treatment relationships identification. These tasks help improve the customized treatment or care delivery (Zhang *et al.*, 2014), define boundaries and classify diseases (Lyalina *et al.*, 2013), predict the health of patients with similar characteristics (Shivade *et al.*, 2014), etc.

Many diseases can be analyzed by considering information from EHR, and more specifically, from clinical notes, with, obesity being one of them. Overweight and obesity are becoming an epidemic affecting much of the industrialized world, particularly minority groups (Aneja *et al.*, 2004; Laing *et al.*, 2015; Kovesdy *et al.*, 2017). These health problems are also accompanied by comorbidities such as: hypertension, coronary heart disease, congestive heart failure, renal diseases, diabetes, asthma, osteoarthritis, cancer, atherosclerosis and obstructive sleep apnea (Aneja *et al.*, 2004; Sutherland *et al.*, 2012; Laing *et al.*, 2015; LaGrotte *et al.*, 2016; Guh *et al.*, 2009).

Obesity has been studied for a long time, and this has led to the conclusion that obesity is a risk factor for some diseases. However, some questions and discrepancies exist because of the variety of obese patient profiles available. For example, few studies have examined the prevalence of obesity in different stages of chronic kidney diseases (Evangelista *et al.*, 2018) or how obesity influences cardiovascular mortality (Laing *et al.*, 2015). Moreover, many studies have analyzed obesity in relationship with a few of its comorbidities instead of all the comorbidities or conditions that can afflict an obese patient.

0.1 Problem statement

The narrative format of clinical notes is peculiar for many reasons. First, the notes contain many abbreviations, acronyms, and misspellings (Chiaramello *et al.*, 2016; Shivade *et al.*, 2014). Secondly, a variety of natural languages are used, depending on the particular health professional or institution (Pereira *et al.*, 2013). These characteristics complicate the extraction of medical entities from a large number of notes or patients (Pradhan *et al.*, 2015; Chiaramello *et al.*, 2016). Furthermore, entity extraction is a time-consuming, labor-intensive, and error-prone endeavor (Shivade *et al.*, 2014; Savova *et al.*, 2010), when done automatically; and worse when done manually.

The patient information contained in clinical notes is scattered and hidden in different parts of the document, meaning that information can be in several sections, and in no particular order. Moreover, the information may be present in different documents, such as discharge summaries, laboratory results, and radiography summaries. All these problems make medical entity extraction and analysis a challenging task. For such cases, the information must be located and correctly interpreted (Alnazzawi *et al.*, 2015).

Although tools like MetaMap and cTAKES have been widely used to extract medical entities (Pradhan *et al.*, 2015; Kovacevic *et al.*, 2013), new findings could appear when they are used in

clinical notes on specific diseases. In fact, (Pradhan *et al.*, 2013) have shown that entity extraction performance varies with text source (e.g., different languages and types of documents). Moreover, no study to date has compared both MetaMap and cTAKES. This comparison could help identify the advantages and disadvantages of both tools and allow recommendations on considerations to be made when selecting a tool for entity extraction. Furthermore, the use of an existing tool is the best option when there is limited time and money for the process of named entity extraction or named entity recognition.

Regarding the analysis of clinical information, and specifically cluster analysis, different diseases have been studied, such as obstructive sleep apnea (Vavougios *et al.*, 2016), asthma (Serrano-Pariente *et al.*, 2015), knee osteoarthritis (van der Esch *et al.*, 2015), chronic heart failure (Ahmad *et al.*, 2014), etc. However, there is a need to analyze and explore new diseases, and obesity is an important disease that deserves deep study.

Furthermore, clusters analysis studies on obesity collected the information manually or from structured EHR data instead of clinical notes. This case must be related with the scarcity of annotated EHR datasets, the sensitivity of the data, and the difficult process of data de-identification (Alnazzawi *et al.*, 2015). Moreover, notwithstanding the presence of multiple comorbidities in obese patients, most works focus on analyzing the relationship between 2 comorbidities.

Various methods, strategies and tools have been applied for cluster analysis: statistical analysis and hierarchical clustering methods are some examples. Other methods and approaches could be applied to explore and analyze relationships between different medical entities. Moreover, considering the need to visualize information either to identify relationships or patterns within clinical documents, graph and network approaches can be exploited. In the biomedical domain, these approaches have been used with large amounts of data to identify relationships such as gene-disease, symptom-disease, and vaccine-gene (Pavlopoulos *et al.*, 2018). Nevertheless,

few works have applied these approaches with unstructured information for cluster analysis and the disease-treatment relationships.

This thesis focuses on the extraction and analysis of information from clinical notes, with obesity as the principal disease to explore. Hence, the research question is:

Can patients' clinical notes provide new insights about a disease?

0.2 Objectives and contributions

Considering the above limitations, the general objective of this research is to analyze clinical notes to extract hidden data and information related to obese patients. The specific objectives are: 1) automatically extract medical entities, and 2) analyze the entities extracted in order to a) identify clusters of patients, and b) identify graphs of patients and disease-treatment relationships.

To achieve the goal of this thesis, three main contributions were made:

1) Automatically extract medical entities from clinical notes. MetaMap and cTAKES have been widely used to extract medical entities. Hence, these tools were used to automatically extract medical entities from obese patient discharge summaries. Next, a comparison of both tools and some important remarks are made. The results were published in the following paper:

Reátegui, R., & Ratté, S. (2018). Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Medical Informatics Decision Making, 2018 (Suppl 3), 74. https://doi.org/10.1186/s12911-018-0654-2

2) Identify clusters of patient based on medical entities extracted from clinical notes. Many research works have been done to identify clusters of obesity, but most of them were focused on structured information, such as ICD codes, or analyzed relationships between one or two

obesity comorbidities. Therefore, a second contribution is an cluster analysis of obesity based on obesity comorbidities extracted automatically from discharge summaries. The results were published in the following paper:

Reátegui, R., Ratté, S., Bautista-Valarezo, E.& Duque, V. Cluster Analysis of Obesity Disease Based on Comorbidities Extracted from Clinical Notes. Journal of Medical Systems, (2019) 43:52. https://doi.org/10.1007/s10916-019-1172-1

3) Analyze and visualize medical entity relationships using a network-based approach. Many approaches, methods and techniques have been used to analyze relationships of medical entities. A network-based approach helps visualize and infer knowledge about medical entity relationships. Hence, this approach was used to identify clusters of patients and disease-treatment relationships from obese patient discharge summaries. The results were published in the following paper:

Reátegui, R., Ratté, S., Bautista-Valarezo, E., & Beltrán, J.F. A network-based analysis of medical information extracted from electronic medical records. International Journal of Medical Informatics. (Under Review)

0.3 Synthesis of links between contributions

We mentioned earlier that the main interest of the present research is to analyze clinical notes to extract hidden data and information related to a specific disease. In order to achieve this goal, two steps are needed: 1) a medical entity extraction process, and 2) an analysis of the entities extracted.

Therefore, our first contribution aims to show that it is possible to automatically extract medical entities from discharge summaries by comparing the result obtained with two tools, MetaMap

and cTAKES. The entities extracted will be the features or variables used in the next contribution.

For the analysis of these extracted entities, we considered two two different approach: a) a cluster analysis, and b) a network-based approach.

The second contribution proposes a cluster-based approach to identify clusters of obese patients; the approach is based on 14 features or obesity comorbidities obtained from the first contribution. The sparse K-means algorithm was applied to identify clusters and a set of features that could explain the main characteristics of each group of patients.

In order, to explore relevant details about the obesity disease, supplementary features (86 diseases and 257 treatments) were automatically extracted following the process described in the first contribution. Then, considering a network approach and the information about the diseases, graphs of patients were found. These bipartite graphs allowed us to visualize association between some diseases and treatments.

0.4 Structure of thesis

This thesis comprises five chapters, as follows.

Chapter 1 presents the main concepts and a review of works related to medical entity extraction, patients clusters analysis, and disease-treatment relationships. A brief description of the dataset used is also presented. **Chapter 2** introduces medical entity extraction using two existing tools, namely, MetaMap and cTAKES. This work was published in the journal BMC Medical Informatics and Decision Making. **Chapter 3** presents an cluster analysis of obesity disease based on comorbidities using Sparse K-means. This work was published in the Journal of Medical Systems. **Chapter 4** presents a network-based approach to identify clusters of patients and disease-treatment relationship. This work was submitted to the International Journal

of Medical Informatics. **Chapter 5** presents a general discussion, the conclusion and some perspectives for future research.

CHAPTER 1

LITERATURE REVIEW

To extract knowledge from clinical notes, two main steps are needed: first, extract medical entities that will become features or variables, and second, analyze the medical entities extracted. Hence, this chapter is divided into two main sections: medical entity recognition and medical entity analysis. This thesis has a special interest in clusters analysis of obese patients.

1.1 Named entity recognition

The problem of automatically extracting relevant concepts from text is known as named entity recognition (NER) (Jonnalagadda *et al.*, 2012). In the medical field, NER refers to the process of identifying medical entities or concepts such as a diseases, treatments, drugs, etc. Some approaches, and a combination thereof, have been used for NER: a dictionary-based approach uses a dictionary or lexicon of the concepts to be extracted; a rule-based approach identifies rules related to the entities to be extracted, while the machine learning approach uses algorithms that require an annotated data training dataset.

Dictionary-based and rule-based approaches were the first to be used for NER, and are still being used in some investigations. In (Alnazzawi *et al.*, 2015), a dictionary-based method using the MetaMap tool, a rule-based approach, and 3 machine learning methods (hidden Markov models (HMMs), maximum entropy Markov models (MEMMs), and conditional random fields (CRFs)) were evaluated to extract some medical entities from clinical notes and articles. The result shows that rules had the highest F-score for both clinical notes and articles. Similarly, (Jonnalagadda *et al.*, 2017) used a rule-based system to identify terms, and consequently, patients with heart failure with preserved ejection fraction. The rules were applied on multiple unstructured notes.

The most popular machine learning methods used to NER are CRFs and support vector machine (SVM). As an example, (Jonnalagadda *et al.*, 2012) extracted medical problems, treat-

ments and tests from clinical notes combining distributional semantics and machine learning algorithms. They used a sliding window and Random Indexing for dimension reduction, after which they worked with a CRF algorithm, adding distributional semantic features to lexicons and linguistic features. The CRF model created during the training phase is used to tag the input sentences with concepts such as medical problems, treatments and tests. In addition, (Tang *et al.*, 2013) worked on two tasks, namely, disorder entity recognition and encoding. For the first task, they used a machine learning approach, and for the second one, a vector space model. In this work, the structural SVM algorithm outperformed CRF in disorder recognition.

Recently, deep learning methods have been applied to improve NER systems in the biomedical and medical fields. (Zhu *et al.*, 2018) implemented a convolutional neural network (CNN) with character embedding and word embedding. They achieved a better performance as compared to the conventional machine learning approach for Biomedical NER; however, their work has some limitations, including the fact that it is not prepared to consider overlapping or disjointed mentions or mentions in tables. Also, it requires a significant amount of training data and is time-consuming. Similarly, (Wu *et al.*, 2017) analyzed CNN and the recurrent neural network (RNN) to extract concepts from clinical texts. They compared both methods with three baseline (CRFs) models and two state-of-the-art clinical NER systems. The results showed that RNN achieved a superior performance for NER.

1.1.1 Limitation of NER approaches

All the above-mentioned approaches have greatly contributed to improving NER in all fields. However, each of them presents some limitations or problems that must be considered before they are selected for a given research work.

Due to the large amount of medical terminology and the continuous increase in vocabulary, the size of a dictionary can become a problem in the performance of dictionary-based approaches (Sun *et al.*, 2018; Zhu *et al.*, 2018). Moreover, as free text, clinical notes include a lot of misspellings, abbreviations and synonyms not covered in the dictionary.

Rules-based approaches raise some problems for consideration: (1) these methods require domain knowledge for manual examination and pattern extraction, which makes them costly and time-consuming; (2) the methods are prone to errors, and scarcity of information in their data can lead to inappropriate or unconsidered rules; (3) the rules identified are valid only for the dataset analyzed, and are thus hard to extrapolate to other domains (Sun *et al.*, 2018; Zhu *et al.*, 2018).

The main problem with a machine learning approach is that it requires labeled or annotated datasets. The amount of labeled data and the features used by learning algorithms are time-consuming and computationally intensive (Zhu *et al.*, 2018).

As we can see, all the NER approaches present some problems and challenges for medical entity extraction. The present work reviews two of the most commonly used NER tools, which enable medical entity extraction without requiring too much time or necessitating expert intervention. The next section introduces these tools.

1.1.2 Medical entity extraction tools

Currently, various tools exist for extracting information from clinical texts created in an unstructured format. Two such tools, both widely known and used in the biomedical field, are MetaMap and cTAKES (Pradhan *et al.*, 2015; Kovacevic *et al.*, 2013).

MetaMap was developed by the National Library of Medicine (NLM) to map biomedical text to concepts in the Unified Medical Language System (UMLS) (Aronso, 2001; Aronson & Lang, 2010). The tool uses a hybrid approach combining natural language processing (NLP), a knowledge-intensive approach and computational linguistic techniques (Aronso, 2001). Metamap execute some tasks such as: tokenization, sentence boundary determination and acronym/abbreviation identification, part-of-speech (POS) tagging, lexical lookup of input words in the SPECIALIST lexicon, shallow parse, variant generation, candidate identification, mapping construction and word sense disambiguation (WSD) (Aronson & Lang, 2010).

Similarly, the Clinical Text Analysis and Knowledge Extraction System (cTAKES) combines rule-based and machine learning techniques to extract information from a clinical text (Savova *et al.*, 2010). cTAKES executes some components in sequence to process the clinical text. Its components include a sentence boundary detector, a tokenizer, a normalizer, a part-of-speech tagger, a shallow parser, and NER with a status and negation annotator. The NER component implements a dictionary lookup algorithm to map a named entity to a concept from the terminology. Initially, a UMLS subset that includes SNOMED-CT and RxNORM vocabularies was used (Savova *et al.*, 2010), but today, it is possible to include other sources and even create custom dictionaries.

Both tools use the Unified Medical Language System (UMLS) to extract and standardize medical concepts. A brief explanation of UMLS is presented below.

1.1.3 Unified medical language system - UMLS

The Unified Medical Language System (UMLS) provides terminology, coding standards, and resources for biomedical and electronic health systems. UMLS was developed by the National Library of Medicine (NLM) in the United States. UMLS has three Knowledge Sources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon & Lexical Tools.

The Metathesaurus is organized by concepts or meanings. A concept has a unique and permanent identifier (CUI) and a preferred name. The concept is a meaning, and a meaning can have different names from different vocabularies or thesauruses (National Library of Medicine (US), 2009). UMLS grabs the meanings of a concept from different sources and links all o that are synonyms. As an example of UMLS coding, hypertension is represented with the CUI "C0020538" and the preference name "hypertension diseases".

The Semantic Network provides (1) a categorization (semantic types) of all concepts represented in the UMLS Metathesaurus; and (2) a set of relationships (semantic relations) between these concepts (National Library of Medicine (US), 2009). The Semantic Network contains 133 semantic types and 54 relationships.

Some of the semantic types defined by UMLS and of interest for this thesis are:

- Antibiotic (antb)
- Clinical drug (clnd)
- Disease or syndrome (dsyn)
- Mental or behavioral dysfunction (mobd)
- Neoplastic process (neop)
- Pathologic function (patf)
- Therapeutic or preventive procedure (topp)
- Pharmacologic substance (phsu)

The SPECIALIST Lexicon is a general English lexicon of biomedical terms. It contains syntactic, morphological, and orthographic information. The Lexical Tools are programs for language processing (National Library of Medicine (US), 2009).

UMLS is based on some electronic thesauruses, classifications, code sets, and controlled terms list such as SNOMED CT and RxNorm (National Library of Medicine (US), 2009). The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) is a multilingual health terminology used for the electronic exchange of clinical health information. In the U.S., SNOMED CT is a international standard for electronic exchange of clinical health information (National Library of Medicine (US), 2016). On the other hand, RxNorm standardizes clinical drug names and links the names to other vocabularies used in pharmacy management and drug interaction software (National Library of Medicine (US), 2014).

1.1.4 Evaluation metrics

The three metrics used to evaluate the result of a NER task are precision (P), recall (R) and F-score.

$$Precision = \frac{TP}{TP+FP} \tag{1.1}$$

$$Recall = \frac{TP}{TP+FN} \tag{1.2}$$

$$F-score = 2.\frac{Precision.Recall}{Precision+Recall} \tag{1.3}$$

where TP is the number of true positives, FN is the number of false negatives, and FP is the number of false positives.

1.2 Medical entities analysis

Once the entities have been extracted, a variety of tasks can be developed. A task of interest in this thesis is patients cluster analysis, and disease-treatment relationships identification.

1.2.1 Patients cluster analysis

This thesis aims to explore clinical notes to identify clusters of patients with similar characteristics or phenotypes through the association of comorbidities and to discover novel insights about obesity-related diseases. Obesity is a heterogeneous disease, which makes it difficult to classify patients using predetermined criteria. Hence, this work aims to explore this disease and the different conditions that can accompany it.

Cluster analysis based on patients medical information might help in personalized treatment management, patient symptom control, boundary definition and disease taxonomies, understanding the heterogeneity of the disease, predicting the health of patients with similar characteristics, predicting future patient risks, and identifying relevant pathophysiology (Bourdin *et al.*, 2014; Lyalina *et al.*, 2013; Shivade *et al.*, 2014; Zhang *et al.*, 2014).

Many studies have worked with cluster analysis. These studies are diverse in methodology and diseases analyzed. As an example, (Joosten *et al.*, 2012) applied K-means algorithms to identify patient with obstructive sleep apnea. They analyzed patient data through a hospital records system. The same disease was studied by (Vavougios *et al.*, 2016), who carried out a cluster analysis using two-step algorithms. Data was collected from 1472 patient records recovered from the University of Larissa Sleep Laboratory database (ULDB). Furthermore, (Bourdin *et al.*, 2014) applied Ward's minimum variance hierarchical clustering method to identify groups of asthma patients. They worked with information from recruitment patients with severe asthma. Also, (Serrano-Pariente *et al.*, 2015) used two-step algorithms to identify and characterize near-fatal ashtma phenotypes. They worked with 84 cases from the Multicentric Life-Threatening Ashtma Study (MLTAS).

K-means was used by (van der Esch *et al.*, 2015) to carry out a cluster analysis and identify and validate established knee osteoarthritis phenotypes. These authors analyzed information on 551 patients from the Amsterdam OA (AMS-OA) cohort, with a unilateral or bilateral diagnosis of knee OA. (Ahmad *et al.*, 2014) used Ward's minimum variance hierarchical clustering method to identify clusters of patients with chronic heart failure. They used the information from Heart Failure: A Controlled Trial Investigating Outcomes of Exercise Training (HF-ACTION). (Chen *et al.*, 2014) worked with Ward's minimum variance hierarchical clustering method to define phenotypes of males with chronic obstructive pulmonary diseases. The features studied were obtained from 377 male recruits.

Furthermore, (Bukhanov *et al.*, 2017) used frequency analysis, association rules mining, and Bayesian network to identify groups of comorbidities in hypertensive patients. Information was extracted from clinical notes and used by an expert to create a vocabulary to represent a classification of encoded diseases. (Antonelli *et al.*, 2013) performed a multiple-level clustering analysis using the DBSCAN algorithm to discover diabetic patients with similar examination histories. The authors used ICD 9-CM codes.

Obesity diseases have also been studied by some researchers. The condition is often accompanied by comorbidities like diabetes, dyslipidemia, hypertension, cardiovascular diseases, asthma, and osteoarthritis (Figueroa & Flores, 2016; Foster *et al.*, 2008; Guh *et al.*, 2009).

Works that make a cluster analysis taking account of obesity-related diseases include that by (Sutherland *et al.*, 2012), which analyzed obesity and asthma. They worked with Ward's minimum-variance hierarchical clustering method. They used information from 250 adults with complete clinical, physiologic and inflammatory data. As well, (Laing *et al.*, 2015) analyzed obesity and atherosclerosis using a statistical method. They used information from 503 patients. Among the features analyzed were extensive family, socio-economic, educational, personal medical history, physical activity, anthropometric measurements, and laboratory data. (LaGrotte *et al.*, 2016) analyzed patients with obstructive sleep apnea, obesity, and excessive daytime sleepiness using statistical analysis. They analyzed the data of 1137 adults, collected as part of a population-based study of sleep disorders

1.2.2 Limitation of patients cluster analysis

A variety of works analyzing diseases carried out a prospective study with patient information gathered through direct measures, interviews or questionnaires. A few examples are: (Ahmad *et al.*, 2014; Chen *et al.*, 2014; Bourdin *et al.*, 2014; Laing *et al.*, 2015; Serrano-Pariente *et al.*, 2015; LaGrotte *et al.*, 2016). The main difficulty with these types of studies lies in their high cost and the time they require. Additionally, the studies call for training of personnel to collect information. On the other hand, retrospective studies which consider existing information in EHR present certain advantages. (Simmons *et al.*, 2016) mention that the use information from EHR is relatively inexpensive because the data is generated as a result of a healthcare process, EHR provides long-term information on a variety of diseases, and the information improves as the record of the same patient grows.

Considering the advantages of information from EHR, a common strategy employed consists in is using structured data from EHR, such as ICD-9 or ICD-10 (Antonelli *et al.*, 2013; Zhang

et al., 2014). However, this approach has been deemed inadequate for describing patient phenotypes (Shivade *et al.*, 2014). Moreover, a new problem emerges where a disease or medical condition does not have a specific code (Anzaldi *et al.*, 2017). To avoid these situations, many studies suggest exploring unstructured information like clinical notes.

Beyond the challenges inherent to the narrative form of clinical notes (explained in the problem statement section), information from EHR poses other problems. In their work, (Simmons *et al.*, 2016) mention difficulties stemming from the legacy, sensitivity and confidential characteristics of these resources. Additionally, when a patient visits a different physician at a different health institution (that does not share the same EHR), the information on the patient's health could be incomplete. Furthermore, the authors mention that prospective studies use a more qualitative data collection method.

The Charlson Comorbidity index and Elixhauser index are patient comorbidity categorization methods. The methods use diagnostic and administrative data to predict mortality and resource usage. However, these indexes consider some comorbidities that need empirical observation and do not scale up for a population with multiple comorbidities (Khan *et al.*, 2018). Many diseases do not appear in isolation. Multiple risk factors must be analyzed comprehensively to understand the effects on health outcomes (Bukhanov *et al.*, 2017).

