EXTRACTION AND REPRESENTATION OF KEY CHARACTERISTICS FROM EPIDEMIOLOGICAL LITERATURE

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2013

By
George Karystianis
School of Computer Science

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ATR	Automated Term Recognition
BD	Big Data
BMI	Body Mass Index
CDC	Centre for Disease, Control and Prevention
CDE	Common Data Elements
CHD	Coronary Heart Disease
CM	Concept Map
CMM	Concept Map Mining
CRF	Conditional Random Fields
CVD	Cardiovascular Disease
DEEL	Detection of Epidemiological Exposures from Literature
DM	Data Mining
EpiTeM	Epidemiological Text Miner
EPM	Epidemiological Text Mining
ExaCT	Extraction of Clinical Trial characteristics
F	F-score
FN	False Negative
FP	False Positive
GUI	Graphical User Interface
ICD	International Classification of Diseases
IE	Information Extraction
IR	Information Retrieval
KMCI	KnowledgeMap Concept Indexer
KR	Knowledge Representation
MEDLINE	Medical Literature Analysis and Retrieval System Online
MeSH	Medical Subject Headings
ML	Machine Learning
NCBI	National Centre for Biotechnology Information
NCCSTS	National Centre for Case Study Teaching in Science
NE	Named Entity
NER	Named Entity Recognition
NHS	National Health Service
NHS CRD	National Health Service Centre for Reviews and Dissemination
NLP	Natural Language Processing
NP	Noun Phrase
OCRe	Ontology of Clinical Research
P	Precision
PoS	Part-of-Speech Tagging
R	Recall
RCT	Randomized Clinical Trial
SNOMED CT	Systematically Organized Computer Processable Collection of Medical Terms
SVM	Support Vector Machine

TM	Text Mining
TP	True Positive
UMLS	Unified Medical Language System
WC	Waist Circumference
WCRF	World Cancer Research Fund
WHO	World Health Organization
WHR	Waist to Hip Ratio
WSD	Word Sense Disambiguation

Abstract

Epidemiological studies are rich in information that could improve the understanding of concept complexity of a health problem, and are important sources for evidence based medicine. However, epidemiologists experience difficulties in recognising and aggregating key characteristics in related research due to an increasing number of published articles. The main aim of this dissertation is to explore how text mining techniques can assist epidemiologists to identify important pieces of information and detect and integrate key knowledge for further research and exploration via concept maps. Concept maps are widely used in medicine for exploration and representation as a relatively formal, easy to design and understand knowledge representation model.

To support this aim, we have developed a methodology for the extraction of key epidemiological characteristics from all types of epidemiological research articles in order to visualise, explore and aggregate concepts related to a health care problem. A generic rule-based approach was designed and implemented for the identification of mentions of six key characteristics, including study design, population, exposure, outcome, covariate and effect size. The system also relies on automatic term recognition and biomedical dictionaries to identify concepts of interests. In order to facilitate knowledge integration and aggregation, extracted characteristics are further normalized and mapped to existing resources. Study design mentions are mapped to an expanded version of the Ontology of Clinical Research (OCRe), whereas exposure, outcome and covariate mentions are mapped to Unified Medical Language System (UMLS) semantic groups and categories. Population mentions are mapped to age groups, gender and nationality/ethnicity, and effect size mentions are normalised with the regards to the used metric and confidence interval and related concept. The evaluation has shown reliable results, with an average micro F-score of 87% for recognition of epidemiological mentions and 91% for normalisation. Normalised concepts are further organised in an automatically generated concept map, which has three sections for exposures, outcomes and covariates.

To demonstrate the potential of the developed methodology, it was applied to a large-scale corpus of epidemiological research abstracts related to obesity. Obesity was chosen as a case study since it has emerged as one of the most important global health problems of the 21st century. Using the concepts extracted from the corpus, we have built a searchable database of key epidemiological characteristics explored in obesity and an automatically generated concept map represented the normalized exposures, outcomes and covariates. An epidemiological workbench (EpiTeM) was designed to enable further exploration and inspection of the

normalized extracted data, with direct links to the literature. The generated results also allow exploration of trends in obesity research and can facilitate understanding of its concept complexity. For example, we have noted the most frequent concepts and the most common pairs of characteristics that have been studied in obesity epidemiology.

Finally, this thesis also discusses a number of challenges for text mining of epidemiological literature and suggests various opportunities for future work.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to thank my (incredible) parents, Sotiri and Eleftheria Karystiani for giving me courage when I needed it most, for always making me feel well (even in times where things were not looking good) and for providing me an exceptional amount of emotional support. Without them, I wouldn't be able to accomplish half of the things I have done in my life and that says a lot!

I would like to thank my good friend Geraint Duck who helped me a lot during some of the most difficult periods of my life as well as giving me useful advice regarding programming languages and computer science concepts that led to the expansion of my knowledge horizons. Thank you also for all these really cool geek conversations we had regarding videogames, movies and tv shows and thanks for introducing me to the awesome show of "Buffy the Vampire Slayer"!

I would like also to acknowledge Xinkai Wang for teaching me the basics of the python programming language, useful bricks of knowledge that were laid in front of me and for assisting me to appreciate programming.

I would like to thank Professor John Keane for his priceless advice, words of wisdom and guidance that helped me understand important things in my environment that I used to ignore and to find my place in a (wider, strange and yet fascinating) world.

I would like to thank my (co-)supervisor, Professor Iain Buchan. The day I went to his office and I was (almost) ready to quit my research, your sentences "The more people grow up, the differences should get smaller between them", "What would your parents like you to do now? Carry on or quit?" were never forgotten, inspired me and kept me focused at my goal.

Finally, I would like to thank the pilot of my research journey, my supervisor Dr Goran Nenadic for his guidance (and patience!) all this rather long! period that I used to ask him a billion questions. I learnt a lot from him including patience, persistence, diligence, perfectionism and how to be a hard working researcher.

Dedication

This thesis is dedicated to my AMAZING parents Sotiri and Eleftheria Karystiani and my beloved grandmother, Eleftheria Karystiani who passed away 5 years ago but still is remembered in my heart (and my manners!)

"Nothing is trivial"

Eric Draven

Chapter 1

Introduction

"*Is is a capital mistake to theorize before one has data.*"
Arthur Conan Doyle, 1887

In recent decades, there has been a constantly rise in the amount of biomedical data (e.g., electronic health records, scientific literature). However, this makes the task of seeking, identifying and analysing important information difficult and time consuming. More specifically, there is a vast collection of available data in the field of epidemiology that can be used for further exploration and analysis. Epidemiology is defined as the study of the distribution and determinants of diseases in health related studies (Last, 2001).

Text mining has been applied to the biomedical domain for the recognition and association of information relevant to a given health care problem from various types of data (e.g., research literature, clinical records) with promising performance (Zweigenbaum et al. 2007; Coden et al. 2009). However, most of the efforts performed in the domain of epidemiological text mining have been focusing on the identification of clinical trial information and certain characteristics only (Fizman et al. 2007; Kiritchenko et al. 2010; Xu et al. 2010). Therefore, the large amounts of epidemiological research despite being available, have yet to be "*processed*" for the identification of relevant information. Knowledge representation and visualization is an important aspect as they enable further analysis and exploration of the text mining results. Numerous approaches usually follow the mapping of text mining results onto a more structured representation model such as semantic networks, ontologies, knowledge bases, mind maps and concept maps. Concept maps are relatively formal, easy to design and to understand knowledge representation models that have been widely used in medicine. They can provide a framework for knowledge exploration through the existing relationships between concepts (Canas et al. 2005; Nesbit et al. 2006; Zubrinic, 2011).

Obesity is one of the most important global health problems of the 21st century (Hossain et al. 2007; Wang et al. 2008; Duncan et al. 2010; Nguyen et al. 2010; WHO, 2013). The rapid and worldwide growth in obesity has affected people of all ages, genders, geographies and ethnicities. It is posing a serious public health and economic challenge with WHO reporting that it is the fifth leading risk for global mortality (Buchan et al. 2007; Ford et al. 2008; Whitlock et al. 2009; Nguyen et al. 2010; WHO, 2013). Despite its fundamental cause being an energy imbalance between calories consumed and calories expended, obesity has been regarded as a multi-dimensional disorder with a complex interaction between major behavioural and environmental determinants, with genetics playing only a minor role (Ogden

et al. 2007; Monasta et al. 2010; CDC, 2013). It has been studied as a risk factor for various chronic disorders (e.g., hypertension, breast cancer) and has been related to an increase in overall mortality. Although its relationships with disease outcomes have been widely reported, the mechanisms of its pathogenesis are only partly understood (de Koning et al. 2007; Canoy, 2008; Whitlock et al. 2009; Canoy, 2010; Ogden et al. 2010; Ryan et al. 2011; CDC, 2012; Ogden et al. 2012).

Growing understanding of obesity has been reflected in the health literature. Published research can be found in clinical databases such as MEDLINE, National Health Service Centre for Reviews and Dissemination (NHS CRD) and the Cochrane Library (Low et al. 2009). In May 2013, there were more than 178,000 articles in MEDLINE related to obesity (Figure 1). Despite the availability of these data, their manual exploration, curation and digestion is impossible and inefficient. Particularly, the number of epidemiological articles related to obesity is increasing yearly with more than 23,000 entries obtained through the search query "*obesity*/*epidemiology*" as a MeSH descriptor (Figure 2).

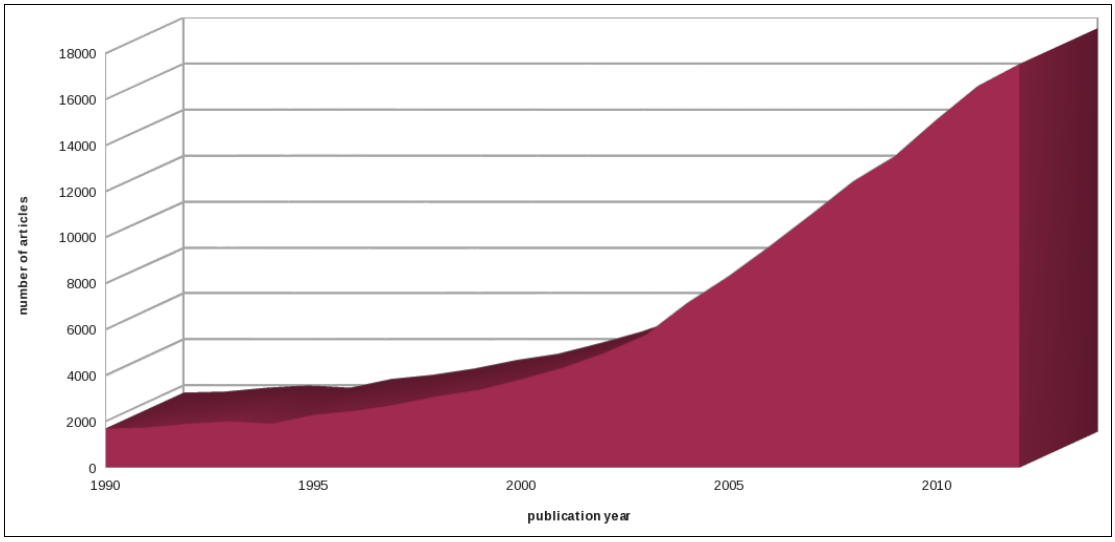


Figure 1: Number of MEDLINE articles for the period 1990-2012 with "*obesity*" as a search term (PUBMED, 2013).

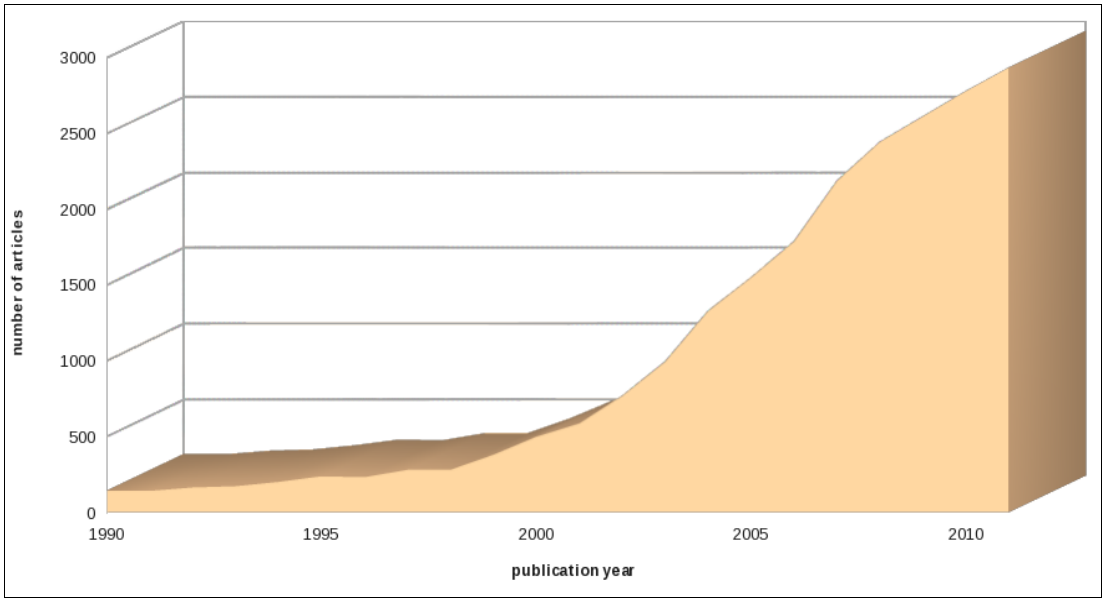


Figure 2: Number of MEDLINE articles for the period 1990-2012 with "*obesity/epidemiology[mesh]*" used as a search term (PUBMED, 2013).

1.1. Research Aim, Hypothesis & Objectives

The main aim of this thesis is to design a framework that enables the construction of concept maps automatically for exploring research questions through the application of text mining to related epidemiological literature. More specifically, we focus on the extraction information for the following characteristics:

1. **study design:** specific plan or protocol that has been followed in the conduct of the study;
2. **population:** demographic details of the individuals (e.g., gender, age, ethnicity, nationality) participating in an epidemiological study;
3. **exposure:** a factor, event, characteristic or other definable entity that brings about change in a health condition, or in other defined characteristics;
4. **outcome:** the consequence from the exposure in the population of interest;
5. **covariate:** a concept that is possibly predictive of the outcome under study;
6. **effect size:** the measure of the strength of the relationship between variables, that relates outcomes to exposures in the population of interest.

The concept map is intended to improve the understanding of the health care concept complexity through the exploration of biomedical knowledge extracted from the relevant literature (in particular epidemiological studies). The aim of the extraction and representation of characteristics from MEDLINE abstracts is to systematically collect detailed epidemiological information, so better modelling of specific and complex health problems can

be performed. The form of a concept map would assist health professionals in the representation, exploration and validation of their expert knowledge, contributing to quality improvement of the health care by understanding potential determinants and outcomes of studies related to health problems. More specifically, for a given health care problem this projects aims to:

1. Design and implement a methodology that would recognise key characteristics (study design, population, exposure, outcome, covariate, effect size) in the scientific epidemiological literature related to the health care problem.
2. Normalize the identified concepts in their respective attributes and classify them under semantic groups in order to enable better manipulation and representation of the information.
3. Automatically organise the normalized and classified concepts into a concept map that will represent the identified epidemiological information for a specific health care problem.
4. Provide a generic framework for the automatic building of concept maps related to health problems from epidemiological studies.

The main hypothesis of this work is that a systematic analysis of epidemiological knowledge related to a particular health care problem through text mining can provide a generic framework for the design of related concept maps by decreasing the amount of time required to inspect targeted information.

1.2. Thesis Contributions

Resulting from the investigation of the questions stated in Section 1.1, this research has proposed, developed and validated a framework that includes text mining techniques for the identification, normalization and representation of key characteristics from epidemiological literature in the form of a concept map for a complex health care problem. More precisely, the main achievements presented are (Figure 3 shows the thesis contributions in a word-cloud format):

1. A methodology that enables the recognition of key characteristics from epidemiological study abstracts by applying rules based on semantic patterns observed in text combined with biomedical dictionaries. This assists in the extraction of key information from all types of epidemiological studies, which is a major challenge considering the nature of certain study types (observational) where specific characteristics are not easy to detect.

2. The expansion of the ontology of clinical research (OCRe) that includes all the main research study design types (observational, experimental, meta-analysis and literature reviews) in a structured hierarchy. This branch was incorporated in the normalization method of recognised study design mentions.
3. A methodology for the normalization of identified key characteristics for their respective attributes. Detailed information about key characteristics is kept as attributes for: study design (attributes depending the study type), population (age, nationality, ethnicity, gender) and effect size (effect size value, confidence interval, related concept, effect size type).
4. The design of a graphical user interface (GUI) that allows the manipulation and inspection of the normalized and classified epidemiological information for each MEDLINE abstract through the interaction of a MySQL database. The GUI enables the user easily to browse and explore large amounts of epidemiological abstracts related to a research question (e.g., a health problem) with specific matching queries for any of the key characteristics. In addition, it provides the opportunity to summarize the recognised and normalized epidemiological information for each abstract in a coherent form, hence making the task of searching for particular characteristic inputs easy and efficient.
5. The automatic creation of a concept map from the normalized epidemiological characteristics (exposure, outcome, covariate). The concept map represents the entire identified and normalized information in an easy to navigate visual form that supports exploration of concepts and the discovery of links between them.
6. The proposed methodologies have been integrated to form a generic framework that enables the recognition, normalization and representation of key characteristics from epidemiological abstracts related to a given health problem into a concept map automatically. This approach provides a-ready-to-inspect visual form that contains all the important concepts related to a health care problem as a start basis for research without having to navigate and integrate any large volumes of scientific documents.
7. The application of this methodology to study the complexity of a particular disease, e.g., obesity, around its various exposures, outcomes and covariates through the automatic generation of a respective concept map.

Parts of the work presented in the thesis have been published in conferences and journals.

1. Karystianis G, Buchan I, Nenadic G. Mining Characteristics of Epidemiological Studies from Medline: A Case Study in Obesity. LBM'2011 [Best Student Paper Award].

2. Karystianis G, Buchan I, Nenadic G. Mining Characteristics of Epidemiological Studies from Medline: A Case Study in Obesity. JBMS [Accepted].
3. Karystianis G, Buchan I, Nenadic G. A Systematic Rule Based Approach for Large Scale Extraction of Key Characteristic from Epidemiological MEDLINE Abstracts: A Case Study in Obesity [In preparation].
4. Karystianis G, Buchan I, Nenadic G. Temporal Epidemiological Research Exploration in Obesity [In preparation].
5. Karystianis G, Buchan I, Nenadic G. A Method to Automatically Generate A Concept Map from Epidemiological Literature [In Preparation].

Figure 3: Thesis contributions section displayed as a word cloud.

1.3. Research Limitations

There are certain limitations of the research and evaluation methodologies presented in this thesis. These are described below:

1. The automatically generated concept map does not name relationships between the concepts. It displays the characteristic in which each concept belongs to (exposure, outcome, covariate) and what semantic group and category has been assigned to it through the application of the MetaMap software. Any exploration for possible relationships among the extracted concepts has to be performed by the relevant health professionals.
2. Currently the system is not able to recognise synonyms. More precisely, the level of normalization is generic and, consequently synonyms (and any potential respective acronyms) are considered separate (and unique concepts) e.g., ("*NAFLD*", "*non-alcoholic fatty liver disease*", "*type 2 diabetes*", "*type 2 diabetes mellitus*"). This could skew the statistical results of the identified and normalized concept and might have an effect at the final form of the concept map.

3. The proposed framework has been tested only on epidemiological abstracts, and not on full-text epidemiological documents. Therefore, despite being applied on text of the same scientific nature, it is not certain that it will perform as well as it has in abstracts, in larger documents.
4. In order to correctly recognise the key characteristics, a training set of 60 abstracts was used and annotated by the author and one other curator of epidemiological expertise. While the agreement was close (61.5% kappa agreement), a third curator with epidemiological expertise would have assisted in making the golden standard potentially more accurate. The abstract sets used for the training, further improvement of the methodology, and its evaluation, were relatively small. In a larger training, development and evaluation set, the implemented rule based approach may have yielded a different set of rules and performance.

1.4. Thesis Structure

The thesis has been organised into 9 chapters (including the Introduction). A brief summary of each of the remaining chapters is given below.

Chapter 2 - Background

This chapter presents an overview of the background knowledge that is required to understand the field of epidemiology, methodologies applied for the recognition of the key epidemiological characteristics and various forms of knowledge visualization. More precisely, a brief introduction is made for the domain of epidemiology. A summarization of the Text Mining field including its definition, goal, challenges and respective techniques is provided. Additionally, an overview of studies that have conducted related research in the field of epidemiological text mining is presented. The main emphasis is on the targeted characteristics for extraction and the applied data. The visualization technique of the concept map is explained, displayed and analysed with its application areas and its potential benefits while being differentiated from other knowledge representation and visualization forms (e.g., semantic networks, ontologies, e/r models and mind maps). Research efforts that automatically created concept maps from text are presented, although the focus was mostly on the data that were used for the generation of the maps, and their ultimate purpose.

Chapter 3 – Research Method Overview

This chapter presents an overview of the method developed for the identification and normalization of key characteristics from epidemiological studies. More specifically, a definition of each of the six key characteristics is provided along with examples in

epidemiological study texts. The creation of the gold standard is explained in detail including the annotation guidelines and discussion is made regarding the cases of disagreement and the characteristics of the training, development and evaluation corpora.

Chapter 4 – Epidemiological Characteristic Extraction

This chapter presents the rule based methodology designed and implemented for the extraction of key characteristics from epidemiological study design abstracts at the document level. Examples of rules for each characteristic are shown along with the respective evaluation. The observed errors generated by the application of the extraction method on the evaluation set are discussed in detail for each characteristic separately and some conclusions are drawn.

Chapter 5 – Epidemiological Characteristic Normalization

This chapter introduces and explains the normalization methods used for the epidemiological characteristics at the document level. Examples for each characteristic are displayed along with the respective evaluation. The errors produced from the application of the normalization approaches in the evaluation set are discussed in detail for each characteristic separately. The classification of exposure, outcome and covariate mentions in semantic groups and categories through the MetaMap is also described. A comparison with the performance of similar studies in epidemiological text mining and concept map mining is carried out and encountered challenges and issues during the implementation of the methodology are reported, while potential solutions are discussed.

Chapter 6 – Automatic Construction of Concept Maps from Epidemiological Text Mining

This chapter shows the method for building a concept map automatically. More specifically, the approach to generate a concept map automatically from the produced results of the rule based method method in a particular format is described. The concept map created from the evaluation set is shown.

Chapter 7 – Extraction of Key Characteristics from Epidemiological Literature on Obesity: a Case Study

This chapters reveals an overview of the background knowledge that is required to understand the clinical complexity of the selected case study. It shows the results generated from the application of the rule based methodology on a large scale corpus related to obesity. The extracted and the normalized results are reported and discussed. Frequency diagrams of the most frequent exposure, outcome and covariate concepts are used and the most frequent

(common) pairs of exposures-outcomes, exposures-covariates, outcomes-covariates are observed and addressed. Parts of the automatically generated concept map representing the concepts of the exposure, outcome and covariate characteristics are displayed. A graphical user interface (EpiTeM) for the manipulation of the extracted and normalized epidemiological characteristics is shown along with related examples.

Chapter 8 – Conclusions and Future Work

The final chapter concludes the thesis with a summary of achievements made during the course of research. Also, it describes further research questions and challenges that were identified and can be further explored in order to improve the extraction and representation of key characteristics from epidemiological abstracts related to a health problem in the form of a concept map.

Chapter 2

Background

In this chapter, we presented an overview of the necessary background knowledge to understand aim of this project. A brief introduction is made for the domain of epidemiology along with a summarization of the Text Mining field and the knowledge representation form of the concept map is explained.

2.1. Epidemiology and Digital Epidemiology

"*It's far more important to know what person the disease has than what disease the person has.*"
Hippocrates, 460 BC-377 BC

Epidemiology is a multidisciplinary field that relies on diverse areas of knowledge such as medicine, statistics, social sciences and geography (Ferreira et al. 2012). The World Health Organisation (WHO) defines the field of epidemiology as "*the study of the distribution and determinants of health related states or events (including diseases), and the application of this study to the control of diseases and other health problems*" (Brandt et al. 2002; WHO, 2013). "*Health related*" could refer to human habits, physical fitness, pregnancy and other physiological (and psychological) phenomena that are not necessarily diseases (Last, 2001). In other words, epidemiology is concerned with the dynamics of health and disease in human populations (Salathe et al. 2012). It has provided a scientific foundation for public health research and guidelines (Xu et al. 2010; Khoury et al. 2013). It has been based on collected and integrated data and resources by public health agencies through clinical centres, hospitals and out in the field (Salathe et al. 2012). The term comes from Greek with "*EPI*", "*DEMIO*" and "*LOGIA*" meaning upon, people and study respectively (study of upon the people) (Calderon, 2000; Salathe et al. 2012).

The goal of epidemiology is the quantitative description and causal information about, associations between exposures and outcomes in human health (Thew et al. 2009). Epidemiologists aim to improve the understanding of disease causes and to prevent them by providing evidence to underpin the development of interventions and policies (Calderon, 2000; Brandt et al. 2002; Thew et al. 2009; Salathe et al. 2012). The combination of numerous individual studies can help obtaining the best estimate of a general effect (i.e., meta-analysis) and clarify the evidence base (Brandt et al. 2002).

There are different ways to form epidemiological hypothesis. These include examining patterns of health in populations and exploring relationships between potential causal factors and their outcomes (WHO, 2013). In other words, epidemiological studies can be classified according to their research design:

- **experimental studies:** scientific experiments that test the effects of interventions in health e.g., clinical trials, preventive trials;
- **observational studies:** study patterns of health without the research making any intervention in health care e.g., case control studies, cohort studies.

The field of epidemiology serves as the framework for public health research and policies (Xu et al. 2010). Through epidemiological studies, determinants can be discovered by observing the targeted populations' health states in systematic ways. Epidemiological studies report rich information that could improve the understanding of the concept complexity of a health problem. Typical key characteristics that detail the followed hypothesis in epidemiological literature were adapted from The Dictionary of Epidemiology (Last, 2011) and are sufficient for the cross-disciplinary nature of this research and defined as:

- **study design:** the type of study used in order to test a given hypothesis in a defined population e.g., "*prospective cohort study*", "*case control study*";
- **population sample:** the number of participants in an epidemiological study selected according to a sampling frame e.g., "*250 Chinese female students*", "*children between 6-10 years old*";
- **exposure:** a factor, event, characteristic or other definable entity that brings about change in a health condition, or in other defined characteristics e.g., "*smoking*", "*physical inactivity*";
- **outcome:** the consequence from the determinant in the population of study e.g., "*breast cancer*", "*stroke*";
- **covariate:** a factor covarying with the outcome of interest which may be on the causal pattern or parallel to it (i.e., confounding) e.g., "*age*", "*gender*";
- **effect size:** the measure of the strength of a determinant in a epidemiological hypothesis e.g., "*odds ratio: 3.71*", "*relative risk, 2.56*".

Despite a large number of epidemiological studies being conducted through traditional methods worldwide, epidemiology is converted into a digital domain, rich with clinical information from the emergence of various data sources. Digital (or e-) epidemiology is defined as "*the science underlying the acquisition, maintenance and application of epidemiological knowledge and information using digital media such as the Internet, mobile*

phones, digital paper and digital TV" (Ekman et al. 2007; Salathe et al. 2012). As a consequence of modern communication (e.g., social networking sites, web queries, mobile devices), the existence of the Internet and an increased use of electronic devices, data are frequently collected directly from individuals (Salathe et al. 2012). Additionally, within the last years, surveys and studies that applied traditional techniques to collect related epidemiological data are now performed on the Internet, increasing the cost efficiency and the convenience for study participants (Ekman et al. 2007; Huybrechts et al. 2010). These types of data are in accessible form and can be used for the identification of meaningful clinical information when harnessed appropriately, while holding unparalleled potential for the domain of digital epidemiology (e.g., providing local and timely information about disease outbreaks and related events) (Salathe et al. 2012). Social networking data are highly contextual and networked and enable epidemiologists (when processed accordingly) to inspect individuals and groups in their life context, to investigate the person-to-person spread of a disease and to observe the social/individual behaviours at their occurring level (Salathe et al. 2012). For example, Figure 4 shows a map that highlights areas that have displayed public tweets (not related with a specific issue).