While obesity and other diseases (e.g., diabetes) appear with multiple comorbidities, most obesity studies, however, focus on the relationship between 2 or 3 comorbidities. Information of patients suffering multiple medical conditions must be considered. Doing so will make it possible to study whether a drug has an effect on the appearance of other medical conditions, if the treatment could lead to a drug interaction, if a treatment for a new patient with similar characteristics could be prescribed, etc.

Cluster analysis studies have been considered features pre-established by the literature or by health experts, but clinical notes contain hidden features and associations that could be unlocked by NER. Clinical notes, as earlier mentioned, are writings from health professionals.

Such notes therefore have more details on the patient's health as well as on the temporal evolution of the disease or treatment.

1.2.3 A network-based approach to medical entity analysis

A strategy used to visualize associations among medical entities and infer medical knowledge is a network-based approach. This method has been used in the biomedical field to understand gene, drug, disease and vaccine, associations and interactions (Pavlopoulos *et al.*, 2018).

Many works have considered clinical data. (Lyalina *et al.*, 2013) used network methods to visualize the association between symptoms or clinical findings related to neuropsychiatric disorders; (Roque *et al.*, 2011) visualized clusters of psychiatric patients; (Chen & Xu, 2014) first extracted association rules for comorbidity patterns of colorectal cancer, after which, a network method was applied to construct a human disease comorbidity network. The work of (Khan *et al.*, 2018) created networks to identify comorbidities and conditions related with type 2 diabetes. They worked with diabetic and nondiabetic cohorts to discover comorbidities exclusive of diabetic patients. Similarly, (Kalgotra *et al.*, 2017) identified comorbidities classified by gender. They worked with information on diagnostics, symptoms, and treatments. Additionally, (Merrill *et al.*, 2015) used a network approach based on information from inpatient and outpatient clinical services to identify care patterns for congestive heart failure. (Zhao et al., 2017) manually annotated medical entities from medical records to construct a network that was used to propose a diagnosis model. (Rotmensch *et al.*, 2017) first extracted diseases and symptoms from structured and unstructured clinical notes. They then constructed a statistical model with the information extracted, and then translated the models learned into knowledge graphs.

The identification of communities in networks helps to uncover unknown topics in information, social communities, patterns or themes in medical information. (Gangopadhyay *et al.*, 2016) divided a network of terms extracted from clinical notes using communities identification. They identified main terms related to each sub-graph, and presented results related to anemia.

The bipartite networks is another important concept. In the biomedical field, one partition could represent genes, proteins, molecules, drugs, or environmental exposures, and the other partition could represent diseases, symptoms, or adverse drug effects (Pavlopoulos *et al.*, 2018). (Goh *et al.*, 2007) used a bipartite graph consisting of diseases and genes nodes to construct a diseasome. A disease and gene were connected if mutations in that gene were involved in the disease. In addition, (Bhavnani *et al.*, 2011) used bipartite networks to represent asthma patients and cytokines as nodes, with the normalized cytokine expression values being the edges. They identified cytokine clusters and their relationship to patient clusters, and drew biological meanings about the patient clusters.

1.2.4 Limitation of network-based approach to medical entities analysis

Recent years have seen an increase in the number of works that analyze clinical information following a network approach. Some works research the relationship between symptoms, diseases and drugs, diseases and genes, and others conduct research on the comorbidities or patterns of a specific disease. However, most works, similarly to the case of patients cluster analysis, use structured information obtained from EHR, or through patient measures, interviews or questionnaires.

Furthermore, works that include genes as features also take the information from databases or catalogues such as the On-line Mendelian Inheritance in Man (OMIN), Genome-Wide Association Studies (GWAS), National Institutes of Health's Genetic Association Database (GAD), Human Genome Organisation (HUGO), and Phenome-Wide Association Studies (PHeWA). All these catalogues comprise structured information.

Unstructured information within clinical notes must therefore be analyzed and explored with a network-based approach.

1.2.5 Sparse K-means

(Bukhanov *et al.*, 2017) describe some of the goals of cluster analysis of patients' information. They include the following:

- To find clusters based on demographic data, risk factors and comorbid diseases.
- To determine recommendations for diagnostics and treatment of diseases base on the cluster information.
- To identify similarities and difference between the clusters.
- To find predictors of treatment responses depending on the patient's profile, clustering results and a combination of recommendations.

Sparse K-means algorithm developed by (Witten & Tibshirani, 2010) helps to simultaneously find clusters and a subset of cluster features. This algorithm assigns a weight to each feature, with the most important ones having the highest values. Sparse K-means can handle either a number of features greater than the number of observations or vice-versa. Also, the algorithm is useful when the dataset has noise variables, and it is characterized by the following criterion:

$$\underset{C_1,\ldots,C_K,\mathbf{w}}{\text{maximize}} \left\{ \sum_{j=1}^{p} w_j \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{i,i',j} \right) \right\} \tag{1.4}$$

subject to $\| w \|^2 \leq 1, \| w \|_1 \leq s, w_j \geq 0 \forall j$, where, w are non-negative weights for the p features and n is the number of observations. The weights will be sparse for an appropriate choice of the tuning parameter s.

Before use sparse k-means, the number of clusters needs to be determined. Gap statistics is a standard method for detecting the number of clusters. The method compares the within-cluster dispersion to its expectation under an appropriate reference null distribution (Tibshirani *et al.*, 2001). These authors define the gap statistics with the following equations:

The sum of the pairwise distances D_r of all elements i,i' in a cluster C_r is as follows:

$$D_r = \sum_{i,i' \in C_r} d_{i,i'} \tag{1.5}$$

The pooled within-cluster sum of squares around the cluster means is defined as W_k:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \tag{1.6}$$

where k is the number of clusters, n represents the number of observation, and r denotes cluster indices.

Then, the gap statistic is defined by the equation:

$$Gap_n(k) = E_n^*\{log(W_k)\} - log(W_k), \tag{1.7}$$

where E_n^* denotes expectation under a sample of size n from the reference distribution.

To evaluate a cluster analysis, the Silhouette coefficient is commonly used. Silhouette allows knowing the intra-cluster cohesion and the inter-cluster separation. (Rousseeuw, 1987) defines this coefficient as follows:

a(i) = average dissimilarity of i to all other objects of a cluster A.

d(i,C) =average dissimilarity of i to all other objects of clusters C. $C \neq A$.

$b(i) = min_{C \neq A} d(i,C)$ b(i) is the minimum average distance from i to other clusters different from the cluster to which i belongs.

The silhouette coefficient s(i) is defined as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{1.8}$$

a(i) reflects the compactness of the cluster to which i belongs. A small value means that the cluster is more compact. The value of b(i) represents the degree to which i is separated from other clusters. Therefore, if b(i) has a large value, then i is more separated from other clusters (Han *et al.*, 2011).

The Silhouette coefficient is a value between -1 and 1. If the Silhouette value is closer to 1, the cluster is compact and well separated from other clusters, and objects in the cluster are well assigned. If the Silhouette value is closer to -1, the cluster is not well separated from other clusters, and objects in it are wrongly assigned. When the Silhouette is closer to zero, it is not clear if the objects have been assigned to the actual cluster or to another cluster (Rousseeuw, 1987).

It is important to note that sparse K-mean is still being improved by some research. For example, (Kondo *et al.*, 2016) worked with a robust sparse K-means (RSKM) to handle outliers. RSKM, as well as Sparse K means, gap statistic, and Silhouette are implemented in R.

1.2.6 Network and graph theory

A network is a structure formed by an ordered pair G = (V , E). V is a set of vertices or nodes. E is a set of edges or connections between the nodes. A node could represent any discrete entity (e.g. an individual or an event), and an edge indicates a relationship between nodes (Merrill *et al.*, 2015; Kalgotra *et al.*, 2017). Therefore, different systems can be represented by networks, such as the transport systems in a city, the relationships between members of a social media platform, the interaction between genes or proteins, the relationship between drugs, patients, etc.

An adjacency matrix A is a way to represent a graph G. A is a square matrix where an element Aij is 1 when there is an edge from node i to node j and 0 when there are no edges. For a

simple graph, the diagonal elements of the matrix A are 0. This graph representation allows a computational analysis.

Due to the huge amount of information that can be represented and visualized by a network, an approach to analyzing this information involves decomposing the networks into highly interconnected communities or set of nodes (Newman, 2006). This community detection, also called network clustering, uses a modularity function with a scalar value between -1 and 1. A positive value indicates the possible presence of a community structure (Blondel *et al.*, 2008; Newman, 2006).

Another concept of interest in this study is the bipartite graph. In this graph, nodes are divided in two non-overlapping sets, and the edges only join two nodes in different sets (Chang & Tang, 2014; Guimera *et al.*, 2007). This network is useful for representing different types of objects. In biomedical fields, a bipartite graph can represent drugs and diseases, genes and diseases, symptoms and diseases, vaccines and gene networks (Pavlopoulos *et al.*, 2018), and so on.

In Gephi, the modularity function is implemented with the algorithm of (Blondel *et al.*, 2008).

1.3 Obesity Dataset

In this thesis, the i2b2 Obesity Dataset was used. The i2b2 (Informatics for Integrating Biology to the Bedside) made a call for an Obesity Challenge in 2008. This challenge is a multi-class, multi-label classification task focused on obesity and its comorbidities. The available dataset consists of 1237 discharge summaries from the Partners HealthCare Research Patient Data Repository. The de-identificated summaries are from patients who were overweight or diabetic and had been hospitalized for obesity or diabetes (Uzuner, 2009).

The dataset has annotations on obesity and fifteen obesity comorbidities: asthma, atherosclerotic cardiovascular disease (CAD), congestive heart failure (CHF), depression, diabetes mellitus (DM), gallstones/cholecystectomy, gastroesophageal reflux disease (GERD), gout, hyper-

cholesterolemia, hypertension (HTN), hypertriglyceridemia, obstructive sleep apnea (OSA), osteoarthritis (OA), peripheral vascular disease (PVD), and venous insufficiency.

The data were annotated by two experts working on two different types of annotations. (Uzuner, 2009) provides a complete description of the annotations. Here is a brief description: (1) a textual annotation where the experts classify each disease as Present, Absent, Questionable, or Unmentioned based on explicitly documented information in the discharge summaries, (2) an intuitive annotation where the experts classify each disease as Present, Absent, or Questionable by applying their intuition and judgement to information in the discharge summaries.

To access the i2b2 Obesity dataset, I signed the Data Use and Confidentiality Agreement from i2b2.

CHAPTER 2

COMPARISON OF METAMAP AND CTAKES FOR ENTITY EXTRACTION IN CLINICAL NOTES

Ruth Reátegui[1, 2], Sylvie Ratté[1]

[1] Département de Génie Logiciel et des Technologies de l'Information, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Departamento de Ciencias de la Computación y Electrónica, Universidad Técnica Particular de Loja (UTPL),
San Cayetano Alto, Loja, Loja, Ecuador 11-01-608

Paper published in the journal BMC Medical Informatics and Decision Making, September 2018.

Abstract

Background: Clinical notes such as discharge summaries have a semi- or unstructured format. These documents contain information about diseases, treatments, drugs, etc. Extracting meaningful information from them becomes challenging due to their narrative format. In this context, we aimed to compare the automatic extraction capacity of medical entities using two tools: MetaMap and cTAKES.

Methods: We worked with i2b2 (Informatics for Integrating Biology to the Bedside) Obesity Challenge data. Two experiments were constructed. In the first one, only one UMLS concept related with the diseases annotated was extracted. In the second, some UMLS concepts were aggregated.

Results: were evaluated with manually annotated medical entities. With the aggregation process the result shows a better improvement. MetaMap had an average of 0.88 in recall, 0.89 in precision, and 0.88 in F-score. With cTAKES, the average of recall, precision and F-score were 0.91, 0.89, and 0.89, respectively.

Conclusions: The aggregation of concepts (with similar and different semantic types) was shown to be a good strategy for improving the extraction of medical entities, and automatic aggregation could be considered in future works.

Keywords: cTAKES; MetaMap; UMLS; Clinical documents.

2.1 Background

Electronic Health Records (EHR) or Electronic Medical Records (EMR) save patients' information in a format that is either structured (e.g., diagnosis codes, laboratory results, medication) or unstructured (e.g., clinical notes). Clinical notes, such as discharge summaries, radiology notes, and progress notes, have an unstructured format with a narrative style. These documents provide a more complete portrait of the patient's health (Roque *et al.*, 2011; Lyalina *et al.*, 2013; Alnazzawi *et al.*, 2015), as well as additional valuable information (e.g., diagnosis, symptoms, medical history, social history, medication, lab tests, treatments, etc.). Unfortunately, unstructured formats complicate information extraction. First, they contain many abbreviations, acronyms, and specialized terms (Chiaramello *et al.*, 2016) . Secondly, a variety of natural languages are used, depending on the particular health professional or institution (Pereira *et al.*, 2013), and may not correspond to a general domain. Furthermore, manual annotations and analysis present in clinical notes can transform extraction into a time-consuming, labor-intensive, and error-prone endeavor (Savova *et al.*, 2010).

Nowadays, various tools exist for extracting information from clinical texts created in an unstructured format. Two such tools, which are widely used and known in the biomedical field, are MetaMap and cTAKES (Pradhan *et al.*, 2015; Kovacevic *et al.*, 2013). MetaMap was developed by the National Library of Medicine (NLM) to map biomedical text to concepts in the Unified Medical Language System (UMLS) (Aronso, 2001; Aronson & Lang, 2010). The tool uses a hybrid approach combining a natural language processing (NLP), knowledge-intensive approach and computational linguistic techniques (Aronso, 2001). The Clinical Text Analysis and Knowledge Extraction System (cTAKES) combines rule-based and machine learning

techniques to extract information from a clinical text (Savova *et al.*, 2010). cTAKES executes some components in sequence to process the clinical text. Both MetaMap and cTAKES use the Unified Medical Language System (UMLS) to extract and standardize medical concepts.

The extraction of medical entities (e.g., diseases, treatments, drugs, etc.) is important for patients and medical research (Chiaramello *et al.*, 2016; Pradhan *et al.*, 2015; Becker & Bockmann, 2016). Moreover, these medical entities form the basis for other tasks such as disease correlation (Roque *et al.*, 2011), disease classification (Yıldırım *et al.*, 2010, 2012) , disease diagnosis (Pereira *et al.*, 2013; Bejan *et al.*, 2012), phenotype identification (Lyalina *et al.*, 2013; Alnazzawi *et al.*, 2015), etc.

Given the significance of medical entity extraction, this paper aims to compare this extraction carried out using two different tools (MetaMap and cTAKES). For this project, we worked with the i2b2 (Informatics for Integrating Biology to the Bedside) Obesity Challenge data. The automated extraction was evaluated against the experts' manual annotations of 14 obesity comorbidities (simultaneous presence of two chronic diseases or conditions in a patient) from discharge summaries.

2.2 Materials and Methods

2.2.1 Dataset

The i2b2 2008 Obesity dataset consists of 1237 discharge summaries of overweight and diabetic patients (Uzuner, 2009). The documents contain two different expert annotations: textual and intuitive. In this work, we use textual annotations where experts classified 15 obesity comorbidities based on the explicit information in discharge summaries. The diseases had four classifications:

- Present: The patient has/had the disease.
- Absent: The patient does not/did not have the disease.

- Questionable: The patient may have the disease.

- Unmentioned: Absence of information of the disease in the discharge summary.

The first column of Table 2.1 shows the 14 comorbidities used. Hypertriglyceridemia was excluded due to a lack of sufficient samples. Out of 1237 summaries, we selected the 412 summaries which had obesity as a comorbidity.

Table 2.1 List of entities or concept

Entities annotated by experts	Entities in the first experiment	Entities or groups in the second experiment
Name of disease	**Preferred name, CUI, Semantic Type**	**Preferred name, CUI, Semantic Type**
Hypertension	Hypertensive disease, C0020538, dsyn	Hypertensive disease, C0020538, dsyn
Diabetes	Diabetes mellitus, C0011849, dsyn	Diabetes mellitus, C0011849, dsyn
		Diabetes mellitus, insulin-dependent, C0011854, dsyn
		Diabetes mellitus, non-insulin-dependent, C0011860, dsyn
Atherosclerotic Cardiovascular Disease (CAD)	Coronary artery disease, C1956346, dsyn	Coronary artery disease, C1956346, dsyn
		Coronary arteriosclerosis, C0010054, dsyn
Congestive Heart Failure (CHF)	Congestive heart failure, C0018802, dsyn	Congestive heart failure, C0018802, dsyn
Hypercholesterolemia	Hypercholesterolemia, C0020443, dsyn	Hypercholesterolemia, C0020443, dsyn
		Hyperlipidemia, C0020473, dsyn
Obstructive Sleep Apnea (OSA)	Sleep apnea obstructive, C0520679, dsyn	Sleep apnea obstructive, C0520679, dsyn
Osteoarthritis (OA)	Degenerative polyarthritis, C0029408, dsyn	Degenerative polyarthritis, C0029408, dsyn
Depression	Mental depression, C0011570, mobd	Mental depression, C0011570, mobd
		Depressive disorder, C0011581, mobd
Asthma	Asthma, C0004096, dsyn	Asthma, C0004096, dsyn
Gastroesophageal Reflux Disease (GERD)	Gastroesophageal reflux disease, C0017168, dsyn	Gastroesophageal reflux disease, C0017168, dsyn
Gallstones/Cholecystectomy	Cholecystectomy procedure, C0008320, topp	Cholecystectomy procedure, C0008320, topp
		Cholecystolithiasis, C0947622, dsyn
		Cholecystitis, C0008325, dsyn
		Cholelithiasis, C0008350, dsyn
Gout	Gout, C0018099, dsyn	Gout, C0018099, dsyn
Peripheral Vascular Disease (PVD)	Peripheral vascular diseases, C0085096, dsyn	Peripheral vascular diseases, C0085096, dsyn
Venous Insufficiency	Venous insufficiency, C0042485, dsyn	Venous insufficiency, C0042485, dsyn
		Postthrombotic syndrome, C0277919, patf

CUI: Concept Unique Identifier
The second experiment grouped together some entities related to the disease annotated by the experts.
dsyn = Disease or Syndrome; mobd = Mental or Behavioral Dysfunction; topp = Therapeutic or Preventive Procedure; patf = Pathologic Function

2.2.2 Unified Medical Language System

The National Library of Medicine Unified Medical Language System (UMLS) provides terminology, coding standards, and resources for biomedical and electronic health systems. UMLS has three Knowledge Sources: the Metathesaurus, the Semantic Network and the SPECIALIST lexicon.

The Metathesaurus is organized by concepts or meanings. A concept has a unique and permanent identifier (CUI) and a preferred name. The concept is a meaning, and a meaning can have different names from different vocabularies or thesauruses (National Library of Medicine (US), 2009). The Semantic Network provides (1) a categorization (semantic type) of all concepts represented in the UMLS Metathesaurus; and (2) a set of relationships (semantic relations) between these concepts (National Library of Medicine (US), 2009). The Semantic Network contains 133 semantic types and 54 relationships.

UMLS is based on some electronic thesauruses, classifications, code sets, and lists of controlled terms like SNOMED CT and RxNorm (National Library of Medicine (US), 2009). The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) is a multilingual health terminology used for the electronic exchange of clinical health information (National Library of Medicine (US), 2016). In the U.S., SNOMED CT is the national standard for electronic exchange of clinical health information (National Library of Medicine (US), 2016). On the other hand, RxNorm standardizes clinical drug names and links the names to other vocabularies used in pharmacy management and drug interaction software (National Library of Medicine (US), 2014).

In this work, the medical entities extracted will be the concepts represented by the CUIs. We worked with SNOMED CT and RxNorm as vocabularies and with four semantic types (henceforth ST):

- Disease or Syndrome
- Mental or Behavioral Dysfunction
- Pathologic Function
- Therapeutic or Preventive Procedure

2.2.3 Automatic Extraction

We used separately MetaMap (version 2015) and cTAKES (version apache-ctakes-3.2) to extract the CUIs related with the 14 obesity comorbidities above mentioned. With each tool, two different experiments were carried out to extract the entities automatically.

In the first experiment, we identified one CUI code related to each comorbidity or disease. The extracted CUI and the preferred name of the concepts are shown in Table 2.1, column 2. In this experiment, diabetes, atherosclerotic cardiovascular disease (CAD), hypercholesterolemia, osteoarthritis, depression, venous insufficiency, and cholecystectomy have low values in the evaluation (see Table 2.2). Therefore, to improve the results for these diseases, a second experiment was performed.

In the second experiment, we worked with two types of aggregations described below. Aggregation has been wide applied in the genetic field. For example, a pathway level is used instead of individual genes to obtain a compact representation or to improve tasks like classification or clustering (Hwang, 2012).

1. Aggregation of CUIs with the same ST. The aggregation of CUIs belonging to the ST "Diseases or Syndromes" allowed us to cover diabetes, coronary artery disease and hypercholesterolemia, while the aggregation of CUIs belonging to the ST "Mental or Behavioral Dysfunction" allowed us to cover mental depression.

2. Aggregation of CUIs with different ST. First, we aggregated CUIs belonging to the ST "Diseases or Syndrome" with CUIs belonging to the ST "Pathologic Function"; this grouping allowed us to recover enough information to better identify venous insufficiency. Second, we aggregated CUIs belonging to the ST "Therapeutic or Preventive Procedure" with CUIs belonging to the ST "Diseases or Syndrome"; this second grouping allowed us to recover the information needed to identify cholecystectomy. Details of the CUIs grouped together are shown in Table 2.1, column 3. Figure 2.1 shows the process for the second experiment and Figure 2.2 shows the aggregation process.