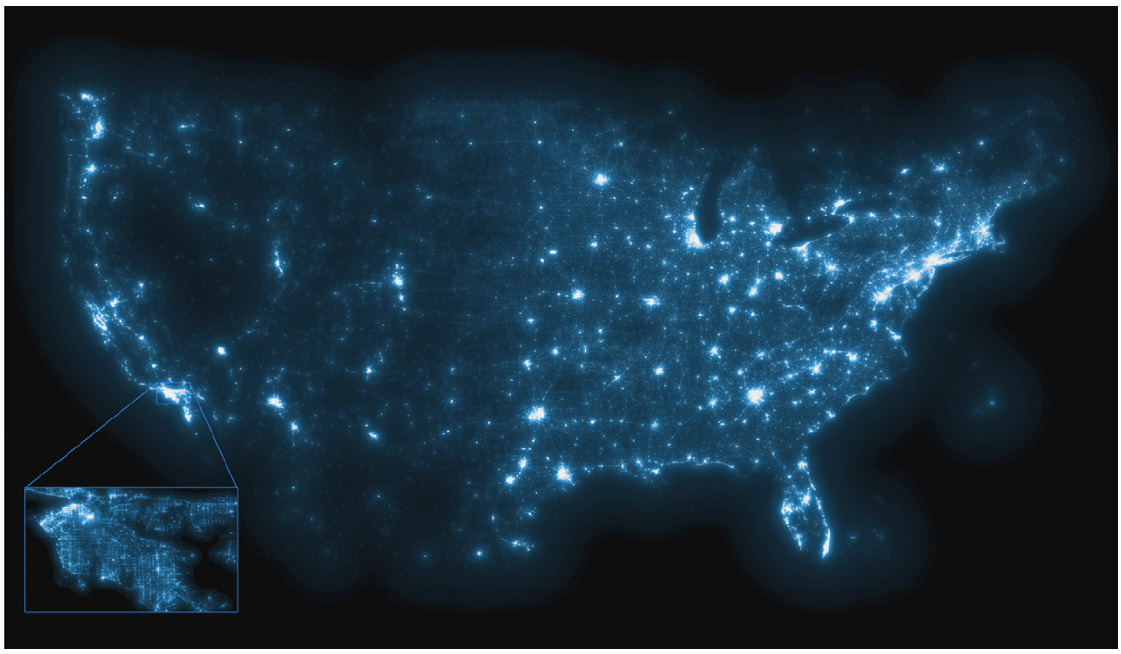


Figure 4: Map from more than 250 million public tweets (collected from Twitter.com). It includes high resolution location information. The inset is the magnified Los Angeles area. The brightness of colour is corresponding to the geographic density of tweets (Salathe et al. 2012).

The reality of the 21st century catches digital epidemiology being a part of the wider health "*Big Data*" (BD) revolution and facing an abundance of information that ranges from genomic and molecular to clinical and environmental (Khoury et al. 2013). BD refers to (digital)

information assemblages that make conventional data or database processing problematic due to a combination of their size, update frequency, and variety (Barrett et al. 2013; Hay et al. 2013). Since there are sheer volumes of information available, there have been BD opportunities for the management of disease surveillance (Hay et al. 2013). The amalgamation of these data can generate such an impact that it can alter the (existing and future) medical and public health decision e.g., assist in research and intervention procedures (Barrett et al. 2013; Khoury et al. 2013). However, such amounts of information are only useful when researchers are able to locate, interpret and access them and currently there is a not an available approach on how to systematically and efficiently tackle the data bulk (Howe et al. 2008). Therefore, the development and implementation of systematic approaches and measures that can vigorously and efficiently manipulate, incorporate, inspect and interpret large volumes of complex data sets is crucial (Howe et al. 2008; Khoury et al. 2013).

2.2. Text Mining of the Biomedical Literature

"Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?"
T.S Eliot, 1934

The above quote from T.S Eliot can be used to summarize the current situation the research community faces with the information overload. Even today, with all the technological achievements in the various fields of science, the primary means of information exchange between experts is through text (Spasic et al. 2005). Given the available state-of-the-art technologies and wide range of digital applications, the volumes of published scientific literature continue to increase, especially those of biomedical nature due to extensive research being conducted (Chapman et al. 2009). Particularly, the 21st century has been characterised by an amazing growth of published and peer-reviewed scientific documents resulting in the rapid expansion of the unstructured biomedical knowledge base. Therefore, making it impossible for the researchers and other professionals to keep abreast with any significant chunk of the literature (Erhardt et al. 2006; Bashyam et al. 2007; Abacha et al. 2011b; Aarts et al. 2012).

PubMed[1] is a primary repository of the biomedical literature. It is a free online database comprised from more than 22 million citations involving biomedical literature from MEDLINE (Medical Literature Analysis and Retrieval System Online - a bibliographical database of life sciences and biomedical information), life science journals and online books (PubMed, 2013). Over 500,000 new citations are added each year and this number is expected

1 http://www.ncbi.nlm.nih.gov/pubmed

to increase even further in the future (Zhou et al. 2010; Rebholz-Schuhman et al. 2012). Figure 5 illustrates the growth of documents in MEDLINE over a 40 year period. A query can return an overwhelming number of results, suggesting that searching for specific entries can be a major problem. It is therefore not unlikely to assume that researchers may miss links between their work and other relevant efforts. These connections have been called by Don R. Swanson "*undiscovered public knowledge*". Swanson himself, after careful reading, discovered eleven associations between magnesium and migraine that previously had not been found, despite being available in the literature (Rodriguez-Esteban, 2009).

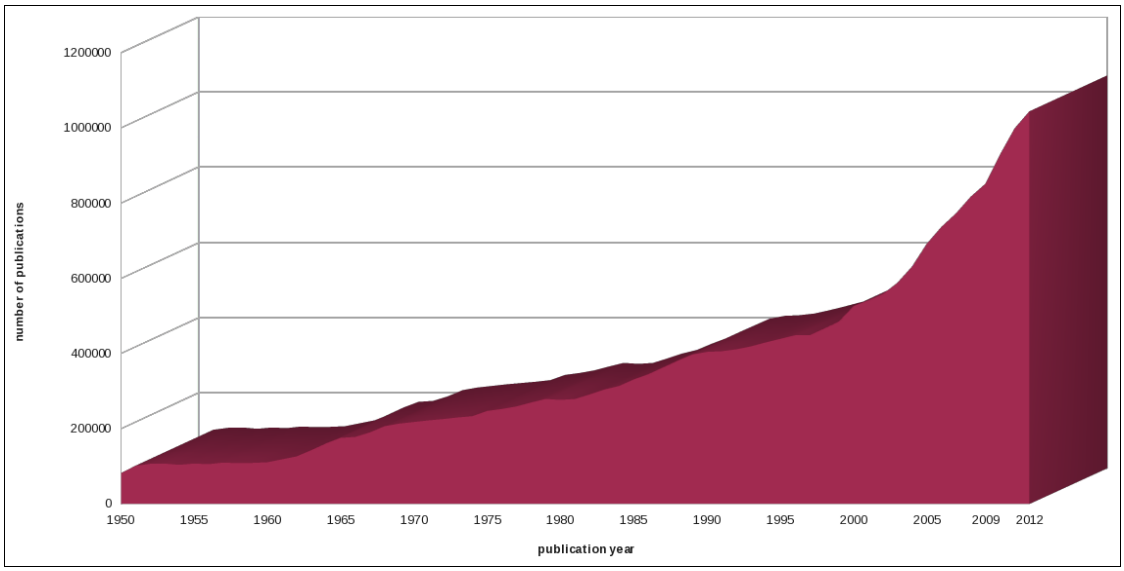


Figure 5: Continuous increase in MEDLINE documents for the period 1965-2012 (PubMed, 2013).

Keeping track of all the updates of new theories, available methods and applications (even if someone is expert in a particular domain) and reviewing the (necessary) amount of literature in order to compose and validate a specific hypothesis are tasks that can be unlikely classified as easy (Shatkay et al. 2003; Cohen et al. 2004; Abacha et al. 2011b). A necessary step to design, perform and interpret the results of any large scale experiment aiming to present accurate and validated results is the examination of the available literature constitutes (Shatkay et al. 2003). However, the number of publications continuously grows making the abundance of accessible literature quite overwhelming (Aarts et al. 2012; Rebholz-Schuhman et al. 2012). Users and most importantly, clinical professionals have to deal with the burden of navigating and manipulating these excessive volumes of text in order to find the information they are looking for (Abacha et al. 2011b). This may result in the inhibition rather than the stimulation of scientific progress and repetition of experiments. The amount of intriguing information and recorded biomedical observations is derived from different and unrelated areas of research

such as epidemiology, medicine and biology and any useful scientific discoveries can go unnoticed due to their implicit existence in scattered research domains (Wang et al. 2008a; Aarts et al. 2012). An additional problem in the biomedical area is the lack of communication between various disciplines. These are required to be highly specialised and divided further into sub-fields. This leads to each one's isolation, narrowing the possibility of comprehending a bigger picture while demitting the discovery of any relationships between different biomedical concepts.

There is now a pressing need to develop technologies that can automatically analyse the scientific literature allowing quick access to important information placed in large volumes of documents (Korhonen et al. 2012). *Natural Language Processing* (NLP) is a subfield of Artificial Intelligence (Erhardt et al. 2006). Its general aim is to achieve a better comprehension of vast amounts of text to a large degree if not completely with the support of computer (Rodriguez-Estaban, 2009). It is concerned with all the aspects and stages of analysing as well as examining spoken, written or printed textual information (Ananiadou et al. 2006). NLP has received an increased attention during the last decade and has been addressing the identification of key biomedical information with language methods applied to a variety of biomedical data such as clinical notes and biomedical journal articles (Mamlin et al. 2003; Schadow et al. 2003; Krallinger et al. 2005; Chapman et al. 2009; Chiang et al. 2010).

NLP methods receive as input unstructured text, which is processed through syntactic and semantic techniques in order to recognise various biomedical concepts (Fiszman et al. 2000; Zeng et al. 2006; Wang et al. 2008b). The combination of NLP, mining tools and semantic constraints is focused on the discovery of new relationships or concepts in the literature and from that perspective it can be said that NLP involves the field of text mining (Shatkay et al. 2003). Nevertheless, so far the application of NLP does not imply a complete understanding of the human language due to its complexity and multiple variations (Erhardt et al. 2006).

The term "*Text Mining*" (TM) (or as "*Text Data Mining*") refers to a multidisciplinary field involving various other sub-fields such as *Information Retrieval* (IR), *Information Extraction* (IE) and *Machine Learning (ML)* (Tan, 1999). Text mining has been introduced to support the scientific reader with necessary tools that can assist in the investigation of the substantial amounts of text (Aarts et al. 2012). In its strict definition, TM involves the automated discovery of new previously, unknown information through an extraction process from large collections of data in unstructured textual form by using automatic or semi-automatic systems (Hearst, 2003; Hotho et al. 2005; Spasic et al. 2005). TM performs the extraction of hidden links among various types of data that may lead to the discovery of new knowledge (Tan,

1999). It connects together the parts of the extracted information, hence creating new facts and hypotheses and reaching new conclusions, all of which can be explored in detail with further research (Hearst, 2003). This distilled knowledge is presented to users in a coherent form (Ananiadou et al. 2006). For example, by analysing the results of biomedical text mining, possible adverse drug interactions, complex underlying diseases, hidden determinants and unconventional covariates linked to a particular topic/concept could be revealed and further explored (Stavrianou et al. 2007).

Text Mining is a relatively new area of computer science and has strong connections with NLP and Data Mining (DM) (Radovanovic et al. 2008). It shares the same techniques with DM, although it is a much more complex task due to the unstructured fuzzy data level it operates with. While NLP is focused mainly on the "*understanding*" of the document as a whole, text mining is dealing with a targeted problem in a specific domain (Cohen et al. 2004). Text mining cannot yet replace humans in complex tasks but it can assist them in the identification and verification of required information in literature more efficiently obscured by the volume of available data (Korhonen et al. 2012).

Biomedical text mining deals specifically with unstructured information resources related to the field of biology, bioinformatics, medical informatics, medicine, and epidemiology. Its role is to help experts in making sense of the large amounts of text by distilling information and extracting facts along with the generation of hypotheses relevant to the user's information need (Zhu et al. 2012). It employs a variety of computational methods (see Section 2.2.1, 2) Term and Named Entity (NE) Recognition) such as machine learning, natural language processing, biostatistics and information theory aiming to discover new hidden pieces of knowledge (Zhu et al. 2012). Currently, text mining techniques offer abstract analysis and full text manipulation with software applications that provide knowledge visualization and improved integration of text mining results with other data resources (Rebholz-Schuhman et al. 2012). In addition, biomedical text mining can greatly decrease the effort spent by researchers on the literature review since it can cover a significantly larger number of data than the average individual in a relatively small amount of time (Erhardt et al. 2006). It can assist in the assimilation process of the high publications rate and it can collect, integrate and discover the knowledge into a coherent and understandable picture outlining ideas, results and suggestions (Shatkay et al. 2003; Ananiadou et al. 2006; Zweigenbaum et al. 2007).

2.2.1. Text Mining Workflow

Text mining typically has four steps (Aarts et al. 2012; Rebholz-Schuhman et al. 2012; Zhu et al. 2012):

1. information retrieval (IR),
2. information extraction (IE),
3. knowledge discovery,
4. hypotheses generation.

Figure 6 reveals the four steps of clinical text mining for the identification of disease and symptoms in documents related to obesity: information retrieval, information extraction, knowledge discovery and hypotheses generation.

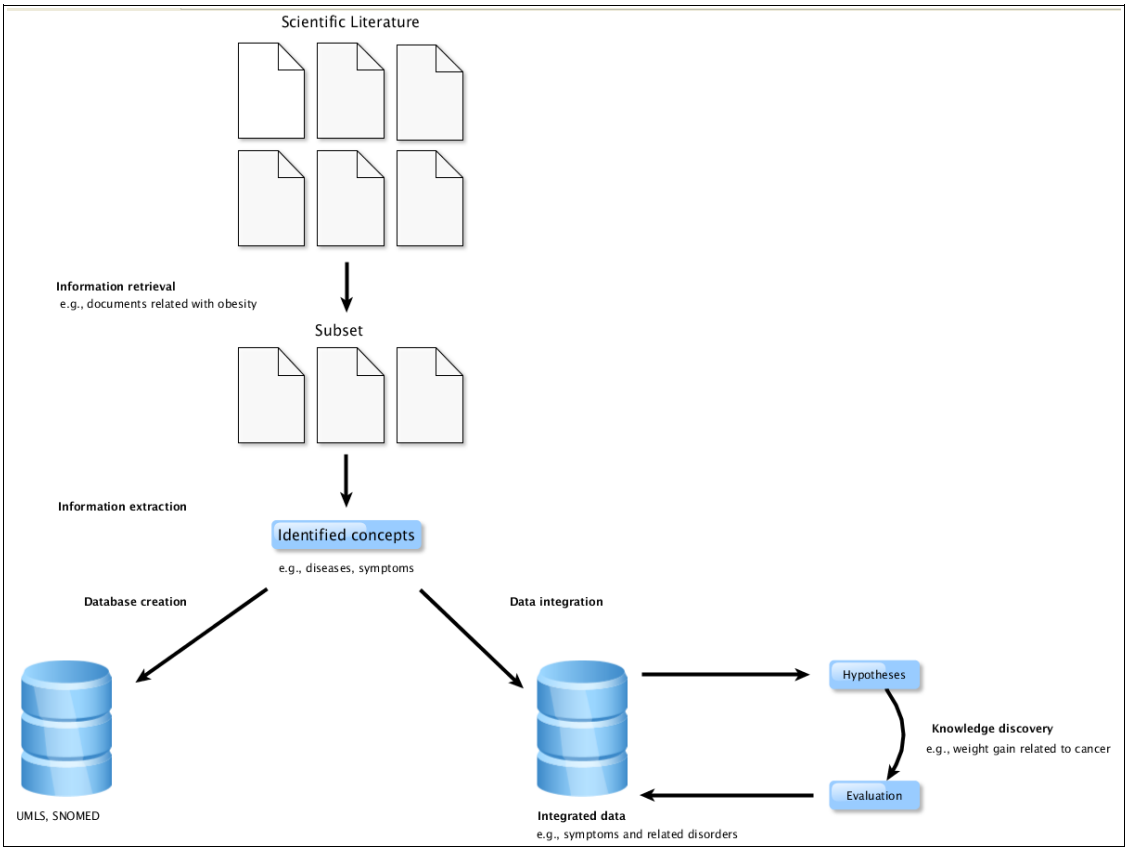


Figure 6: Typical TM pipeline. It aims to identify medical information from the scientific literature related to a specific disease e.g., obesity. More specifically, a subset of clinical documents associated with the concept of obesity will be retrieved from the scientific literature based on a user's query and the targeted blocks of information (e.g., diseases, symptoms) will be extracted. The recognized information will be used to build a related database (e.g., UMLS) and may lead to knowledge discovery through the generation and the evaluation of various new hypotheses.

Information Retrieval

Information retrieval is the process of identifying data or documents from a collection as a result to the submission of a user's query aiming to reduce the number of data/documents for analysis (Aarts et al. 2012; Rebholz-Schuhman et al. 2012). There is a high possibility of not retrieving the related documents (because of variability) or retrieve irrelevant ones (because of ambiguity) and therefore, failure to obtain the necessary information may occur (Spasic et al. 2005).

Users apply document retrieval to seek background information related to a research question. Usually, these research questions can be answered through specific search and retrieval engines such as PubMed or UK PubMed Central as well as providers such as Google Scholar (Rebholz-Schuhman et al. 2012). The most common way of representing a user query is the boolean approach - a query that includes more than one terms with boolean operators such as "*and"*. Consequently, a number of documents is returned if they satisfy the query's conditions. This approach is typically for the Web and is supported by many IR systems like PubMed, which have implemented search engines such as the Entrez system[2] (Shatkay et al. 2003). Most biomedical experts use the PubMed information-retrieval system, which is available at the National Centre for Biotechnology Information (NCBI[3]) and runs on the MEDLINE database. It incorporates Boolean query searches based on indexed look-up techniques for the manipulation of the scientific literature such as descriptors belonging to the controlled vocabulary MeSH[4] (Medical Subject Headings), and a document-similarity search engine based on word-frequency similarities (Krallinger et al. 2005). Table 1 shows a number of (bio)clinical information retrieval tools of scientific literature that are available for use online (PubMed, GoPubMed, RefMED, UK PubMed Central, PolySearch).

In an IR system, each term has a reference to all documents that contain it. As a result, when a user asks a specific query involving a certain word, the index is being searched and the documents including it are retrieved (Shatkay et al. 2003). Index is the representation of a document in terms of a set of index terms. These terms are often single words or word stems (normalized words). Index terms can be weighted on the basis of their frequency in order to rank retrieved documents. Using single words as index terms generally has good exhaustivity, but poor specificity due to the existence of ambiguous words (Takenobu et al. 2000). Many of the directly related to the search topic documents may not be retrieved at all due to ambiguity and synonyms issues (Shatkay et al. 2003).

2 http://www.ncbi.nlm.nih.gov/sites/gquery
3 http://www.ncbi.nlm.nih.gov/
4 http://www.nlm.nih.gov/mesh/

Table 1: Examples of biomedical information retrieval tools (Rebholz-Schuhman et al. 2012).

Name	Content	Input	Description	URL
PubMed	Abstracts	Standard query	Retrieves abstracts of scientific publications from MEDLINE according to the user query. Results are provided as a list and can be further filtered with MeSH terms and an advanced search functionality.	http://www.ncbi.nlm.nih.gov/pubmed
GoPubMed	Abstracts	Standard query	Retrieves publications from MEDLINE and additional functionality by classifying publications according to Gene Ontology concepts to allow improved screening of results.	http://www.gopubmed.com/web/gopubmed
RefMED	Any text	Standard query	Allows user to submit feedback and consequently learns how to search PubMed for relevant articles according to feedback provided.	http://dm.postech.ac.kr./refmed
UK PubMed Central (UKPMC)	Full text	Standard query	Retrieves full-text documents from PubMed and mines the documents for mentions of genes, drugs and Gene Ontology concepts using the Whatizit infrastructure.	http://ukpmc.ac.uk
PolySearch	Abstracts, databases	Standard query	Retrieves information (such as documents and database entries) according to particular patterns of queries. Supports 50 different classes of queries.	http://wishart.biology.ualberta.ca/polysearch/index.htm

Information Extraction

Information Extraction generates structured results or knowledge bases from unstructured textual data in order to extract terms (concepts) (such as disease names and medications) as well as to identify of complex relationships between those entities e.g., disease and drug interactions while associating them to the subject of interest (Muslea, 1999; Aarts et al. 2012; Rebholz-Schuhman et al. 2012). IE can be seen as a restricted form of a full natural language understanding where it is known what kind of information has to be found (Hotho et al. 2005). It can be based on patterns, machine learning techniques, statistical analyses or automated reasoning (Rebholz-Schuhman et al. 2012). The purpose of the extracted information is the provision of well targeted data for further analysis and mining and perhaps the discovery of new knowledge (Hearst, 2003; Shatkay et al. 2003). GATE[5] and Minor Third[6] are examples of systems that provide an information extraction framework. While Gate is rule based focused, MinorThird utilizes both rules and machine learning methods.

5 http://gate.ac.uk/ie/
6 http://sourceforge.net/apps/trac/minorthird/wiki

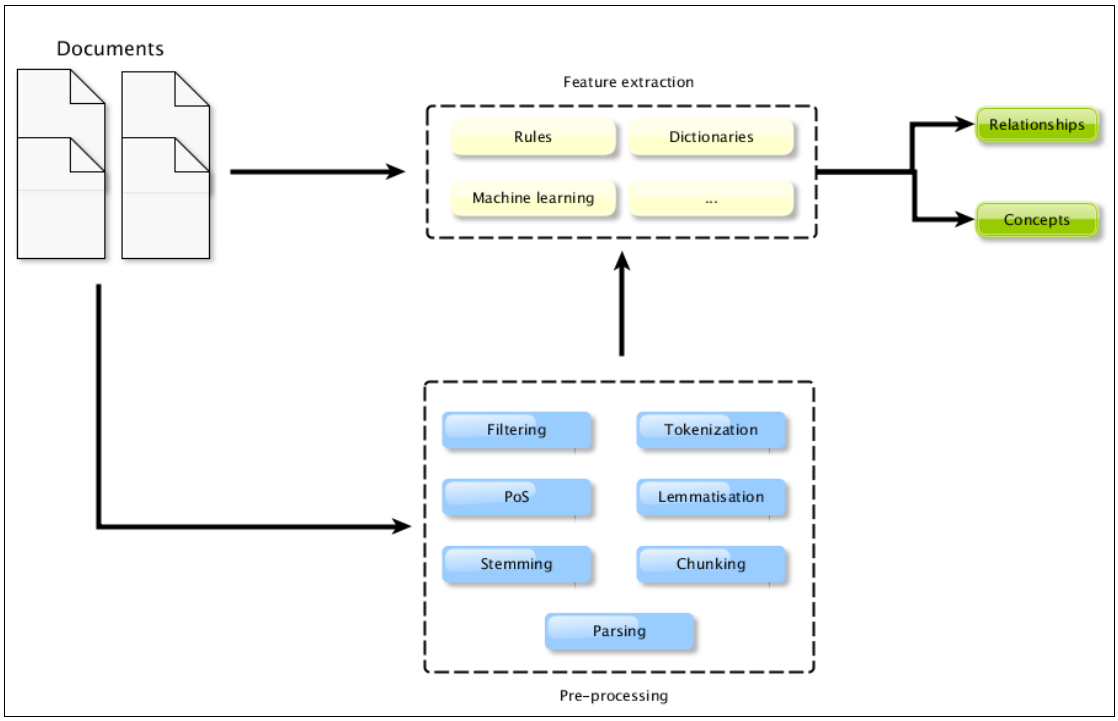


Figure 7: Overview of the IE step. For the detection of the targeted information, various methods can be used: rule based, machine learning based and dictionaries. Documents can be pre-processed before the extraction of the targeted features. This is an optional procedure and may include stop-word filtering, tokenization, part-of-speech tagging (PoS), lemmatisation, stemming, parsing and world sense disambiguation.

There are typical steps that are applied to the data for the extraction of information (Figure 7):

- pre-processing,
- feature extraction,
- Named Entity Recognition (NER) / Automated Term Recognition (ATR).

Their order can be different depending on the task's nature as well as the user's requirements.

1) Pre-Processing

TM often requires the conversion of fuzzy and unstructured textual data into a coherent format which can be suitable for the implementation of information extraction techniques (Hong et al. 2010). There is a variety of methods that are applied during this stage such as tokenization, stop-word filtering, part-of-speech tagging (PoS), lemmatization, stemming, parsing and world sense disambiguation.

Tokenisation

Tokenisation is typically the first step in analyzing text but it is not obligatory (Stavrianou et al. 2007). It is a process required to differentiate all the words (or tokens) used in a document

(Hotho et al. 2005; Ananiadou et al. 2006). Tokens are elementary linguistically plausible units, or - in other words - primitive blocks (Ananiadou et al. 2006; Erhardt et al. 2006). Tokenization can occur in a number of different levels depending on the user and the target application (Shatkay et al. 2003). The delimiters can vary and the most common one is the space or the tab between the words. A variety of algorithms are used to identify boundaries of the words and sentences by using rules or statistical models or even both (Seasr, 2010). The tokenized representation then is used for further analysis and the set of different words emerged from tokenization is named the *dictionary* of that document collection (Hotho et al. 2005). However, this process can be challenging due to the existence of hyphenations, multiple formats (e.g., numbers, dates, addresses) and sentence boundary detections. Examples can be seen below. The medical abbreviation "*b.d*" (stands for "*bis in die*" in latin meaning "*twice daily*") utilized in clinical medication prescriptions can be tokenized in three ways with the "." considered a part of the single token or a token itself, while the genetic disease "*Prader-Willi syndrome*" can tokenized in a similar way due to the existence of a hyphenation. Table 2 reveals the ways in which these two examples can be tokenized into.

Table 2: Examples that can be tokenized in more than one way due to the existence of hyphenations and acronyms.

examples	tokens		number of tokens
b.d	"b","." , "d"		3
	"b.", "d"	"b", ".d"	2
	"b.d"		1
Prader-Willi syndrome	"Prader", "-", "Willi", "syndrome"		4
	"Prader", "-Walli", "syndrome"	"Prader-","Walli", "syndrome"	3
	"Prader-Willi", "syndrome"		2
	"Prader-Willi syndrome"		1

However, both examples should be considered as one token rather than multiple ones. Tokenizers are included in a variety of established NLP libraries e.g., GATE, MinorThird, CTAKES. The performance of the cTAKES tokenizer has been reported to be 94.9% on clinical free text (Savova et al. 2010).

Stop-word Filtering

Usually by applying the tokenization process, the size of the dictionary ends up being large. In order to reduce its size as well as the dimensionality of the documents within the collection, methods such as stop-word filtering are used (Hotho et al. 2005). Any words that bare limited, irrelevant or in very rare cases no information at all (such as articles) are removed and called

stop-words. Specifically, words appearing frequently or rarely are likely to have no specific statistical relevance and can be removed as well or considered candidates for filtering (Hotho et al. 2005). These types of terms create noise that may lead to less distinguishable texts, decreasing the quality of text mining methods (Seasr, 2010). A commonly used stop-word list was created by Fox (1989) and contains four hundred twenty one words. So far, there is not a specific clinical stop-word list that can be used, hence most researchers manually create one.

Part-of-Speech Tagging

PoS tagging is the annotation of words with the appropriate Part-of-Speech tags based on their context (Shatkay et al. 2003). Each word in the input tokenized stream is mapped to the most common set of categories named tags (articles, nouns, verbs, adjectives, prepositions, numbers and proper nouns) (Jurafsky et al 2000; Erhardt et al. 2006; Rebholz-Schuhman et al. 2012). PoS tagging is challenging because of word ambiguity, particularly within short text, where words carry limited context. More specifically, a word can be assigned to multiple part-of-speech tags depending on the context that it is used in. For example, the word "*report*" can be tagged either as a common noun or a verb depending the sentence:

1. **Verb:** We report the following results...
2. **Noun:** A medical report was written in order to...

or in e.g., medical prescriptions where words can be assigned with more than one tag ("*use two sprays every day*" - with "*use*" being considered as both a verb and as a noun).

There is a variety of PoS tagging tools used in a wide range of fields. Most commonly applied taggers include the cTAKES[7] and the Stanford Tagger[8]. cTAKES has been designed to be used in clinical free text and has revealed encouraging results (accuracy 93.6%) (Savova et al. 2010). The Stanford Tagger is used in general domain texts e.g., news-wire and has better overall performance (96.8% accuracy) (Toutanova et al. 2000). General PoS taggers do not usually perform well on biomedical text since the lexical characteristics of biomedical documents are considerably different from those of newspaper articles, which are often used as the training data for a general-purpose tagger.

Lemmatisation and Stemming

Lemmatisation and stemming refer to the conversion of any word to its canonical form (Stavrianou et al. 2007). Lemmatization methods try to map words to their base forms e.g., verb forms to the infinite tense and nouns to their respective singular form (Hotho et al. 2005;

7 http://ctakes.apache.org/
8 http://nlp.stanford.edu/software/tagger.shtml

Spasic et al. 2005). Lemmatisation requires the words to be tagged with their part-of-speech. Stemmers determine the morphological root of a particular word by removing affixes like '*ing*' from the verbs or the plural '*s'* from nouns. More specifically, the stem of a word excludes the derivational elements that indicate gender, person, plurality, tense, etc. Particularly, the strings for example, '*cancers*' or '*cancer-like*' should be identified as variations of the word '*cancer*', which is the root (Erhardt et al. 2006). A widely used stemmer is the Porter Stemmer[9].

Word Sense Disambiguation

Word Sense Disambiguation (WSD) involves the problem of specifying the particular meaning of a multiple meaning word in a given sentence by finding the most probable one (Agirre et al. 2006; Erhardt et al. 2006). It includes the resolution of word and sentence ambiguity. This can be solved by considering the context in which a particular word is found and it may include obtaining the word's grammatical category (Stavrianou et al. 2007). Available extensive knowledge sources such as the Unified Medical Language System (UMLS) and WordNet[10] are widely utilized to tackle WSD problems (Xiong et al. 2013).

As an example, consider the words "*cut*" and "*patch*" (when entered as an input in the MetaMap (Aronson et al. 2010):

1. **cut**: incised wound from injury or poisoning
   cut: cut in medical device material.

2. **patch:** can be regarded as a finding for skin lesions (plaque)
   patch: can be referring to the biomedical or dental material (human patch material)
   patch: referred to the transdermal patch
   patch: can be considered as the surgical patch
   patch: body tissue patch material

The most probable meanings of "*cut*" and "*patch*" in clinical text are incised wound and transdermal patch. However, since both words have multiple meanings, there is a need to specify their meaning depending the context they are used in the related sentences.

Parsing

Parsing is a part of syntactic analysis and usually follows the recognition of words and their tagging. It is a process that analyses the sentence in order to define its grammatical structure according to a given formal grammar or statistical rules by producing a representation that

9 http://tartarus.org/~martin/PorterStemmer/index.html
10 http://wordnet.princeton.edu/

delivers components and dependencies (Hotho et al. 2005; Ananiadou et al. 2006). According to the detail of the produced sentence structure, parsing can be shallow or deep.

Shallow parsing (chunking) deals with the recognition of chunks e.g., noun phrases, verb phrases, etc in each sentence based on the possible phrase boundary positions. Phrases are a series of non overlapping word sequences and as a result grammatically related words are grouped together and each is tagged by one of the predefined grammatical tags such as Noun Phrase (NP), Verb Phrase (VP), Propositional Phrase (PP), Adverb Phrase (AP), Adjective Phrase (AP), Conjunction Phrase (CP) and Subordinated clause (SB) and List Marker (LM) (Shatkay et al. 2003). Shallow parsing does not produce a full parse tree that can reveal all the aspects of phrasal attachment. In the following example:

"The medications cured the symptoms"

the shallow parse output of the sentence would be (NP "*the medications*") (VP "*cured*") (NP "*the symptoms*") since it searches for sequences of word types of even specific words without the generation of a parse tree. The benefit of implementing shallow parsing can be seen in the speed and robustness of the process overall including less ambiguity (no "*attachment*" to verbs, prepositions, etc) although the depth and the granularity of the analysis are being compromised (Hotho et al. 2005). Shallow parsers include cTAKES with an overall performance of 92.4% on clinical free text.

Full (deep) parsing looks for the relationships between words along with their respective functions in the sentence, therefore producing the full parse tree. More specifically, from the parse tree, the relation of each word to all the others in one sentence can be found along with its function (Hotho et al. 2005). The parser algorithm usually takes a sequence of tokens as input data and builds a parse tree based on these tokens (Erhardt et al. 2006). The full parser is associated with grammar-either manually constructed or machine-learned from an annotated corpus which consists of a set of rules to specify syntactic structures in text. The output is a syntax tree and its leafs represent the individual words of the text while the internal nodes are the syntactic structures identified by grammatical tags such as noun or verb, phrase (Shatkay et al. 2003). Figure 8 shows an example of a parsing tree.

However, full parsing is challenging because it provides the relationships within and between the different phrases. A deep parser is the Stanford Parser[11] trained on general domain.

11 http://nlp.stanford.edu/software/lex-parser.shtml

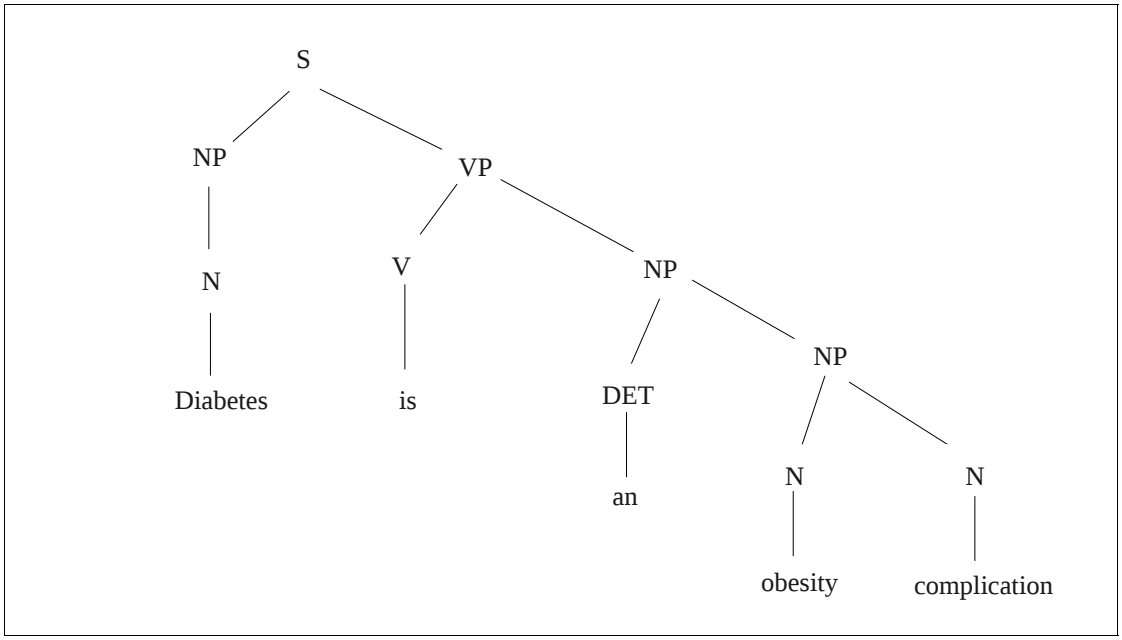


Figure 8: A parsing tree example.

2) Term and Named Entity Recognition

Most of IE applications are designed around biomedical terms. A term can be a single linguistic unit or a combination of units used by field experts to describe and refer to specific well defined concepts in a particular domain (Nenadic et al. 2004; Rebholz-Schuhman et al. 2012). They are the preferred designator of concepts in text by using any of the surface forms that are variants of the corresponding preferred term (Nenadic et al. 2004). An example of a term can be "*cancer*" (disease). Various (and many) terms related to a particular domain or field are grouped together and compose what is called a terminology (Ananiadou et al. 2006; Rodriguez-Esteban, 2009). On the other hand, a Named Entity (NE) is a word or a phrase that refers to a concrete object and can be for example, a disease ("*parkinson's disease*"), symptom ("*fever*"), etc (Li et al. 2008; Wang, 2009; Suakkaphong et al. 2011).

Named Entity Recognition (NER) has been an area of interest in NLP for many years (Krallinger et al. 2005; Zhu et al. 2012). It refers to the task of identifying mentions of entities that belong to a specific semantic class (Cohen et al. 2005; Leaman et al. 2008; Li et al. 2008; Mansouri et al. 2008; Meystre et al. 2008; Wang et al. 2008b; Aramaki et al. 2010; Chowdhury et al. 2010). The main idea behind NER in IE is that if the entities are known and presented in a consistent and normalized form that would facilitate the mining of their relationships or enhance semantic text interpretation (Cohen et al. 2005; Gurlingappa et al. 2010). Due to the importance of the biomedical terms discovery, text mining researchers have created various algorithms designed to recognize them. Examples include Whatizit[12], a tool

12 http://www/ebi.ac.uk/webservices/whatizit/info.jsf

able to recognize several types of entities, and LINNAEUS[13], used for the identification of species in general text (Gerner et al. 2010).

The C-value method for Automated Term Recognition (ATR) is an approach to the extraction of technical terms from special language corpora. It assigns a numeric value named C-value to each extracted term from text, indicating its significance in the whole corpus (Frantzi et al. 1997). ATR considers statistical information such as the frequency of occurrence and string nestdness identifying terms that are relevant for the whole document set (Krauthammer et al. 2004). Multi-word ATR uses linguistic information in the form of a grammar that mainly allows noun phrases to be extracted as candidate terms. The grammar applied may be different: from sequences of nouns only to accept a small number of prepositions as a part of the extracted terms (Frantzi et al. 1997). The method was improved by adding term variation into the ATR process (Krauthammer et al. 2004). Figure 9 shows an example of extracted multi-word terms from epidemiological abstracts related to obesity.

Extracted multi-word terms	C-value	Normalized version
1" body-mass-index | Body mass index | body-mass index | body mass index | BMI "	388.315813	"body mass index"
2" MeS | Metabolic Syndrome | MetS | MES | MS | Metabolic syndrome | metabolic syndrome "	289.708333	"metabolic syndrome"
3" Risk Factors | Risk factors | risk factor | risk factors "	184.246753	"risk factor"
4" weight losses | weight loss | Weight loss | weight Loss "	168.578313	"weight loss"
5" Adipose Tissue | adiposte tissues | Adipose tissue | AT | adipose tissue "	105.931507	"adipose tissue"
6" Childhood Obesity | Childhood obesity | childhood obesity "	94.833333	"childhood obesity"
7" Diabetes mellitus | DM | Diabetes Mellitus | diabetes mellitus "	89.888889	"diabetes mellitus"
8" Blood Pressure | Blood pressure | BP | blood pressure "	89.852941	"blood pressure"

Figure 9: Multi-word terms from obesity related MEDLINE abstracts after applying C-value ATR. The first column shows the identified multi-word terms, the second the respective C-value assigned and the third reveals the normalized version of the recognized multi-word terms.

NER can be performed with (mainly) three approaches: dictionary based, rule based and machine learning (and statistical) based (Krauthammer et al. 2004; Mansouri et al. 2008; Rebholz-Schuhman et al. 2012; Zhu et al. 2012).

Dictionary based NER methods

In the dictionary based approach, a dictionary is being created manually through the input of domain experts or existing terminological resources (Hsiao et al. 2009). More specifically, an entity mentioned in the text is matched to the best match from the dictionary resource and then linked to a database entry (Rebholz-Schuhman et al. 2012). However, a number of existed terms are failed to be recognized when a straightforward dictionary match is used due to potential limitations regarding synonyms (Nenadic et al. 2004; Hsiao et al. 2009; Gurulingappa et al. 2011; Zhu et al. 2012). Examples of dictionary approach NER tools include:

13 http://linnaeus.sourceforge.net/

- **LINNAEUS** (Gerner et al. 2010): uses an internal dictionary of the most frequently mentioned species in MEDLINE, or can be used to match any domain-specific terms using an appropriately compiled dictionary.
- **cTAKES NER component** (Savova et al. 2010): cTAKES uses a terminology-agnostic dictionary, which is a subset of UMLS that includes concepts belonging to one of the following categories: diseases, signs, procedures, drugs and anatomy.

Rule based NER Methods

Ruled based NER focuses on the development of a rule set that can include grammatical, syntactic and orthographic features. This set can describe common naming structures for certain term classes and sometimes it can be combined with dictionaries (Krauthammer et al. 2004). This approach is capable of detecting complex entities that other models have difficulty to do so. However, the lack of portability, the time consuming process of manual rule curation and the disability of not adapting well to new domains are all considered important limitations (Muslea, 1999; Mansouri et al. 2008; Zhu et al. 2012). Examples of rule based NER tools include:

- **Whatizit:** is used tor the recognition of entities in biological text.
- **PROPER** (Fukuda et al. 2010)**:** uses simple lexical and orthographic features to automatically identify protein names in text.
- **Gooch et al. (2011):** developed a rule based system through GATE JAPE grammar in order to recognize clinical terms in MEDLINE abstracts.

Machine Learning NER Methods

The machine learning and statistical based approaches look for patterns and relationships into the unstructured text by identifying and classifying nouns and noun-phrases into particular classes (Mansouri et al. 2008). While statistical approaches may address the recognition of general terms such as keywords, the machine learning ones are aiming towards more specific terms (Krauthammer et al. 2004). The main challenges they face is the proper selection of discriminating features that can be used to extract precisely term instances as well as the detection of term boundaries and the existence of pre-annotated training corpora (Krauthammer et al. 2004). Most machine learning methods are probabilistic in nature, thus they can identify entities only when standard (and large) annotated training data sets are used, which usually take as much time as rule based ones (Wang, 2009; Spasic et al. 2010; Rebholz-Schuhman et al. 2012; Zhu et al. 2012). However, due to limited access in clinical information for confidentiality issues and the expensive cost of training, often only small corpora are

available for machine learning training (Wang et al. 2009). Examples of machine-learning based NER tools include:

- **ABNER** (Settles, 2005): automatically recognises protein and genes names in biological free text through the application of conditional random field (CRF).
- **ManTime** (Filannino et al. 2013): A NER tool that identifies temporal expressions from general domain texts by using CRFs.
- **BANNER** (Leaman et al. 2008): is a machine learning system based on CRFs aiming to recognise biomedical named entities in raw text.

Hybrid NER Methods

Another approach can be a combination of the above, (usually) between rule based and machine learning based methods. This approach ties together the strengths from the types it is consisted of. Despite the fact it can get better results than others, its weakness comes from the rule based features that remain the same in case there is a change in the data domain (Mansouri et al. 2008). Examples of hybrid NER approach tools include:

- **ExaCT** (Kiritchenko et al. 2010): utilizes a machine learning algorithm (Support Vector Machine, SVM) combined with rules in order to automatically extract clinical trial characteristics from journal publications.
- **Patrick et al. (2010):** designed and implemented a hybrid system based on the machine learning techniques of CRF and SVM combined with the creation of rule set for the identification of medical problems, tests and treatments in clinical records.

Normalization of NE mentions

The next step in the recognition of biomedical named entities from text is their mapping to their unique identifier that can be found in a database (Solt et al. 2010). For example, MetaMap (Aronson et al. 2010) enables the mapping of biomedical text phrases to the UMLS Metathesaurus as well as to detect Metathesaurus concepts in text. Through MetaMap, each concept returns a list of Metathesaurus candidates, intermediate results consisting of Metathesaurus strings that are matched partially or wholly with a respective score and a list of MetaMap mappings (combinations of candidates matching as much of the phrase as possible). The MetaMap mapping with the highest score is selected by default assuming that the higher the score, the easier to determine the nature of the concept can be. Figures 10 and 11 show the mapping of the "*depression*" and "*leptin*" entities through MetaMap into the UMLS Metathesaurus.

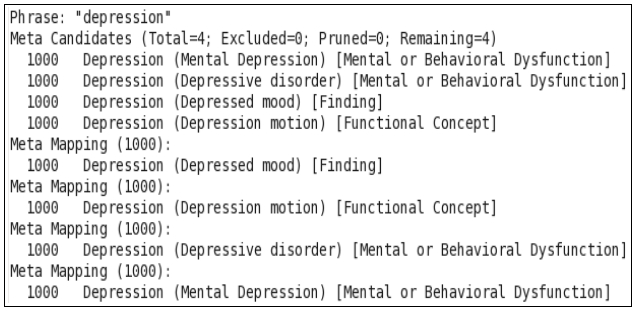


Figure 10: Mapping of the concept "*depression*" into the UMLS Metatheusarus.

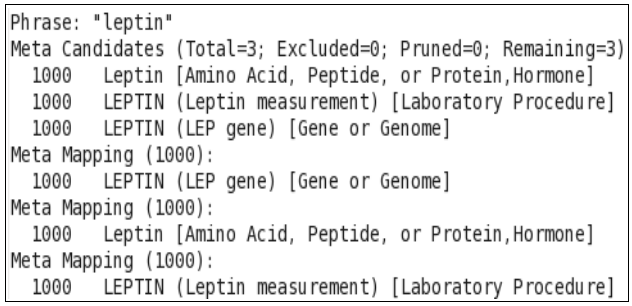


Figure 11: Mapping of the concept "*leptin*" into the UMLS Metatheusarus.

2.2.2. Challenges in Biomedical Text Mining

The automatic processing of biomedical data remains a challenging task (Aramaki et al. 2009; Cano et al. 2009; Mykowiecka et al. 2009; Wang, 2009; Li et al. 2010; Gurulingappa et al. 2011). The complexities of the human language result in grammatical ambiguities, existence of synonyms, short-hand expressions, abbreviations, misspellings and specific text format, hence making the identification of entities a difficult and laborious task (Cano et al. 2009; Coden et al. 2009; Leaman et al. 2009; Hamon et al. 2010; Heung-Seon et al. 2011).

- **Complex content and expressions:** The literature contains specialized terms and implicit background information that may not be directly linked to the paper's main point since they are not addressing the usual public reader (Shatkay et al. 2003). In addition, biomedical data do not follow a traditional structure as it exhibits signs of informal and formal linguistic style (Li et al. 2008; Rodriguez-Esteban, 2009). Most of them (particularly those of clinical nature) are in the form of fragmented free text and are often composed of short, telegraphic and unorthodox grammatical sentences missing punctuation marks (Xu et al. 2004; Uzuner et al. 2007; Meystre et al. 2008; Cano et al. 2009; Friedman, 2009; Wang, 2009; Li et al. 2010). Differences in style among various health professionals suggest a wide variation in documentation (Sirohi et al. 2005; Wang, 2009; Li et al. 2010). Consequently, these textual sources are difficult to process due to their explicit complex format and sentence structure. Furthermore, the structure of medical records is vastly different compared to the

common scientific article, any sequence annotations and even health guidelines (Rodriguez-Esteban, 2009). The above add a level of complexity to the biomedical mining process of literature with standard tools (Shatkay et al. 2003). The non-textual form of the information needed from the user poses another problem (Hearst, 2003). There are cases where important information was coded as a graph and other representations instead of being described in words. This leads to the requirement of design and implement better algorithms that can break the diagrams into smaller information parts and extract the necessary knowledge of interest.

- **Term ambiguity and acronyms:** A common feature of the natural language and biomedical vocabulary is the existence of term ambiguity (polysemy), which makes the recognition of the biomedical concepts difficult (French et al. 2009; Wang, 2009). Polysemy originates from the Greek word "*polisimia*" and refers to a word that has more that one meanings in different contexts (Hamon et al. 2010). This problem is so vivid in the biomedical sciences that many words have a cluster of meanings depending the field they are used on. The presence of local "*dialects*" in science is not uncommon and may cause further complications to existing TM tools (Rodriguez-Esteban, 2009). Biomedical text is overloaded with shorthands (abbreviations, acronyms) as well as local dialectal phrases (Meystre et al. 2008). In MEDLINE only in 2004, almost 63,000 acronyms were newly introduced (Ananiadou et al. 2006). Their use adds extra complexity in the ambiguity problem (Aramaki et al. 2009; Cano et al. 2009; Gurulingappa et al. 2011; Zhu et al. 2012). The probability of having the same classification code for a disease and a drug is very high. A text mining system should be able to distinguish different concepts with the same name and to understand if that name refers to a completely different entity as well (Erhardt et al. 2006).
- **Variation of language expression:** Another problem is the terminological and language expression variation (e.g., synonyms) (Abacha et al. 2011b). Synonyms are derived from the Greek word "*synonimia*", which refers to a variety of different words (or sentences) expressing the same meaning. A biomedical term may have several differently written forms (e.g., s*wine influenza = swine flu = pig flu*) (Deleger et al. 2010; Zhu et al. 2012). Sometimes synonyms can be found in an extended compound form involving complicated sentences instead being composed from only one word (Ananiadou et al. 2006). Therefore, it is hard to understand if experts are referring to the same concept since they are using a range of specific words (e.g., *hay fever = allergic inflammation of the nasal airways =allergic rhinitis)*. It has been reported that the probability of two experts to use the same word for the same concept is less than 20% (Spasic et al. 2005).

- **Access and lack of annotated corpora:** Information access is a key obstacle for the development of better search tools (Dickman, 2003). Currently, text processing is filled with multiple "*dead-ends*" as well as "*short circuits*" in the information flow among biomedical scientists since everything is almost private. Besides that, asking for permission to view every article that contains results from conducted studies, experiments and surveys slows down the pace of research. If not for these restrictions, tools could be more powerful if they were based on full text and had unlimited access to the required data (Dickman, 2003). PubMed and PubMed Central enable the users to access a variety of published biomedical articles. However, PubMed offers access only to abstract text, whereas PubMed Central provides full text articles but only to a small sub set of the total available literature. The lack of annotated corpora also contributes to the difficulty of performing text mining in the biomedical field. The process of manually curating biomedical documents is time-consuming and slow, therefore it results to the generation of a small sets of annotated texts. Consequently, especially machine learning text mining systems are unable to train since they require large clusters of data and their produced results could be misleading.

2.2.3. Evaluation of Biomedical Text Mining

Precision and Recall

When applying text mining tools, it is important to be able to know how reliable the results are. The assessment of the performance quality is a rather difficult task (Shatkay et al. 2003). This typically requires a corpus which contains manually annotated or tagged text items that are targets for IE and constitute the golden standard, and a measure that reflects the performance of the text analysis system (Shatkay et al. 2003). It is difficult to compare the performance between various text mining tools if different datasets are used for the same tasks (Erhardt et al. 2006). A common way to measure the performance of a text mining task is through the calculation of the performance measures recall (R) and precision (P) (Shatkay et al. 2003; Hotho et al. 2005). They rely on concepts of:

- **true positive:** a number of relevant items recognised as relevant,
- **false positive:** a number of irrelevant items recognised as relevant,
- **false negative:** a number of relevant items recognised as irrelevant,
- **true negative:** a number of irrelevant items recognised as irrelevant.

Precision measures the proportion of the correct entities or relations that were returned correctly by the system in total of the retrieved entities. It produces the accuracy of the system in recognizing desirable terms (Ananiadou et al. 2006). Precision is calculated as:

$$P = \frac{TP}{TP + FP}$$

On the other hand, recall measures the number of correct entities returned compared to the total number of relevant entities. It produces the coverage of the system and is calculated as (Ananiadou et al. 2006):

$$R = \frac{TP}{TP + FN}$$

F-score

Precision and recall have inverse relationship between them. For example, when an increase occurs in precision, a simultaneous decrease is observed in recall and vice versa. Therefore, another measure used for evaluating the performance of IE systems is the *F-score* (F) (Hotho et al. 2005). F-score is the harmonic mean of precision and recall (Rodriguez-Esteban, 2009). It is calculated as:

$$F = \frac{2 \cdot R \cdot P}{P + R}$$

2.2.4. Text Mining in Epidemiological Data

"Far and away the best prize that life has to offer is the chance to work hard at work worth doing"

Theodore Roosevelt, 1903

In the last decade, a significant amount of research has been conducted for the extraction of information in the biomedical field, especially in the identification of biological (e.g. Cohen et al. 2005; Meystre et al. 2008) and clinical concepts (e.g. Aramaki et al. 2010; Chowdhury et al. 2010) in the literature. TM has been applied to the field of biology for almost 20 years now (Zweigenbaum et al. 2007). The biology, molecular biology and medicine as well as the bioinformatics area require the application of bio-entity recognition (specific terms corresponding to concepts of biological interest) and are characterized by a high degree of interdisciplinary nature (Krallinger et al. 2005; Spasic et al. 2005). Text mining in the field of medicine can generate significant benefits involving the quality of the providing health care (Fiszman et al 2000; Cano et al 2009; Friedman 2009; Wang 2009; Meystre et al 2010).

So far, TM applications are necessary in the procedures of analysis, classification and information extraction from huge amounts of medical documents, data and scientific abstracts (Aramaki et al. 2009; Cano et al. 2009; Gupta et al. 2009; Abacha et al. 2011b; Heung-Seon et al. 2011). The discovery of hidden links from various medical records can lead towards a better and personalized medicine while various connections between diseases and treatments can uncover adverse drug events, a neglected aspect in the medical research (Yang et al. 2009). Medical text mining can predict accurately and efficiently diseases statutes from clinical discharge summaries (Li, 2008; Yang et al. 2009; de Bruijin et al. 2011). The establishment of electronic health records makes it beneficial to perform text mining since all the knowledge about individuals and their personal health care is stored in one big database. These results combined with those from the discipline of biology can generate new hypotheses and solutions to long lasting health problems such as cancer or diabetes. (Spasic et al. 2005).