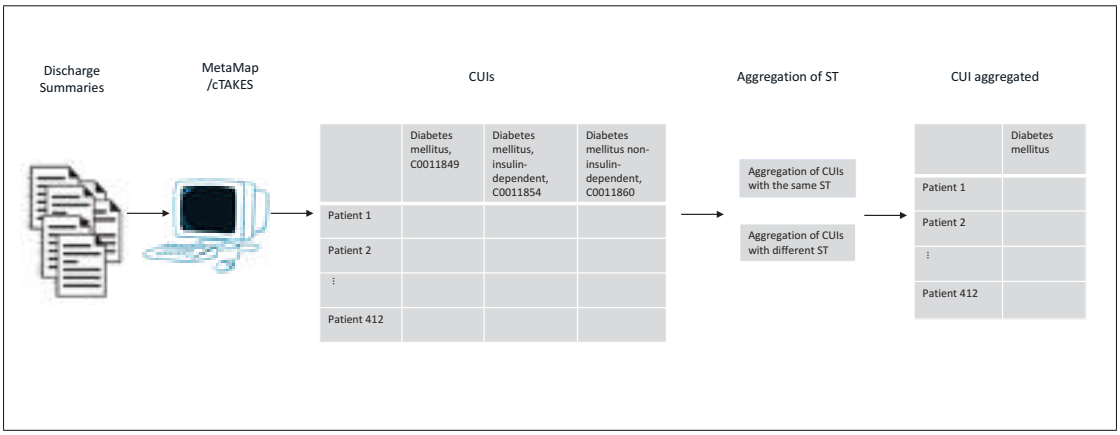


Figure 2.1 Process for the second experiment. Discharge summaries were analyzed with MetaMap or cTAKES to extract CUIs. Then some CUIs were aggregated to obtain the 14 comorbidities related with obesity.

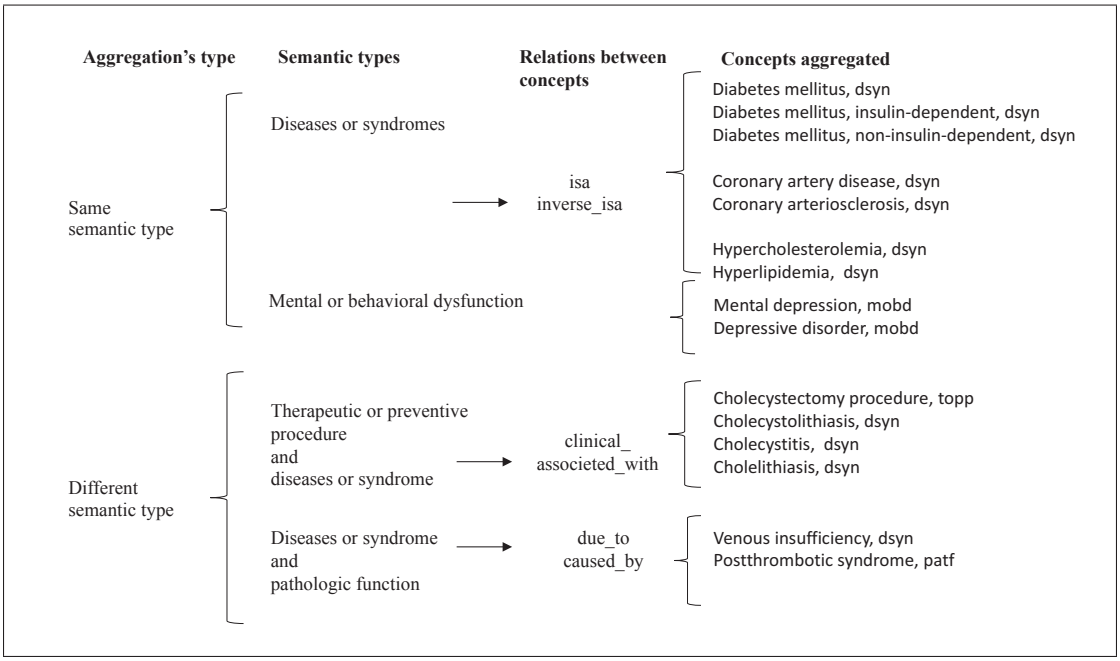


Figure 2.2 Aggregation process

2.2.4 Evaluation Metrics

We considered the experts' annotations as a gold standard in evaluating the automatic extraction. Only the "Present" annotation was taken into account in identifying whether the patient

has or had the diseases. We used the recall (or sensitivity), precision and F-score to evaluate the results:

$$Recall = \frac{TP}{TP+FN} \tag{2.1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.2}$$

$$Fscore = 2\frac{Precision.Recall}{(Precision+Recall)} \tag{2.3}$$

where TP is the number of true positives of the CUIs mentioned, FN is the number of false negatives of the CUIs mentioned, and FP is the number of false positives of the CUIs mentioned.

2.3 Results

In the first experiment (see Table 2.2), the averages for the recall, precision and F-score with MetaMap were 0.78, 0.91, and 0.82, respectively. With cTAKES, the averages for the same measures were 0.82, 0.91, and 0.84, respectively. MetaMap showed a low recall value for diabetes (0.65), CAD (0.45), hypercholesterolemia (0.59), and venous insufficiency (0.29). Cholecystectomy presents a satisfactory recall value (0.73) although much lower than the overall results. Also, cTAKES had low recall values for hypercholesterolemia (0.51), osteoarthritis (0.67), and venous insufficiency (0.29).

In the second experiment (see Table 2.3), we achieved better results. MetaMap had an average of 0.88 in recall, 0.89 in precision, and 0.88 in F-score. With cTAKES, the averages for recall, precision and F-score were 0.91, 0.89, and 0.89, respectively. That means that aggregation improves the results. For example, in the first experiment, diabetes had a recall value of 0.65 (MetaMap) and 0.83 (cTAKES), but in the second experiment, these values increased to 0.89

(MetaMap) and 0.92 (cTAKES). The same can be said about hypercholesterolemia. In the first experiment, this disease had a recall value of 0.59 (MetaMap) and 0.51 (cTAKES), but in the second experiment, these values improved to 0.88 and 0.81.

Table 2.2 Summary of first experiment

Diseases	Number of patients			Evaluation					
	Annotations	MetaMap	cTAKES	MetaMap			cTAKES		
				Recall	Precision	F-score	Recall	Precision	F-score
Hypertension	325	336	340	0.99	0.96	0.98	0.99	0.95	0.97
Diabetes *	259	186	235	**0.65**	0.91	0.76	0.83	0.91	0.87
Atherosclerotic Cardiovascular Disease (CAD) *	181	95	199	**0.45**	0.86	0.59	0.92	0.84	0.88
Congestive Heart Failure (CHF)	172	175	183	0.89	0.87	0.88	0.92	0.86	0.89
Hypercholesterolemia *	172	108	92	**0.59**	0.94	0.73	**0.51**	0.95	0.66
Obstructive Sleep Apnea (OSA)	127	105	102	0.78	0.94	0.85	0.76	0.94	0.84
Osteoarthritis (OA) *	87	76	61	0.76	0.87	0.81	**0.67**	0.95	0.78
Depression *	83	105	116	0.89	**0.70**	0.79	0.99	**0.71**	0.82
Asthma	81	83	92	0.93	0.90	0.91	1.00	0.88	0.94
Gastroesophageal Reflux Disease (GERD)	76	83	85	0.97	0.89	0.93	0.99	0.88	0.93
Gallstones/Cholecystectomy *	74	54	58	**0.73**	1.00	0.84	0.78	1.00	0.88
Gout	56	58	58	0.98	0.95	0.96	0.98	0.95	0.96
Peripheral Vascular Disease (PVD)	37	37	32	0.97	0.97	0.97	0.84	0.97	0.90
Venous Insufficiency *	21	6	6	**0.29**	1.00	0.44	**0.29**	1.00	0.44
AVERAGE				0.78	0.91	0.82	0.82	0.91	0.84
Disease with low evaluation. The lowest values for recall and precision are in bold.									

Table 2.3 Summary of second experiment

Diseases	Number of patients			Evaluation					
	Annotations	MetaMap	cTAKES	MetaMap			cTAKES		
				Recall	Precision	F-score	Recall	Precision	F-score
Hypertension	325	336	340	0.99	0.96	0.98	0.99	0.95	0.97
Diabetes *	259	254	266	**0.89**	**0.91**	**0.90**	**0.92**	0.89	**0.91**
Atherosclerotic Cardiovascular Disease (CAD) *	181	130	205	**0.60**	0.83	**0.69**	0.92	0.81	0.87
Congestive Heart Failure (CHF)	172	175	183	0.89	0.87	0.88	0.92	0.86	0.89
Hypercholesterolemia *	172	159	146	**0.88**	**0.96**	**0.92**	**0.81**	**0.96**	**0.88**
Obstructive Sleep Apnea (OSA)	127	105	102	0.78	0.94	0.85	0.76	0.94	0.84
Osteoarthritis (OA)	87	76	61	0.76	0.87	0.81	0.67	0.95	0.78
Depression *	83	109	116	**0.93**	**0.706**	**0.802**	0.99	0.71	0.82
Asthma	81	83	92	0.93	0.90	0.91	1.00	0.88	0.94
Gastroesophageal Reflux Disease (GERD)	76	83	85	0.97	0.89	0.93	0.99	0.88	0.93
Gallstones/Cholecystectomy *	74	65	68	**0.865**	0.99	**0.92**	**0.89**	0.97	**0.93**
Gout	56	58	58	0.98	0.95	0.96	0.98	0.95	0.96
Peripheral Vascular Disease (PVD)	37	37	32	0.97	0.97	0.97	0.84	0.97	0.90
Venous Insufficiency *	21	27	30	**0.905**	0.704	**0.792**	**1**	0.7	**0.824**
AVERAGE				0.88	0.89	0.88	0.91	0.89	0.89
* Diseases formed by two or more UMLS concepts. The values improved are in bold.									

CAD is a special case which illustrates the difference between both tools. For a sentence like (1) below, cTAKES recognized, among many other clues, the abbreviation “CAD”, but MetaMap did not. Consequently, the number of patients with this disease was lower in MetaMap; however, this notwithstanding, the recall increased from 0.45 to 0.6, which is a direct consequence of the aggregation of ST.

1. “Conditions, Infections, Complications, Affecting Treatment/Stay HTN, CAD, High cholesterol, OSA, OA, Depression, and Anxiety”
2. “ST depression in the inferior leads and V5-V6”
3. “was found to be in atrial flutter with a 2:1 block and 2-3 mm lateral ST depressions in V4-V6”

Depression is another interesting case. In the first experiment, it was the disease with the lowest precision in both tools, 0.70 in MetaMap, and 0.71 in cTAKES. Sentences (2) and (3) above illustrate the problem. For both sentences, MetaMap and cTAKES consider that the word “depression” refers to the disease, which is clearly not the case. In both sentences, “depression” refers to a part that is lower than the surrounding area, not to the disease. This problem increased the number of false positives. Consequently, the aggregation of ST, used in the second experiment, did not significantly increase precision. However, the aggregation of ST allowed MetaMap to increase the recall from 0.89 to 0.93.

In the first experiment, we considered the cholecystectomy procedure, but in order to know other ways to identify the presence of gallstones, we added information referring to diseases and syndromes such as cholecystolithiasis, cholecystitis, and cholelithiasis. Therefore, the second experiment increased the recall from 0.73 to 0.87 (for MetaMap), and from 0.78 to 0.87 (for cTAKES).

Venous insufficiency increased its recall from 0.29 to 0.9 (for MetaMap), and from 0.29 to 1 (for cTAKES). To improve the venous insufficiency result, we added the postthrombotic syn-

drome which corresponded to the ST pathologic function. Osteoarthritis or degenerative polyarthritis presented a low recall with cTAKES, bringing us to review the automatic extraction of the disease. In many cases, health professionals use the abbreviation OA for this disease, an abbreviation which is not recognized by cTAKES; consequently, the number of patients with this disease was low as compared to MetaMap. In some cases, MetaMap mapped this disease to a precise CUI such as C0409959 (Degenerative joint disease of knee), but in other cases, when the experts classified the disease as “OA”, MetaMap and cTAKES generalized it using the general concept “arthritis”. Since osteoarthritis is a specific type of arthritis, we decided not to proceed, in that specific case, with the aggregation of all CUIs under “arthritis”.

2.4 Discussion

Considering the results shown in Table 2.2 (first experiment), it is not surprising that previous authors chose to combine both tools to secure better results (Tang *et al.*, 2013). In this work, we avoid that combination because we intended to compare the results of each tools. The results in Table 2.3 (second experiment) show that at least two types of relationships have to be taken into account to obtain, with both tools, better results.

1. Aggregation of CUIs with the same ST (e.g., “Disease or Syndrome” and "Mental or Behavioral Dysfunction"): This form of aggregation takes into account the "isa/inverse_ isa ” relations between concepts in the Metathesaurus. This relation, allowed us to group under “diabetes mellitus”, both “insulin-dependent-diabetes” and “non-insulin-dependent-diabetes”. Similarly, “coronary arteriosclerosis” was grouped with “coronary artery disease”, “hyperlipidemia” with “hypercholesterolemia”, and “depressive disorder” with “mental depression”.

2. Aggregation of CUIs with different ST: An example here is using the Metathesaurus relation “due_ to/caused_ by” to combine venous insufficiency disease with the postthrombotic syndrome pathologic function. Also, we noted that for many forms of gallstones, the clinical notes mentioned the cholecystectomy procedure instead of the specific disease

(e.g., cholecystolithiasis). Using the relation “clinically_ associated_ with”, we were able to connect the cholecystectomy procedure with the cholelithiasis disease, and then with the cholecystolithiasis and cholecystitis diseases, among others.

Tables 2.2 and 2.3 show the results of the first and second experiments. Overall, the aggregations carried out in the second experiment increased the F-score by 7.3% for MetaMap, and by 6% for cTAKES. The recall values increased by 12.8% for MetaMap and by 11% for cTAKES, while the precision values decreased slightly in both tools, -2.2% for both MetaMap and cTAKES.

As we mentioned above, clinical notes contain many abbreviations, acronyms, and specialized terms that renders difficult the extraction of patient information. Abbreviations such as CHF and PVD were identified by both tools, but CAD and OA were not. It means that the results are sensitive to abbreviations used in the clinical notes. To resolve this problem, MetaMap allows users to define a list of abbreviations and acronyms. On the other hand, cTAKES does not have such a list (Jonnagaddala *et al.*, 2016). In this work, we did not use any list of abbreviations with the aim to keep the same configuration for both tools, but the use of this option could help MetaMap improve its results.

In the annotations made by the experts, they used general names or maybe a preferred name to denote a comorbidity. For that reason, in the second experiment, we had to look for some UMLS concept to identify one annotated comorbidity (e.g. we matched 3 UMLS diabetes mellitus concepts). In other cases, we worked with different semantic types such as pathological function and therapeutic or preventive procedures to referred to a comorbidity mentioned by the experts (e.g. venous insufficiency and gallstones).

2.5 Future Works

In future works, we will consider the automatic aggregation of concepts or CUIs using the relations between the concepts described in the Metathesaurus and the semantic relation present in the Semantic Network.

Also, while clinical notes hold information on many medical entities, some of them are in negative contexts (e.g., “The patient does not have diabetes”). In this work, we did not use algorithms like NegEx (Chapman *et al.*, 2001) that permit a recognition of entities in negative contexts. Moreover, for the extraction of medical entities, all sections were considered, including the parts such as family history, which can describe diseases that the patient does not have. Therefore, these characteristics can be taken into account to decrease the rate of false positives and improve precision.

2.6 Conclusion

In this paper, we compared the automatic extraction of 14 obesity comorbidities using MetaMap and cTAKES. Automatic extraction was compared to manual annotation by experts. The result of the experiments we conducted proved that cTAKES slightly outperforms MetaMap, but this situation could change considering other configuration options that each tool has such as the abbreviations list in the MetaMap tool. Moreover, we worked with two types of aggregations: aggregation of CUIs with the same semantic type and aggregation of CUIs with different semantic types. These groups improve the results. Hence, the use of cTAKES or even MetaMap, using the proposed aggregations, can represent a good strategy to replace the manual extraction of medical entities.

Finally, it should be noted that both tools are constantly improving the quality of their results. However, we believe that the combination of both, along with the aggregations, might even permit to cover complementary cases where both tools give different results.

CHAPTER 3

CLUSTER ANALYSIS OF OBESITY DISEASE BASED ON COMORBIDITIES EXTRACTED FROM CLINICAL NOTES

Ruth Reátegui[1, 2], Sylvie Ratté[1], Estefanía Bautista-Valarezo[3], Victor Duque[3]

[1] Département de Génie Logiciel et des Technologies de l'Information, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Departamento de Ciencias de la Computación y Electrónica, Universidad Técnica Particular de Loja (UTPL),
San Cayetano Alto, Loja, Loja, Ecuador 11-01-608
[3] Departamento de Ciencias de la Salud, Universidad Técnica Particular de Loja (UTPL),
San Cayetano Alto, Loja, Loja, Ecuador 11-01-608

Paper published in the Journal of Medical Systems, January 2019.

Abstract: Clinical notes provide a comprehensive and overall impression of the patient's health. However, the automatic extraction of information within these notes is challenging due to their narrative style. In this context, our goal was to identify clusters of patients based on fourteen comorbidities related to obesity, automatically extracted with the cTAKES tool from the i2b2 Obesity Challenge data. Furthermore, results were compared with clusters obtained from experts' annotated data. The sparse K-means algorithms were used in both experiment at two levels: at the first level, three clusters were found, and at the second, new clusters were found by applying the same algorithm to each of the clusters from the former level. The results show that three types of clusters could be identified based on the number of comorbidities and the percentage of patients suffering from them. Diabetes, hypercholesterolemia, atherosclerotic cardiovascular diseases, congestive heart failure, obstructive sleep apnea, and depression were the diseases with the highest weights contributing to the cluster distribution.

Keywords: Obesity, clinical notes, cTAKES, cluster analysis.

3.1 Introduction

Patients' information, including diseases, symptoms, treatments, drugs, etc., can be derived from clinical notes such as discharge summaries. These documents have a narrative format, which allows the health professional to write in a flexible manner. These notes contain local dialectal phrases, negations, acronyms, abbreviations, misspellings and typing errors, which all make it difficult to automatically extract patients' information from them (Shivade *et al.*, 2014; Bukhanov *et al.*, 2017).

Manual extraction of patient information is carried out by experts, and is laborious and time-consuming. Even automatic extraction is extremely difficult because the information sought is hidden within significant amounts of data residing in clinical notes (Shivade *et al.*, 2014). The process of getting structured medical information requires extracting named entities or concepts and then mapping them to codes according to controlled vocabulary or medical standards (Bukhanov *et al.*, 2017). Two standards used to map biomedical concepts are the Unified Medical Language System (UMLS) (National Library of Medicine (US), 2009) and the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) (National Library of Medicine (US), 2016). Clinical features, such as comorbidities (simultaneous presence of two diseases or conditions in a patient) related to a specific disease, are important features and are at the root of other tasks such as cluster analysis.

In the medical field, cluster analysis helps in identifying and tailoring treatment or care delivery, defining boundaries and disease taxonomies, understanding the heterogeneity of the disease, identifying subsets of patients with similar characteristics, identifying relevant pathophysiologies, etc. (Bourdin *et al.*, 2014; Chen *et al.*, 2014; Rocha & Rocha, 2014; van der Esch *et al.*, 2015). Many authors have applied cluster analysis to various conditions, such as obstructive sleep apnea (Joosten *et al.*, 2012; Vavougios *et al.*, 2016), asthma (Bourdin *et al.*, 2014; Serrano-Pariente *et al.*, 2015), knee osteoarthritis (van der Esch *et al.*, 2015), chronic heart failure (Ahmad *et al.*, 2014), and chronic obstructive pulmonary diseases (Chen *et al.*, 2014).

Overweight and obesity are a global health problem that is becoming an epidemic in both children and adults (Poirier *et al.*, 2006). Obesity is often accompanied by other health risks or comorbidities such as diabetes, dyslipidemia, hypertension, cardiovascular diseases, asthma and osteoarthritis (Figueroa & Flores, 2016; Foster *et al.*, 2008; Guh *et al.*, 2009).

Cluster analysis studies that take into account diseases related to obesity include that by Sutherland et al. (Sutherland *et al.*, 2012), which identified clusters of patients suffering from obesity and asthma simultaneously. Laing et al. (Laing *et al.*, 2015) analyzed the relationship between obesity and atherosclerosis, while LaGrotte et al. (LaGrotte *et al.*, 2016) focused on patients with obstructive sleep apnea, obesity, and excessive daytime sleepiness.

Notwithstanding the presence of multiple comorbidities in obese patients, most of the related works focus on analyzing the relationship between 2 comorbidities instead of 14. Furthermore, all the features in the above works were collected manually or from structured EHR data despite the significant amount of information inside clinical notes. The motivation for this work was therefore to apply cluster analysis to obesity comorbidities in order to gain insights into the different types of obese patients that can exist according to the number of comorbidities they have.

Based on the above explanation, the goal of our work is to identify clusters of obese patients based on obesity comorbidities extracted from clinical notes automatically. In addition, a cluster analysis based on the comorbidities annotated by experts from the same dataset was developed in order to allow a comparison with the cluster analysis result from the extracted data. The i2b2 (Informatics for Integrating Biology to the Bedside) Obesity Challenge data was used.

3.2 Materials and Methods

In this section, we will describe the dataset used and the process for expert annotation and automatic extraction. We will also explain the cluster analysis.

3.2.1 Dataset

We used the i2b2 2008 Obesity dataset. This dataset consists of 1237 discharge summaries of overweight and diabetic patients (Uzuner, 2009). The documents in the dataset contain expert annotations that classify 15 obesity comorbidities as present, absent, questionable or unmentioned (Uzuner, 2009). Table 3.1, column 1, shows the 14 comorbidities (known as diseases hereinafter) used. Hypertriglyceridemia does not have sufficient samples, and was therefore excluded. Out of 1237 summaries, 412 summaries which had obesity and at least one of the 14 diseases were selected. The last preselection was made to avoid samples with 0 in all the columns, because in the dataset obtained with the automatic extraction, there were some cases where the patients showed obesity without another comorbidity. Also, in this work, we wanted to keep the same patients that were selected in our previous work (Reátegui & Ratté, 2018b).

3.2.2 Experts' Annotation and Automatic Entity Extraction

In our previous work (Reátegui & Ratté, 2018b), the cTAKES and MetaMap tools were compared in the extraction process of 14 obesity comorbidities. Also, experts' textual annotations were used as a gold standard. The comorbidities were treated as dichotomy variables or features (values of 0 or 1, respectively depicting the non-existence or existence of the disease in discharge summaries). The results showed cTAKES slightly outperforming MetaMap. In this work, therefore, we decided to use the results obtained with cTAKES, together with the expert annotations to identify clusters of obese patients. These results also consider an aggregation process and some semantics types. Table 3.1 shows the expert annotations and the cTAKES results. The average for recall, precision and F-score are 0.91, 0.89, and 0.89, respectively. More details of the extraction and aggregation processes are given in (Reátegui & Ratté, 2018b).