Table 3: Overview of studies in epidemiological text mining. Table includes the respective values of precision, recall, F-score in associated data with the targeted elements for extraction and the utilized information extraction approach.

study	data	targeted element	approach	precision	recall	F-score
Fizman et al. 2007	MEDLINE citations	risk factors	rules	67.0%	53.0%	59.1%
Hara et al. 2007	MEDLINE abstracts	patient population	SVM + CRF	80.3%	79.4%	79.8%
	MEDLINE abstracts	compared treatments	SVM + CRF	81.7%	79.3%	80.4%
de Bruijin et al. 2008	RCT publications	23 key trial elements	SVM + rules	65.0%	75.0%	69.6%
Hansen et al. 2008	RCT publications	number of trial participants	SVM	-	-	84.0%
Chung 2009a	RCT publications	intervention	CRF	87.0%	80.0%	83.0%
	RCT publications	participants	CRF	76.0%	40.0%	52.0%
	RCT publications	outcome measures	CRF	82.0%	85.0%	84.0%
Chung 2009b	RCT publications	coordinating constructions	CRF + MaxEntropy	88.0%	81.0%	84.0%
Kiritchenko et al. 2010	MEDLINE RCT reports	21 key trial elements	SVM + rules	93.0%	91.0%	91.9%
Xu et al. 2010	titles of epidemiological articles	exposure-related terms	rules	61.0%	69.0%	69.7%
Luo et al. 2011	clinical trial eligibility criteria	temporal constraints	CRF + rules	83.1%	78.9%	79.8%
Luo et al. 2012	clinical trial eligibility criteria	common data elements	Apriori algorithm	82.3%	79.7%	81.0%

In the subfield of epidemiology however, a small number of studies has been performed focusing on the recognition of epidemiological information from related unstructured data mostly in the area of Randomized Clinical Trials (RCTs). Most of these studies are aiming to identify various trial characteristics (e.g., primary outcome, method of intervention) from respective data while their scope is limited to clinical trials rather than other epidemiological research designs e.g., observational studies. Additionally, there is little research on the automatic recognition of key characteristics such as exposures and covariates that can be found

in both observational studies and experimental research (e.g., clinical trials) (Binder, 2010). Table 3 shows an overview of studies in epidemiological text mining.

Clinical Trial Characteristics Extraction

Most work on the information extraction studies involved the identification of key characteristics from clinical trial publications with a variety of approaches. More specifically, Hara et al. (2007) performed information extraction in clinical trials. They recognised information related to (phase III) clinical trials such as patient population and compared treatments from MEDLINE abstracts. They focused on the determination of base NPs. Then, a class label was attached to each base NP recognised (Table 4) and the MEDLINE abstracts were transformed into a simpler set of sentences. They utilised SVM and CRF to perform automatic base NP chunking and categorization. Simple regular patterns were designed in order to extract the characteristics (patient population and compared treatments) of phase III trials (Table 5). Their precision for population and compared treatments was 80.0% and 82.0% respectively with recall being 79.4% and 79.3%.

Table 4: Class labels for recognized base NPs (Hara et al. 2007).

class label	covered concept	example
disease	disease, symptom, pain, complication	"metastatic breast cancer"
treatment	drug, placebo, therapy, surgery	"doxorubicin"
patient	participants in clinical trials	"patients"
study	clinical trial	"a randomized controlled trial"
others	other than the above	"the efficacy and safety"

Table 5: Regular expression patterns used for the recognition of phase III trials elements (Hara et al. 2007).

IE target	regular expression pattern
Patient population	/Patient with Disease/
	/Treatment (for|of|in) Disease/
Compared treatments	/compar.* Treatment.* Treatment/
	/Treatment.* (versus|vs|or|compared with) .*Treatment/

Most studies when aiming to recognise clinical trial characteristics from journal publications performed a two step approach: a) sentence selection that contains possibly the targeted information and b) trial element extraction. Particularly, de Bruijn et al. (2008) performed information extraction for a number of key elements in clinical trial publications. More

specifically, they aimed to identify key trial design elements using a machine learning approach in RCTs. Their method included a SVM text classifier that was trained to recognise the most promising sentences in which the information element is present with each sentence being represented by a bag of terms (words as well as multi-word phrases). For each information element (trial characteristic) present, extraction patterns relying on "*weak*" extraction rules (based on regular expressions) were applied. More specifically, "weak" rules were manually crafted based on the idea that a simple extraction pattern is possible to be accurate when it is applied to the right context. An example can be seen for the characteristic "*Start date of enrolment*". "*Date ... "* could mean many things if detected anywhere in a RCT article (e.g., date of visit, date of medication prescription, date of birth, etc). However, if it is spotted in a sentence classified as relevant to the start date of enrolment, it is implied that this "*date*" could be the "*enrolment start date*". They reported an overall precision and recall of 65.0% and 75.0% respectively based on an evaluation set of 10 previously unseen RCT articles. The list of the 23 key trial design characteristics that were targeted for extraction can be seen in Table 6 below with the respective performance of the extraction system as a whole and separately for the classification and extraction modules.]

Table 6: Key elements existing in RCT publications. N is the number of data points, P represents precision and R stands for recall. Evaluation results are presented separately for the two modules (SVM classifier, extraction module). The final column represents the evaluation results of the system as a whole.

Information Element	Classifier			Extractor			Entire System		
	n	P	R	n	P	R	n	P	R
Author name (first author only)	N/A	N/A	N/A	N/A	N/A	N/A	10	1.00	1.00
Date of publication	N/A	N/A	N/A	N/A	N/A	N/A	10	1.00	1.00
DOI (Digital Object Identifier)	N/A	N/A	N/A	N/A	N/A	N/A	10	1.00	1.00
Dose (multiple dosages possible)	46	0.53	0.91	144	0.80	0.92	21	0.90	0.90
Duration of the treatment	41	0.30	0.59	72	0.80	0.88	17	0.94	1.00
Early stopping	5	0.06	1.00	5	1.00	0.80	1	1.00	1.00
Eligibility criteria	77	0.90	0.92	N/A	N/A	N/A	37	0.69	0.54
End date of enrolment	54	0.68	1.00	63	0.98	0.98	8	0.80	1.00
Frequency of treatment	40	0.42	0.83	68	0.91	0.88	16	0.76	1.00
Funding organization name	77	0.94	0.96	177	0.84	0.94	21	0.11	0.33
Funding number (grant number)	34	0.39	0.91	78	0.98	1.00	5	0.56	1.00
Make of device	2	0.01	0.50	5	0.67	0.40	1	0.50	1.00
Manufacturer of device	7	0.08	0.86	9	0.75	0.89	2	0.17	1.00
Name of control treatment	69	0.75	0.86	158	0.91	0.50	11	1.00	0.82
Name of experimental treatment	78	0.82	0.83	343	0.89	0.28	16	0.67	0.38
Primary outcome name	77	0.87	0.90	N/A	N/A	N/A	10	1.00	1.00
Primary outcome – time point	58	0.48	0.66	106	0.78	0.84	8	0.46	0.75
Registration identifier of trial	57	0.68	0.95	64	1.00	1.00	11	1.00	1.00
Route of treatment	20	0.23	0.90	45	0.92	0.98	5	1.00	1.00
Sample size	78	0.43	0.44	95	0.81	0.80	10	0.62	0.80
Secondary outcome name	50	0.57	0.90	N/A	N/A	N/A	9	0.70	0.78
Secondary outcome - time point	21	0.20	0.76	41	0.71	0.85	14	0.91	0.71
Start date of enrolment	71	0.89	0.99	84	1.00	0.99	10	1.00	1.00
Overall	**962**	**0.51**	**0.84**	**1557**	**0.87**	**0.82**	**263**	**0.65**	**0.75**

At the later stages of this thesis, Kiritchenko et al. (2010) expanded de Bruijin's approach by designing and implementing a system named ExaCT (Extraction of Clinical Trial characteristics[14]). ExaCT focuses in the recognition of 21 diverse information elements from

14 http://exactdemo.iit.nrc.ca/

full text journal publications of human RCT (Table 7) excluding other types of primary research such as cluster randomized and cross over trials.

Table 7: Trials elements identified from the ExaCT system in RCT publications (Kiritchenko et al. 2010).

element	description
eligibility criteria	logical conditions for being included in the trial, usually split into inclusion and exclusion criteria
sample size	the total number of participants actually enrolled (randomized) in the trial
start date of enrolment	date the enrolment actually started, including, day, month, year or as much as presented
end date of enrolment	date the enrolment actually ended, including day, month, year or as much as presented
name of experimental treatment	name of experimental intervention
name of control treatment	name of control intervention
dose	dosage of experimental/control intervention
frequency of treatment	frequency of administration of experimental/control intervention
route of treatment	route of administration of experimental/control intervention
duration of treatment	duration of administration of experimental/control intervention
primary outcome name	the outcomes(s) of greatest importance, where outcome is a “component of a participant's clinical and functional status after an intervention has been applied, that is used to assess the effectiveness of an intervention”
primary outcome time point	point in time when a primary outcome was assessed
secondary outcome name	outcome(s) used to evaluate additional effects of the intervention deemed a priori as being less important than the primary outcomes
secondary outcome time point	point in time when a secondary outcome was assessed
funding organization name	name of a funding source
funding number	funding grant number
early stopping	whether the trail was stopped earlier
registration identifier of trial	trial registration ID, often ClinicalTrials.gov NCT number
author name	first and last name of the first author
date of publication	year the article was published
DOI	digital object identifier for the publication

A two step approach includes sentence classification based on the SVM algorithm and information extraction. A text classifier is selecting the sentences that are most likely to contain a particular piece of information (one of the 21 trial characteristics) by seeking text excerpts that most closely describe the trial information elements of interest. More specifically, for the sentence classification, the machine learning component based on SVM learns a statistical model from articles that a field expert manually annotated and then a separate statistical model is created for each information element; each model is being applied to all sentences to discover which ones are most similar to the training examples for this particular element. A total number of five best promising sentences that contain the target information candidate is produced for each element with decreasing order of confidence and with the term of interest highlighted. Through the application of regular expression rules (that have been

manually crafted), the exact term excerpts are recognised from these sentences. Particularly, the extracted terms are used to fill in a pre-defined template that includes 21 information elements based on the CONSORT statement.

Figure 12 reveals an example of the template filled in with the extracted information from a RCT publication abstract. They reported an overall precision and recall for the information extraction component 93.0% and 91.0% respectively for exact matches. Table 8 shows the performance of the information extraction component of the ExaCT system in RCT journal publications. However, they focused only on randomised controlled drug treatment trials instead of more general study designs.

Table 8: Precision and recall values from the ExaCT system in RCT journal publications (Kiritchenko et al. 2010).

Information element	# of expert's fragments	exact match		partial match	
		precision	recall	precision	recall
Eligibility criteria	103	1.00	1.00	1.00	1.00
Sample size	46	0.89	0.87	0.89	0.87
Start date of enrolment	32	1.00	1.00	1.00	1.00
End date of enrolment	31	1.00	1.00	1.00	1.00
Name of experimental treatment	54	0.72	0.54	0.97	0.72
Name of control treatment	55	0.83	0.80	0.89	0.85
Dose	103	0.91	0.90	0.96	0.97
Frequency of treatment	70	0.91	0.87	0.99	0.93
Route of treatment	53	0.94	0.92	0.94	0.92
Duration of treatment	45	0.84	0.91	0.86	0.93
Primary outcome name	38	0.97	0.97	0.97	0.97
Primary outcome time point	33	0.90	0.79	0.97	0.85
Secondary outcome name	43	0.93	0.88	1.00	1.00
Secondary outcome time point	25	0.72	0.72	0.92	0.92
Funding organization name	45	0.90	0.98	0.90	0.98
Funding number	7	1.00	1.00	1.00	1.00
Early stopping	2	1.00	1.00	1.00	1.00
Registration identifier of trial	29	1.00	1.00	1.00	1.00
Author name	49	1.00	1.00	1.00	1.00
Date of publication	49	1.00	1.00	1.00	1.00
DOI	47	1.00	1.00	1.00	1.00
Micro-average	959	**0.93**	**0.91**	**0.96**	**0.94**
Macro-average		**0.93**	**0.91**	**0.96**	**0.95**

Panickar J, Lakhanpaul M, Lambert PC, Kenia P, Stephenson T, Smyth A, Grigg J: **Oral Prednisolone for Preschool Children with Acute Virus-Induced Wheezing.** N Engl J Med. 2009 Jan 22;360(4):329-38

Background Attacks of wheezing induced by upper respiratory viral infections are common in preschool children between the ages of 10 months and 6 years. A short course of oral prednisolone is widely used to treat preschool children with wheezing who present to a hospital, but there is conflicting evidence regarding its efficacy in this age group.
Methods We conducted a randomized, double-blind, placebo-controlled trial comparing a 5-day course of oral prednisolone (10 mg once a day for children 10 to 24 months of age and 20 mg once a day for older children) with placebo in 700 children between the ages of 10 months and 60 months. The children presented to three hospitals in England with an attack of wheezing associated with a viral infection; 687 children were included in the intention-to-treat analysis (343 in the prednisolone group and 344 in the placebo group). The primary outcome was the duration of hospitalization. Secondary outcomes were the score on the Preschool Respiratory Assessment Measure, albuterol use, and a 7-day symptom score.
Results There was no significant difference in the duration of hospitalization between the placebo group and the prednisolone group (13.9 hours vs. 11.0 hours; ratio of geometric means, 0.90; 95% confidence interval, 0.77 to 1.05) or in the interval between hospital admission and signoff for discharge by a physician. In addition, there was no significant difference between the two study groups for any of the secondary outcomes or for the number of adverse events.
Conclusions In preschool children presenting to a hospital with mild-to-moderate wheezing associated with a viral infection, oral prednisolone was not superior to placebo. (Current Controlled Trials number, ISRCTN58363576 [controlled-trials.com] .)

Trial design template:

Eligibility criteria	• children between the ages of 10 months and 60 months; • the children presented to three hospitals in England with an attack of wheezing associated with a viral infection
Sample size	700
Start date of enrolment	
End date of enrolment	
Name of experimental treatment	prednisolone
Name of control treatment	placebo
Dose	• 10 mg for children 10 to 24 months of age • 20 mg for older children
Frequency of treatment	once a day
Route of treatment	oral
Duration of treatment	5-day
Primary outcome name	the duration of hospitalization
Primary outcome time point	
Secondary outcome name	• the score on the Preschool Respiratory Assessment Measure; • albuterol use; • a 7-day symptom score
Secondary outcome time point	7-day
Funding organization name	
Funding number	
Early stopping	
Registration identifier of trial	ISRCTN58363576
Author name	Panickar J
Date of publication	2009 Jan 22
DOI	

Figure 12: Example of an RCT publication and the corresponding template filled with the identified trial elements. A number of slots are left empty since the information is not present in the abstract (Kiritchenko et al. 2010).

Hansen et al. (2008) worked on RCT journal articles that identify the total number of trial participants before any exclusions or different allocations. An overview of the method can be seen in Figure 13. Particularly, the number of trial participants was determined from the application of a binary classifier that assigned an integer to one of two classes; a positive class for the number of trial participants, and a negative class for all the other candidate numbers. The RCT articles are transformed into data ready to be used for classification by selecting features based on the analysis of how the number of trial participants appeared in RCT abstracts (Table 9). Hansen et al. (2008) recognised a feature only if the number is an integer and it is not followed by a unit that is being defined in a closed set of words. Supervised classification was then performed with SVM with its input based on a set of feature vectors

(features with *). 56 out of a total of 94 features formed the basis for evaluating additional ones with 38 added to the feature set after their testing with the training set. Additionally, they performed post-processing in order to reduce the number of false positives by selecting only one number per abstract (one with the maximum value). They reported an F-score of 84.0% .

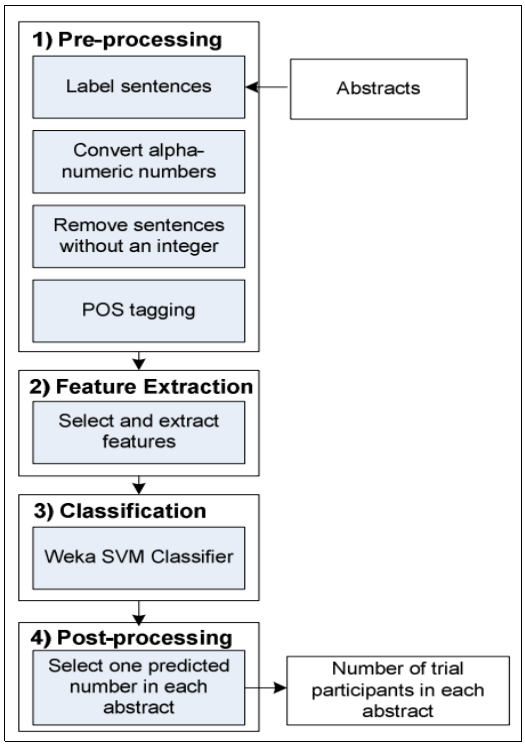


Figure 13: A machine learning approach for the recognition of participant number in clinical trials (Hansen et al. 2008).

Table 9: 14 different types of features used for classification (a total of 94 features). The (*) indicates features that form the basis for evaluation of the additional applied features (Hansen et al. 2008) (continuing on next page).

evaluation features	feature
(*)	int > 10
(*)	len(int) < 5
	otherIntegers
(*)	patientGroup
	patientWindow
(*)	n=
	patientGroupBeforeN

Table 9: 14 different types of features used for classification (a total of 94 features). The (*) indicates features that form the basis for evaluation of the additional applied features (Hansen et al. 2008).

	POSbefore
(*)	POSafter
(*)	wordBefore
	wordAfter
	verbInSentence
	sentenceNumber
	label

Other related work has focused on identification of sentences from RCT articles. Chung (2009a) implemented a machine learning approach (CRF) on RCT journal articles for the automatic categorization of sentences that refer to intervention, participants and outcome measures. An extension of a previous approach in which the sentences are labelled under four roles (aim, method, results and conclusion) is used. More specifically, for labeling with intervention, participation and outcome measure, a first order linear-chain CRF is built for each problem. This results to five states where one state represents the label in question and the rest the four rhetorical roles. Therefore, the vector for each state is derived from the observed sentence data and their syntactic features and the ordering of the intervention, outcome measures and participants sentences are modelled by the states in relation to the rhetorical roles. Chung reported an F-score of 83.0%, 84.0% and 52.0% for the recognition of sentences that describe intervention, outcome measures and participants respectively.

In further work, Chung (2009b) discussed and implemented techniques for the recognition of experimenting details of RCTs (Figure 14). The author focused on the automatic recognition of coordinating constructions that describe intervention arms (therapies) in MEDLINE RCT reports that could assist in the exact identification of intervention trial arms. Two or more constituents of coordinated phrases, linked by coordinating conjunctions, prevalent in the expression of intervention comparisons. Through the implementation of a full sentence parser, verb, subject and object information is recognised in sentences. With the addition of linguistic rules, the syntactic structure is converted into a set of predicate argument structures – normalized forms representing syntactic relations. An example of the parser tree can be seen in Figure 15. All the coordinating sentences related to each sentence are extracted and each coordinating phrase is identified individually from the parse tree. The study was aiming to identify the pharmaceutical interventions and non-pharmaceutical ones were ignored. The P, R and F-score were respectively 88.0%, 81.0% and 84.0%.

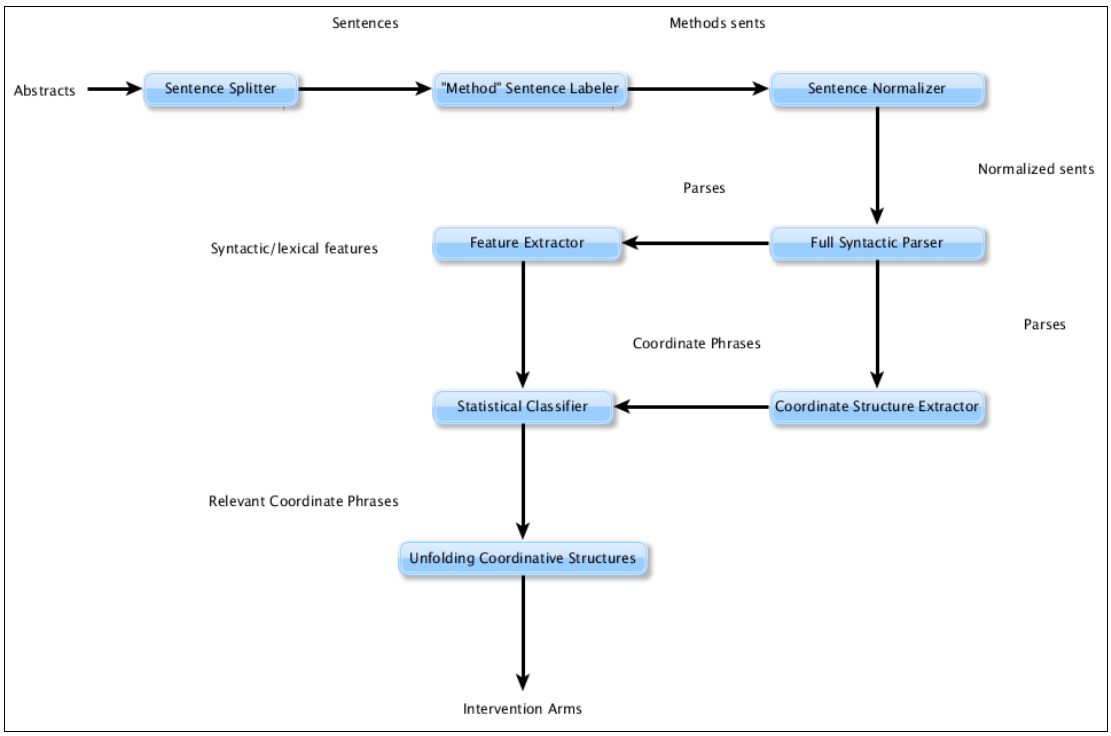


Figure 14: The architecture system proposed by Chung (2009b) for the identification of coordinating constructions relevant to intervention arms of RCTs.

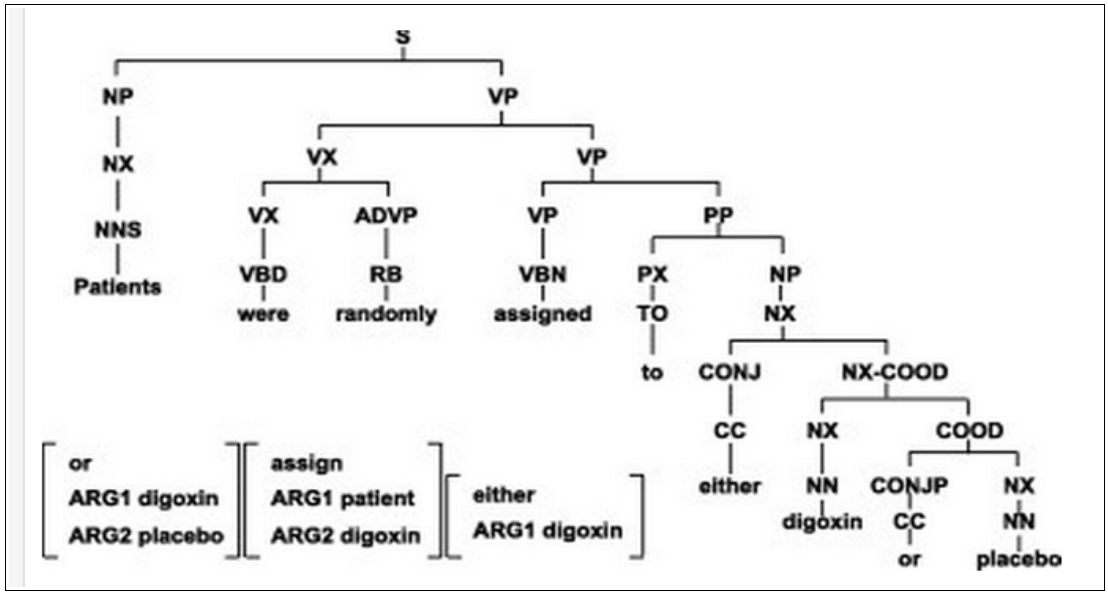


Figure 15: An example of the parse tree for the following sentence: "*Patients were randomly assigned to either digoxin or placebo*" (Chung 2009b).

Recently, while this thesis was in its late stages, Luo et al. (2011) reported on an approach to identify temporal constraints from randomly selected eligibility criteria of clinical trials with a machine learning approach. More specifically, they used a semantic lexicon to extract UMLS clinical terms in the eligibility criteria free-text. Because the performance on eligibility criteria in previous studies was relatively low, it is interesting that Luo et al. (2011) tried to improve their detection in clinical trial text. An annotator was applied for the generation of the machine learning features, trained with a parser that was developed in order to extract temporal elements of interest according to an ontology model of temporal constraints. They used CRFs to train the parser from manually annotated eligibility criteria and defined three types of features:

- words/terms themselves identified by a semantic annotator aiming to recognise UMLS concepts,
- common time-related terms such as numerals, months, week days and time units highlighted by a rule based feature identification program using the Context Scanning Strategy,
- contextual information in which they used for any given term its previous features and successor term's features as context.

They reported a mean F-score of 80.0% from the returned F-scores of all classes (Table 10). Figure 16 reveals the architecture of the system used for the identification of temporal constraints from clinical trial eligibility criteria.

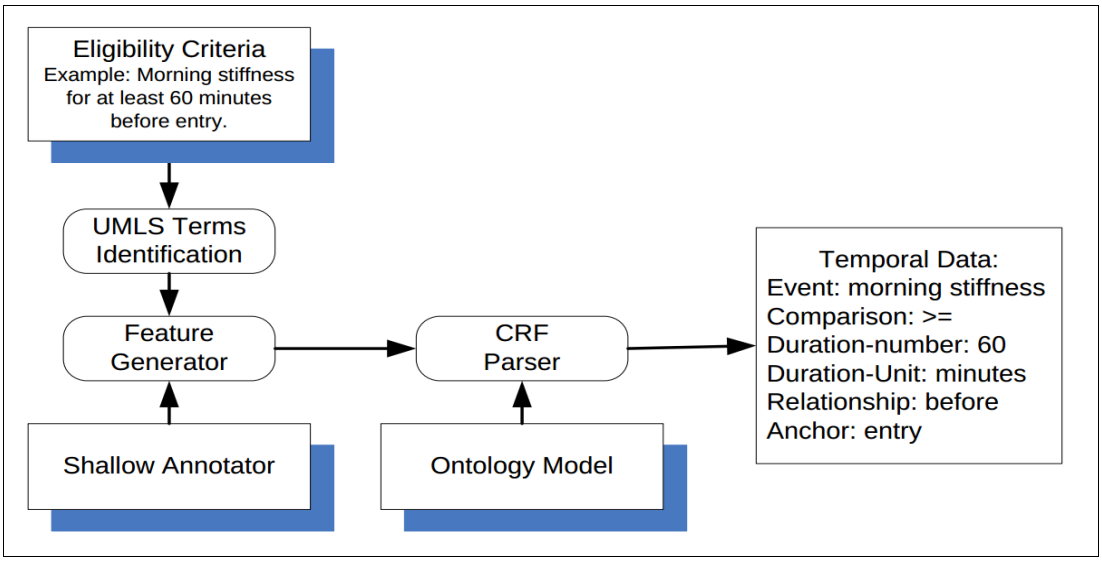


Figure 16: An overview of the system used for the recognition of temporal constraints in eligibility criteria.

Table 10: Evaluation results for the identification of temporal constraints in clinical trial eligibility criteria (Luo et al. 2011).