Table 3.1 Diseases annotated by experts and extracted with MetaMap. Reátegui, R., Ratté, S. (2018) Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC medical informatics and decision making 18 (Suppl 3):74. doi:10.1186/s12911-018-0654-2

Diseases	**Experts' annotation**		**cTAKES Extraction**		**Evaluation**		
	NP (*)	**%(*)**	**NP**	**%**	**Recall**	**Precision**	**F-score**
Hypertension	325	79	340	83	0.99	0.95	0.97
Diabetes	259	63	266	65	0.92	0.89	0.91
CAD	181	44	205	50	0.92	0.81	0.87
CHF	172	42	183	44	0.92	0.86	0.89
HCL	172	42	146	35	0.81	0.96	0.88
OSA	127	31	102	25	0.76	0.94	0.84
OA	87	21	61	15	0.67	0.95	0.78
Depression	83	20	116	28	0.99	0.71	0.82
Asthma	81	20	92	22	1.00	0.88	0.94
GERD	76	18	85	20	0.99	0.88	0.93
CCY	74	18	68	17	0.89	0.97	0.93
Gout	56	14	58	14	0.98	0.95	0.96
PVD	37	9	32	8	0.84	0.97	0.90
VI	21	5	30	7	1	0.7	0.824
Evaluation Results Average:					**0.91**	**0.89**	**0.89**

(*) Number and percentage of patients with the disease.
CAD: atherosclerotic cardiovascular diseases; CHF: congestive heart failure;
HCL: hypercholesterolemia; OSA: obstructive sleep apnea; OA: osteoarthritis;
GERD: gastroesophageal reflux disease; CCY: cholecystectomy;
PVD: peripheral vascular disease; VI: venous insufficiency.

3.2.3 Cluster Analysis

Sparse K-means clustering developed by (Witten & Tibshirani, 2010) was chosen to conduct two experiments: a cluster analysis using the automatic extracted data and a cluster analysis using the experts' annotated data. The sparse K-means has the advantage of allowing an accurate identification of the groups and providing interpretable results following the identification of the most relevant clustering features (Witten & Tibshirani, 2010). This algorithm assigns a weight to each disease (used as features by the algorithm), with the diseases that contribute the most to a cluster having the highest values. Before using sparse K-means, we applied gap

statistics (Tibshirani *et al.*, 2001) to estimate the number of clusters. Cluster analysis was performed at two levels in both experiments: in the data extracted and in the annotated data. At the first level, sparse K-means was applied to all 412 patients, resulting in three clusters. At the second level, the same algorithm was applied to each of the three clusters of the first level. The second level gave us a total of 11 clusters. Fig. 3.1 shows their distribution and the cluster equivalence between both experiments.

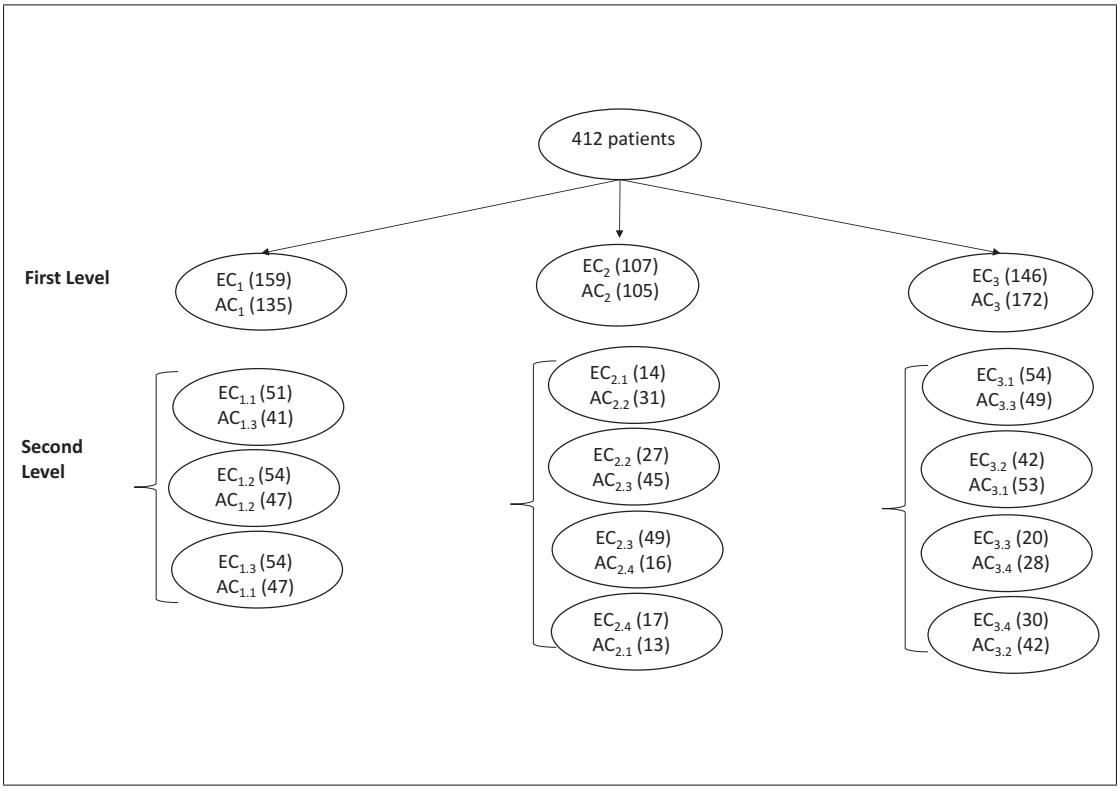


Figure 3.1 Cluster analysis by levels. The numbers in parentheses are the patients in each cluster. EC clusters are from the extracted data and AC clusters are from annotated data.

3.3 Results

In this section, we will detail the results of the cluster analysis. Fig. 3.1 shows the distribution of the clusters of both experiments, along with the correspondence between the clusters according to the highest percentage of patients suffering from a disease.

3.3.1 Cluster Analysis with Extracted Data

The first level had 3 clusters (See Table 3.2), and the diseases with the highest weights were hypercholesterolemia and diabetes. These clusters had the following characteristics: EC1 had 159 patients, all of them had diabetes, and 79% had hypertension. A moderate percentage had CAD (52%) and CHF (43%). There were no patients with hypercholesterolemia. EC2 had 107 patients, 80% had hypertension, and a moderate percentage had CHF (38%). There were no patients with diabetes and hypercholesterolemia. EC3 had 146 patients, all of them had hypercholesterolemia, 88% had hypertension, 73% had diabetes, 66% had CAD, and a moderate percentage had CHF (51%). In these big groups, it is easy identify obese patients with hypertension, diabetes and without hypercholesterolemia (EC1), obese patients with hypertension and without diabetes and hypercholesterolemia (EC2), and obese patients with hypertension, diabetes, CAD, and hypercholesterolemia (EC3). Other diseases are present in moderate and low rates. At the second level from EC1, 3 clusters were obtained: EC1.1, EC1.2, and EC1.3. The diseases with the highest weights were CAD and depression. From EC2, 4 clusters were obtained: EC2.1, EC2.2, EC2.3 and EC2.4. The diseases with the highest weights were CHF and OSA. From EC3, 4 clusters were obtained: EC3.1, EC3.2, EC3.3 and EC3.4. The diseases with the highest weights were CAD and CHF. Table 3.3 shows the results at the second level.

Table 3.2 Clusters from the first level

	CLUSTERS WITH EXTRACTED DATA					CLUSTERS WITH ANNOTATED DATA				
DISEASES	**NUM PATIENTS**	**DW**	**EC1 (HC)**	**EC2 (MC)**	**EC3 (HC)**	**NUM PATIENTS**	**DW**	**AC1 (HC)**	**AC2 (MC)**	**AC3 (HC)**
Hypertension	340	0	**79**	**80**	**88**	325	0	**73**	**72**	**87**
Diabetes	266	0.23	**100**	0	**73**	259	0.23	**100**	0	**72**
CAD	205	0	52	25	**66**	181	0	45	17	59
CHF	183	0	43	38	51	172	0	44	34	45
HCL	146	0.97	0	0	**100**	172	0.97	0	0	**100**
OSA	102	0	24	29	23	127	0	30	42	24
OA	61	0	9	20	17	87	0	11	30	23
Depression	116	0	32	24	27	83	0	28	20	14
Asthma	92	0	18	30	22	81	0	17	27	17
GERD	85	0	17	20	25	76	0	14	16	23
CCY	68	0	13	21	17	74	0	14	22	19
Gout	58	0	15	14	13	56	0	15	14	12
PVD	32	0	11	2	9	37	0	13	6	8
VI	30	0	6	7	8	21	0	3	5	7
Patients per Cluster			159	107	146			135	105	172

The cluster results are represented in percentages according to the number of patients in each group.
The Disease Weights (DW) show the weights that sparse K-means assigns to each disease.
Bold numbers are the highest values.
HC: High Comorbidity; MC: Medium Comorbidity
CAD: atherosclerotic cardiovascular diseases; CHF: congestive heart failure; HCL: hypercholesterolemia;
OSA: obstructive sleep apnea; OA: osteoarthritis; GERD: gastroesophageal reflux disease; CCY: cholecystectomy;
PVD: peripheral vascular disease; VI: venous insufficiency.

Table 3.3 Clusters from the second level with extracted data

DISEASES	**DW**	**EC1.1 (HC)**	**EC1.2 (HC)**	**EC1.3 (MC)**	**DW**	**EC2.1 (HC)**	**EC2.2 (HC)**	**EC2.3 (LC)**	**EC2.4 (HC)**	**DW**	**EC3.1 (HC)**	**EC3.2 (HC)**	**EC3.3 (HC)**	**EC3.4 (HC)**
Hypertension	0.01	**86**	**72**	**78**	0	**93**	**93**	**73**	**71**	0	**80**	**93**	**90**	**97**
Diabetes	0	**100**	**100**	**100**	0	0	0	0	0	0	**76**	**69**	**75**	**73**
CAD	0.61	55	**100**	0	0	29	37	22	12	0.46	**100**	**100**	0	0
CHF	0.05	37	59	31	0.97	**100**	**100**	0	0	0.89	**100**	0	**100**	0
HCL	0	0	0	0	0	0	0	0	0	0	**100**	**100**	**100**	**100**
OSA	0.01	31	20	20	0.23	**100**	0	0	**100**	0	26	17	30	20
OA	0.01	8	15	6	0	21	19	27	0	0	13	17	30	17
Depression	0.79	**100**	0	0	0	36	33	10	41	0	20	43	20	20
Asthma	0	18	19	17	0	43	33	27	24	0	24	21	20	20
GERD	0.02	27	11	13	0	29	15	22	12	0	26	19	35	27
CCY	0	18	9	13	0	0	41	14	24	0	17	21	10	17
Gout	0.01	22	13	11	0	29	19	10	6	0	22	5	10	10
PVD	0.01	4	19	9	0	0	4	2	0	0	11	12	5	3
VI	0	2	7	9	0	21	7	2	12	0	9	2	15	10
Patients per Cluster		51	54	54		14	27	49	17		54	42	20	30

The cluster results are represented in percentages according to the number of patients in each group.
The Disease Weights (DW) show the weights that sparse K-means assigns to each disease.
Bold numbers are the highest values.
HC: High Comorbidity; MC: Medium Comorbidity; LC: Low Comorbidity.
CAD: atherosclerotic cardiovascular diseases; CHF: congestive heart failure; HCL: hypercholesterolemia;
OSA: obstructive sleep apnea; OA: osteoarthritis; GERD: gastroesophageal reflux disease; CCY: cholecystectomy;
PVD: peripheral vascular disease; VI: venous insufficiency.

3.3.2 Cluster Analysis with Annotated Data

A cluster analysis using the experts' annotated data was carried out to compare the results obtained with the automatic extracted data. As in the above experiment, the first level had 3 clusters, AC1, AC2, and AC3. The diseases with the highest weights were hypercholesterolemia and diabetes. These new clusters compared with the clusters from the extracted data (at the first level) show a small difference in the percentage of the patients suffering from a specific disease. See Table 3.2 for more details. At the second level from AC1, 3 clusters were obtained: AC1.1, AC1.2, and AC1.3. The diseases with the highest weights were CAD and OSA. From AC2, 4 clusters were obtained: AC2.1, AC2.2, AC2.3 and AC2.4. The diseases with the highest weights were hypertension and OSA. From AC3, 4 clusters were obtained: AC3.1, AC3.2, AC3.3 and AC3.4. The diseases with the highest weights were CAD and CHF. The difference between these new clusters and those obtained with the extracted data lies in the fact that with the new ones, we have a new distribution of clusters AC1 based on OSA diseases and of clusters AC2 based on hypertension diseases. Table 3.4 shows the results at the second level.

Table 3.4 Clusters from the second level with annotated data

DISEASES	DW	AC1.1 (MC)	AC1.2 (HC)	AC1.3 (HC)	DW	AC2.1 (MC)	AC2.2 (HC)	AC2.3 (MC)	AC2.4 (LC)	DW	AC3.1 (HC)	AC3.2 (HC)	AC3.3 (HC)	AC3.4 (HC)
Hypertension	0.04	64	**70**	**88**	0.23	0	**100**	**100**	0	0	**96**	**90**	**78**	**82**
Diabetes	0	**100**	**100**	**100**	0	0	0	0	0	0	60	64	**90**	**75**
CAD	0.64	0	**100**	34	0	8	6	22	31	0.23	**100**	0	**100**	0
CHF	0.05	32	60	39	0	54	26	44	6	0.97	0	0	**100**	**100**
HCL	0	0	0	0	0	0	0	0	0	0	**100**	**100**	**100**	**100**
OSA	0.76	0	0	**100**	0.97	**100**	**100**	0	0	0	19	14	33	36
OA	0.01	6	11	17	0	54	26	24	38	0	21	26	18	32
Depression	0.02	26	21	39	0	31	35	11	6	0	11	12	14	21
Asthma	0.03	11	11	32	0	15	32	24	31	0	6	24	20	25
GERD	0.04	15	2	27	0	31	16	13	13	0	17	31	24	21
CCY	0.01	19	11	12	0	15	13	29	25	0	19	19	27	4
Gout	0	11	17	17	0	23	19	13	0	0	6	10	22	11
PVD	0.05	2	28	7	0	0	0	7	19	0	8	7	10	7
VI	0	2	2	5	0	15	6	2	0	0	2	10	6	14
Patients per Cluster		47	47	41		13	31	45	16		53	42	49	28

The cluster results are represented in percentages according to the number of patients in each group.
The Disease Weights (DW) show the weights that sparse K-means assigns to each disease.
Bold numbers are the highest values.
HC: High Comorbidity; MC: Medium Comorbidity; LC: Low Comorbidity.
CAD: atherosclerotic cardiovascular diseases; CHF: congestive heart failure; HCL: hypercholesterolemia;
OSA: obstructive sleep apnea; OA: osteoarthritis; GERD: gastroesophageal reflux disease; CCY: cholecystectomy;
PVD: peripheral vascular disease; VI: venous insufficiency.

3.3.3 Cluster Classification

Based on an observation of the clusters, and considering the number of diseases (comorbidities) and the percentage of patients with them, three types of comorbidities were identified:

- High comorbidity: Occurs when a cluster has one of the following characteristics: (1) Three or more diseases with a high percentage of patients (67% to 100%); (2) Two diseases with a high percentage of patients (67% to 100%) and one or more diseases, with 33% to 66% of the patients suffering from them.
- Medium comorbidity: Occurs when a cluster has one of the following characteristics: (1) Two diseases with a high percentage of patients (67% to 100%); (2) One disease with a high percentage of patients (67% to 100%) and one or more diseases, with 33% to 66% of patients suffering from them.
- Low comorbidity: Occurs when a cluster has one of the following characteristics: (1) One disease with a high percentage of patients (67% to 100%); (2) One or more diseases in the 33% to 66% range or between 0% to 32% .

Given the above explanations, and considering the clusters obtained with the extracted data, at the first level, EC1 and EC3 have a high comorbidity, and EC2 has a medium comorbidity. At the second level, EC1.1 and EC1.2 have a high comorbidity, and EC1.3 has a medium comorbidity; EC2.1 has a high comorbidity, EC2.2 and EC2.4 have high comorbidity, and EC2.3 has a low comorbidity, and EC3.1, EC3.2, EC3.3, and EC3.4 have a high comorbidity. Considering the clusters obtained with the annotated data, at the first level, AC1 and AC3 have a high comorbidity, and AC2 has a medium comorbidity. At the second level, AC1.2 and AC1.3 have a high comorbidity, and AC1.1 has a medium comorbidity; AC2.2 has a high comorbidity, AC2.1 and AC2.3 have medium comorbidity, and AC2.4 has a low comorbidity; AC3.1, AC3.2, AC3.3, and AC3.4 have a high comorbidity.

3.4 Discussion

To get a medical interpretation of the clusters obtained with the extracted data, two physicians were asked to express their opinion about Table 3.2 and Table 3.3. They provided the comments below. Fig. 3.2 shows the diseases with a high percentage of patients in each sub-cluster.

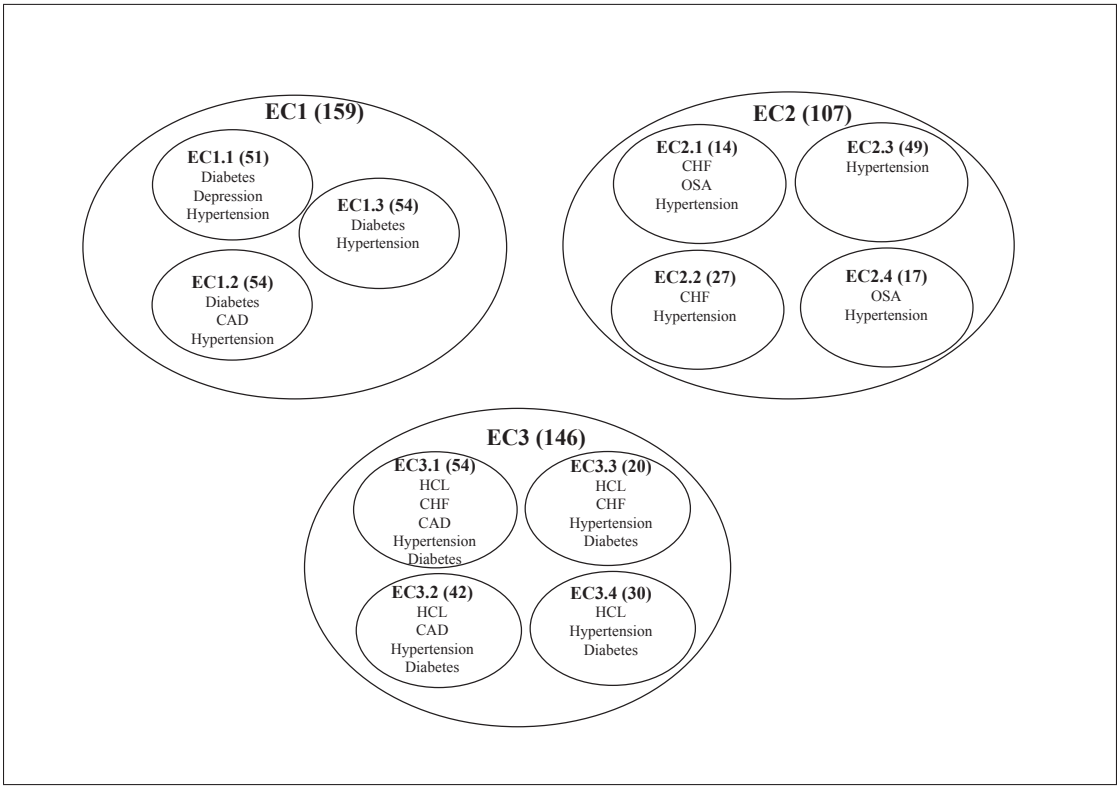


Figure 3.2 Diseases with a high percentage (67 to 100%) of patients in each sub-cluster.

Considering the percentage of patients with a comorbidity, hypertension, diabetes, CAD, CHF, HCL and OSA have the highest values. These results agree with previous works such as (Bruce *et al.*, 2011; Guh *et al.*, 2009; Poirier *et al.*, 2006) which show the common comorbidities related to obesity and overweight people.

In cluster EC1, all patients have diabetes. Also, all EC3 incorporate more than 70% of diabetic patients. Although the experiment does not detail the type of diabetes, we could mention the high association of type 2 diabetes with obesity. More than 80% of cases of type 2 diabetes can

be attributed to obesity that causes insulin resistance, and ultimately, hyperglycemia (Willett *et al.*, 1999). Although metformin (one of the drugs most commonly used in the treatment of type 2 diabetes) causes weight reduction, some drugs used in type 2 diabetes can cause moderate increases in weight (Leslie *et al.*, 2007).

In cluster EC2, all the patients do not have diabetes and HCL. This cluster has the highest percentage (29%) of patients with OSA. OSA is a disorder that occurs during sleep, in which the patient experiences repetitive episodes of apnea (stops breathing) or a reduction of airflow due to an obstruction of the upper airway. Obesity is the most potent risk factor in the development of OSA, and its relative risk increases as the body mass index (BMI) increases (Peppard *et al.*, 2013). In EC2.1 and EC2.4, all the patients have OSA with comorbidities such as hypertension and CHF. These diseases, among others, were identified as comorbidities in OSA patients in the work of (Vavougios *et al.*, 2016). The prevalence of cardiovascular risk factors such as hypertension and type 2 diabetes is substantially higher in patients with OSA. OSA is a clear risk factor for cardiovascular events, although it is not entirely clear whether it is due to its high association with obesity and other coexisting factors (Wolf *et al.*, 2007). This could explain why not all patients with OSA are present in this cluster, while 24% and 23% of OSA patients are in EC1 and EC3, which have patients with other OSA comorbidities such as diabetes and CAD. Also, EC2 has the highest percentage of hypertensive patients (80%); this can be explained by the fact that hypertension is independently associated with OSA.