Temporal Elements	Precision	Recall	F-Score
DURATION-UNIT	100.00%	100.00%	100.00%
DATE(CALENDAR)	100.00%	100.00%	100.00%
DURATION-COMPARISON	96.97%	94.12%	95.52%
DURING	100.00%	91.30%	95.45%
DURATION-NUMBER	98.31%	90.63%	94.31%
AFTER	100.00%	82.76%	90.57%
BEFORE	100.00%	81.82%	90.00%
EVENT	85.30%	79.07%	82.07%
FREQUENCY-RECURRENCE	100.00%	66.67%	80.00%
CONJUNCTION	73.68%	58.33%	65.12%
MODIFIER	54.17%	50.00%	52.00%
ANCHOR	40.00%	58.82%	47.62%
DURATION	32.35%	73.33%	44.90%
Average	**83.14%**	**78.99%**	**79.81%**

Additionally, Luo et al. (2012) aimed to recognise Common Data Elements (CDEs) in eligibility criteria of multiple clinical trials related to a particular disease (breast cancer and cardiovascular disease). More specifically, clinical trials that are focusing on one disease quite frequently employ common variables that determine the patient's eligibility e.g., for diabetes trials, blood glucose levels are defined as inclusion criteria. These variables are referring to CDEs. They implemented a semi-automatic approach that identifies eligibility criteria as a set of UMLS-recognizable terms; however, if there is mapping to more than one UMLS term, then the one that works best in the context of the clinical trials is selected. An association rule-learning algorithm is applied in order to discover CDEs that are associated with each specific disease. They have defined association rules as links between a head X and a body Y implicating X=> Y with X being the set of terms that is representing the patient characteristics in disease Y trials. The Apriori algorithm was applied outputing a set of rules that reports how often CDEs are present. Due to the large number of results, a filtering process is occurring according to preferred UMLS semantic types in order to improve information relevance by manual selection of a total of 48 semantic classes used as groups and filters for CDEs. Since the results are randomly ranked, the CDEs are grouped by string similarity (through Dice Coefficient) to measure the similarity between pairs of strings. An overview of the proposed approach can be seen in Figure 17. The average F-score was 81.0% with precision and recall 82.3% and 79.7% respectively.

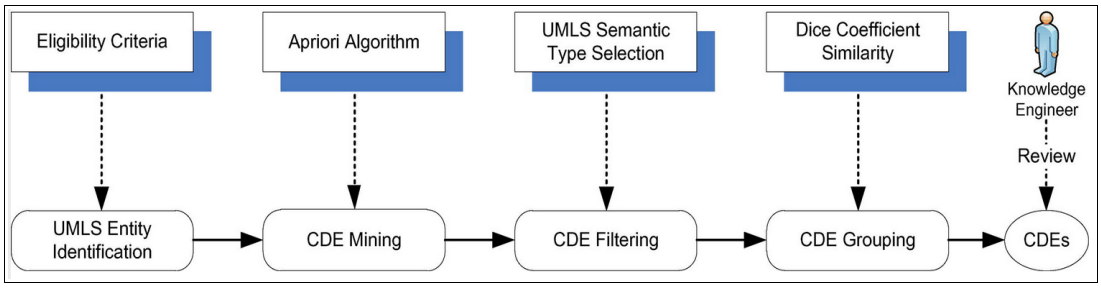


Figure 17: The semi-automated approach of Luo et al. (2012) for the recognition of CDEs from clinical trial eligibility criteria.

Epidemiological Characteristics Extraction

Research work on the identification of key epidemiological characteristics in text included the extraction of risk factors and exposure terms from MEDLINE citations and titles of epidemiological articles respectively. Particularly, Fizman et al. (2007) focused on the recognition of risk factors for metabolic syndrome in MEDLINE citations. More specifically, they used existing semantic processors, specifically SemRep – to recognise semantic predications in biomedical text, and automatically identified risk factors for metabolic syndrome and its predisposed by it diseases. Through the PubMed query with the MeSH heading "*Risk Factors*", indicator rules for disease risk factors were found. Sentences that included disease risk factors were isolated for linguistic analysis. Hence providing the syntactic patterns for the basis of additional indicator rules in SemRep (32 of total rules were assimilated). This step was followed by analysis of the risk factor sentences with the addition of semantic types as subject semantic arguments of the predication. They reported 67.0% precision and 53.0% recall. Figure 18 shows patterns for indication rules.

a. RiskF {be}*risk factor for/of* Disorder
b. RiskF *risk for/of* Disorder
c. RiskF *predict* Disorder
d. RiskF *marker of* Disorder
e. RiskF *determinant of/for* Disorder
f. RiskF *contribute* Disorder
g. RiskF *promote* Disorder

Figure 18: Patterns for indicator rules for disease risk factors (Fizman et al. 2007).

Xu et al. (2010) aimed to extract exposure-related terms (factors that can affect and alter the health status of an individual) as noun phrases from titles of epidemiological articles rather than full text or abstracts in the American Journal of Epidemiology through the implementation of an existing biomedical NLP system, the KnowledgeMap Concept Indexer (KMCI) and a score based classifier. The combination of the classifier and the KMCI generated the Detection of Epidemiological Exposures from Literature system (DEEL) (Figure 19).

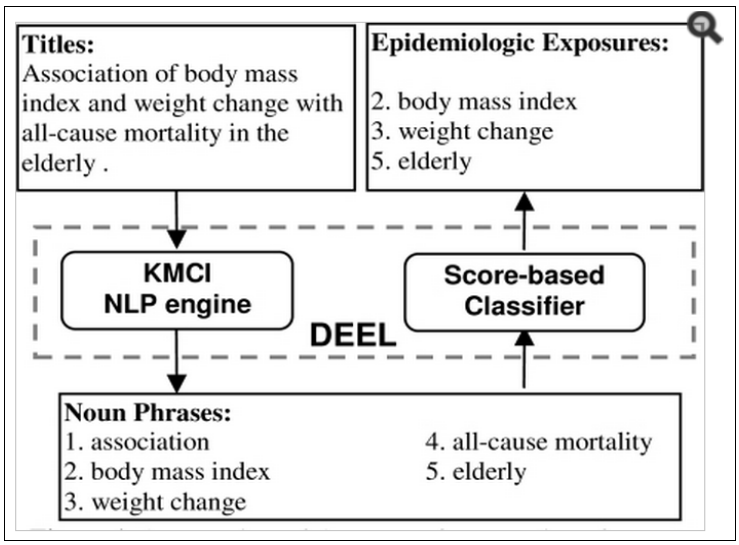


Figure 19: An overview of the system's processing flow for the identification of exposures from literature (Xu et al. 2010).

KMCI identifies findings (noun phrases) from biomedical documents based on regular expressions that implement POS tags and encodes them into UMLS concepts. They developed a score based classifier that applies heuristic rules (30 in total) and selects the appropriate terms that contain epidemiological exposures. To determine if a highlighted noun phrase is an exposure or not, the classifier considers the evidence from the phrase itself i.e., UMLS semantic type or from the context around it such as any neighbouring words and syntactic patterns (see Figure 20 with example). However, this approach was based and developed only on titles coming from one journal. The reported precision and recall for this method were 61% and 69% respectively.

Example: "Association of body mass index and weight change with all-cause mortality in the elderly."

Noun Phrases	KMCI identified Semantic Type
Association	Mental Process
body mass index	Diagnostic Procedure
weight change	Finding
all-cause mortality	Finding
elderly	Age Group

Examples of Scoring Rules:

1. IF NP is "Finding", THEN score -1
2. IF NP is "Diagnostic Procedure", THEN score +2
3. IF NP is "Age Group", THEN score +3
4. IF NP matches the pattern "association of ... NP ... with Finding/Disease or Symptom", THEN score +2
5. IF NP matches the pattern "association of ...with NP", THEN score -2

Noun Phrases	Rules: 1	2	3	4	5	Total
Association						0
body mass index	+2			+2		+4
weight change	-1			+2		+1
all-cause mortality	-1				-2	-3
elderly			+3			+3

Figure 20: An example set of scoring rules for DEEL. It demonstrates how the final score is calculated for candidate epidemiological exposure noun phrases for the example sentence "*Association of body mass index and weight change with all-cause mortality in the elderly*".

Additionally, Blake (2004) applied a text mining approach for the extraction of candidate risk factors of breast cancer. The author developed a prototype system (Multi-User Extraction for Information Synthesis – METIS) that identifies 19 facts (minimum and maximum age of the affected individuals, location and time frame of the study and risk factor exposure among others) from MEDLINE breast cancer scientific articles. These facts are used as parameters to an external database in order to estimate a control population. Through a randomized meta analysis, the rate of the exposure of each article is compared with the estimated population control group leading to the generation of the related effect size and to a graphical result summarization.

2.3. Knowledge Representation and Concept Maps

"You can't depend on your eyes when your imagination is out of focus"
Mark Twain, 1889

A variety of definitions has been given to knowledge, but generally it can be defined as information with purpose (Canas et al. 2005). The primary elements of knowledge are concepts and their relationships (Novak 2008). New knowledge is being constructed with the use of new propositions between connected concepts with a linking relationship. This is called a semantic unit (Canas et al. 2005). Knowledge representation is defined as the use of formal representations to transfer knowledge between at least two individuals (de la Villa et al. 2012). Several knowledge representation (KR) approaches have been developed, including concept maps, semantic networks, ontologies, E/R diagrams and mind maps. We first focus on the definition and description of concept maps and then compare them to other systems.

2.3.1. Definition and Aim of Concept Maps

The term "*concept map*" (CM) has been associated to a wide range of diagrammatic knowledge representation models. More specifically, CM has a character of pedagogical nature and is a special type of a propositional semantic network that is flexible and oriented to humans rather than for machine's consumption (Zubrinic, 2011). It is essentially a graphical tool often used in learning that organizes and represents knowledge of learners' conceptual understanding of information in a specific area (Novak et al. 2008; Kinchin et al. 2010; Lin et al. 2010). Particularly, a concept map is defined as a two dimensional display of knowledge comprised of concepts connected by directed arcs that encode relationships between the concept pairs. It promotes the visualization of concept relationships through connecting lines and intersecting figures and it may include flowcharts, timelines and tables (Nesbit et al. 2006; Safayeni et al. 2010). Therefore, it can be seen as a graphical tool that enables the users to express (their) knowledge in a comprehensible form, less complex than it is presented in other KR models e.g., ontologies (Canas et al. 2005; Wafula, 2006). Additionally, individual concepts can be linked to various types of resources (e.g., pdf documents or databases) in order to complement the information represented by the respective concepts (Willemsen et al. 2008).

Concept maps were developed by Joseph D. Novak in 1972 at Cornell University aiming to follow and understand any changes in the student's knowledge in the field of science. He based his work on the theories of David Ausubel, who underlined the significance to learn new concepts from already existing knowledge (Plotnick, 1997). Concept maps have been already used as a cognitive tool back in the early 1980's for representing and tracking the students'

concept understanding and any possible changes in the students' understanding respectively. This led to the creation of cognitive maps as a practical means for organising graphically the concept structure (Dabbagh, 2001). This new visual tool was used by several studies in order to prove its assistance in the field of education (Hall et al. 1992; Lambiotte et al. 1992; Soyibo, 1995; Markow et al. 1998; Kinchin, 2000). In addition, concept maps proved beneficial in knowledge representation issues and helpful in the organization and understanding of new subject matter (Dabbagh, 2001). Therefore, their application started expanding in other areas such as expert knowledge representation, initiating their wide use (Gaines et al. 1995).

A concept map is constituted from concept nodes and links (Lin, 2010).

1. Concepts have been defined by Novak as *"perceived regularities in events or objects, or records of events or objects, designated by labels"* (Canas et al. 2004; Kinchin et al. 2010). The nodes of a concept map can contain a variety of concept formats such as words that tend to be nouns or noun phrases and even images, usually visually enclosed in cycles, squares and various other structures (Novak et al. 2008; Zubrinic, 2011; de la Villa et al. 2012; Zubrinic et al. 2012). It is possible for a concept map to have clusters or segments of concepts grouped together that express a particular aspect of the concept map.
2. Proposition is the connection between two concepts with a linking phrase. This is called a "*unit of meaning*" or a "*meaningful statement*" (Figure 21) (Safayeni et al. 2010; Zubrinic 2011; Zubrinic et al. 2012). Propositions can include two or more concepts in order to form a meaningful statement (Kinchin et al. 2010). Propositions (or links) are unique characteristics that differentiate the concept map from other similar graphical tools such as mind maps (Canas et al. 2004; Safayeni et al. 2010). The links represent the relationships between the nodes. They can be non, uni or bi-directional (Novak et al. 2008). Although they can be labeled in order to provide sufficient explanation for the nature of these relationships, it is not obligatory (Novak et al. 2008). The linking phrases can express any type of relationship and they are not restricted to a defined set like in other representation techniques such as semantic networks (Canas et al. 2005). For example, links can be describing casual or temporal relationships between concepts (Plotnick, 1997). A concept map can also include cross links, relationships between concepts situated in the different segments of the concept map[15]. CMs can have specific instances of concepts that may help in the clarification of a complex concept meaning (Novak et al. 2008). An example of a concept map with directional and non-directional links between its nodes for the diagnosis and the treatment of a patient can be seen in Figure 22.

15 http://www.shiftn.com/obesity/Full-Map.html

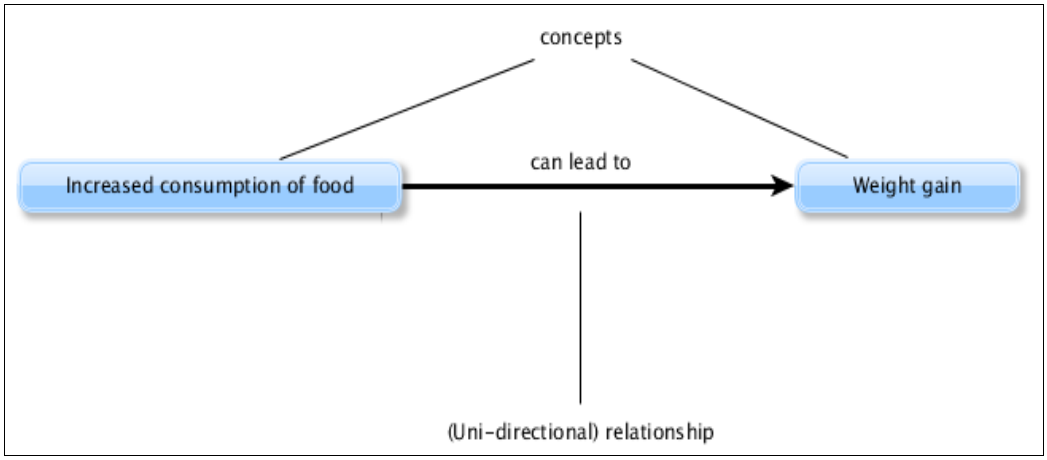


Figure 21: A unit of meaning example. Two nodes (concepts) are forming the unit of meaning in a concept map. This unit of meaning is read by following the direction of the relationship between the two concepts as "*increased consumption of food*" leads to "*weight gain*".

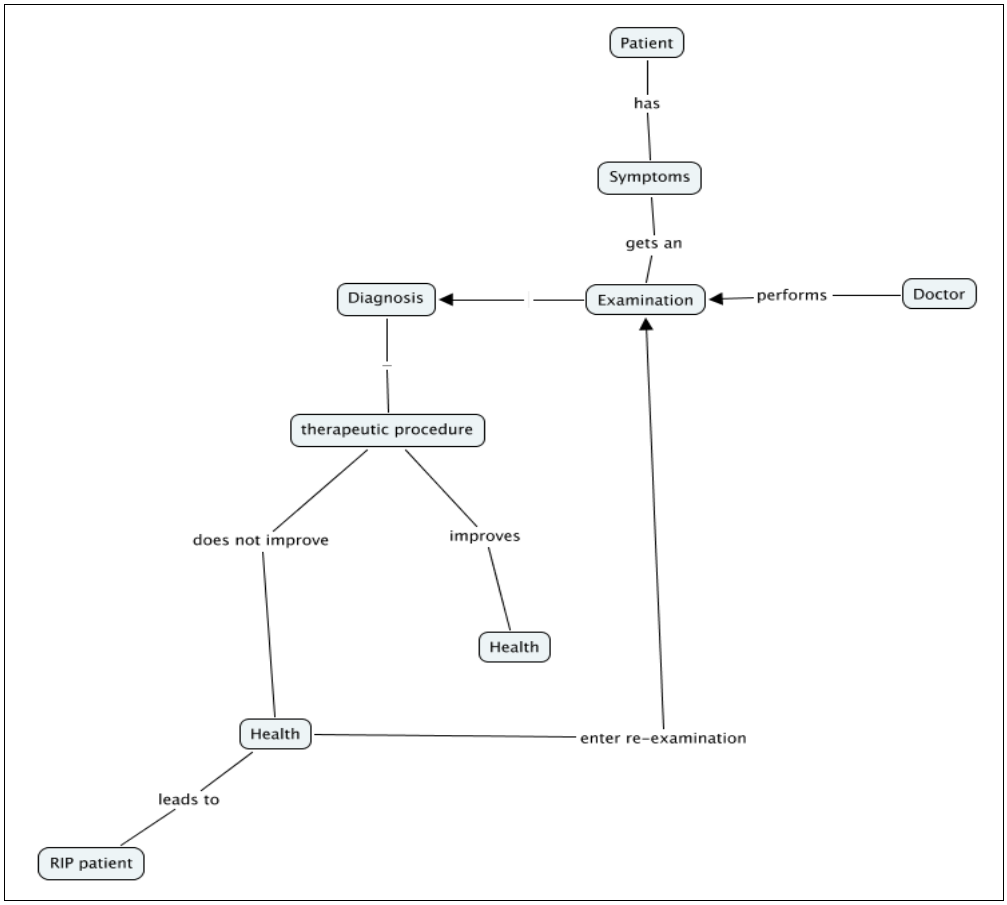


Figure 22: A concept map example. The map includes the diagnosis and treatment outcome of a patient. It contains directional and non-directional links representing the relationships between its concepts. No instances are present in this map.

Concept maps promote meaningful learning. More precisely, a concept map gathers, shares and explores information around one or more topics or problems that a user may be interested to understand (Lin, 2010; GraphicOrg, 2013). Since its role is mainly the visualization of data and information in a simple form, the focus lays on the existing relationships among the concepts and the discovery of new ones. A concept map can be used to generate, communicate and spread pioneering ideas and views (brainstorming). Through this process, previously unknown concepts may be introduced and taken under consideration, enabling the alteration of the perspectives a problem or an issue could be examined under (Lanzing, 1997).

2.3.2. Designing Concept Map

Concept maps are generally similar to an organization chart or flow diagram. Their topology can take a variety of forms that range from hierarchical to non-hierarchical (e.g., brainstorming) (Zubrinic, 2011). The more common between these two is the hierarchical approach (SmallStock, 2013; de la Villa et al. 2012). Novak emphasized the importance of hierarchy because keeping concept maps hierarchical with a single central node, makes it easier for the learner to grasp how concept maps are constructed. Its designing starts from the top and then it expands towards the bottom, with the most general/inclusive concepts placed higher while the generality is progressively reduced at the lower levels (Safayeni et al. 2010). This approach has always a root node that is placed at the top of the map containing the main theme or problem (Villalon et al. 2008). It is the most useful form in terms of teaching and learning (Novak et al. 2008). Concept maps (including their propositions) generally are "*read*" from top to bottom (Canas et al. 2004; Safayeni et al. 2010).

An hierarchical concept map can be defined as (Villalon et al. 2008; Zubrinic et al. 2012) a set CM = {C, R, T} where:

1. $C=\{c_0,c_1,...,c_{n-1}\}$ is a set of concepts. Each concept $c_i \in C$; $0 \le i \le n$ is a word or phrase and is unique in C.

2. $R=\{r_0,r_1,...,r_{m-1}\}$ is a set of relationships among concepts. Each relationship $r_j \in R = (c_p,c_q,l_j)$; $p \neq q$; $0 \le p < n$; $0 \le q \le n$; $0 \le i < m$, connects two concepts, $c_p,c_q \in C$. Label l_j is a term that labels relationship r_j and represents a conceptual relationship between coupled concepts.

3. $T = \{t_0,t_1,...,t_{s-1}\}$; $t_{k-1} < t_k < t_{k+1}$; $0 < k < s-1$ is a sorted set of hierarchical levels in a CM. Each element $t_k \in T = \{c_0,c_1,...,c_{n-1}\}$; $0 \le r < n$ corresponds to a set of concepts that share the same level of generalization in a CM.

In the second approach, a concept map can be designed with the root node placed in the centre and then the addition of concepts and propositions expand outwards. This approach is called brainstorming. It is useful when groups of people work together on a specific problem or theme. They begin generating ideas resulting in the creation of a concept map (SmallStock, 2013). Here, the arrows do not follow a specific structure. They rather exist to connect and reveal the special relationships between various concepts nodes (Novak et al. 2008).

Certain characteristics, however, indicate a good structure in a concept map (Canas et al. 2005). First, concepts and linking phrases should be as short as possible while the structure should be hierarchical. The root node of the map usually should be at the top of the map, being the most generic concept, which directly represents the theme (or problem) of the concept map. Additionally, two concepts connected through a linking phrase should form a standalone proposition that can be read independently of the map while maintaining its meaning (Canas et al. 2005). Every concept depends on another through one or more relationships with other concepts. Thus, there is no concept disconnected from others. It should be noted that the number of ingoing and outgoing links of a concept may provide additional information regarding a concept's role in the map (Canas et al. 2005). Concept maps are highly idiosyncratic, thus those that have similar or same topics will vary to each other due to different perspectives and opinions (Canas et al. 2005; Wafula, 2006; Novak, 2008).

The structure of a concept map depends on its context. CMs' strength relies heavily on the ability to measure a person's knowledge about a certain theme, main idea or domain (Canas et al. 2005). If the opposite occurs, then it violates the nature of the concept map that involves knowledge exploration and not (completely) faithful organization and representation.

In order to create properly a concept map there are certain steps that need to be followed (Lanzing, 1997; Hammarlund et al. 2011):

1. A selection of the main theme(s) along with the identification of related key words or phrases. People involved in the designing process of the concept map should be aware of the central theme or problem for which the map needs to be created and implemented (Lanzing, 1997). Also, they must be able to understand if this problem can be represented or should be included in the concept map as a single concept or through the application of many concepts (GraphicOrg, 2013).
2. Figuring out exactly the words and questions associated with the particular topic including the reason why and how these ideas have been chosen and related to. Any irrelevant concepts may lead to presentation of misinformation and (possibly) wrong conclusions (GraphicOrg, 2013).

3. The concepts should be ranked from the most general to the most specific ones and along with their relationships they should be sorted according to their conceptual similarities.
4. Concepts should be clustered (into segments) based on the similar level of abstraction and close association (Novak et al. 2008; Hammarlund et al. 2011).
5. Concepts should be diagrammatic represented with the addition of possible links and labels in order to enable potential exploration of the represented knowledge (Hammarlund et al. 2011).

Most of the design of concept maps is being done manually despite the existence of various computer-based tools. The actual input still relies solely on expert's contribution (Chen et al. 2008).

2.3.3. Applications of Concept Maps

Concept maps have been often applied to education and research. There has been an increasing use of them over the past decade, particularly in the field of medical education and they have been characterised as an effective learning tool for clinical professionals and other field experts (Torre et al. 2007; Zubrinic, 2011). Their use has started expanding in various other areas such as Artificial Intelligence, Philosophy, Biology and Management (Lambiotte et al. 1992; Markham, 1994; Gaines et al. 1995; Schmidt, 2004; Burke et al. 2005; Watson et al. 2005; Willemsen et al. 2008; Kinchin et al. 2010; 'Miller et al. 2012). Particularly, researchers have been interested in the application of concept maps as a representation means for creativity, hypertext design, learning and evaluation of knowledge (Soyibo, 1995; Plotnick, 1997; Markow et al. 1998; Kinchin 2000; GraphicOrg, 2013; Safayeni et al. 2010). Various concept map designing tools exist and are available for use (see Appendix B).

Concepts maps are mainly used for instruction and learning as well as knowledge organization tools exploring changes in meaning frameworks (Novak, 1990; Lambiotte et al. 1992; Canas et al. 1999). Also, they allow domain experts to quickly evaluate represented parts (Willemsen et al. 2008). They are considered by educational psychologists as an important tool to enhance learning by understanding knowledge and using it effectively (Oliveira et al. 2001; Chen et al. 2008). They assist students in concept understanding since they are being asked to actively construct concept maps based on certain topics (Novak, 1990; Wafula, 2006). It has been shown that knowledge organization can have a positive influence in the efficiency and effectiveness of problem solving around a certain domain, especially if it is of biomedical nature (Schmidt, 2004). Several studies suggest a positive effect of concept map application in (especially medical) education labeling them as a valid and reliable medium to represent

student's understanding of complex knowledge as well as to describe and interpret information (Lambiotte et al. 1992; Soyibo, 1995; Plotnick 1997; Markow et al. 1998; McClure et al. 1999; Kinchin, 2000; Burke et al. 2005; Watson et al. 2005; Torre et al. 2007; Gonzalez et al. 2008; Molaison et al. 2008; Willemsen et al. 2008; Villano et al. 2008; Kinchin et al. 2010; Miller et al. 2012). Through the application of concept maps in the learning process, the learner is required to pay attention to the concept relationships since new knowledge can be integrated into existing one (Plotnick, 1997).

Concept maps provide an alternative to natural language as a means of communication (Gaines et al. 1995). Their implementation in education can produce multiple benefits such as:

- **Comprehensive Summarization & Exploration:** When the choice of concepts and linking words is being performed carefully, concept maps can become a powerful tool for observing different aspects of meaning while at the same time investigating new ones (Canas et al. 2005). Concept maps have the ability of bringing together the various meanings and relationships between a wide range of ideas, therefore making the process of learning and exchanging information more active rather than passive (Novak et al. 2008).
- **Creative Thinking & Meaningful Learning:** Concept maps facilitate and sharpen creative thinking through time as well as promoting meaningful learning by enabling the user to make relationships between concepts (Wafula, 2006; Chen et al. 2008; Gonzalez et al. 2008; Molaison et al. 2008). Visual symbols and diagrams can be recognized quicker and more easier than just plain text. They allow the development of a holistic understanding which words only by themselves can not convey (Plotnick, 1997). Additionally, any targeted analysis of concept maps can (possibly) reveal different patterns of understanding in the same topic that may also lead into meaningful thinking and learning (Kinchin et al. 2010).
- **Easy Manipulation and Navigation:** The representation of knowledge in the visual format of the concept map offers the ability to overview a domain of knowledge allowing a number of users to look through large volumes of information in a small amount of time (Plotnick, 1997; Canas et al. 1999; Nesbit et al. 2006; Zubrinic, 2011). Concept maps make economical use of text and as a result it is easy to scan for a word or phrase (Plotnick, 1997). They can easily be extended with less need for re-organization than lists and outlines (Nesbit et al. 2006).

2.3.4. Criticism of Concept Maps

While adopting concept maps as a knowledge representation model, there are issues that should be taken under consideration before and during their implementation (Schmidt, 2004):

- **Informal Design:** Although concept maps have a widespread use as knowledge elicitation tools and are considered an efficient method for generating domain knowledge models, still, it is not widely accepted as a knowledge representation tool when it is being used in its "*pure*" form as Novak describes it (Canas et al. 2004). This is due to the user's freedom in the concept and relationship selection. Consequently, the concept map is being regarded from the scientific community (particularly the Artificial intelligence one) as an interesting knowledge visualization means when some "*formalized*" changes in its structure are applied (Canas et al. 2004). Additionally, concept maps should be seen only as a simple tool that can be created and manipulated by users without necessary having any expert knowledge, whereas other more detailed knowledge representation models such as semantic networks and ontologies are not easily understood by a person outside the scientific community (see Section 2.3.7).
- **Subjectivity:** It is not clear if individuals can generate the same maps as experts when they are instructed to perform the same task without guidelines. Any produced concept map from a learner less familiar with the subject can appear incoherent, inconsistent and unreliable compared to the one of an expert (Schmidt, 2004). Hence, there is a matter of subject reliability. Additionally, more complexity in the concept map - there is a risk that concept maps will constantly gain information resulting in complex maps that are difficult to interpret - does not necessarily mean better problem solving and comprehension within a particular domain (Schmidt, 2004; Willemsen et al. 2008). It is a matter of importance for the concept maps to stay cohesive. Therefore, the use of concept maps do not enhance deep learning (Laight, 2004).

2.3.5. Concept Maps and other Knowledge Representation Methods

There are a number of knowledge representation forms besides concept maps including semantic networks, ontologies, entity-relationship (E/R) diagrams and mind maps. We researched these models in terms of their aim and the reasons they are used for, in order to comprehend their design process, the level of knowledge representation detail they include and the users they are addressed to along with their typical respective application areas.