In EC3, we find 100% of patients with HCL. Furthermore, this cluster has the highest percentage of patients with hypertension and CAD. The risk of CAD is higher in obese people. Most experts attribute part of this relationship to the coexistence of risk factors, although the American Heart Association considers obesity as an independent risk factor for CAD (Poirier *et al.*, 2006). In the same cluster, for example, 88% of patients have hypertension, 73% have diabetes, and 100% have HCL. These are precisely the main risk factors for CAD that usually exist in patients with obesity. Previous works (Canto *et al.*, 2011; Mamudu *et al.*, 2016) showed that diabetes, hypertension, hypercholesterolemia, etc., are related to CAD patients, and these diseases are present in our clusters as well. It is important to note that the sub-

clusters EC3.1, EC3.2, and EC1.2 include patients a high percentage of whom have three or more comorbidities. These clusters have a high percentage of patients (over 50%) with at least two of the mentioned risk factors. This reinforces the theory that the relationship between obesity and CAD is mainly due to the coexistence of many acting morbidities as risk factors for cardiovascular events.

This study also allows us to see how a pathology behaves in the different clusters. With respect to CHF and obese patients, (Ahmad *et al.*, 2014) found 4 clusters with phenotypes related to these diseases. Some important comorbidities they identified are hypertension, diabetes, hyperlipidemia, etc. Our clusters (EC2.1, EC2.3, EC3.1) show the presence of the same comorbidities as well.

In EC2.3, no patient has diabetes, CHF, HCL, OSA, and a low percentage have CAD and other diseases. This cluster give us an idea of the multifactorial and multimorbid character generally associated with obesity. We can say this the healthiest group with the lowest mortality risk diseases. In medicine, the term "metabolically healthy obese" is used to refer to obese patients who do not have cardio-metabolic abnormalities associated with adipocytes. Studies have shown that these patients have an increased risk of mortality compared to normal weight and metabolically healthy individuals (Kramer *et al.*, 2013). As well, the patients in that cluster could have lived with a risk factor for a short time period (e.g., few overweight years); they could also be patients with low BMI. One hypothesis that could be further explored is that this group seems to contain patients who have been suffering from obesity for a few years or whose BMI is not too high (overweight or low-grade obesity). The present study does not have sufficient data to test such a hypothesis.

Another case to analyze is depression. There are diverse opinions respecting the association of obesity with depression. For example, (Dixon *et al.*, 2003) suggests that depression is associated with severe obesity, especially among young women that have a poor body image. A prospective study by (Roberts *et al.*, 2003) indicates that obesity increases the risk of depres-

sion, but that the inverse is not true, that is, depression does not appear to increase the risk of gaining weight.

On the other hand, a meta-analysis of 15 prospective observational studies published in 2010 similarly shows that the risk of developing depression among obese patients and among depressed patients is the same; in other words, having either one of these pathologies predisposes a patient to the other (Luppino *et al.*, 2010).

This may account for 20% (83 patients in the annotated data) of the 412 patients considered in the present work having depression. In cluster EC1.1, all the patients have depression, while in EC2.1, EC2.2, EC2.4, and EC3.2, only a moderate percentage of them presents this disease Asthma is present among a moderate percentage of patients in EC2.1, EC2.2. The relationship between depression and asthma was addressed in a meta-analysis of 8 prospective studies published by (Gao *et al.*, 2015). This meta-analysis establishes that the risk of developing asthma increases by 43% among patients who have depression as compared to those who do not, although asthma does not appear to increase the risk of depression.

3.5 Conclusion and Future Work

Considering 14 obesity comorbidities, clustering analysis at 2 levels was applied. The first level provides a general idea of the prevalent diseases afflicting obese patients, as well as the type of comorbidity (HC and MC) they have. At the second level, groups of patients were identified, with more details provided about their comorbidities. Most of the clusters present a high comorbidity with common diseases mentioned by experts in the literature. Furthermore, despite the differences in the weights assigned to diseases in the second level, the extracted and annotated data present some equivalence in the clusters found in both experiments. This shows that the automatic extraction of medical entities and cluster analysis allow to discover groups of patients with similar characteristics. These clusters help doctors to gain insights into the variety of patient phenotyping characterizing a disease such as obesity. The present work has some limitations that should be covered in future studies. For example, 14 obesity comorbidi-

ties based on the Obesity Challenge promoted by i2b2 were used, but other diseases present in the discharge summaries could be considered. Moreover, this work did not distinguish between diabetes types, as mentioned above; knowing which patients have type 2 DM can help physicians confirm the relationship between obesity and this disease.

CHAPTER 4

A NETWORK-BASED ANALYSIS OF MEDICAL INFORMATION EXTRACTED FROM ELECTRONIC MEDICAL RECORDS

Ruth Reátegui[1, 2], Sylvie Ratté[1], Estefanía Bautista-Valarezo[3], Juan Francisco Beltrán[3]

[1] Département de Génie Logiciel et des Technologies de l'Information, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Departamento de Ciencias de la Computación y Electrónica, Universidad Técnica Particular de Loja (UTPL),
San Cayetano Alto, Loja, Loja, Ecuador 11-01-608
[3] Departamento de Ciencias de la Salud, Universidad Técnica Particular de Loja (UTPL),
San Cayetano Alto, Loja, Loja, Ecuador 11-01-608

Paper submitted to the journal International Journal of Medical Informatics, November 2018.

Abstract

Background: Clinical notes constitute a rich source of medical entities that could be useful in identifying graphs of patients with similar characteristics. Network-based approaches permit to visualize associations between medical entities and to infer medical knowledge. This paper aims to apply such an approach to identify the graphs of obese patients as well as relationships between diseases and treatments extracted from discharge summaries.

Methods: Two experiments were designed. In the first experiment, a 412-node graph representing patients was constructed to identify patient groups. Graphs were obtained with the modularity function. In the second, some bipartite graphs were constructed to identify diseases-treatments relationships from patient graphs.

Results: The results were congruent in both experiments. Patient graphs corresponding to obese patients with diseases derived from a metabolic problem were identified; some had infectious diseases, while others had diseases derived from a mechanical problem. Furthermore, groups of diseases and treatments related to obesity could be observed.

Conclusions: This work identified obesity-patient graphs and relationships between diseases and treatments based on a network approach, which took into account information extracted from clinical notes.

Keywords: Electronic medical records, network, graph, obesity.

4.1 Introduction

Obesity and overweight are abnormal or excessive fat accumulation that negatively affect health (World Health Organization, 2019). These health problems are becoming an epidemic affecting adults and children. Obesity is accompanied by comorbidities such as: hypertension, coronary heart disease, stroke, renal diseases, diabetes, gallbladder disease, respiratory problems, sleep apnea, asthma, osteoarthritis, and cancer (Aneja *et al.*, 2004; Guh *et al.*, 2009; Apovian, 2016)

Patient information from Electronic Medical Records (EMR) constitutes an important source of information to analyze health problems such as obesity and overweight. Structured information (e.g. International Classification of Diseases [ICD] codes) from EMR often do not reflect the diagnosis of a population (Shivade *et al.*, 2014; Pantalone *et al.*, 2017), but unstructured information like clinical notes save a more complete profile of patients' health. Clinical notes are a rich source of medical entities (diseases, treatments, drugs) that help classify diseases, predict patients' health, and have a better understanding of diseases, treatments, and so on (Lyalina *et al.*, 2013; Shivade *et al.*, 2014; Zhang *et al.*, 2014; Sutherland *et al.*, 2012).

Obesity is a heterogeneous disease that deserves to be analyzed considering all the possible comorbidities and health problems that a clinical note can show.

A network-based approach has been used to visualize associations between medical entities and infer medical knowledge (Bauer-Mehren *et al.*, 2013; Zhao *et al.*, 2017). For instance, (Lyalina *et al.*, 2013) visualize phenotype-phenotype association extracted from medical records related to three neuropsychiatric disorders. Also, (Roque *et al.*, 2011) visualize clusters of patients

constructed with International Classification of Diseases (ICD)10 codes extracted from medical records of psychiatric patients. Furthermore, (Gangopadhyay *et al.*, 2016) applied graphs to identify the main patterns related to anemia within clinical notes. In their work, (Khan *et al.*, 2018) extracted ICD codes from type-2 diabetes patients' admission history and created networks to identify the comorbidities and conditions related to the disease. Similarly, using ICD codes, (Kalgotra *et al.*, 2017) identified comorbidities classified by gender. Also, (Merrill *et al.*, 2015) used a network approach based on information pertaining to inpatient and outpatient clinical services to identify care patterns for congestive heart failure.

Bipartite graph is other concept used to analyze relations between biomedical information. (Goh *et al.*, 2007) used a bipartite graph to relate disorders and disease genes. The list of disorders, genes and associations between them was obtained from Online Mendelian Inheritance in Man (OMIM). Also, (Bhavnani *et al.*, 2011) used bipartite graphs to represent asthma patients and cytokines relationships. They worked with the data from a secondary analysis of cytokine profiles collected in a consortium-wide study (Brasier *et al.*, 2008). Moreover, (Zhou *et al.*, 2014) based on biomedical literature constructed a symptom-disease network to identify relationships between clinical manifestations and molecular interaction.

Therefore, different diseases have been analyzed with a network approach, but to date has not been a work that uses this approach to explore obesity disease based on clinical notes. The work of (Gangopadhyay *et al.*, 2016) is close to our idea of using a network approach with information from clinical notes, but whereas they identified patterns of diseases, our work is focused on identifying groups of patients with similar characteristics. Also, to our knowledge, there has not been work that uses bipartite graphs in order to analyze relationships between diseases and the treatment used in specific group of patients.

Based on the above observations, and given the importance of clinical notes as a source of information, two goals were defined for this work. First, we will identify graphs or groups of patients with similar characteristics based on diseases extracted from discharge summaries.

Secondly, we will examine the association between diseases and treatments considering the graphs identified earlier.

4.2 Methodology

4.2.1 Automatic Extraction and Aggregation of Medical Entities

From the i2b2 Obesity dataset (Uzuner, 2009), 412 discharge summaries were used, much like in our previous work (Reátegui & Ratté, 2018b). Also, in that previous work, we showed that MetaMap is a good strategy to extract medical entities, therefore this tool was used to automatically extract medical entities from the summaries. The concepts extracted correspond to the following Unified Medical Language System (UMLS) (National Library of Medicine (US), 2009) semantic types: disease or syndrome (dsyn), mental or behavioral dysfunction (mobd), neoplastic process (neop), therapeutic or preventive procedure (topp), pharmacologic substance (phsu), antibiotic (antb), and clinical drugs (clnd). For ease of writing, dsyn, modb and neop are going to be jointly called diseases hereafter, while topp, phsu, antb, clnd will collectively be called treatments.

An aggregation was also performed to reduce and group some diseases and treatments. Following that, we decided to eliminate the features that were present in less than 10 patients; that led to 343 features corresponding to 86 diseases and 257 treatments. The extraction and aggregation process were based on our previous works (Reátegui & Ratté, 2018b,a).

In the first experiment, we worked with the diseases as features, and in the second, we considered the patients suffering from specific diseases and some relevant treatments.

4.2.2 Graph Representation

Graphs or networks represent interactions between nodes or elements (Kalgotra *et al.*, 2017). A node represents any discrete entity (e.g., an individual or an event), and an edge indicates a relationship between nodes (Merrill *et al.*, 2015). The algorithm of Blonde et al., implemented

in Gephi, was used to calculate modularity partitions (Blondel *et al.*, 2008). The modularity with a positive value indicates the possible presence of a community structure (Blondel *et al.*, 2008; Newman, 2006); therefore, a positive value will help to evaluate the partitions found. Furthermore, bipartite graphs were constructed to identify relationships between diseases and treatments. In a bipartite network, nodes are divided into two non-overlapping sets, and the edges only join two nodes in different sets (Chang & Tang, 2014; Guimera *et al.*, 2007). In our cases, one set represents diseases and the other, treatments.

The two experiments were constructed as follows:

- First experiment: Considering the information of 412 patients and 86 diseases, we constructed an adjacency matrix with 412 nodes that represent the patients. This process was used in our previous work (Reátegui & Ratté, 2018a). The undirected edges represent the co-occurrence of diseases between two patients. Using Gephi and the modularity function some sub-graphs were obtained. These sub-graphs are groups of patients with similar diseases. Hence, we named these graphs "patient graphs". In this experiment, the modularity function was used at two levels. At the first level, the algorithm was applied over the graphs constructed from the adjacency matrix (412 nodes), and at the second, the same algorithm was used in each sub-graph obtained at the first level.

- Second experiment: To analyze and visualize the treatment for some groups of patients (specially the sub-graphs obtained from the first experiment), we constructed a bipartite graph to relate the diseases and treatments in each group. Before obtaining the bipartite graphs, some previous steps were taken: (1) An adjacency matrix was created from a matrix where the rows represented the patients to analyze, and the columns represented the features extracted. (2) The diseases and the relevant treatment to analyze were filtered. (3) Using Gephi and the Event Graph Layout plugin, the bipartite graph was created. The undirected edges represent relationships between diseases and treatments. We named these graphs "treatment graphs".

4.3 Results

4.3.1 First Experiment: Patient Graphs

The graphs or communities of Figure 4.1 were obtained. At the first level, 3 graphs (SG0, SG1, SG2) were obtained. The percentage of patients suffering from some of the 30 relevant diseases in each graph are shown in Table 4.1. At the second level, eight new graphs were found. Table 4.2 shows details of the 30 most relevant diseases in each graph.

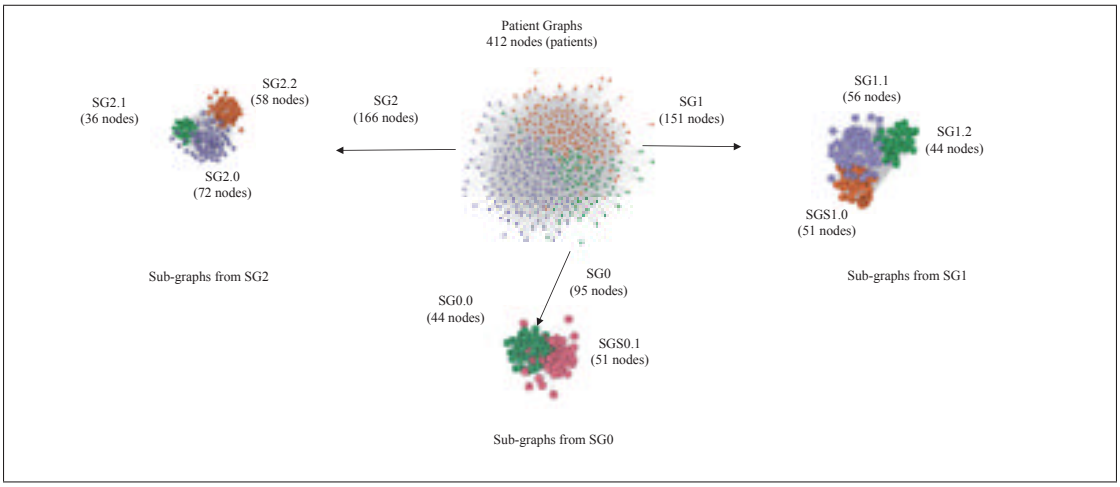


Figure 4.1 Graphs obtained in the first experiment. An the first level, 3 sub-graphs were obtained (SG0, SG1, SG2). At the second level, 8 sub-graphs were obtained: SG0.0 and SG0.1 from SG0; SG1.0, SG1.1 and SG1.2 from SG1; SG2.0, SG2.1 and SG2.2 from SG2.

4.3.2 SecondExperiment: Treatment Graphs

Table 4.3 shows details of the 30 relevant treatments in each bipartite graph analyzed in this experiment. Also, Figure 4.2 presents the bipartite graphs.

Table 4.1 Details of the 30 prevalent diseases from patient graphs in the first level

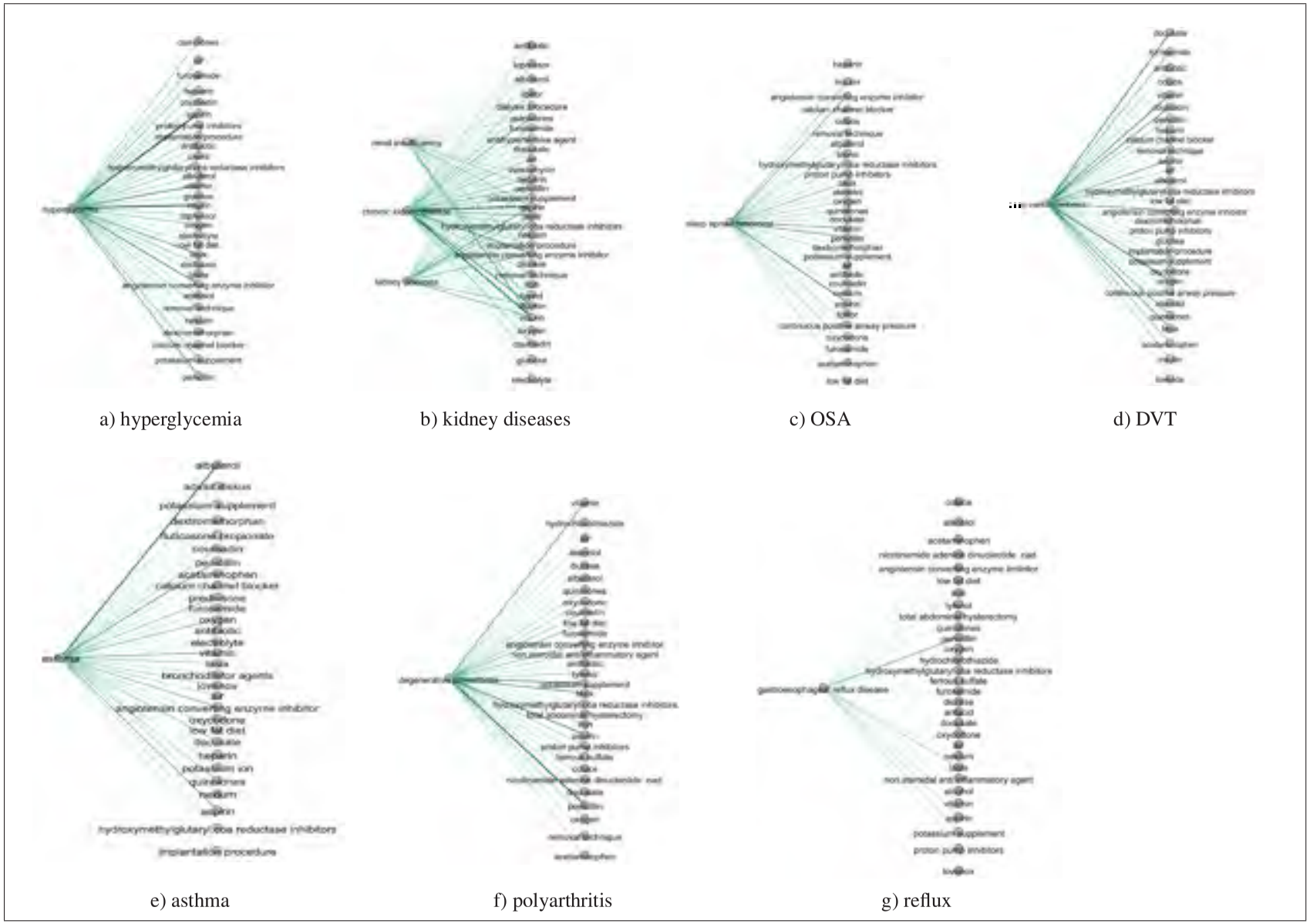
Diseases extracted automatically				**Sub-graphs**		
Num	Diseases	Total	%	SG0	SG1	SG2
1	Hypertensive disease	323	78	**69**	**81**	**81**
2	Hyperglycemia	250	61	**81**	**66**	**43**
3	Congestive heart failure	158	38	**45**	**38**	**35**
4	Hyperlipidemia	155	38	**34**	**54**	25
5	Sleep apnea syndrome	138	33	8	3	**75**
6	Heart diseases	129	31	16	**68**	7
7	Morbid obesity	102	25	15	8	**46**
8	Communicable disease	95	23	**55**	9	17
9	Cardiomyopathy	88	21	7	**47**	6
10	Deep vein thrombosis	80	19	14	12	**30**
11	Asthma	79	19	8	10	**34**
12	Fibrillation	78	19	25	22	13
13	Chronic kidney disease	78	19	**38**	17	10
14	Gastroesophageal reflux disease	78	19	6	15	**30**
15	Depressive disorder	74	18	14	11	27
16	Left ventricular hypertrophy	73	18	22	18	15
17	Anemia	70	17	**30**	10	16
18	Degenerative polyarthritis	70	17	13	11	25
19	Erythema	70	17	27	12	16
20	Chronic obstructive airway disease	68	17	11	13	22
21	Urinary tract infection	67	16	**38**	7	12
22	Lung diseases	59	14	14	9	19
23	Cancer	59	14	17	15	12
24	Gout	58	14	15	8	19
25	Vascular disease	57	14	14	21	8
26	Kidney diseases	53	13	**30**	11	4
27	Renal insufficiency	49	12	22	7	11
28	Anxiety	47	11	16	7	13
29	Pneumonia	45	11	17	11	8
30	Cerebrovascular accident	44	11	11	15	6
			Patients per sub-graph	95	151	166

The results of the sub-graphs are represented in percentages according to the number of patients in each group. Bold numbers represent diseases present in more than 30% of the population in each sub-graph.

Table 4.2 Details of the 30 prevalent diseases from the patient graphs in the second level

Diseases	**Sub-graphs from SG0**		**Sub-graphs from SG1**			**Sub-graphs from SG2**		
	SG0.0	SG0.1	SG1.0	SG1.1	SG1.2	SG2.0	SG2.1	SG2.2
Hypertensive disease	**84**	**57**	**80**	**91**	**70**	**78**	**72**	**30**
Hyperglycemia	**75**	**88**	**65**	**73**	**59**	**36**	8	25
Congestive heart failure	**73**	22	**98**	7	7	10	**50**	19
Hyperlipidemia	**45**	25	**53**	**64**	**41**	13	19	15
Sleep apnea syndrome	9	8	4	2	5	**76**	**56**	29
Heart diseases	18	14	**63**	**64**	**77**	7	3	3
Morbid obesity	18	12	10	2	14	**74**	**58**	1
Communicable disease	**34**	**75**	6	9	14	21	8	6
Cardiomyopathy	7	8	**55**	0	**98**	6	3	3
Deep vein thrombosis	11	16	24	4	9	**39**	19	8
Asthma	7	10	8	11	11	**33**	22	14
Fibrillation	**30**	22	**47**	11	7	11	28	2
Chronic kidney disease	**48**	29	**33**	5	14	13	6	3
Gastroesophageal reflux disease	11	2	14	18	14	10	**42**	16
Depressive disorder	11	16	10	7	16	24	22	12
Left ventricular hypertrophy	**41**	6	20	21	11	11	8	8
Anemia	**45**	18	10	11	9	10	19	7
Degenerative polyarthritis	16	10	10	14	9	6	**92**	2
Erythema	2	**49**	6	14	16	28	8	2
Chronic obstructive airway disease	9	14	8	13	20	14	17	12
Urinary tract infection	**57**	22	4	7	11	17	8	3
Lung diseases	9	18	12	2	16	21	25	5
Cancer	11	22	18	11	18	19	8	2
Gout	23	8	10	5	9	7	19	12
Vascular disease	11	16	25	16	20	11	0	3
Kidney diseases	**45**	18	20	5	9	4	0	2
Renal insufficiency	**32**	14	14	4	2	8	3	6
Anxiety	18	14	8	4	9	8	17	6
Pneumonia	2	29	10	7	16	4	3	5
Cerebrovascular accident	16	8	22	9	16	6	6	2
Patients per sub-graph	44	51	51	56	44	72	36	58

The results of the sub-graphs are represented in percentages according to the number of patients in each group. Bold numbers represent diseases present in more than 29% of the population in each sub-graph.



a) hyperglycemia b) kidney diseases c) OSA d) DVT

e) asthma f) polyarthritis g) reflux

Figure 4.2 Bipartite graphs obtained in the second experiment

Table 4.3 Details of the 30 relevant treatments in the bipartite graphs from the second experiment

Hyperglycemia	**Renal diseases**	**OSA**	**Asthma**	**Deep vein thrombosis**	**Degenerative polyarthritis**	**Gastroesophageal reflux**
Aspirin	Aspirin	Aspirin	Albuterol	Oxygen	Penicillin	Penicillin
Insulin	Insulin	Penicillin	Penicillin	Vitamin	Aspirin	Lasix
Angiotensin	Vitamin	Angiotensin	Vitamin	Docusate	Potassium supplement	Aspirin
Vitamin	Angiotensin	Vitamin	Aspirin	Lasix	Vitamin	Potassium supplement
Lasix	Lasix	Lasix	Oxygen	Coumadin	Lasix	Vitamin
Penicillin	Penicillin	Insulin	Angiotensin	Aspirin	Angiotensin	Iron
Furosemide	Quinolones	Albuterol	Docusate	Angiotensin	Low fat diet	Hydrochlorothiazide
Potassium supplement	Heparin	Oxygen	Prednisone	Quinolones	Coumadin	Total abdominal hysterectomy
Quinolones	Air	Potassium supplement	Lasix	Albuterol	Oxygen	Proton pump inhibitors
Oxygen	Coumadin	Furosemide	Quinolones	Furosemide	Proton pump inhibitors	Acetaminophen
HMG-CoA Reductase	HMG-CoA Reductase	Low fat diet	Lovenox	Penicillin	Oxycodone	NSAIDs
Docusate	Removal technique	Continuous positive airway pressure	Coumadin	Acetaminophen	Quinolones	Ferrous.sulfate
Removal technique	Potassium supplement	Quinolones	Air	Air	Tylenol	Furosemide
Air	Toprol	Proton pump inhibitors	Low fat diet	Insulin	Docusate	Nexium
Lopressor	Implantation procedure	HMG-CoA Reductase	Fluticasone propionate	Antibiotic	Acetaminophen	Alcohol
Implantation procedure	Lipitor	Docusate	Bronchodilator agents	Heparin	NSAIDs	Angiotensin
Heparin	Dialysis procedure	Atenolol	Furosemide	Low fat diet	Ferrous.sulfate	Low fat diet
Toprol	Docusate	Removal technique	Acetaminophen	Oxycodone	Furosemide	Oxygen
Coumadin	Furosemide	Calcium channel blocker	Oxycodone	Potassium supplement	Air	Oxycodone
Antibiotic	Diurese	Dextromethorphan	Potassium supplement	HMG-CoA Reductase	Colace	Air
Albuterol	Glucose	Coumadin	HMG-CoA Reductase	Implantation procedure	Iron	Colace
Glucose	Antibiotic	Heparin	Implantation procedure	Atenolol	NAD	NAD
Lipitor	Lopressor	Air	Calcium channel blocker	Removal technique	Hydrochlorothiazide	Quinolones
Low fat diet	Oxygen	Nexium	Dextromethorphan	Colace	Albuterol	Tylenol
Nexium	Nexium	Acetaminophen	Nexium	Continuous positive airway pressure	Antibiotic	Docusate
Atenolol	Vancomycin	Antibiotic	Electrolyte	Lovenox	HMG-CoA Reductase	HMG-CoA Reductase
Dextromethorphan	Antihypertensive	Lipitor	Potassium ion	Calcium channel blocker	Removal technique	Diurese
Calcium channel blocker	Iron	Oxycodone	Advair diskus	Proton pump inhibitors	Diurese	Atenolol
Electrolyte	Albuterol	Colace	Antibiotic	Glucose	Total abdominal hysterectomy	Lovenox
Proton pump inhibitors	Electrolyte	Toprol	Heparin	Dextromethorphan	Atenolol	Antacid

HMG-CoA Reductase = Hydroxymethylglutaryl.coa reductase inhibitors
NSAIDs = Non.steroidal anti.inflammatory agent
NAD = Nicotinamide adenine dinucleotide

4.4 Discussion

Using a network-based approach, we found patient groups and some relationships between diseases and treatments confirming the comorbidities affecting the obese patients.

Next, we will present the analysis of our experiments.

4.4.1 First Experiment: Patient Graphs

SG0 has 95 patients. This graph is composed mostly of non-morbidly obese people who have developed hypertension (69%) as a direct complication. The main characteristics of this graph is that they are obese with metabolic problems such as hyperglycemia (81%). Possibly, they are people with insulin resistance that is manifested before the onset of hyperglycemia (Beck-Nielsen & Groop, 1994). Also, several studies show obesity as a metabolic risk for type 2 diabetes, observing that over 80% of such cases can be attributed to obesity (Mokdad *et al.*, 2003).

SG0 has a significant percentage of patients with kidney pathologies. As an example, SG0.0 has 48% of the patients with chronic kidney diseases, 45% with kidney diseases, and 32% with renal diseases. The Hypertension Detection and Follow-Up Program and the Multiphasic Health Testing Services Program suggest that obesity can be an independent risk factor for developing chronic renal failure, but on the other hand, the Framingham Offspring Study states that obesity is not an independent risk factor for developing CRF, and that it must be associated with diseases such as diabetes or hypertension, tobacco use, and so on (Kramer *et al.*, 2005; Foster *et al.*, 2008).

Hyperglycemia is directly related with obesity and other pathologies of infectious origin (associated with social relationships and dangerous, risky or disordered lifestyles). This may explain the high frequency of signs such as erythema among 49% of the patients in SG0.1, and other less obvious signs such as anemia, present in 45% in SG0.0. Obese patients, independently of the coexistence of other pathologies, are more susceptible to developing infectious

processes (Falagas & Kompoti, 2006). In obesity conditions, physio-pathological phenomena affect the immunological processes. This impacts the stability of the organic protection mechanisms against infectious aggressions. The relationship between obesity and an increased frequency of infectious processes is clearly described in the medical literature. This is not a cause-consequence relationship, but rather, a factor that modifies the capacity of individuals with body weight disorders to respond to infection attacks.

SG1 has 151 patients. This graph is also composed of non-morbidly obese people. 81% of them have hypertension; 66%, hyperglycemia, and 54%, hyperlipidemia. These patients have developed cardiovascular complications. For example, 98% of the patients in SG1.0 have congestive heart failure, 63% have heart disease and 55% have cardiomyopathy. Also, these patients have had cerebrovascular accidents.

A "metabolic" mechanism is a common factor of the diseases in SG1, which can be derived from overweight and obesity. The association of deleterious changes in the lipid metabolism of obese patients is manifested through high concentrations of cholesterol, low-density lipoprotein cholesterol (LDL), very low-density lipoprotein cholesterol (VLDL), triglycerides and a reduction of high-density lipoprotein cholesterol (HDL). The reduction of HDL has been observed to present a higher relative risk of developing coronary heart disease (The Emerging Risk Factors, 2011).

Several studies show a linear relationship between obesity and the incidence of coronary heart disease, in addition to a gradual positive relationship between the body mass index (BMI) and risk factors for coronary heart disease, such as dyslipidemia, hypertension and diabetes (Poirier *et al.*, 2006).

SG1 describes the association between obesity and metabolic diseases, which in this case, is explained as a direct impact of increased body weight on metabolic phenomena. Patients in SG1 may have the worst prognosis because they present a high percentage of heart and cerebrovascular disease (Daniels, 2012).

SG2 has 166 patients. This graph comprised a significant percentage of patients with morbid obesity (46%), obstructive sleep apnea (75%), gastroesophageal reflux disease (30%), asthma (34%), and poly-arthritis (25%). Looking for a common pathophysiological denominator, SG2 could be explained from a "mechanical" point of view. The mechanical effect of increased body mass (overweight and obese) directly affects the development of disorders such as obstructive sleep apnea syndrome (OSAS) and gastrointestinal reflux (GERD).

Obesity is a risk factor for gastrointestinal (GI) disease, which includes GERD, erosive esophagitis, esophageal adenocarcinoma and gastric cancer (body fatness). For example, SG2.1 has 42% of the patients with GERD. A “mechanical” process (amplified intragastric pressure), increases the risk of GERD (Cook *et al.*, 2008). GERD is in the differential diagnosis of OSAS because it can present similar symptoms due to the irritation of the upper airway by the arrival of stomach acids. As an aside, GERD would seem to benefit from nocturnal oxygen therapy (OSAS treatment). In addition to the association by similarity of symptoms, an independent association between both pathologies has been found (Gilani *et al.*, 2016). We can see this in SG2.1, where OSA is present in 56% of the patients and GERD in 42%. Also, GERD is precursor to the appearance of asthma and other related problems, such as chronic obstructive pulmonary disease (COPD) (Cook *et al.*, 2008; Peppard *et al.*, 2000).

OSAS is strongly associated with the existence of obesity (e.g., SG2.0, SG2.1). Obesity acts as a more powerful risk factor in the development of OSAS, and its relative risk is higher when the BMI increases (Peppard *et al.*, 2013). OSAS is a clear risk factor for cardiovascular events (e.g., SG2.1), although it is not entirely clear whether it is due to its high association with obesity or other coexisting factors (Wolf *et al.*, 2007).

Moreover, asthma is present in these patients (e.g., SG2.0). Several studies have identified a higher prevalence of asthma among obese individuals than in those with normal BMI (Cook *et al.*, 2008). Asthma and OSA usually coexist in patients by having shared desiccating factors such as nasal irritation and obesity (Devouassoux *et al.*, 2007). In SG2.0, 76% of the patients have OSA and 33% have asthma.

Another significant problem present in SG2 is deep vein thrombosis, with 39% of the patients in SG2.0 having this disease. In this case, its development would be explained by mechanical effects, derived from overweight, which affect the venous return (Delluc *et al.*, 2009).

Depression is another pathology representative of this graph (e.g., SG2.0). It has been observed that depression is more associated with severe obesity (Dixon *et al.*, 2003).

Degenerative poly-arthritis or osteoarthritis (OA) is a disease present in SG2.1, with a big percentage of patients (92%) having it. Osteoarthritis is a disease characterized by the progressive degeneration of articular joints. Obesity and overweight are primary risk factors for OA, and mechanical factors increase the risk its progression (Francisco *et al.*, 2018). Because of excess loads on any weight-bearing joint, the surface of the cartilage suffers tear and wear (Berenbaum *et al.*, 2013).

Finally, it should be stated that hypertension is a pathology with high percentages in all subgraphs. Hypertension cannot be classified in a specific group due to the strong relationship between obesity and hypertension and the pathophysiological mechanisms with which this disease occurs. For this reason, hypertension in patients with high BMI is treated through weight loss (Aneja *et al.*, 2004; Daniels, 2012).

4.4.2 Second Experiment: Treatments Graphs

Is should be recalled that our dataset is characterized by patients suffering from obesity, with a high presence of hypertension and hyperglycemia. Therefore, the “treatments graphs”, certain drugs such as aspirin, angiotensin-converting enzyme inhibitors, insulin and laxative drugs are present, along with a low-fat diet.

In addition, as we can see in the graphs obtained in the first experiment, obese patients have certain comorbidities, and as a result, the treatment for a specific disease is accompanied with treatments for other diseases. This notwithstanding, and depending on the group of patients,

some treatments appear with more frequency for specific diseases present in the graph. Next, we will analyze some cases.

Hyperglycemia diseases: This bipartite graph was created considering the information of all patients (250) that have hyperglycemia. The graph shows treatments such as insulin and glucose. Treatment with insulin is used in different types of diabetes mellitus. All patients with type 1 diabetes need insulin, while for those with type-2 diabetes, insulin is only required when hyperglycemia persists against oral agents, or when the patient has a severe metabolic disturbance such as plasma glucose greater than 250 mg/dl or HbA1C > 9.5. Among the complications of the use of insulin is an increase in weight and the hypoglycemia. The presence of hypoglycemia justifies the use of glucose in this group of patients. Hypoglycemia is found to be more prevalent in type 1 diabetics than in type 2 (UK Hypoglycaemia Study Group, 2007). This bipartite graph also comprises a variety of treatments for infection, lipid reduction, anticoagulants, reflux treatments, and so on, which shows the diversity of diseases present in the patients.

Renal diseases: This bipartite graph was created taking into account the patients (126) having at least one of these diseases: chronic kidney disease, kidney disease, and renal insufficiency. Dialysis, Toprol, furosemide, and diurese are the treatments shown on this graph. As well, the graph also presents treatments for infection, anemia, hypertension, and diabetes. As we can see in SG0.0, these diseases appear together with renal diseases.

OSA diseases: This bipartite graph was created considering the information of all patients who have OSA. 138 patients suffer from this disease. The graph shows treatments for OSA, such as continuous positive airway pressure, oxygen, and air. It also shows a variety of treatments for diseases such as gastroesophageal reflux disease, asthma, and deep vein thrombosis. These diseases are pathologies with physio-pathological phenomena, and with a mechanical function resulting from a high body weight. In this sense, the evidence is clear that obesity triggers alterations in the function of multiple organs and systems. Therefore, it is at the root of several diseases which further complicate the health of obese patients.

The pharmacological therapy used in this group strengthens the hypothesis that this group is represented by obese patients with mechanical respiratory and gastrointestinal problems. Hence, in this graph, we have drugs such as bronchodilators, antitussives, and antihistamines, which are specifically for pathologies such as asthma. We also have drugs such as proton-pumps inhibitors and antacids, which are used to treat reflux.

Asthma and deep vein thrombosis: SG2.0, with 72 patients, is an OSA graph with a moderate percentage of patients with asthma and deep vein thrombosis. The bipartite graph for asthma in SG2.0 shows albuterol, advair diskus, and antihistamine drugs that are specifically for diseases such as asthma. For deep vein thrombosis, the bipartite graph shows heparin and Coumadin, which are drug used to treat this disease.

Degenerative poly-arthritis and GERD: SG2.1, with 36 patients, is also an OSA graph with a high percentage of patients with degenerative poly-arthritis (92%). The bipartite graph that represents the relationship between degenerative poly-arthritis and treatments shows oxycodone, Tylenol, acetaminophen and a nonsteroidal anti-inflammatory agent, all of which are used to relieve pain. SG2.1 has 42% of patients with gastroesophageal reflux diseases. Therefore, the bipartite graph that represents the relationship between GERD and the treatments shows inhibitors of proton pumps, Nexium, and antacids, which are used to treat reflux.

Cardiomyopathy: This bipartite graph was created considering all the patients (88) with cardiomyopathy. Some of the treatments related to this disease shown in the graph include: furosemide, Toprol, heparin, Coumadin, nitroglycerin, atenolol and calcium channel blockers. SG1.0 and SG1.2. are examples of graphs with high percentages of patients with this disease.

Heart disease: SG1.1 does not have patients with cardiomyopathy, but it has 64% of patients with heart disease. We created a bipartite graph for this disease, taking into account the 56 patients in SG1.1. The bipartite graph shows drugs such as statins and aspirin, used as treatment and to prevent cardiac complications. Also, we can see medications for cardiovascular complications such as beta blockers and antiarrhythmic, as well as procedures such as cardiac bypass.

The SG1 graph has a high percentage of patients with hyperlipidemia and hyperglycemia, which is why the bipartite graph for cardiomyopathy and heart disease also shows treatments that control these diseases with a metabolic origin.

4.5 Conclusion

Based on of 86 diseases and 257 treatments extracted from discharge summaries of 412 obese patients, and using a network-based approach, we could analyze and visualize information related with obesity diseases. We identify three large groups or graphs that could contribute to a new classification of obesity disease. The classification is as follows: 1. Obese patients with infections problems (SG0), 2. Obese patients with metabolic problems (SG1), and 3. Obese patients with a mechanical problem (SG2). Also, it was possible to visualize associations between diseases and treatments.

This work has some limitations. We did not consider information such as BMI, gender, race or other features, which could help refine the sub-graph. Also, adding more information to our extracted features could help us identify possible tailored treatments for each graph found in the first experiment. In biomedicine, current works focus on stratified medicine, with the aim of finding the best therapy for a patient graph, and the future will see personalized medicine aimed at ensuring decisions, practices and therapies tailored to individual patients (Holzinger, 2016).

In future work, we would like to experiment with the introduction of new features as aforementioned. Also, the time when diseases and treatments appear could offer insights regarding the progress of a patient's health.

CHAPTER 5

GENERAL DISCUSSION

This thesis addresses two main tasks: medical entity extraction and medical entity analysis.

Regarding the medical entity extraction or recognition task, MetaMap and cTAKES were used to extract 14 obesity comorbidities from discharge summaries. A comparison of both tools was made, and showed cTAKES slightly outperforming MetaMap; however, this could change since both tools constantly release new versions and configuration options. Such changes could lead to new results, even in the same datasets.

It is important to consider that health professionals usually write an entity name in different ways (e.g., depressive disorder, mental depression) or they may describe a medical procedure instead of a disease (e.g., cholecystectomy instead of cholecystolithiasis). The different ways of writing or describing medical entities also means that terminology such as UMLS could recognize the same entity with different codes belonging to the same or different semantic types. Therefore, an aggregation process considering semantic types and relationships defined in the UMLS could improve the results.

Moreover, methods such as rules-based, machine learning and deep learning have performed well in medical entity extraction, but they require experts, annotated datasets, and a considerable amount of information. Therefore, the methods are time consuming and expensive, which then makes using MetaMap and cTAKES a good medical entity extraction strategy.

Figure 5.1 presents a general view of the medical entity extraction and aggregation process. The steps are detailed below:

- From discharge summaries, extract diseases (medical entities) using Metamap or cTAKES.
- Create a matrix where a row represents a patient, and the columns are the diseases patient has.

- Based on UMLS (terminology, thesaurus) and the entities extracted, identify semantic types and concepts relationships.
- Aggregate de diseases or medical entity considering the previous step.

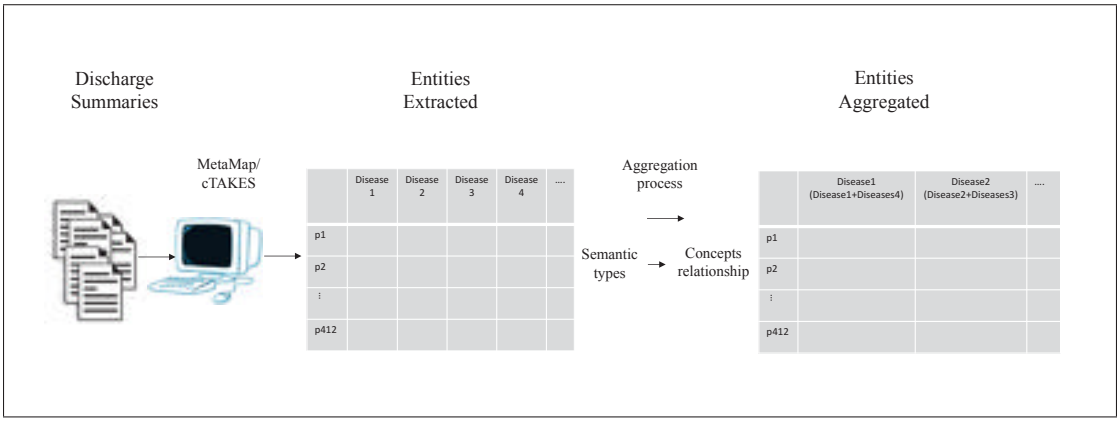


Figure 5.1 Process to extract and aggregate medical entities

Two approaches to patients cluster analysis were also described. In the first approach, a cluster analysis with a sparse K-mean algorithm was applied at two levels. This approach helps clearly identify the main features of each cluster. Works such as (Sutherland *et al.*, 2012; Laing *et al.*, 2015; LaGrotte *et al.*, 2016) consider two or three comorbidities in identifying obesity clusters. In contrast, this thesis analyzed 14 obesity comorbidities. The i2b2 obesity dataset contains 15 comorbidities annotated by experts; 14 of them (hyperglycemia was excluded) were used as features in the cluster analysis. Hence, to the best of our knowledge, this is the first attempt at identifying clusters of patients based on comorbidities as features. Moreover, the 14 comorbidities were automatically extracted from discharge summaries, while other studies obtain features mainly from patient measurements or through interviews or questionnaires.

Figure 5.2 shows the patients clusters analysis process using sparse K-means. The steps are detailed below:

- From discharge summaries, extract medical entities using a NER tool such as cTAKES or Metamap.