Semantic Networks

Semantic networks are a graphic structure for representation of knowledge using interconnected nodes and arcs (Sowa, 1992). Sometimes the terms semantic network and concept map are used more or less in the same way by a certain group of experts without noting the exact difference (Hartley et al. 1997). Figure 23 gives an example of a semantic network regarding the health care pathway for a patient involving the procedures of diagnosis and treatment. The main difference between a semantic network and a concept map is in the definition of the relationship set. Relations in the semantic network are presented in a more formal way than those in concept maps (Hartley et al. 1997).

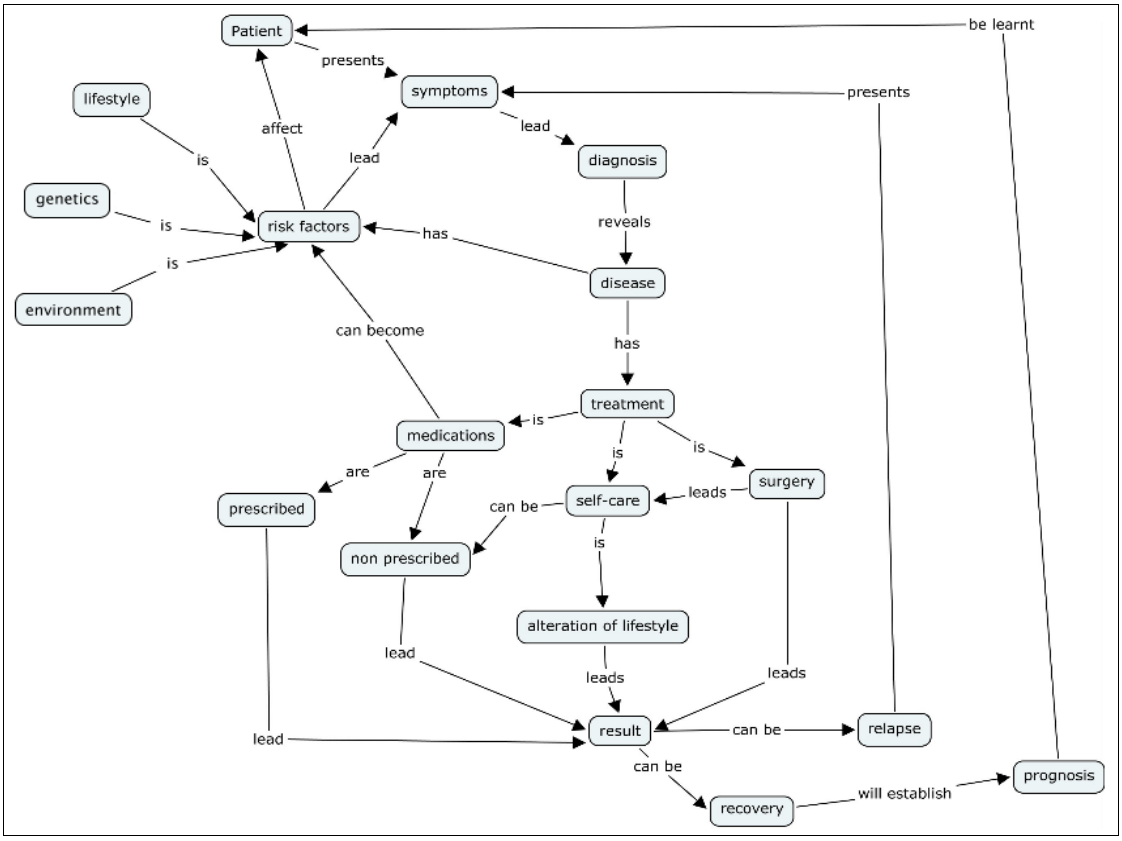


Figure 23: A semantic network example. It contains a clinical health care pathway regarding the diagnosis and treatment of a patient.

A concept map has no limit in the use of linking phrases. They can express any type of relationship and they are not restrained to a defined set, whereas semantic networks are. This freedom in the construction of linking phrases prevents the concept map to be a formal representation and this is where the role of the semantic network begins in the scientific fields (Canas et al. 2005). Additionally, a concept map visualises knowledge in an informal way with its main aim being the knowledge gathering and exploration. That explains the lack of explicit detail in the relationships among its concepts. Therefore, if a concept map has too much detailed concept relationship representation, it becomes too big and, hence too difficult to

navigate, leading to decrease of its usefulness and violating its main aim (Hartley et al. 1997). Consequently, semantic networks have been extensively used in the field of Artificial Intelligence but they have also been applied to the fields of Cognitive Psychology (Sowa, 1992), Machine Translation (Gaines et al. 1995), Philosophy and Linguistics (Harrigton, 2010; Navigli et al. 2010; Moro et al 2012). Users are mostly computer scientists due to the nature of semantic networks.

Ontologies

One of the most widely applied knowledge representation models to distribute and standardize knowledge in the last decade are ontologies, particularly after the raise of the Semantic Web (Zubrinic, 2011; de la Villa et al. 2012). An ontology is an explicit formal representation of a particular domain's concepts (or terms) as well as the relations among them (Noy, 2001; Zubrinic, 2011). An ontology provides a source of formally defined terms while aiming to deliver a commonly agreed vocabulary for a specific domain in order to share and annotate expert knowledge (Noy, 2001; Sure et al. 2002; Jarrar et al. 2002; Rebholz-Schuhman et al. 2012). An ontology conceptualizes the domain's concepts (also called classes, types or terms) and interrelationships in a declarative way (Jarrar et al. 2002). The "main" relationships are "is-a" and "part-of" and besides those, they model also domain-specific relations, for instance "*part-of*" relations between anatomical parts of an organism (Spasic et al. 2005; Rebholz-Schuhman et al. 2012). Ontologies can be considered as the reflection of the structure of a specific domain along with the constraint of potential term interpretations (Spasic et al. 2005). An ontology follows a top-down approach with the most general classes being placed on top while the most specific ones are found towards the bottom (Noy et al. 2001).

Concept mapping can be seen as a first step in ontology building (Chen et al. 2008). However, there is a vast difference between these two. Ontologies are more formal and expressive with attributes, values and restrictions (Zubrinic, 2011). Although both are used for knowledge representation in a variety of areas, in ontologies there are various components besides the existence of nodes (classes) and relationships (W3C, 2013). Slots (attributes of the classes), facets (slot restrictions), subclasses and instances are all parts of an ontology (Noy et al. 2001). On the other hand, a concept map does not require concept attributes and can only be described through the presence of the concepts and their relationships. However, the similarities between ontologies and concept maps is in the hierarchical relationships among classes (de la Villa et al. 2012).

In contrast to the users of ontologies that are researchers, experts (or professionals) in a particular domain, since concept maps are "informal" representations, they are meant to communicate knowledge between individuals and not necessarily experts or machines.

Therefore they are not sufficiently and formally specified to be used as a common vocabulary by experts and automated reasoners (Reichherzer et al. 2005). Ontologies have been criticized because they scale poorly and when they are large enough to capture most of the possible relationships among concepts, they become unmaintainable (Dickman, 2003). Ontology-based applications are seen in Artificial Intelligence, Computer Science, Semantic Web, Software Engineering, Biomedical Informatics, Library and Information Sciences and their main use varies depending the field in which they are applied to e.g., decision support, data integration, knowledge management (Noy et al. 2001; de la Villa et al. 2012).

Most of the biomedical ontologies are formalized in the OBO flatfile format as well as the Web Ontology Language (OWL) (Rebholz-Schuhman et al. 2012). Examples of biomedical ontologies are:

1. **Disease Ontology[16]:** a formal ontology of human diseases with associated medical codes;
2. **BioPAX[17]:** an ontology for the exchange and interoperability of biological pathway data;
3. **SNOMED CT (Systematic Nomenclature of Medicine Clinical Terms):** a collection of medical terms including codes, terms, synonyms and definitions applied to clinical documentation and reporting.

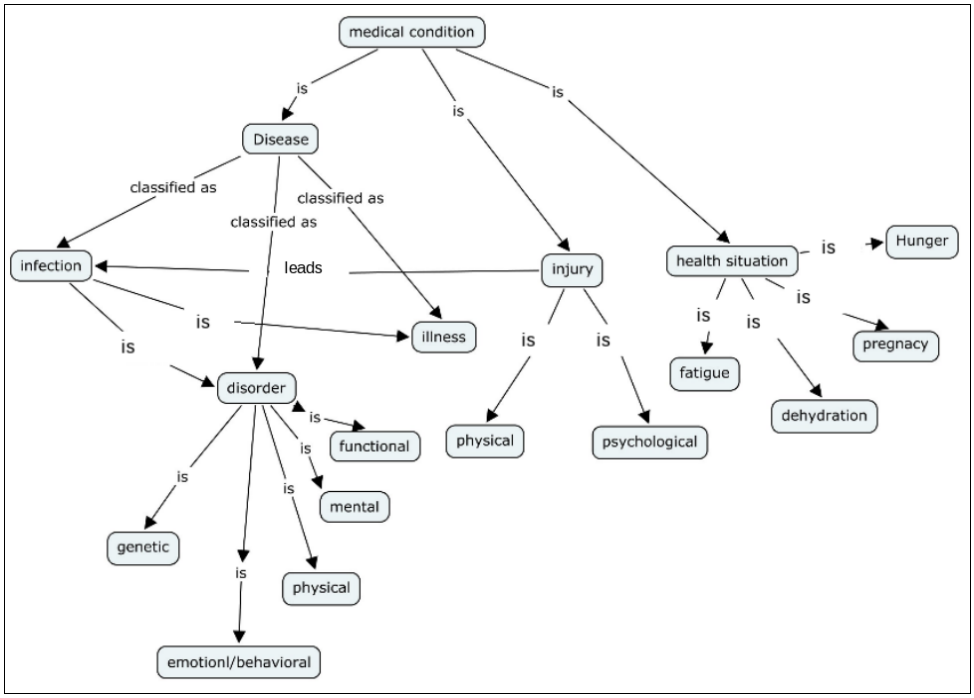


Figure 24: A manually created ontology example for disease categorization by the author.

16 http://disease-ontology.org/
17 http://www.biopax.org/

These include a wide range of biomedical concepts and information involving the type or class they belong to and how they are related. We present a (manually created by the author) ontology example regarding disease categorization can be seen in Figure 24.

Entity-Relationship Diagrams (E/R)

An entity-relationship diagram (E/R) is a specialized graph that illustrates the interrelationships between entities (Chapple, 2010). Its purpose is often to create and implement a database while its patterns help to focus on how the database works with the interactions and how the data flows. It assists in the design, optimization and debugging of database programs. Furthermore, an E/R diagram can be used as a unification of different views of data since it adopts the most natural view of the real world that is constituted from (concepts) entities and relationships (Chen, 1976; Chen, 2002). An E/R diagram has three types of symbols to represent three types of information (Chen, 1983; Lanzing, 1997; Chapple, 2010):

1. boxes for entities;
2. diamonds for relationships;
3. ovals for attributes.

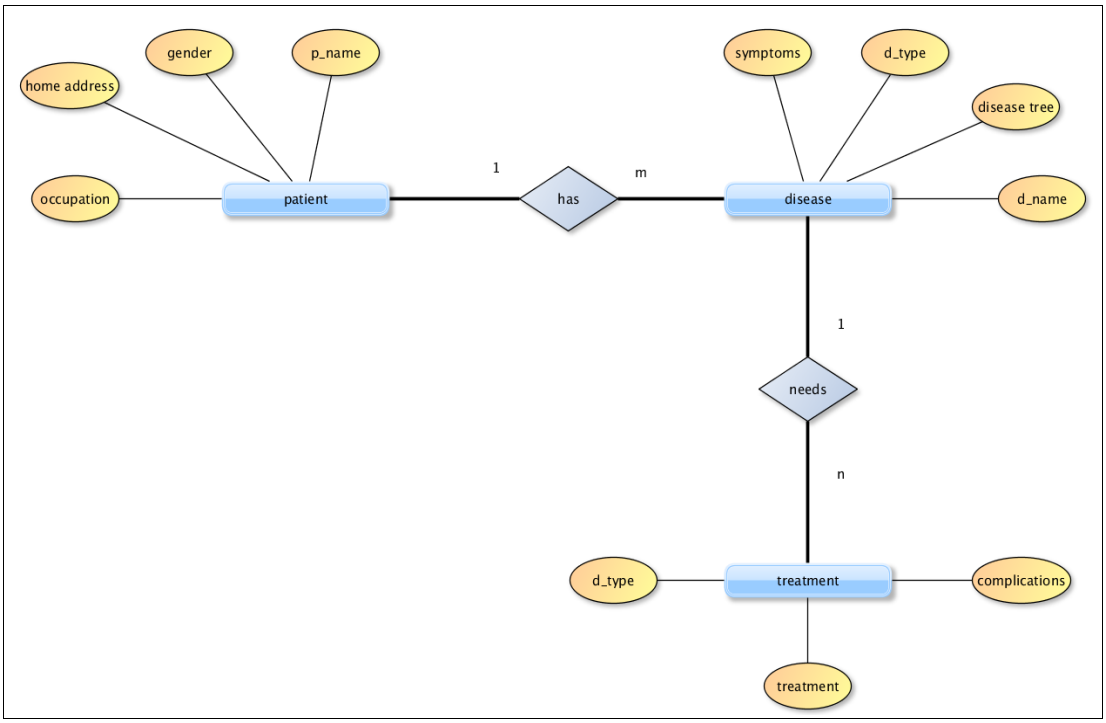


Figure 25: Example of an E-R diagram. There are three entities in this model (patient, disease, treatment). Each one has its respective attributes and there are two relationships (has, needs). The degree of the first relationship is 1:m. This means that one patient can have more than one diseases. The degree for the second relationship (needs) is 1:n. This shows that one disease can have more than treatments.

The difference between a concept map and an E/R diagram is the existence of attributes that are necessary to exist in order to define the concepts. The relationships are defined according to the number of the entities associated with. An example of an E-R diagram is seen in Figure 25.

In this type of knowledge presentation, it is important to define the degree (o*ne to many, many to one, many to many* and *one to one)* of a relationship among concepts (Chen, 1976). Since E/R model's design is based on the creation of a database, entity keys are also defined. Nevertheless, not every possible relationship is being recorded in the diagram since only the entities and the relations that are going to enter the designing process of the database are being presented (Sauter, 2000). It is up to the preference of the engineer and researcher to decide which ones to include in the diagram. The E-R model has the ability to achieve a high degree of data independence (Chen et al. 1976).

Mind Maps

A mind map is an organization and knowledge representation model, similar to that one of concept map. Mind maps were developed by Tony Buzan and he states that mind maps can be applied to every aspect of life where improved learning and cleared thinking can and will enhance human performance (Mento, 1999). Purpose of the mind map is to organize knowledge in a simple way without exploring all the possible relationships among concepts (Mind-Mapping, 2013). It is a non-linear knowledge presentation tool and it reflects visually what can be thought about a certain topic (Mento, 1999; de la Villa et al. 2012). An example of a mind map can be seen in Figure 26.

A mindmap can be used also for the support of creative thinking through brainstorming, in order to solve a problem. A mind map contains a single central word or concept, around which maximum 5 to 10 main ideas are being drawn related to that theme and for every one of these ideas, maximum 5 to 10 new ones are attached creating a tree-alike structure. Consequently, mind maps are centered around a certain topic, with all the subtopics branching from it.

Concept maps have been wrongfully considered to be the same as mind maps. Although mind maps and concept maps share common attributes such as easy understanding by second parties, there is a main difference between these two (Mind-Mapping, 2013; de la Villa et al. 2012). A concept map can have more than one theme or idea whereas a mind map has only one maintaining the level of information detail basic due to its simplicity and easy way to design (SmallStock, 2013). A mind map can be a concept map but not vice versa (Mind-Mapping, 2013).

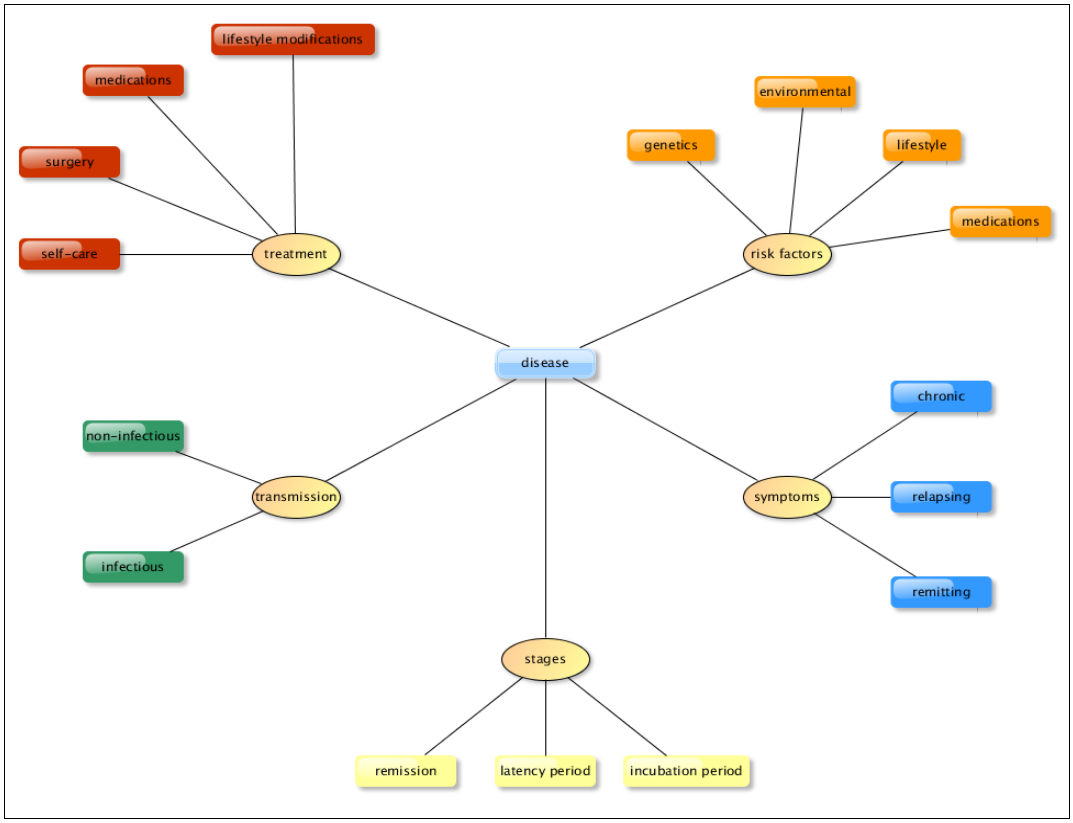


Figure 26: A mind map example for the general concept of "*disease*". "*Disease*" is the central theme of this mind map with five branches representing its key aspects (treatment, transmission, risk factors, symptoms, stages).

Mind maps are easy to create since their purpose is to help in the organization of thoughts and ideas around a certain topic and not to see any possible relationships between the objects. Mind maps are a means of note taking, creative thinking or brainstorming and outlining main ideas, hence assisting people to learn more effectively. A mind map has a tree-like structure (Mind-Mapping, 2013; SmallStock, 2013). Mind maps have been used in a wide range of fields such as those of education, business and medical education (Mento, 1999; de la Villa et al. 2012).

Figure 27 illustrates knowledge representation types in terms of detail, complexity and formality. Table 11 contains the summary from the analysis above. Particularly, the least complex KR model is mind map. As it can be seen from Figure 27, the level of complexity and detail increases towards the ontology, Additionally, concept maps be seen as a relatively formal KR method but not to the extend of ontologies or the informality that mind maps bear. Concept maps can be detailed but the more information that a map can carry, the process of reading and comprehending the map the more difficult it can be.

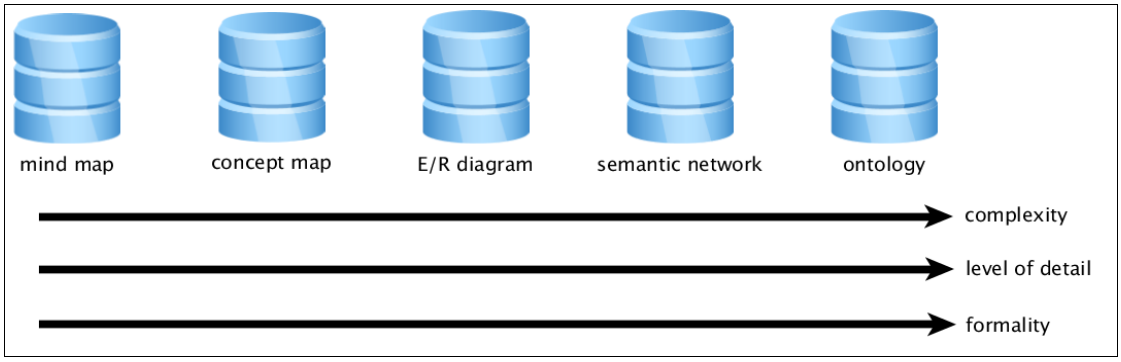


Figure 27: Organizing the five types of knowledge representation in terms of complexity, detail and formality.

Table 11: Characteristics (aim, structure, level of detail, users, area) of various knowledge representation models.

	Knowledge representation model				
	Concept Map	**Semantic Network**	**Ontology**	**E/R Diagram**	**Mind maps**
Aim	Gather, understand, explore knowledge.	Knowledge representation in systems (computer based).	Definition of a common vocabulary for researchers, experts to share information in a domain.	Contribution in databases design.	Organisation, representation of knowledge.
Structure	Nodes, links. 2 types: hierarchical or brainstorming. Use of different colours, shapes and dimensions. Network-like structure	Nodes, links. Hierarchical approach Use of different colours, shapes and dimensions. Network-like structure	Classes, subclasses, properties, facets, instances. 3 structure types: from Top to down, from down to top or combination of the two above Use of different colours, shapes and dimensions. Network-like structure	Entities, relationships, attributes. Hierarchical approach. Three types of representing symbols: boxes for entities, diamonds for relationships and ovals for attributes. Network-like structure	Main node (theme). 5 to 10 branches (nodes) around the main theme, and 5 to 10 branches from each node, etc. Use of different colours, shapes and dimensions. Tree-like structure
Level of detail	Medium detailed.	Very detailed.	Explicitly detailed.	Very detailed.	Not detailed
Users	Researchers, scientists, engineers, students.	Computer scientists.	Researchers, experts in various domains.	Researchers, engineers.	Wide range of users, mainly public.
Area	Education, Medicine, Computer Science	Artificial Intelligence, Cognitive Psychology, Machine Translation, Philosophy, Linguistics	Artificial Intelligence, Semantic Web, Software Engineering, Biomedical Informatics, Library, Information Sciences, etc.	Various fields depending where and what database will be created.	Various fields from Education to Business.

2.3.6. Concept Map Mining from Text

The generation of concept maps from documents belonging to a particular field is a process that can be performed manually, semi-automatically or automatically, although most of the efforts have focused on the manual approach (Chen et al. 2008; Lee et al. 2009; Lin, 2010; Zubrinic, 2011). The task of manually creating a concept map is difficult and time consuming (Lee et al. 2009). In addition, concept maps aiming to represent personal perspectives of a concerned subject will necessarily vary from one individual to another. Thus, semi-automatic or automatic construction of concept maps can assist in reducing this problem (Zubrinic, 2011; Zubrinic, 2012). In the semi-automatic process, the system detects and indicates some CM elements, but the user has to complete the map manually by utilizing provided information (Wafula, 2006). In the automatic construction, the user's assistance is not mandatory since the map is generated automatically from available resources (Zubrinic, 2012).

Concept map mining (CMM) is defined as the process of extracting information from one or more documents in order to automatically create a CM relevant to a particular domain or topic (Villalon et al. 2008; Lin, 2010). Its goal is the production of a CM that is an accurate, coherent and generic visual abstract of a source text. A CM can be created from a single document or from multiple ones. A form of a document can vary e.g., abstracts, full academic papers, student essays, theses, medical diagnosis reports (Zubrinic, 2011). A number of studies has focused on automatic generation of CMs (Oliveira et al. 2001; Chen et al. 2008; Zubrinic et al. 2012). They have suggested that CMM can assist in concept learning and exploration (particularly in medical/clinical education) and in the production of a starting model that can speed up the process of CM creation that later could be refined either by a person or by another automatic process (Oliveira et al. 2001; Villalon et al. 2008; Zubrinic et al. 2012).

CMM typically involves three steps (Figure 28) (Leake et al. 2008; Villalon et al. 2008; Zubrinic, 2011):

1. extraction of concepts of interest;
2. extraction of relationships that link the concepts;
3. summarization of the extracted set and creation of a subset that contains relevant concepts, relationships and topological information of the map.

However, CMM comes with a particular challenge; the lack of an objective evaluation framework. Doubts exist about how sufficient and informative a concept map generated by CMM can be and what should be the threshold for selected concepts and relationships in order

to avoid reaching sizes of a semantic network and ontology (Villalon et al. 2008; Zubrinic, 2011). The evaluation of automatically created CMs is mostly subjective and conducted by one or few experts and even then, they examine from their own perspective the source content. However, this can be reduced by including an inter-annotator agreement among skilful human evaluators who have to be familiar both with the CMM process and requirements of created CMs (Zubrinic, 2011).

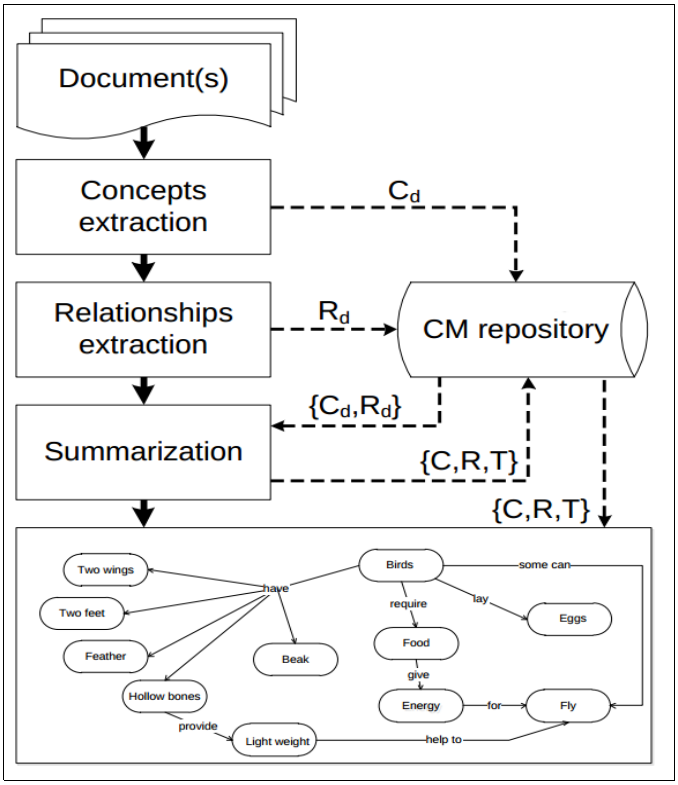


Figure 28: An overview of the general CMM steps. d refers to a document which the concept map will represent; C_d a set of all the concepts; R_d a set of all relationships that can be recognised from the document; every term used to describe a concept or a relationship must appear in the document, therefore the document itself defines all potential words or phrases that could become a part of the generated CM; C, R, T refer to concepts, relationships and the map's topology respectively after the summarization process (Villalon et al. 2008; Zubrinic, 2011).

Previous research efforts in CMM have focused on the automatic generation of concept maps from various types of text (e.g., articles (Oliveira et al. 2001), clinical notes (Watson et al. 2005), on-line documents (Leake et al. 2006)) rather than a specific type of data (e.g., epidemiological study abstracts). Their aims for creating a concept map automatically from text vary through from users being able to manipulate more efficiently large amounts of clinical data (Watson et al. 2005), to the provision of basis for educational applications

(Villanon et al. 2008) and to the representation of information in such a way that is understandable by both humans and machines (de la Villa et al. 2012). Table 12 shows an overview of studies that are associated with mining concept maps from text.

Table 12: Overview of studies in concept map mining. Studies with "-" in the fourth column did not specify the aim of the constructed concept map.

study	data	aim	concept map aim
Oliveira et al. 2001	articles, manuals, educative texts	learning and understanding from domain independent text	-
Watson et al. 2005	pathology case notes	manipulation of large amounts of clinical data	manipulation of large amounts of clinical data
Leake et al. 2006	on-line documents	generation of a preliminary version of a concept map from text	-
Chen et al. 2008	e-Learning Journal, conference articles	construction of e-Learning domain concept maps	reference for new researchers in the e-learning field, study related issues, design of adaptive learning materials, understand the whole picture of e-learning domain knowledge
Villanon et al. 2008	-	educational applications	educational applications
Bai et al. 2009	-	generate concepts maps in adaptive learning systems	-
Lee et al. 2009	-	generate concepts maps in adaptive learning systems	-
Chen et al. 2010	-	generate concepts maps in adaptive learning systems	-
Zubrinic 2011	unstructured textual resources	human quality generation of a CM from documents	-
de la Villa et al. 2012	medical term lists	sharing of different biomedical knowledge resources, assist in the representation of the information in a human and machine understandable way	representation of information in a human and machine understandable way

Villanon et al. (2008) has provided a strict definition for CMM (Figure 29) and proposed a generic evaluation framework for future studies due to its subjective nature and meaning. They suggested a gold standard creation constituted from concept maps that have been extracted and generated by human experts from textual data. A human-machine agreement will provide the quality measure of the automatically generated concept maps while the inter human agreement will provide a baseline. A difference between human and machine generated concept maps with reference to an inter-human agreement will therefore be revealed. The evaluation gold standard should be a set of concept maps created by two or more human annotators who are capable of identifying relevant propositions.

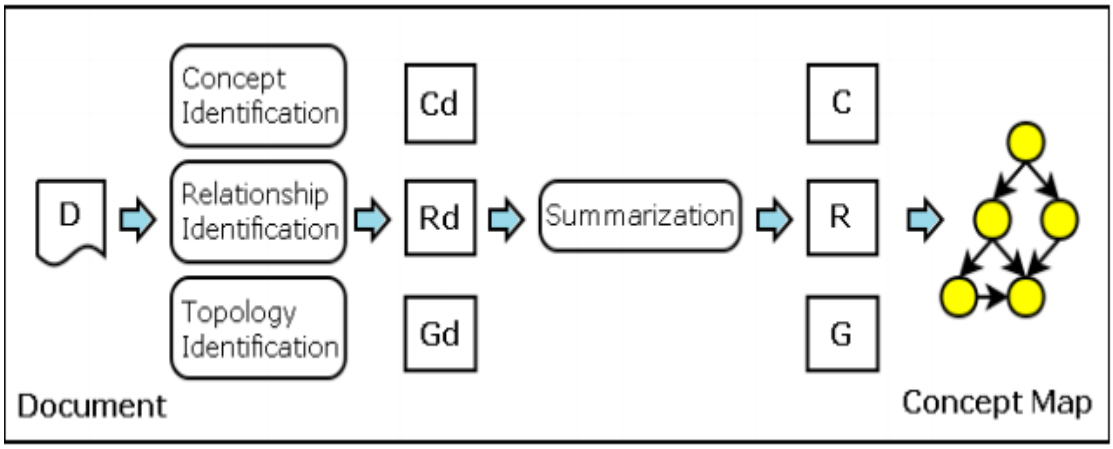


Figure 29: CMM process. Cd, Rd, and Gd refer to the identified concepts, relationships and a generalization of the set of concepts from a document D (Villalon et al. 2008).

Zubrinic (2011) proposed a CMM process from unstructured texts (Figure 30). More specifically, the proposal was made for the retrieval and preparation of the necessary data with the removal of uninformative elements and the storage of enriched clean text with semantic information. The procedure of tokenization is applied and (during linguistic pre-processing), any stopwords are removed. Each token should be associated with its lexical role using the POS tagger, and anaphora resolution could be performed on POS tagged sentences and for problematic pronouns, corresponding ones could be determined. Tokenization of longer documents however, results in a higher dimensionality of the set, which can be reduced by lemmatization and stemming. Consequently, the result of normalization is a reduced set of terms normalized to their basic forms. Through the usage of a previously created dictionary of key terms (by extracting words and phrases from a learning set of documents based on their frequency) in a particular domain, concept candidates are chosen from a set of tokens. A set of rules are created and implemented and all the terms whose base forms are found in the dictionary are directly marked as candidates. Relationship candidates are the words semantically connected to extracted concepts. With the usage of POS tagged verb phrases, the extraction of less ambiguous relationships could be possible. For each pair of concept candidates, all words positioned around their area are temporarily saved as candidates for a link and a set of rules is created for link candidate extraction from the temporary storage, based on the frequency of their appearance in the concepts area. A set of propositions is created then through the combination of normalized concepts and their respective links while the most important propositions are chosen from the set by their statistical significance.

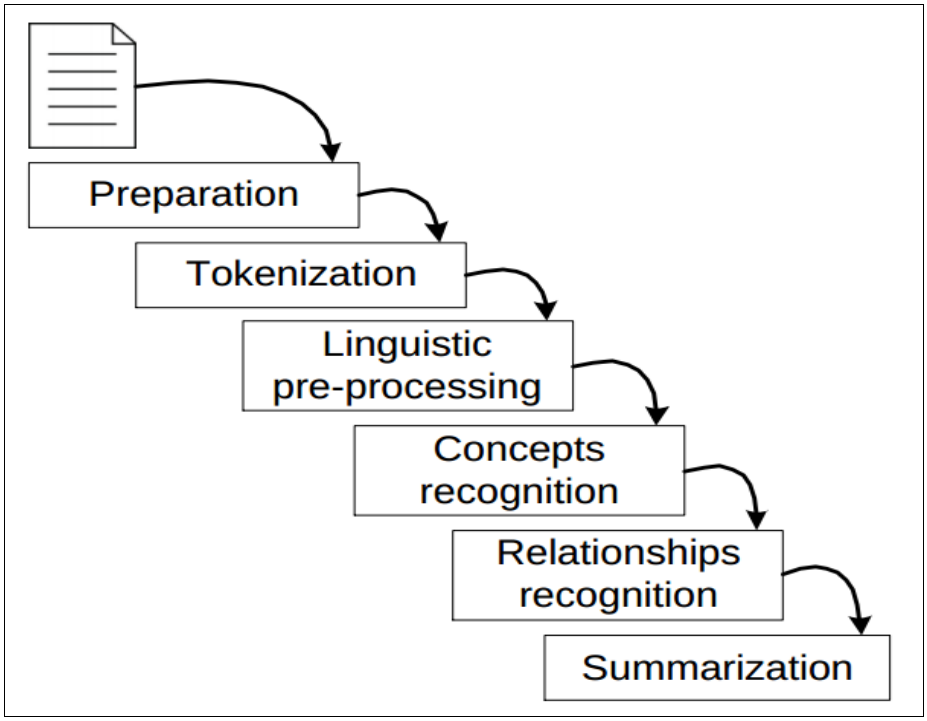


Figure 30: General CMM process applied to unstructured textual data (Zubrinic, 2011).

Oliveira et al. (2001) presented a framework that extracts and creates CMs semi-automatically from a small set (a total of 21) of various document types (articles, manuals and educative texts), which are interactively completed further by the user's input. They developed two modules named TextStorm and Clouds (see Figure 31). TextStorm is a natural language tool processing that receives text as input that is then tagged through the use of an external lexicon (WordNet) and parsed using an augmented grammar with its parameters being verbs (representing the correlation between two objects) and adjectives (representing the notion of property). They apply anaphora resolution and co-reference disambiguation module only if the concept extraction in a dependent sentence is not clear enough due to ambiguity issues. The lack of initial data in the knowledge base of TextStorm is leading to the formalization of the text concepts that depends entirely on the context where they are found. However, their approach does not include acronyms, abbreviations etc. and WordNet's use is limited in the supplement of lexical verification of words that are present in sentences. The (binary) predicate extraction is initialized by applying syntactic and discourse knowledge, and any relations between two concepts from the parsing of sentences are mapped. These predicates (examples in Table 13) are used as input in Clouds, a concept map building tool that uses machine learning inspired algorithms to complete the map through user feedback by asking questions about new possible concepts and relations that it may detects. The generated concept map (an example can be seen in Figure 32) consists of a set of binary predicates representing relations between the document's concepts.

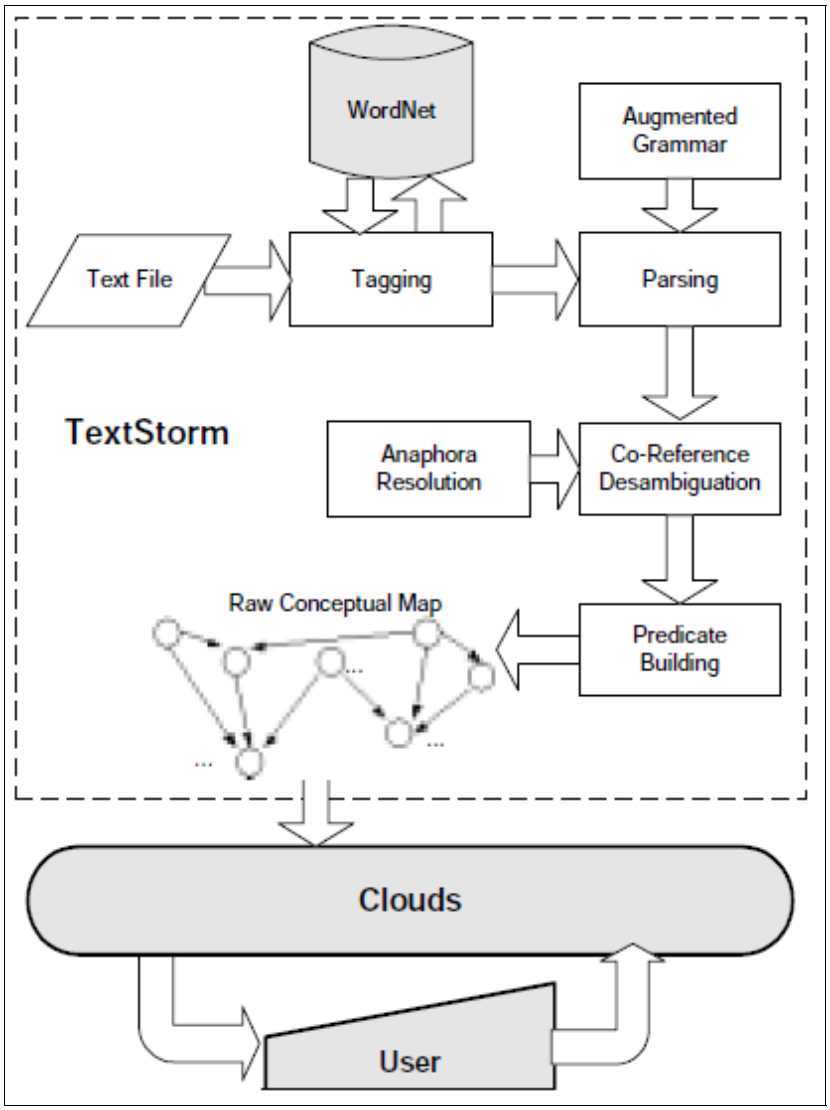


Figure 31: The architecture of the system that Oliviera et al. (2001) used for semi-automatic construction of concept maps from text.

Table 13: An example of the identification of possible binary predicates.

sentence	Cows, as well as rabbits, eat only vegetables, while humans eat also meat
predicate1	eat (*cow*, *vegetables*)
predicate2	eat (*rabbit*, *vegetables*)
predicate3	eat (*human*, *vegetables*)
predicate4	eat (*human*, *meat*)

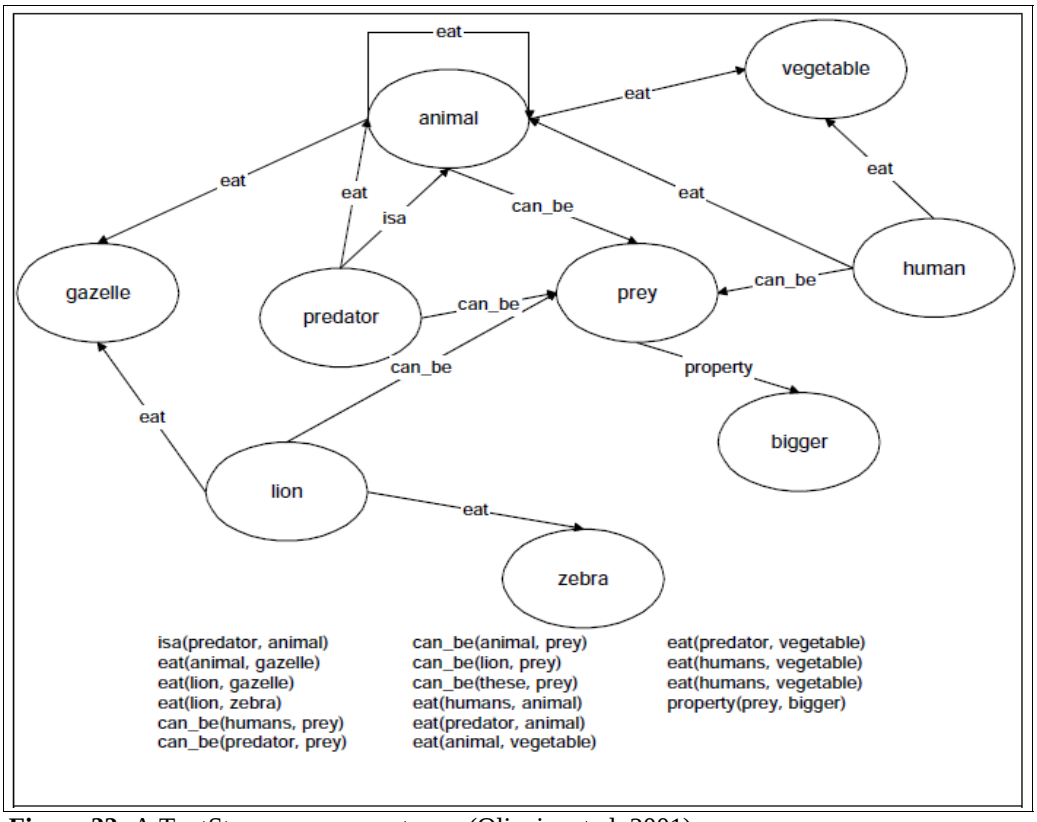


Figure 32: A TextStorm raw concept map (Oliveira et al. 2001).

Watson et al. (2005) developed Leximancer, a tool that uses semantic mapping to extract concept maps from text. More specifically, Leximancer was applied to 421 educational-based pathology case notes for the identification of unique relationships. It employs semantic and then relational co-occurrence information extraction to recognise links among concepts in the source documents. The approach employs non-linear dynamics and machine learning by automatically extracting the most important concepts in a document set (such as patient records or case studies). An example of Leximancer output can be seen in Figure 33. Each concept has a coloured circle indicating the co-occurrences of the concept relative to others, while its size suggests the centrality of the concepts in the text. The grey hand drawn circles indicate concepts that are in similar semantic groups.

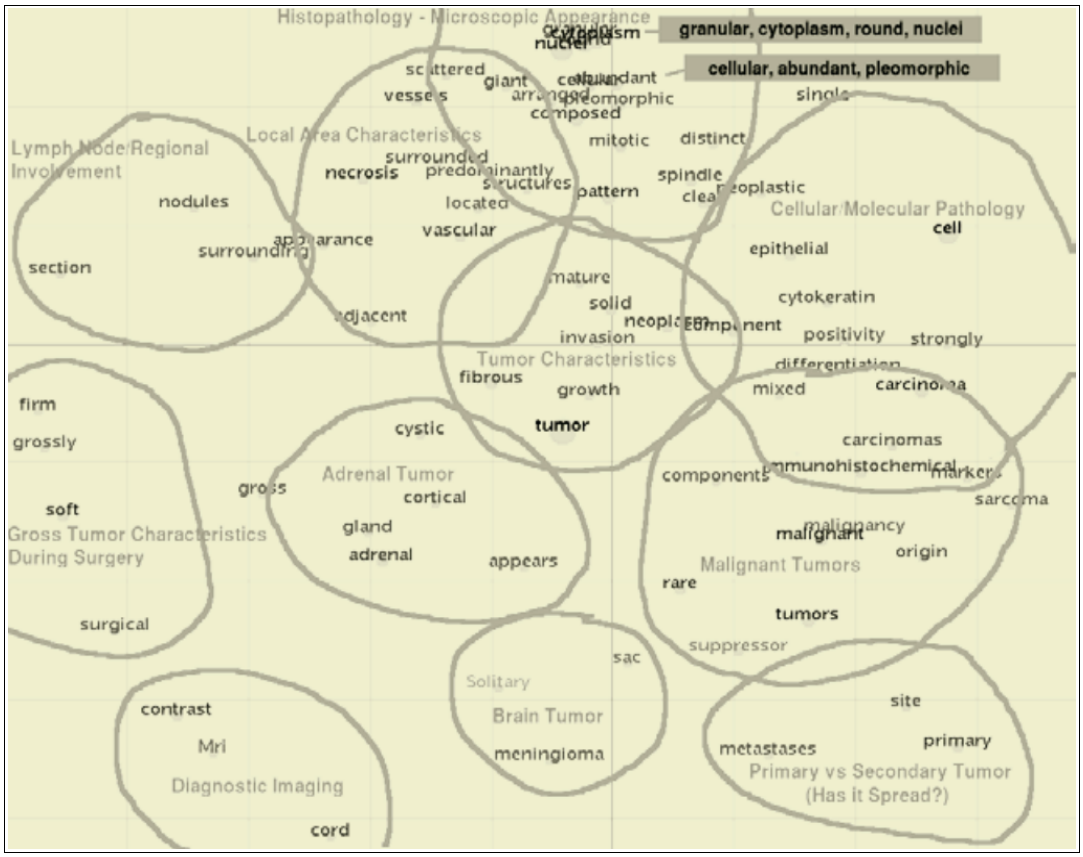


Figure 33: A Leximancer generated concept map for tumour (Watson et al. 2005).

Leake et al. (2006) presented a method that automatically creates preliminary concept maps from on-line documents. Their work focused on the generation of a single concept map from a single document with the assumption that the input document is clearly written and contains the description of concepts. They use sentence syntactic structure and dependency information to find relations between concepts (Figure 34). More specifically, for the extraction of concepts, they applied a sentence boundary detection algorithm based on regular expressions, while deep syntactic analysis is performed on each sentence in order to select noun phrases of interest. In order to rank the identified concepts according to their relevance to the respective source document, a basic term weighting approach is used through simple term document frequency since no additional context information is available. Nouns and adjectives were considered since these are the most common parts of speech in concept describing phrases. From parsed set and normalized word set, a set of noun phrases classified as candidate concepts is generated and ranked by using an ordered list of nouns and adjectives. For the linkage phrase detection, all pairs of concepts with an indirect dependency link through verb phrases are extracted and the information is gathered from tagged parse trees to assist in the construction of the concept map. Figure 35 shows an example of the system output from processing a short document.

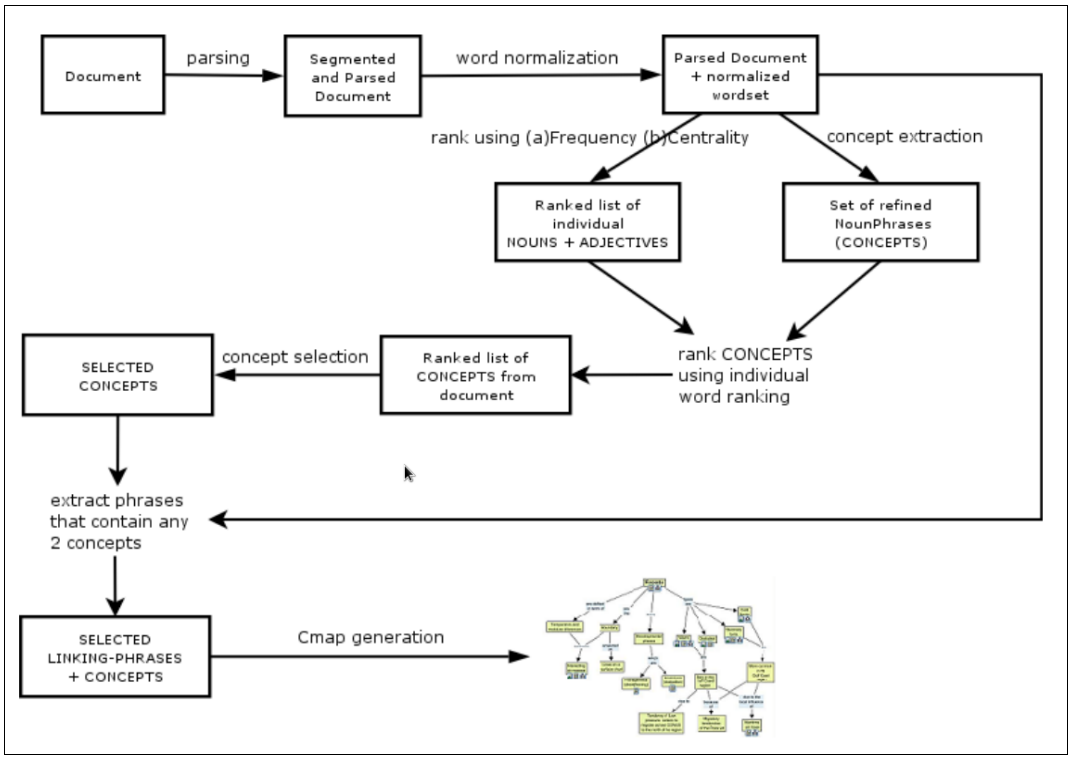


Figure 34: Detailed procedure to construct automatically a concept map from text (Leake et al. 2006).

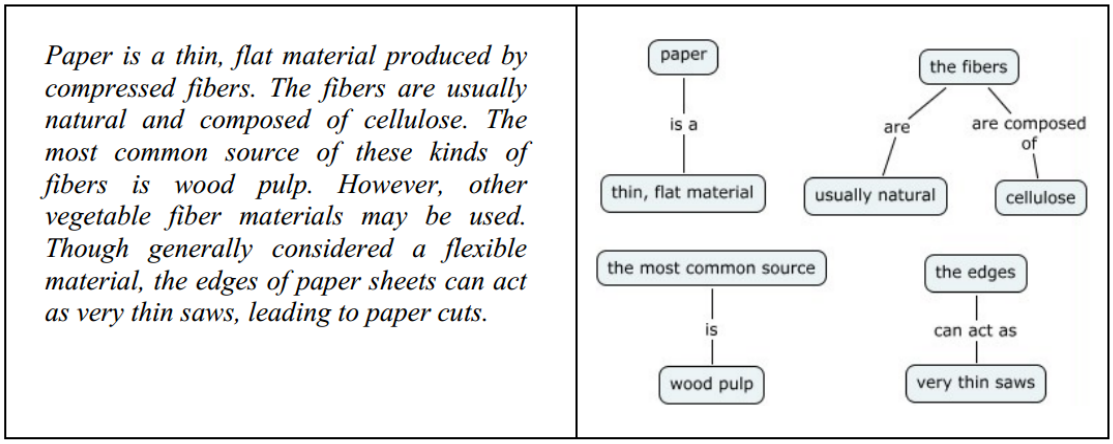


Figure 35: A concept map produced by processing a small document with isolated concepts removed for legibility (Leake et al. 2006).

Chen et al. (2008) designed and developed a system that extracts concept maps from academic articles in the e-Learning domain without relationship labeling. The system includes four main steps (Figure 36):

- information retrieval from articles;
- concept extraction;
- research keyword indexing;
- calculation of "*relation strength*" between two keywords.

Users can utilize query parameters to generate concept maps. Keywords targeted for extraction, are classified into appropriate groups, thus reducing the total number of terms. A thesaurus is created to store these grouped terms in topic databases and the highest frequency term is used to name each group. The weights associated with each keyword are comprehensive index values calculated through the application of the keyword indexing method. Table 14 and Figure 37 show the top 15 keywords extracted from an e-Learning domain from a conducted query holistic based query and the corresponding concept map.

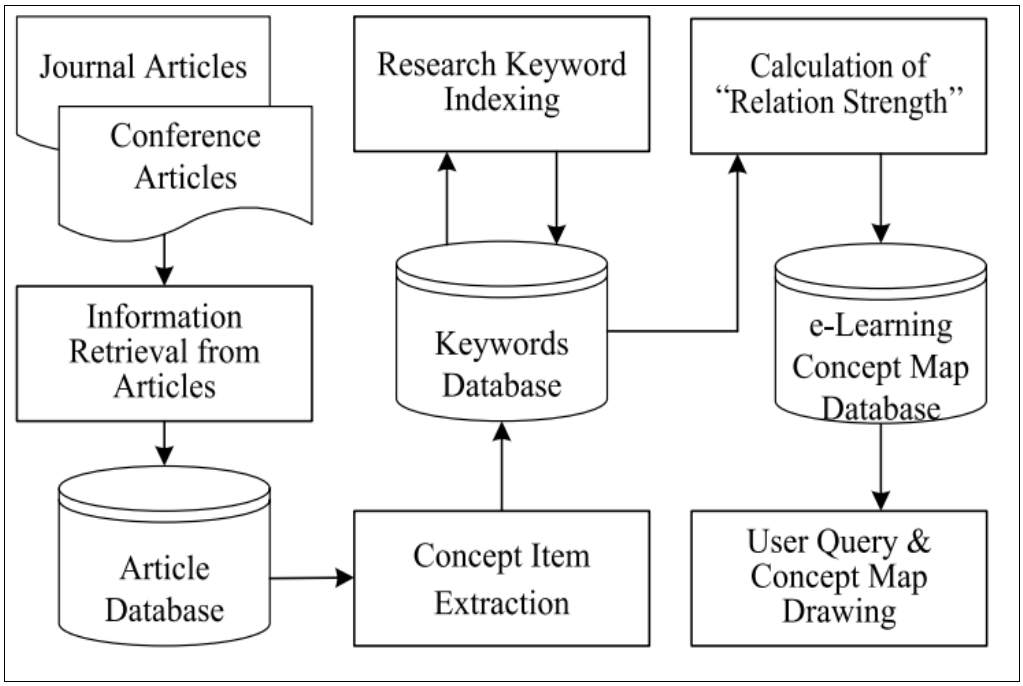


Figure 36: An overview of the fours steps (information retrieval, concept extraction, research keyword indexing, calculation of "relation strength") in the system designed and implemented by Chen et al. (2008) for the automatic generation of concept maps in the e-Learning field.

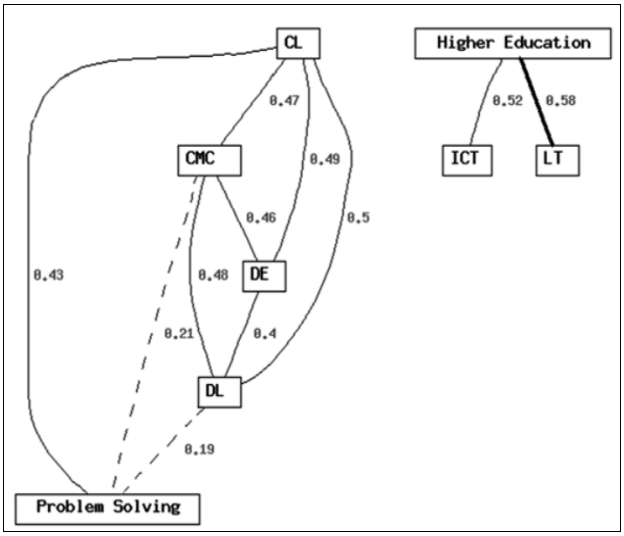


Figure 37: The generated concept map based on a holistic query representing the top 15 keywords extracted from the e-Learning domain.