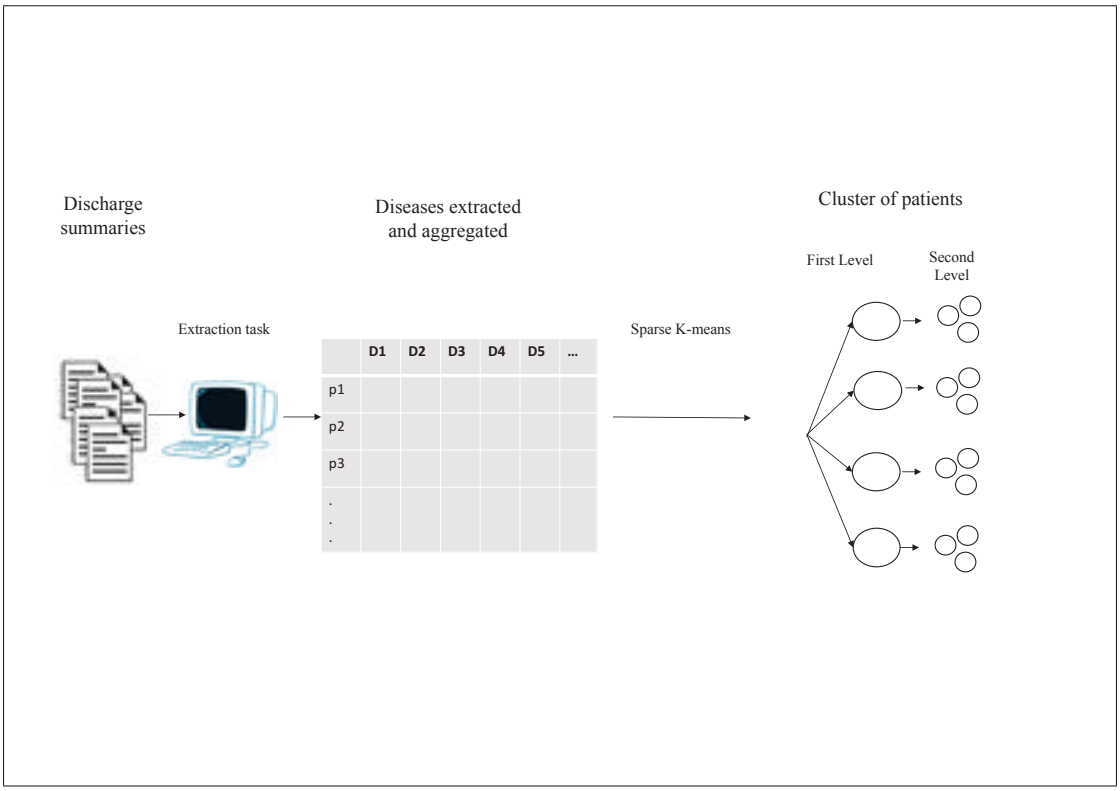


Figure 5.2 Patients cluster analysis process with sparse K-means

- Create a matrix where a row represents a patient, and the columns are the medical entities extracted and aggregated.
- Use a two-level sparse K-means algorithm. At the first level, the algorithm is applied to the number of patients to be analyzed. At the second, the algorithm is applied to each cluster identified in the first level.

After analyzing the 14 comorbidities mentioned above, and with the aim of exploring more information from discharge summaries, 86 diseases and 257 treatments were extracted. A network-based approach was considered to visualize and explore the medical entities. The modularity function was applied at two levels to identify patient communities or clusters. The results show three main clusters of patients representing obese patients with infection problems, obese patients with diseases derived from a mechanical problem, and obese patients with

diseases derived from a metabolic problem. The clusters found provide new insights into a new obese patient classification based on the diseases afflicting patients.

In addition, the results show that information inside unstructured medical notes save more details about a patient's health. This information enables a better understanding of a disease. To date, the works of (Roque *et al.*, 2011; Gangopadhyay *et al.*, 2016) come closest to our idea of identifying patient clusters and disease-treatment relationships using unstructured data from EHR with a network-based approach. On the one hand, (Roque *et al.*, 2011), worked with information on patients with mental problems and used a similarity cosine to identify patients clusters; for this thesis, I worked with obese patients and a modularity function to identify patient clusters or communities. Unlike the work of (Gangopadhyay *et al.*, 2016) which used a modularity function to extract patterns of a disease from EHR, this thesis worked with patient clusters.

The network approach plays an important role in information visualization. Therefore, after identifying patient clusters and disease-treatment relationships, some clusters were visualized and analyzed. Although most of the relationships were from a specific disease or group of related diseases (e.g., renal insufficiency, chronic kidney disease, kidney diseases) to some treatments, the bipartite graphs shows treatments related with other diseases. This confirms that some comorbidities affect obese patients.

Figure 5.3 shows the patient clusters and disease-treatment relationship identification process using a network-based approach. The steps are detailed below:

- From discharge summaries, extract medical entities using a NER tool such as cTAKES or Metamap.
- Create a matrix where a row represents a patient, and the columns are the medical entities extracted and aggregated.

- Use a two-level modularity function. At the first level, the function is applied to the number of patients to be analyzed. At the second level, the function is applied to each cluster identified in the first level.

- Based on the patients treatments extracted and the clusters found in the former step, use bipartite graphs to identified diseases-treatments relationships.

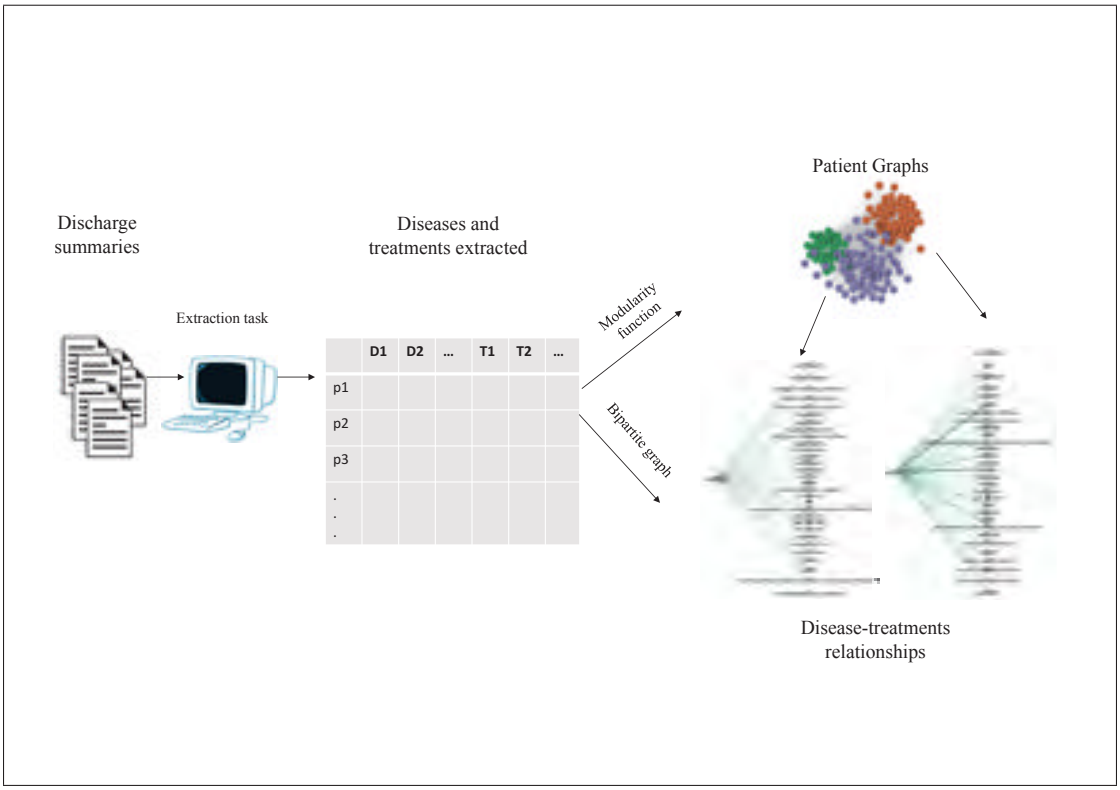


Figure 5.3 Network-based approach to medical entities exploration

In general, this thesis addresses a new option for characterizing obese patients based on co-morbidities and other diseases suffered by such patients. Furthermore, this work describes a new methodology for analyzing medical information from clinical notes that could be applied to explore new diseases. The main steps (see Figure 5.4) developed are:

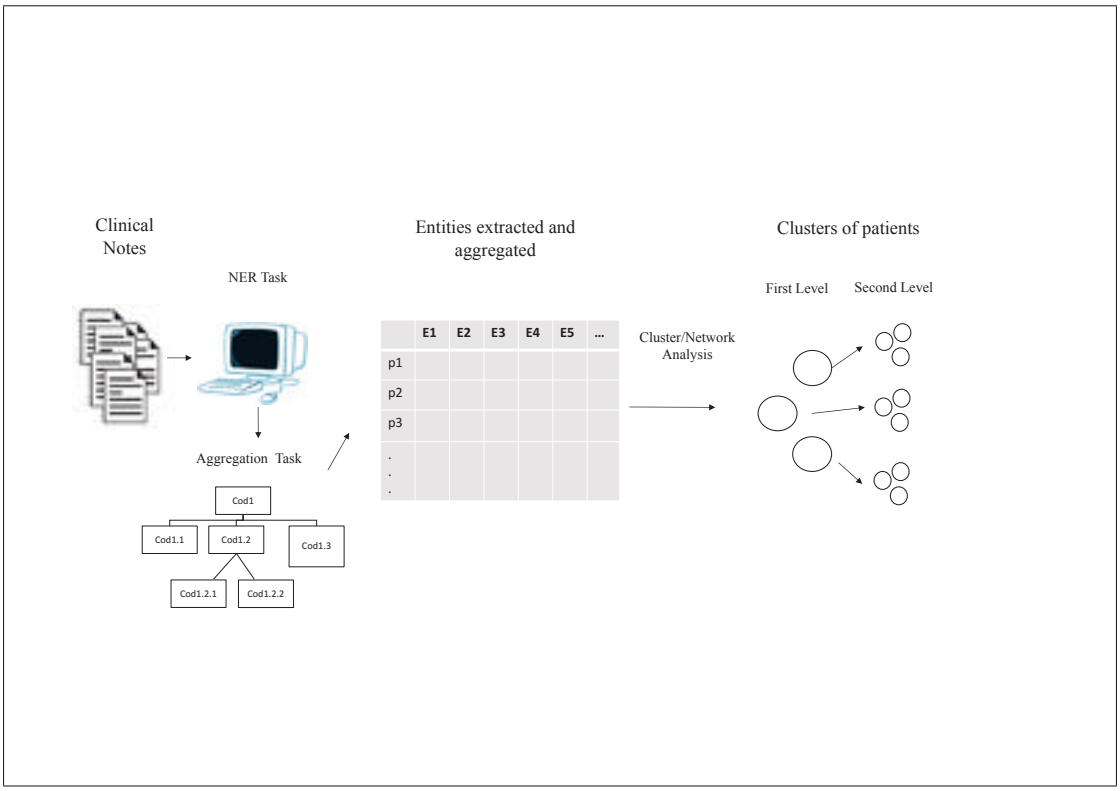


Figure 5.4 General methodology to identified clusters of patients at two levels

- Automatically extract medical entities from clinical notes. Today, we have tools that could be used for this task. A combination of some tools such as MetaMap and cTAKES could constitute a good strategy.

- Aggregate medical entities considering different semantic types and relationships. Specialized vocabularies and standards such as UMLS and SNOMED-CT feature a medical entity classification that enables an aggregation process.

- Create a matrix where a row represents a patient, and the columns are the medical entities extracted and aggregated.

- Identify patient clusters, considering cluster analysis or network approach. Considerate a two-level analysis. At the first level, where the method is applied to the entire data (patients), the clusters show characteristics that provide a general understanding of the diseases

affecting patients in that groups. At the second level, where the method is applied to clusters from the first level, the results show more detailed information about the patient groups.

CONCLUSION AND RECOMMENDATIONS

The main objective of this thesis was to analyze clinical notes in order to extract hidden data and information related to obese patients. This objective was achieved through three main contributions. First, we extracted medical entities from existing tools. Second, clusters of patients were identified using a cluster analysis based on 14 obesity comorbidities. Third, a network-based approach was used to visualize and explore relationships between diseases and relationships between diseases and treatments.

14 obesity comorbidities were extracted with MetaMap and cTAKES tools. This extraction allowed a comparison of these tools and some remarks about the medical entity aggregation process. In the dataset selected, the result showed that cTAKES slightly outperforms MetaMap. The bottom line is that either of these tools could be selected for the extraction task. However, the aggregation process plays an important role in improving results, and terminologies such as UMLS could be used for this process.

Two-level sparse k-means algorithms were used to find clusters of patients based on 14 obesity comorbidities. At the first level, three types of clusters (low, medium and high comorbidity) were identified based on the number of comorbidities and the percentage of patients suffering from them. At the second level, clusters were identified, with more details provided about their comorbidities. Using the sparse K-means enables identification of features, or in these cases, comorbidities that are more representative (based on the weights) for each cluster.

To visualize and analyze the association between medical entities extracted from clinical notes, a network-based approach was applied. Based on 86 diseases, three main obesity clusters were identified. Although the process of extracting medical entities did not require the intervention of experts, physicians have a crucial role in the validation of the clusters and in naming the clusters: patients with metabolic problems, patients with infections problems, and patients with a mechanical problem. Furthermore, disease-treatment associations were found in some

of the clusters. Compared to the experiment in chapter 3, a network-based approach applied on clinical note information provides a way to visualize entity relationships and infer new insights about comorbidities and treatments for specific diseases.

Through the above-mentioned contributions, this thesis shows that clinical notes provide health professionals and researchers with new insights about diseases.

Future works

This thesis has some limitations that should be covered in future studies.

A comparison of MetaMap and cTAKES was done, but a combination of both tools could represent a good strategy for NER in the medical domain. Also, comparisons using other datasets and new versions of the tools could help discover new insights about the tools themselves and their configuration options applied to new datasets. Furthermore, the use of other methods like rule-based approach could be used to compare and validate the result.

A technical aspect to take into account is that, for now, there is no tool or application that can automatically aggregate extracted entities. UMLS has a list of semantic types and relation that could be used to automatically aggregate concepts that correspond with the same or different semantic type.

All experiments in this thesis did not consider the the temporal aspect of the problem, e.g. when the diseases or treatment appears. For future works, the inclusion of time could help understand the influence of a comorbidity or treatment in the progress of a disease. Also, the inclusion of a different medical entity (e.g., laboratory test), features such as sex, age, geographical distribution, corporal measurements (e.g., height, weight, BMI), etc., could help discover new characteristics about patients' health. Similarly, in this work, an aggregation of

some diseases was made, but different types of a disease (e.g., types 1 and 2 diabetes) can help in finding more details about a disease.

Furthermore, the exploration of new datasets on other diseases or even in other languages could be considered. Of special interest for future works is a medical dataset in Spanish that could allow the exploration of medical entity associations and other medical circumstances in underdeveloped countries. That is a big challenge because most datasets for research are only available in English.

Moreover, to date, there is not tool that allow the automatically extraction of entities from clinical notes written in Spanish. The scarcity of this type of tools opens the possibility of new researches and works for the extraction of entities in new languages. The use of dictionaries, rule-based and machine learning approaches have been explored in the past for this task. Also, deep learning has been employed in recent years to identify a medical condition or phenotype (Gehrmann *et al.*, 2018; Cheng *et al.*, 2016), to represent and predict the patients' health (Miotto *et al.*, 2016), etc. Hence, these methods are complementary to our works, and therefore could be applied in future work.

In this thesis, the opinion of experts were taken into account to validate the result obtained with the cluster analysis and network approaches, but a comparison with "rule-based systems created by experts" could be covered in future researches.

APPENDIX I

PUBLICATIONS

Journal Article

Reátegui, R., & Ratté, S. (2018). Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Medical Informatics Decision Making, 2018 (Suppl 3), 74. https://doi.org/10.1186/s12018-0654-2

Reátegui, R., Ratté, S., Bautista-Valarezo, E. & Duque, V. Cluster Analysis of Obesity Diseases Based on Comorbidities Extracted from Clinical Notes. Journal of Medical Systems, (2019) 43:52. https://doi.org/10.1007/s10916-019-1172-1

Reátegui, R., Ratté S, Bautista-Valarezo, E. & Beltrán, J.F. A network-based analysis of medical information extracted from electronic medical records. International Journal of Medical Informatics. (Under Review)

Conference Proceedings

Reátegui, R., & Ratté, S. (2018). Automatic Extraction and Aggregation of Diseases from Clinical Notes. In: Rocha Á., Guarda T. (eds) Proceedings of the International Conference on Information Technology & Systems (ICITS 2018). ICITS 2018. Advances in Intelligent Systems and Computing, vol 721. Springer, Cham.

Reátegui, R., Ratté, S., Bautista-Valarezo, E. & Duque, V. (2018). Obesity Cohorts Based on Comorbidities Extracted from Clinical Notes. In: Rocha Á., Adeli H., Reis L., Costanzo S. (eds) Trends and Advances in Information Systems and Technologies. WorldCIST'18 2018. Advances in Intelligent Systems and Computing, vol 746. Springer, Cham.

Reátegui, R., & Ratté, S. (2019). Analysis of Medical Documents with Text Mining and Association Rule Mining. Paper accepted at the ICITS 19 International Conference on Information Technology & Systems, Quito, Ecuador.

BIBLIOGRAPHY

Ahmad, T., Pencina, M. J., Schulte, P. J., O'Brien, E., Whellan, D. J., Pina, I. L., Kitzman, D. W., Lee, K. L., O'Connor, C. M. & Felker, G. M. (2014). Clinical Implications of Chronic Heart Failure Phenotypes Defined by Cluster Analysis. *Journal of the American College of Cardiology*, 64(17), 1765-1774. doi: 10.1016/j.jacc.2014.07.979.

Alnazzawi, N., Thompson, P., Batista-Navarro, R. & Ananiadou, S. (2015). Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. *Bmc Medical Informatics and Decision Making*, 15, 1-10. doi: Artn S3 10.1186/1472-6947-15-S2-S3.

Aneja, A., El-Atat, F., McFarlane, S. & Sowers, J. R. (2004). Hypertension and obesity. *Recent Progress in Hormone Research, Vol 59*, 59, 169-205. Consulted at <GotoISI>://WOS: 000189490300009.

Antonelli, D., Baralis, E., Bruno, G., Cerquitelli, T., Chiusano, S. & Mahoto, N. (2013). Analysis of diabetic patients through their examination history. *Expert Systems with Applications*, 40(11), 4672-4678. doi: DOI 10.1016/j.eswa.2013.02.006.

Anzaldi, L. J., Davison, A., Boyd, C. M., Leff, B. & Kharrazi, H. (2017). Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. *BMC geriatrics*, 17(1), 248. Consulted at <GotoISI>://MEDLINE:29070036.

Apovian, C. M. (2016). Obesity: definition, comorbidities, causes, and burden. *The American journal of managed care*, 22(7 Suppl), s176-85. Consulted at <GotoISI>://MEDLINE: 27356115.

Aronso, A. (2001). Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *AMIA Annu Symp Proc 2001*, 17-21.

Aronson, A. R. & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *JAMIA*, 17(3), 229-236. doi: 10.1136/jamia.2009.002733.

Bauer-Mehren, A., LePendu, P., Iyer, S. V., Harpaz, R., Leeper, N. J. & Shah, N. H. (2013). Network analysis of unstructured EHR data for clinical research. *AMIA Summits on Translational Science Proceedings*, 2013, 14-18. Consulted at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3845760/.

Beck-Nielsen, H. & Groop, L. C. (1994). Metabolic and genetic characterization of prediabetic states. Sequence of events leading to non-insulin-dependent diabetes mellitus. *J Clin Invest*, 94(5), 1714-21. doi: 10.1172/JCI117518.

Becker, M. & Bockmann, B. (2016). Extraction of UMLS (R) Concepts Using Apache cTAKES (TM) for German Language. *Health Informatics Meets Ehealth*, 223, 71-76. doi: 10.3233/978-1-61499-645-3-71.

Bejan, C. A., Xia, F., Vanderwende, L., Wurfel, M. M. & Yetisgen-Yildiz, M. (2012). Pneumonia identification using statistical feature selection. *JAMIA*, 19(5), 817-823. doi: 10.1136/amiajnl-2011-000752.

Berenbaum, F., Eymard, F. & Houard, X. (2013). Osteoarthritis, inflammation and obesity. *Current Opinion in Rheumatology*, 25(1), 114-118. Consulted at <GotoISI>://WOS: 000311944600019.

Bhavnani, S. K., Victor, S., Calhoun, W. J., Busse, W. W., Bleecker, E., Castro, M., Ju, H., Pillai, R., Oezguen, N., Bellala, G. & Brasier, A. R. (2011). How cytokines co-occur across asthma patients: from bipartite network analysis to a molecular-based classification. *Journal of biomedical informatics*, 44 Suppl 1, S24-30. Consulted at <GotoISI>://MEDLINE:21986291.

Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics-Theory and Experiment*, 1-12. doi: Artn P10008 10.1088/1742-5468/2008/10/P10008.

Bourdin, A., Molinari, N., Vachier, I., Varrin, M., Marin, G., Gamez, A.-S., Paganin, F. & Chanez, P. (2014). Prognostic value of cluster analysis of severe asthma phenotypes. *Journal of Allergy and Clinical Immunology*, 134(5), 1043-1050. doi: 10.1016/j.jaci.2014.04.038.

Brasier, A. R., Victor, S., Boetticher, G., Ju, H., Lee, C., Bleecker, E. R., Castro, M., Busse, W. W. & Calhoun, W. J. (2008). Molecular phenotyping of severe asthma using pattern recognition of bronchoalveolar lavage-derived cytokines. *Journal of Allergy and Clinical Immunology*, 121(1), 30-37. Consulted at <GotoISI>://WOS:000252372000005.

Bruce, S. G., Riediger, N. D., Zacharias, J. M. & Young, T. K. (2011). Obesity and obesity-related comorbidities in a Canadian First Nation population. *Prev Chronic Dis*, 8(1), A03. Consulted at https://www.ncbi.nlm.nih.gov/pubmed/21159215.

Bukhanov, N., Balakhontceva, M., Krikunov, A., Sabirov, A., Semakova, A., Zvartau, N. & Konradi, A. (2017). Clustering of comorbidities based on conditional probabilities of diseases in hypertensive patients. *Procedia Computer Science*, 108, 2478-2487. doi: https://doi.org/10.1016/j.procs.2017.05.073.

Canto, J. G., Kiefe, C. I., Rogers, W. J., Peterson, E. D., Frederick, P. D., French, W. J., Gibson, C. M., Pollack, C. V., Ornato, J. P., Zalenski, R. J., Penney, J., Tiefenbrunn, A. J., Greenland, P. & Investigators, N. (2011). Number of Coronary Heart Disease Risk Factors and Mortality in Patients With First Myocardial Infarction. *Jama-Journal of the American Medical Association*, 306(19), 2120-2127. doi: 10.1001/jama.2011.1654.

Chang, C. & Tang, C. (2014). Community detection for networks with unipartite and bipartite structure. *New Journal of Physics*, 16, 1-27. Consulted at <GotoISI>://WOS: 000342050300001.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. (2001). A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5), 301-310. doi: http://dx.doi.org/10.1006/jbin.2001.1029.

Chen, C.-Z., Wang, L.-Y., Ou, C.-Y., Lee, C.-H., Lin, C.-C. & Hsiue, T.-R. (2014). Using Cluster Analysis to Identify Phenotypes and Validation of Mortality in Men with COPD. *Lung*, 192(6), 889-896. doi: 10.1007/s00408-014-9646-x.

Chen, Y. & Xu, R. (2014). *Network analysis of human disease comorbidity patterns based on large-scale data mining*. Conference Proceedings presented in Proceedings of the International Symposium on Bioinformatics Research and Applications (pp. 243-254).

Cheng, Y., Wang, F., Zhang, P. & Hu, J. (2016). Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining* (pp. 432-440). doi: 10.1137/1.9781611974348.49.

Chiaramello, E., Paglialonga, A., Pinciroli, F. & Tognola, G. (2016). Attempting to Use MetaMap in Clinical Practice: A Feasibility Study on the Identification of Medical Concepts from Italian Clinical Notes. *Stud Health Technol Inform*, 228, 28-32. Consulted at https://www.ncbi.nlm.nih.gov/pubmed/27577335.

Cook, M. B., Greenwood, D. C., Hardie, L. J., Wild, C. P. & Forman, D. (2008). A systematic review and meta-analysis of the riskof increasing adiposity on Barrett's esophagus. *American Journal of Gastroenterology*, 103(2), 292-300. doi: 10.1111/j.1572-0241.2007.01621.x.

Daniels, S. R. (2012). Obesity, Vascular Changes, and Elevated Blood Pressure. *Journal of the American College of Cardiology*, 60(25), 2651-2652. doi: https://doi.org/10.1016/j.jacc.2012.09.030.

Delluc, A., Mottier, D., Le Gal, G., Oger, E. & Lacut, K. (2009). Underweight is associated with a reduced risk of venous thromboembolism. Results from the EDITH case-control study. *Journal of Thrombosis and Haemostasis*, 7(4), 728-729. Consulted at <GotoISI>://WOS:000264373800033.

Devouassoux, G., Levy, P., Rossini, E., Pin, I., Fior-Gozlan, M., Henry, M., Seigneurin, D. & Pepin, J. L. (2007). Sleep apnea is associated with bronchial inflammation and continuous positive airway pressure-induced airway hyperresponsiveness. *Journal of Allergy and Clinical Immunology*, 119(3), 597-603. doi: 10.1016/j.jaci.2006.11.638.

Dixon, J. B., Dixon, M. E. & O'Brien, P. E. (2003). Depression in association with severe obesity - Changes with weight loss. *Archives of Internal Medicine*, 163(17), 2058-2065. doi: DOI 10.1001/archinte.163.17.2058.

Evangelista, L. S., Cho, W. K. & Kim, Y. (2018). Obesity and chronic kidney disease: A population-based study among South Koreans. *Plos One*, 13(2), 1-13. doi: ARTN e0193559 10.1371/journal.pone.0193559.

Falagas, M. & Kompoti, M. (2006). *Obesity and infection*. doi: 10.1016/S1473-3099(06)70523-0.

Figueroa, R. L. & Flores, C. A. (2016). Extracting Information from Electronic Medical Records to Identify the Obesity Status of a Patient Based on Comorbidities and Body-weight Measures. *Journal of Medical Systems*, 40(8), 1-9. Consulted at <GotoISI>: //WOS:000380697200008.

Foster, M. C., Hwang, S. J., Larson, M. G., Lichtman, J. H., Parikh, N. I., Vasan, R. S., Levy, D. & Fox, C. S. (2008). Overweight, obesity, and the development of stage 3 CKD: the Framingham Heart Study. *Am J Kidney Dis*, 52(1), 39-48. doi: 10.1053/j.ajkd.2008.03.003.

Francisco, V., Perez, T., Pino, J., Lopez, V., Franco, E., Alonso, A., Gonzalez-Gay, M. A., Mera, A., Lago, F., Gomez, R. & Gualillo, O. (2018). Biomechanics, obesity, and osteoarthritis. The role of adipokines: When the levee breaks. *Journal of Orthopaedic Research*, 36(2), 594-604. Consulted at <GotoISI>://WOS:000426733700008.

Gangopadhyay, A., Yesha, R. & Siegel, E. (2016). Knowledge Discovery in Clinical Data. In Holzinger, A. (Ed.), *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges* (pp. 337-356). Cham: Springer International Publishing. doi: 10.1007/978-3-319-50478-0_17.

Gao, Y. H., Zhao, H. S., Zhang, F. R., Gao, Y., Shen, P., Chen, R. C. & Zhang, G. J. (2015). The Relationship between Depression and Asthma: A Meta-Analysis of Prospective Studies. *Plos One*, 10(7), 1-12. doi: ARTN e0132424 10.1371/journal.pone.0132424.

Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., Foote,John, J., Moseley, E. T., Grant, D. W., Tyler, P. D. & Celi, L. A. (2018). Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, 13(2), e0192360. Consulted at <GotoISI>://MEDLINE:29447188.

Gilani, S., Quan, S. F., Pynnonen, M. A. & Shin, J. J. (2016). Obstructive Sleep Apnea and Gastroesophageal Reflux: A Multivariate Population-Level Analysis. *Otolaryngology-Head and Neck Surgery*, 154(2), 390-395. doi: 10.1177/0194599815621557.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. & Barabasi, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21), 8685-90. Consulted at <GotoISI>://MEDLINE:17502601.

Guh, D. P., Zhang, W., Bansback, N., Amarsi, Z., Birmingham, C. L. & Anis, A. H. (2009). The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis. *Bmc Public Health*, 9, 1-20. doi: Artn 88 10.1186/1471-2458-9-88.

Guimera, R., Sales-Pardo, M. & Amaral, L. A. N. (2007). Module identification in bipartite and directed networks. *Physical Review E*, 76(3), 1-15. Consulted at <GotoISI>://WOS: 000249785900009.

Han, J., Kamber, M. & Pei, J. (2011). *Data Mining: Concepts and Techniques* (ed. Third Edition). Morgan Kaufmann Publishers Inc.

Holzinger, A. (2016). Machine Learning for Health Informatics. In Holzinger, A. (Ed.), *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges* (pp. 1-24). Cham: Springer International Publishing. doi: 10.1007/978-3-319-50478-0_17.

Hripcsak, G. & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*, 20(1), 117-21. doi: 10.1136/amiajnl-2012-001145.

Hwang, S. (2012). Comparison and evaluation of pathway-level aggregation methods of gene expression data. *Bmc Genomics*, 13. doi: Artn S26 10.1186/1471-2164-13-S7-S26.

Jackson, R. G., Patel, R., Jayatilleke, N., Kolliakou, A., Ball, M., Gorrell, G., Roberts, A., Dobson, R. J. & Stewart, R. (2017). Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open*, 7(1), e012012. doi: 10.1136/bmjopen-2016-012012.

Jonnagaddala, J., Jue, T. R., Chang, N. W. & Dai, H. J. (2016). Improving the dictionary lookup approach for disease normalization using enhanced dictionary and query expansion. *Database (Oxford)*, 2016, 1-14. doi: 10.1093/database/baw112.

Jonnalagadda, S., Cohen, T., Wu, S. & Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *J Biomed Inform*, 45(1), 129-40. doi: 10.1016/j.jbi.2011.10.007.

Jonnalagadda, S. R., Adupa, A. K., Garg, R. P., Corona-Cox, J. & Shah, S. J. (2017). Text Mining of the Electronic Health Record: An Information Extraction Approach for Automated Identification and Subphenotyping of HFpEF Patients for Clinical Trials. *J Cardiovasc Transl Res*, 10(3), 313-321. doi: 10.1007/s12265-017-9752-2.

Joosten, S. A., Hamza, K., Sands, S., Turton, A., Berger, P. & Hamilton, G. (2012). Phenotypes of patients with mild to moderate obstructive sleep apnoea as confirmed by cluster analysis. *Respirology*, 17(1), 99-107. doi: 10.1111/j.1440-1843.2011.02037.x.

Kalgotra, P., Sharda, R. & Croff, J. M. (2017). Examining health disparities by gender: A multimorbidity network analysis of electronic medical record. *Int J Med Inform*, 108, 22-28. doi: 10.1016/j.ijmedinf.2017.09.014.

Khan, A., Uddin, S. & Srinivasan, U. (2018). Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression. *Int J Med Inform*, 115, 1-9. doi: 10.1016/j.ijmedinf.2018.04.001.

Kondo, Y., Salibian-Barrera, M. & Zamar, R. (2016). RSKC: An R Package for a Robust and Sparse K-Means Clustering Algorithm. *Journal of Statistical Software*, 72(5), 1-26. Consulted at <GotoISI>://WOS:000389072200001.

Kovacevic, A., Dehghan, A., Filannino, M., Keane, J. A. & Nenadic, G. (2013). Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association*, 20(5), 859-866. doi: 10.1136/amiajnl-2013-001625.

Kovesdy, C. P., Furth, S. L., Zoccali, C. & on behalf of the World Kidney Day Steering, C. (2017). Obesity and Kidney Disease: Hidden Consequences of the Epidemic. *Canadian Journal of Kidney Health and Disease*, 4, 2054358117698669. doi: 10.1177/2054358117698669.

Kramer, C. K., Zinman, B. & Retnakaran, R. (2013). Are metabolically healthy overweight and obesity benign conditions?: A systematic review and meta-analysis. *Ann Intern Med*, 159(11), 758-69. doi: 10.7326/0003-4819-159-11-201312030-00008.

Kramer, H., Luke, A., Bidani, A., Cao, G., Cooper, R. & McGee, D. (2005). Obesity and Prevalent and Incident CKD: The Hypertension Detection and Follow-Up Program. *American Journal of Kidney Diseases*, 46(4), 587-594. doi: https://doi.org/10.1053/j.ajkd.2005.06.007.

LaGrotte, C., Fernandez-Mendoza, J., Calhoun, S. L., Liao, D., Bixler, E. O. & Vgontzas, A. N. (2016). The relative association of obstructive sleep apnea, obesity, and excessive daytime sleepiness with incident depression: A longitudinal, population-based study. *Int J Obes*, 1-8. doi: 10.1038/ijo.2016.87.

Laing, S. T., Smulevitz, B., Vatcheva, K. P., Rahbar, M. H., Reininger, B., McPherson, D. D., McCormick, J. B. & Fisher-Hoch, S. P. (2015). Subclinical atherosclerosis and obesity phenotypes among Mexican Americans. *J Am Heart Assoc*, 4(3), e001540. doi: 10.1161/jaha.114.001540.

Leslie, W. S., Hankey, C. R. & Lean, M. E. J. (2007). Weight gain as an adverse effect of some commonly prescribed drugs: a systematic review. *Qjm-an International Journal of Medicine*, 100(7), 395-404. doi: 10.1093/qjmed/hcm044.

Luppino, F. S., de Wit, L. M., Bouvy, P. F., Stijnen, T., Cuijpers, P., Penninx, B. W. & Zitman, F. G. (2010). Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. *Arch Gen Psychiatry*, 67(3), 220-9. doi: 10.1001/archgenpsychiatry.2010.2.

Lyalina, S., Percha, B., LePendu, P., Iyer, S. V., Altman, R. B. & Shah, N. H. (2013). Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. *JAMIA*, 20(e2), e297-e305. doi: 10.1136/amiajnl-2013-001933.

Mamudu, H. M., Paul, T. K., Wang, L., Veeranki, S. P., Panchal, H. B., Alamian, A., Sarnosky, K. & Budoff, M. (2016). The effects of multiple coronary artery disease risk factors on subclinical atherosclerosis in a rural population in the United States. *Preventive Medicine*, 88, 140-146. doi: 10.1016/j.ypmed.2016.04.003.

Merrill, J. A., Sheehan, B. M., Carley, K. M. & Stetson, P. D. (2015). Transition Networks in a Cohort of Patients with Congestive Heart Failure A novel application of informatics methods to inform care coordination. *Applied Clinical Informatics*, 6(3), 548-564. Consulted at <GotoISI>://WOS:000362263300010.

Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific reports*, 6, 26094. Consulted at <GotoISI>://MEDLINE:27185194.

Mizyed, I., Fass, S. S. & Fass, R. (2009). Review article: gastro-oesophageal reflux disease and psychological comorbidity. *Aliment Pharmacol Ther*, 29(4), 351-8. doi: 10.1111/j.1365-2036.2008.03883.x.

Mokdad, A. H., Ford, E. S., Bowman, B. A. & et al. (2003). Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA*, 289(1), 76-79. doi: 10.1001/jama.289.1.76.

National Library of Medicine (US). (2009, September). UMLS Reference Manual [Web Page]. Consulted at http://www.ncbi.nlm.nih.gov/books/NBK9676/.

National Library of Medicine (US). (2014). RxNORM [Web Page]. Consulted at https://www.nlm.nih.gov/research/umls/rxnorm/.

National Library of Medicine (US). (2016). Overview of SNOMED CT [Web Page]. Consulted at https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html.

Newby, C., Heaney, L. G., Menzies-Gow, A., Niven, R. M., Mansur, A., Bucknall, C., Chaudhuri, R., Thompson, J., Burton, P., Brightling, C. & British Thoracic Soc, S. (2014). Statistical Cluster Analysis of the British Thoracic Society Severe Refractory Asthma Registry: Clinical Outcomes and Phenotype Stability. *Plos One*, 9(7), 1-11. doi: 10.1371/journal.pone.0102987.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23), 8577-8582. doi: 10.1073/pnas.0601602103.

Pang-Ning, T. & Steinbach,M. Kumar, V. (2006). *Introduction to data mining*. Addison-Wesley.

Pantalone, K. M., Hobbs, T. M., Chagin, K. M., Kong, S. X., Wells, B. J., Kattan, M. W., Bouchard, J., Sakurada, B., Milinovich, A., Weng, W., Bauman, J., Misra-Hebert, A. D., Zimmerman, R. S. & Burguera, B. (2017). Prevalence and recognition of obesity and its associated comorbidities: cross-sectional analysis of electronic health record data from a large US integrated health system. *BMJ open*, 7(11), e017583. Consulted at <GotoISI>://MEDLINE:29150468.

Pavlopoulos, G. A., Kontou, P. I., Pavlopoulou, A., Bouyioukos, C., Markou, E. & Bagos, P. G. (2018). Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience*, 7(4), 1-31. Consulted at <GotoISI>://MEDLINE:29648623.

Peppard, P. E., Young, T., Palta, M., Dempsey, J. & Skatrud, J. (2000). Longitudinal study of moderate weight change and sleep-disordered breathing. *JAMA*, 284(23), 3015-21. Consulted at https://www.ncbi.nlm.nih.gov/pubmed/11122588.

Peppard, P. E., Young, T., Barnet, J. H., Palta, M., Hagen, E. & Hla, K. M. (2013). Increased Prevalence of Sleep-Disordered Breathing in Adults. *American Journal of Epidemiology*, 177(9), 1006-1014. doi: 10.1093/aje/kws342.

Pereira, L., Rijo, R., Silva, C. & Agostinho, M. (2013, Oct). Using text mining to diagnose and classify epilepsy in children. *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*, pp. 345-349. doi: 10.1109/HealthCom.2013.6720698.

Poirier, P., Giles, T. D., Bray, G. A., Hong, Y., Stern, J. S., Pi-Sunyer, F. X. & Eckel, R. H. (2006). Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss. *Arterioscler Thromb Vasc Biol*, 26(5), 968-76. doi: 10.1161/01.ATV.0000216787.85457.f3.

Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W. & Savova, G. (2015). Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J Am Med Inform Assoc*, 22(1), 143-54. doi: 10.1136/amiajnl-2013-002544.

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y. & Zhong, Z. (2013). *Towards Robust Linguistic Analysis using OntoNotes* (pp. 143-152). Consulted at http://www.aclweb.org/anthology/W13-3516.

Reátegui, R. & Ratté, S. (2018a). Automatic Extraction and Aggregation of Diseases from Clinical Notes. *Proceedings of the International Conference on Information Technology & Systems (ICITS 2018)*, pp. 846–855.

Reátegui, R. & Ratté, S. (2018b). Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Med Inform Decis Mak*, 18(Suppl 3), 74. doi: 10.1186/s12911-018-0654-2.

Roberts, R. E., Deleger, S., Strawbridge, W. J. & Kaplan, G. A. (2003). Prospective association between obesity and depression: evidence from the Alameda County Study. *International Journal of Obesity*, 27(4), 514-521. doi: 10.1038/sj.ijo.08022204.

Rocha, A. & Rocha, B. (2014). Adopting nursing health record standards. *Inform Health Soc Care*, 39(1), 1-14. doi: 10.3109/17538157.2013.827200.

Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T. & Brunak, S. (2011). (2011). Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS Computational Biology*, 7(8), 1-10.

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S. & Sontag, D. (2017). Learning a Health Knowledge Graph from Electronic Medical Records. *Scientific reports*, 7(1), 5994. Consulted at <GotoISI>://MEDLINE:28729710.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65. doi: http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5), 507-13. doi: 10.1136/jamia.2009.001560.

Serrano-Pariente, J., Rodrigo, G., Fiz, J. A., Crespo, A., Plaza, V. & High Risk Asthma Res, G. (2015). Identification and characterization of near-fatal asthma phenotypes by cluster analysis. *Allergy*, 70(9), 1139-1147. doi: 10.1111/all.12654.

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B. & Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *JAMIA*, 21(2), 221-230. doi: 10.1136/amiajnl-2013-001935.

Simmons, M., Singhal, A. & Lu, Z. (2016). Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health. *Advances in experimental medicine and biology*, 939, 139-166. Consulted at <GotoISI>://MEDLINE:27807747.

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S. & Wang, G. (2018). Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *J Healthc Eng*, 2018, 4302425. doi: 10.1155/2018/4302425.

Sutherland, E. R., Goleva, E., King, T. S., Lehman, E., Stevens, A. D., Jackson, L. P., Stream, A. R., Fahy, J. V., Leung, D. Y. M. & Asthma Clin Res, N. (2012). Cluster Analysis of Obesity and Asthma Phenotypes. *Plos One*, 7(5), 1-7. doi: 10.1371/journal.pone.0036631.

Tang, B., Wu, Y., Jiang, M., C., D. J. & Xu, H. (2013). Recognizing and Encoding Discorder Concepts in Clinical Text using Machine Learning and Vector Space Model. *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, 23 - 26.

The Emerging Risk Factors, C. (2011). Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. *Lancet*, 377(9784), 1085-1095. doi: 10.1016/S0140-6736(11)60105-0.

Tibshirani, R., Walther, G. & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 63, 411-423. doi: Doi 10.1111/1467-9868.00293.

UK Hypoglycaemia Study Group. (2007). Risk of hypoglycaemia in types 1 and 2 diabetes: effects of treatment modalities and their duration. *Diabetologia*, 50(6), 1140-1147. doi: 10.1007/s00125-007-0599-y.

Uzuner, O. (2009). Recognizing Obesity and Comorbidities in Sparse Data. *JAMIA*, 16(4), 561-570.

van der Esch, M., Knoop, J., van der Leeden, M., Roorda, L. D., Lems, W. F., Knol, D. L. & Dekker, J. (2015). Clinical phenotypes in patients with knee osteoarthritis: a study in the Amsterdam osteoarthritis cohort. *Osteoarthritis and Cartilage*, 23(4), 544-549. doi: 10.1016/j.joca.2015.01.006.

Vavougios, G. D., Natsios, G., Pastaka, C., Zarogiannis, S. G. & Gourgoulianis, K. I. (2016). Phenotypes of comorbidity in OSAS patients: combining categorical principal component analysis with cluster analysis. *Journal of Sleep Research*, 25(1), 31-38. doi: 10.1111/jsr.12344.

Willett, W. C., Dietz, W. H. & Colditz, G. A. (1999). Guidelines for healthy weight. *N Engl J Med*, 341(6), 427-34. doi: 10.1056/NEJM199908053410607.

Witten, D. M. & Tibshirani, R. (2010). A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association*, 105(490), 713-726. doi: 10.1198/jasa.2010.tm09415.

Wolf, J., Lewicka, J. & Narkiewicz, K. (2007). Obstructive sleep apnea: An update on mechanisms and cardiovascular consequences. *Nutrition Metabolism and Cardiovascular Diseases*, 17(3), 233-240. doi: 10.1016/j.numecd.2006.12.005.

World Health Organization. (2019). Health Topics [Web Page]. Consulted at https://www.who.int/topics/obesity/en/.

Wu, Y., Jiang, M., Xu, J., Zhi, D. & Xu, H. (2017). Clinical Named Entity Recognition Using Deep Learning Models. *AMIA Annual Symposium Proceedings*, 2017, 1812-1819. Consulted at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977567/.

Yıldırım, P., Çeken, c., Çeken, K. & Tolun, M. (2010). Clustering Analysis for Vasculitic Diseases. In Zavoral, F., Yaghob, J., Pichappan, P. & El-Qawasmeh, E. (Eds.), *Networked Digital Technologies* (vol. 88, ch. 5, pp. 36-45). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-14306-9_5.

Yıldırım, P., Çeken, c., Çeken, K. & Tolun, M. (2012). Prediction of Similarities Among Rheumatic Diseases. *J. Med. Syst.*, 36(3), 1485-1490. doi: 10.1007/s10916-010-9609-6.

Zhang, P., Wang, F., Hu, J. & Sorrentino, R. (2014). Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics. *AMIA Summits on Translational Science Proceedings*, 2014, 132-136. Consulted at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4333693/.

Zhao, C., Jiang, J. C., Xu, Z. M. & Guan, Y. (2017). A study of EMR-based medical knowledge network and its applications. *Computer Methods and Programs in Biomedicine*, 143, 13-23. doi: 10.1016/j.cmpb.2017.02.016.

Zhou, X., Menche, J., Barabasi, A. L. & Sharma, A. (2014). Human symptoms-disease network. *Nat Commun*, 5, 4212. doi: 10.1038/ncomms5212.

Zhu, Q., Li, X., Conesa, A. & Pereira, C. (2018). GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*, 34(9), 1547-1554. doi: 10.1093/bioinformatics/btx815.