Table 14: Top 15 research keywords in the e-Learning domain from 1999 to 2004.

Rank	Keyword	Weight
1	Information Technology (IT)	27238.5
2	Intelligent Tutoring Systems (ITS)	6200.92
3	World Wide Web (WWW)	4541.00
4	Information and Communication Technologies (ICT)	3107.74
5	Web Based	30.11.72
6	On line	2836.39
7	Collaborative Learning (CL)	2694.93
8	Distance Education (DE)	2636.04
9	Learning Technologies (LT)	2557.93
10	Higher Education	1649.58
11	Computer Mediated Communications (CMC)	1625.32
12	Distance Learning (DL)	1561.40
13	Learning Object Systems (LOS)	1428.60
14	Problem Solving	1369.68
15	Learning Process	1107.98

Only recently, de la Villa et al. (2012) designed and implemented a tool that automatically constructs concept maps and represents information from different ontologies and knowledge bases. Their tool comprises of two components accessible through a web interface based on Freebase data that enhances the understanding of users for clinical cases through concept visualization and semantic graph visualization. The first component is an information access system that retrieves the Freebase topic's names such as symptoms, treatments and diseases in order to annotate concepts from an input clinical text and shows rich information about these concepts while the second deals with generation of a corresponding concept map with the discovered concepts as an input. Particularly for the medical concept recognition, the proposed system pre-processes a set of medical terms compiled into lists used by the Gazetter component in the GATE distribution. The generated graph for a concept can be viewed through the results of Freebase and UMLS terminology services queries that returned related information e.g., treatments, causes, risk factors etc. For evaluation of recognised concepts, they reported a recall of 62%. An example of a case abstract from the National Centre for Case Study Teaching in Science (NCCSTS) for the disease of "*stroke*" can be seen in Figure 38 with the related concepts (risk factors, causes, prevention factors, notable people, treatment and symptoms) being retrieved through Freebase.

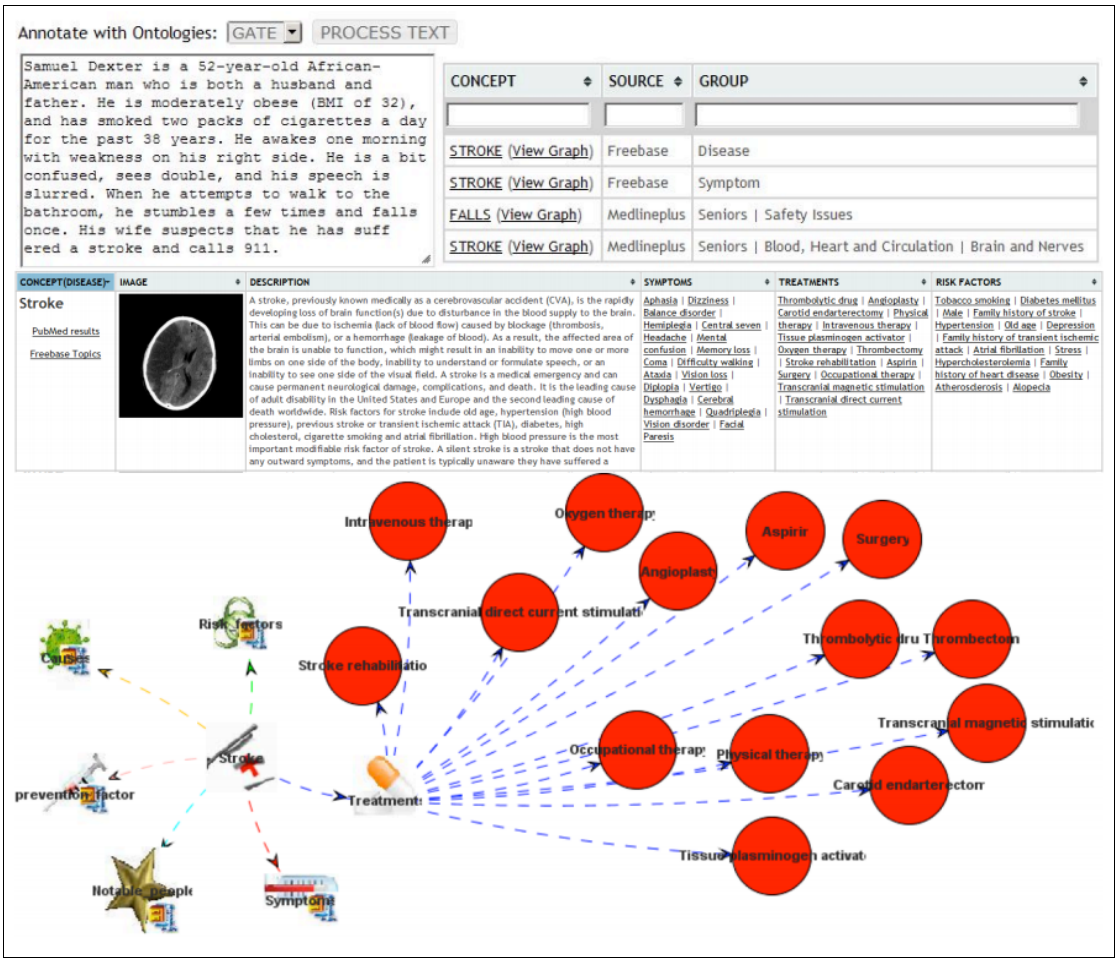


Figure 38: An automatically generated concept map for the disease "*stroke*". The node "treatment" expanded after receiving as an input one case abstract from the NCCSTS archives. The retrieved information was acquired from the Freebase database (de la Villa et al. 2012).

Bai et al. (2008) presented an approach that performs CMM on students' testing records for adaptive learning systems. Their method is based on fuzzy reasoning techniques which are based on fuzzy rules. Lee et al. (2009) applied an algorithm (Apriori for Concept Map) to develop a concept diagnostic system (ICDS) of an automatically constructed concept map of learning in order to guide learners and assist in the increasing learning performance of students. Finally Chen et al. (2010) presented a CMM methodology for adaptive learning systems based on data mining techniques. However, these studies are focusing in the educational aspect of CMM with techniques and implementation that are exceeding the aim of this research project so a detailed and fair comparison could not performed.

However, these studies are not using epidemiological data as source texts and they are not aiming to reproduce automatically a concept map that summarizes key characteristics from epidemiological studies related to a particular health problem through the application of text mining techniques.

2.3.7. Summary

Epidemiological studies contain rich information that could improve the understanding of a health problem and underpin the development of preventive measures and health care policies. Text mining has been applied successfully in various biomedical domains for the identification of important information. Its techniques can assist epidemiologists to identify pieces of information in epidemiological studies by reducing the amount of time that is required to detect and integrate key epidemiological knowledge for further research and exploration.

Concept maps are a knowledge organization and representation form involving a specific topic or theme that promote meaningful learning and are mostly used in the field of medical education. Their application enhances the user's understanding of complex concepts while new associations can be easily made and observed through their existing links between their concepts. Therefore, the utilization of concept map can represent and integrate the extracted epidemiological information from studies in a simple and coherent figure. Additionally, the concept map could reveal a concept overview of the health problem and suggest potential new hypotheses for exploration.

Still, concept maps can be used as a knowledge representation model that aims at knowledge exploration, rather than knowledge consolidation, which is the main aim of ontologies. Given a huge number of potential concepts that could be linked to a particular concept e.g., disease and the dynamic nature of medical research, concept maps appear to be a suitable model that facilitate understanding of complexity for a specific health care issue.

However, limited research has been conducted for the recognition of key characteristics from epidemiological studies. Previous work has focused on the identification of either specific characteristics from particular study types such as clinical trials either on one key characteristic such as risk factors and exposure-related terms. Furthermore, any characteristics that have been recognised by the current approaches are not normalized to the individual attributes they carry (e.g., exposure-related terms classified to their UMLS semantic group and category classification). Epidemiological data offer a variety of characteristics that could be potentially targeted for recognition. Despite the availability of vast amounts of epidemiological data, previous methods have not performed a large scale text mining procedure in epidemiological literature (i.e., epidemiological study abstracts) related to a complex health care problem.

There have been a number of efforts for the automatic or semi-automatic generation of concept maps from text. Only a few have focused on clinical data and not for the representation of a health problem overview. Research work for the automatic creation of concept maps from text have been relying on the representation of textual knowledge in general (revealing concept

maps from various fields such as the e-Learning domain and pathological notes) rather than providing an overview of a health problem's related and classified (normalized) concepts. More specifically, previous work has not addressed the representation of semantically rich clinical information (e.g., epidemiological characteristics) and any efforts in the biomedical domain for the generation of concept maps from patient records have put emphasis in the identification of unique relationships between the concepts. Research has been conducted in the understanding of medical concepts from clinical notes (Watson et al. 2005) by generating a concept map but the approach has not focused on more structured data such as epidemiological studies and aimed to be used as a learning tool rather than a base for concept exploration.

To the best of our knowledge, there has not been a method for the automatic extraction and normalization of key epidemiological characteristics from all study types of epidemiological research (observational and experimental studies) related to a particular health problem and their automatic representation into the form of a concept map. Due to the implementation of text mining methods in large scale (clinical) data, there is a need to be able to manipulate the recognised results more efficiently. The visual representation that concept maps offer, enable us to navigate large amounts of recognised and classified data more efficiently while suggesting new hypothesis through further data exploration. The inclusion of normalized information that reveals the nature of each concept can assist in the discovery and understanding of potential concepts clustered together in the automatically generated concept map that may share similar attributes and functions.

Chapter 3

Research Method Overview

In this thesis, we follow a rule based approach for the identification and normalization of key epidemiological characteristics related to a given health care problem. The rule based method was chosen due to the lack of large annotated corpora and the selection of appropriate features that are required for the training of machine learnings techniques. The manual inspection of the related data was performed by the author only, thus the limited human resources could pose a problem when it comes to time constraints. In addition, it was considered that since epidemiological texts contain a relative structure, it would be more efficient to use rules that could enable and identify common syntactical expressions presented in text that suggest the presence of a key characteristic of interest. In this chapter, we provide an overview of the methodology, including the explanation of the target entities and data resources used.

3.1. Definition of Epidemiological Characteristics

Epidemiology is a relatively structured field with its own dictionary and reporting style, deliberately written in a typical semi-structured format in order to standardize and improve study design, communication and collaboration internationally. The standard characteristics in most epidemiological studies include (Last, 2001):

- **study design**: a specific plan or protocol that has been followed for the conduction of the study; it allows the investigator to translate their conceptual hypothesis into an operational one e.g., "*prospective birth cohort study*", "*triple blind randomized clinical trial*". Table 15 shows more examples of epidemiological study design.

Table 15: Epidemiological literature examples for study design.

example	study design
This was an ***observational study***.	observational study
Methods: a **case-control study** that included ...	case-control study
Methodin a **prospective study** of Taiwanese adults aged a(c)y60 years (n=3922) between 1989 and 1999 ...	prospective study
This **cross-sectional study** in 2006 determined the prevalence and demographic characteristics ...	cross-sectional study
Methods/design: this is an **observational cohort study** and involves ...	observational cohort study
Research design and methods This study was a **randomized double-blind placebo-controlled trial** consisting of 100 severely obese children ...	randomized double-blind placebo-controlled trial

- **population**: details of the individuals (e.g., gender, age, ethnicity, nationality) participating in an epidemiological study are important in comprehending the context of the study – this is referred to as "*the defined population*" (see Table 16 for population examples).

Table 16: Epidemiological literature examples for population.

example	population				
	size	**age**	**nationality**	**ethnicity**	**gender**
We studied a population sample of **1,516 men aged from 50-70 years old**.	1,516	50-70 years old	-	-	male
Socioeconomic, demographic and family dysfunction related to obesity in **6 to 9 year-old children**.	-	6 to 9 year-old children	-	-	-
Methods: a total of **15,061 adults at 35 years old or over** were surveyed.	15,061	35 years old or over	-	-	-
This was a cross-sectional study of **214 overweight/obese and 47 normal-weight Mexican children 6-12 years old.**	214, 47	6-12 years old	Mexican	-	-
Methods: this epidemiological study involved **490 Chinese college students (aged 15-25 years, mean 18.9-/+1.2 years),** ...	490	aged 15-25 years, mean 18.9-/+1.2 years	Chinese	Chinese	-
The authors tested the relationship between bmi at baseline and the 5-year incidence of periodontal disease in a sample of **2787 males and 803 females**.	2787	-	-	-	male/female

- **exposure**: a factor, event, characteristic or other definable entity that brings about change in a health condition, or in other defined characteristics. Typical exposures include environmental, social and behavioural factors or specific biological and clinical ones. Table 17 reveals some example of exposures in epidemiological text.

Table 17: Epidemiological literature examples for exposure (continuing on next page).

example	exposure(s)
Sedentary lifestyle is linked to obesity.	sedentary lifestyle
The analysis of multiple logistic regression revealed that **age** (or=1.06), **alcohol consumption**, **obesity** (or=3.12) and **levels of triglyceride** (or=1.30) and **cholesterol** (or=1.32) and **serum glucose** (or=1.41) were risk factors of hypertension.	age
	alcohol consumption
	obesity
	levels of triglyceride
	cholesterol
	serum glucose
Conclusions: Lifestyle factors including **heavy alcohol consumption**, **heavy smoking**, **metabolic disorders**, and **hiatal hernia** increased the risk of erosive esophagitis.	heavy alcohol consumption
	heavy smoking
	metabolic disorders
	hiatal hernia
Impact of **cigarette smoking** on onset of nonalcoholic fatty liver disease over a 10-year period.	cigarette smoking

Table 17: Epidemiological literature examples for exposure.

Body mass index is an established risk factor for post-menopausal breast cancer.	body mass index
Although no association was seen between **family dysfunction** and obesity, ...	family dysfunction

- **outcome**: the consequence from the exposure in the population of interest; it could be of biological or health in nature. Table 18 provides some examples of outcomes in epidemiological text.

Table 18: Epidemiological literature examples for outcome.

example	**outcome(s)**
Sedentary lifestyle is linked to **obesity**.	obesity
Prevalence and anthropometric predictors of **high blood pressure** in schoolchildren from Joao Pessoa ...	high blood pressure
Objective: to determine if sleep difficulties are associated with **overweight/obesity status** among preadolescents	overweight/obesity status
Coffee consumption is inversely associated with **type 2 diabetes** in Chinese	type 2 diabetes
Risk factors for **erosive esophagitis**: a cross-sectional study of a large number of Japanese males.	erosive esophagitis
Relationship of physical activity and eating behaviour with **obesity** and **type 2 diabetes mellitus**:	obesity
	type 2 diabetes mellitus

- **covariate**: a factor that is possibly predictive of the outcome under study. It could be of a direct interest to the study or may be a confounding variable affecting the outcome of the study itself without knowing the exact effect. A covariate could be a concept of various nature (from biological to clinical and environmental). Examples of covariate mentions are shown in Table 19 in epidemiological literature.

Table 19: Epidemiological literature examples for covariate (continuing on next page).

example	**covariate(s)**
The population was studied after adjusting for confounders such as **age and sex**.	age
	sex
After adjusting for **gender** and **age**, overweight was significantly associated with increased odds of having high triglycerides ...	gender
	age
The increase in weight over time remained statistically significant after being controlled in multivariate analysis for **socio-economic status** and **race.**	socio-economic status
	race
All multivariate models control for **age**, **sex**, **race**, **smoking**, and **socio-economic resources**.	age
	sex
	race
	smoking
	socio-economic resources

Table 19: Epidemiological literature examples for covariate.

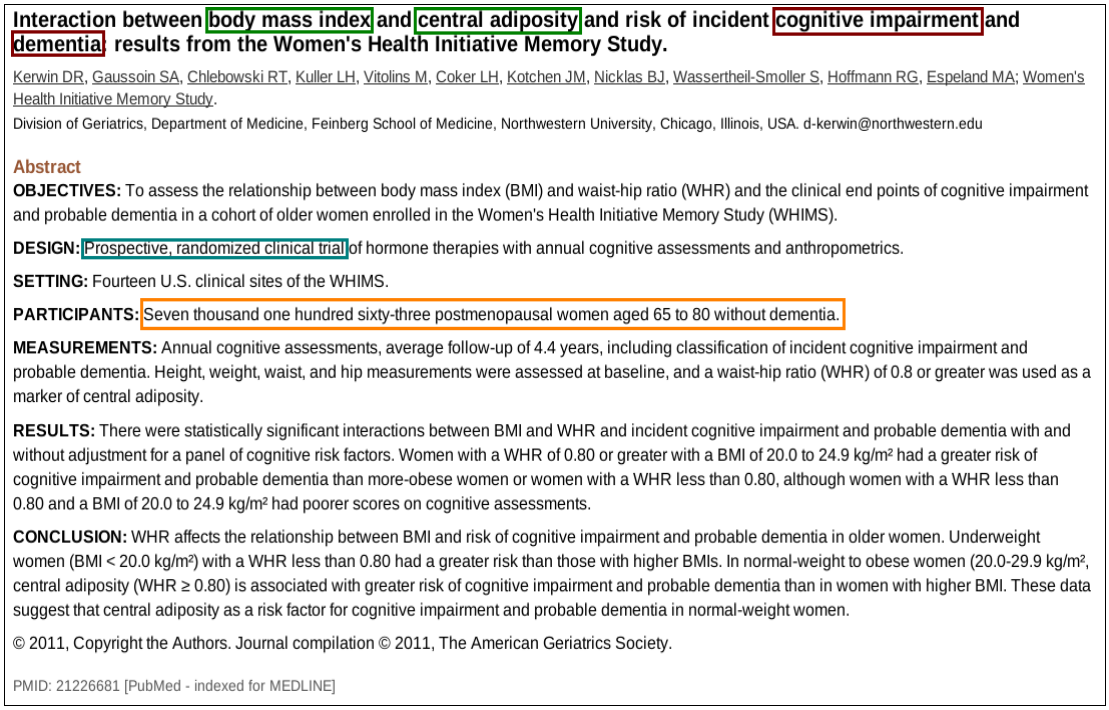
All models were adjusted for **age**, **education**, **physical activity**, **self-rated health**, **employment**, **diet**, **smoking**, and **alcohol intake**.	age
	education
	physical activity
	self-rated health
	employment
	diet
	smoking
	alcohol intake
... using multivariable analysis, controlling for confounders such as **maternal age**, **fertility treatments**, and **ethnicity**, ...	maternal age
	fertility treatments
	ethnicity

- **effect size**: the measure of the strength of the relationship between two variables in the studied population. Standard measures include hazard ratio, (adjusted) odds ratio, relative risk, prevalence and incidence. Table 20 demonstrates effect size examples in epidemiological text.

Table 20: Epidemiological literature examples for effect size.

example	effect size(s)			
	type	**value**	**confidence interval**	**concept**
The **hazard ratio** for being **overweight** was found to be **0.91**.	hazard ratio	0.91	-	overweight
The overall **prevalence** of **overweight** was **19.0%** and of **obesity** was **23.3%**.	prevalence	19.0%	-	overweight
	prevalence	23.3%	-	obesity
The annual **incidence** of **depression** was **12%** in this cohort.	incidence	12%	-	depression
An association risk factor exposure and disease was documented (**odds ratio**, [95% confidence limits]) for **smoking** (**2.47 [1.68-3.64]**), **laboratory dyslipidemia** (**1.92 [1.33-2.77]**), **low hdl-c** (**2.12 [1.31-3.42]**), **careless diet** (**4.46 [2.88-6.90]**) and **sedentary lifestyle** (**1.79 [1.22-2.62]**).	odds ratio	2.47	[1.68-3.64]	smoking
	odds ratio	1.92	[1.33-2.77]	laboratory dyslipidemia
	odds ratio	2.12	[1.31-3.42]	low hdl-c
	odds ratio	4.46	[2.88-6.90]	careless diet
	odds ratio	1.79	[1.22-2.62]	sedentary lifestyle
Mortality risk was higher for persons with **severe obesity** (**relative risk = 1.571, 95% confidence interval = 1.335-1.849, p <.001**).	relative risk	1.571	95% [1.335-1.849]	severe obesity
... infants in the highest quintile of pwv had strongly increased risks of **overweight/obesity** at the age of 4 years (**odds ratio (95% confidence interval):15.01 (9.63, 23.38)**).	odds ratio	15.01	95% (9.63, 23.38)	overweight/obesity

Not all abstracts unfortunately, include all six of the above mentioned characteristics despite the standardized efforts occurring in the field of epidemiology. Figures 39 and 40 show examples of MEDLINE study design abstracts with their epidemiological characteristics (if existing) highlighted.



Interaction between body mass index and central adiposity and risk of incident cognitive impairment and dementia: results from the Women's Health Initiative Memory Study.

Kerwin DR, Gaussoin SA, Chlebowski RT, Kuller LH, Vitolins M, Coker LH, Kotchen JM, Nicklas BJ, Wassertheil-Smoller S, Hoffmann RG, Espeland MA; Women's Health Initiative Memory Study.

Division of Geriatrics, Department of Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA. d-kerwin@northwestern.edu

Abstract

OBJECTIVES: To assess the relationship between body mass index (BMI) and waist-hip ratio (WHR) and the clinical end points of cognitive impairment and probable dementia in a cohort of older women enrolled in the Women's Health Initiative Memory Study (WHIMS).

DESIGN: Prospective, randomized clinical trial of hormone therapies with annual cognitive assessments and anthropometrics.

SETTING: Fourteen U.S. clinical sites of the WHIMS.

PARTICIPANTS: Seven thousand one hundred sixty-three postmenopausal women aged 65 to 80 without dementia.

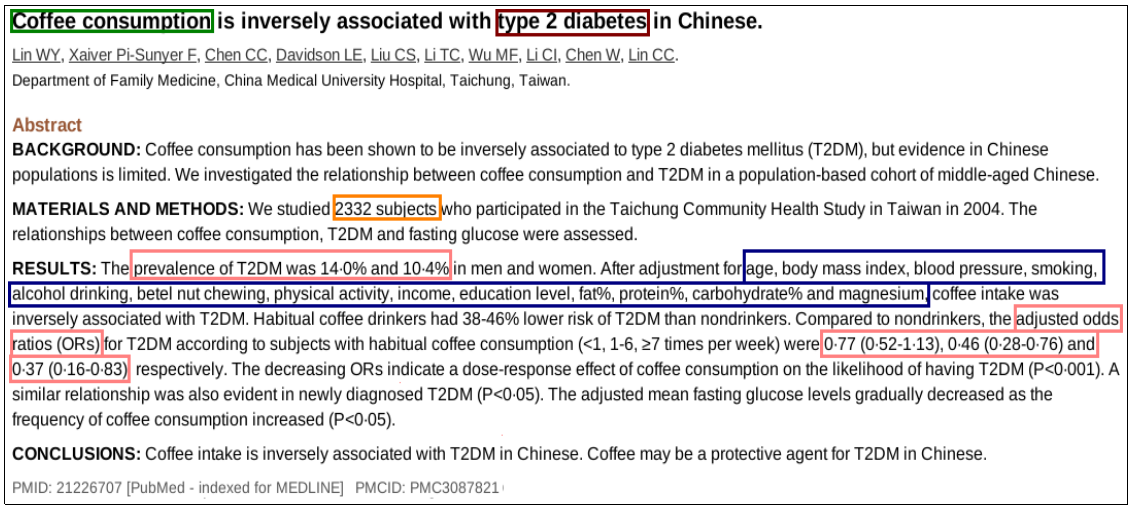
MEASUREMENTS: Annual cognitive assessments, average follow-up of 4.4 years, including classification of incident cognitive impairment and probable dementia. Height, weight, waist, and hip measurements were assessed at baseline, and a waist-hip ratio (WHR) of 0.8 or greater was used as a marker of central adiposity.

RESULTS: There were statistically significant interactions between BMI and WHR and incident cognitive impairment and probable dementia with and without adjustment for a panel of cognitive risk factors. Women with a WHR of 0.80 or greater with a BMI of 20.0 to 24.9 kg/m² had a greater risk of cognitive impairment and probable dementia than more-obese women or women with a WHR less than 0.80, although women with a WHR less than 0.80 and a BMI of 20.0 to 24.9 kg/m² had poorer scores on cognitive assessments.

CONCLUSION: WHR affects the relationship between BMI and risk of cognitive impairment and probable dementia in older women. Underweight women (BMI < 20.0 kg/m²) with a WHR less than 0.80 had a greater risk than those with higher BMIs. In normal-weight to obese women (20.0-29.9 kg/m², central adiposity (WHR ≥ 0.80) is associated with greater risk of cognitive impairment and probable dementia than in women with higher BMI. These data suggest that central adiposity as a risk factor for cognitive impairment and probable dementia in normal-weight women.

PMID: 21226681 [PubMed - indexed for MEDLINE]

Figure 39: An example of a MEDLINE study design abstract with highlighted key characteristics at the document level. "*Body mass index*" and "*central adiposity*" are the exposures of the study (highlighted with green) whereas the "*cognitive impairment*" and "*dementia*" are the outcomes (highlighted with red). The design of this epidemiological study is "*prospective randomized clinical trial*" (highlighted with blue) while the population participating in the study is "*seven thousand one hundred sixty-three postmenopausal women aged 65 to 80 without dementia*" (highlighted with orange). No covariate and effect size mentions were observed.



Coffee consumption is inversely associated with type 2 diabetes in Chinese.

Lin WY, Xaiver Pi-Sunyer F, Chen CC, Davidson LE, Liu CS, Li TC, Wu MF, Li CI, Chen W, Lin CC.

Department of Family Medicine, China Medical University Hospital, Taichung, Taiwan.

Abstract

BACKGROUND: Coffee consumption has been shown to be inversely associated to type 2 diabetes mellitus (T2DM), but evidence in Chinese populations is limited. We investigated the relationship between coffee consumption and T2DM in a population-based cohort of middle-aged Chinese.

MATERIALS AND METHODS: We studied 2332 subjects who participated in the Taichung Community Health Study in Taiwan in 2004. The relationships between coffee consumption, T2DM and fasting glucose were assessed.

RESULTS: The prevalence of T2DM was 14·0% and 10·4% in men and women. After adjustment for age, body mass index, blood pressure, smoking, alcohol drinking, betel nut chewing, physical activity, income, education level, fat%, protein%, carbohydrate% and magnesium, coffee intake was inversely associated with T2DM. Habitual coffee drinkers had 38-46% lower risk of T2DM than nondrinkers. Compared to nondrinkers, the adjusted odds ratios (ORs) for T2DM according to subjects with habitual coffee consumption (<1, 1-6, ≥7 times per week) were 0·77 (0·52-1·13), 0·46 (0·28-0·76) and 0·37 (0·16-0·83) respectively. The decreasing ORs indicate a dose-response effect of coffee consumption on the likelihood of having T2DM (P<0·001). A similar relationship was also evident in newly diagnosed T2DM (P<0·05). The adjusted mean fasting glucose levels gradually decreased as the frequency of coffee consumption increased (P<0·05).

CONCLUSIONS: Coffee intake is inversely associated with T2DM in Chinese. Coffee may be a protective agent for T2DM in Chinese.

PMID: 21226707 [PubMed - indexed for MEDLINE] PMCID: PMC3087821

Figure 40: An example of a MEDLINE study design abstract with highlighted key characteristics at the document level. "*Coffee consumption*" is the exposure of the study (highlighted with green) whereas the "*type 2 diabetes*" is the outcome (highlighted with red).The population participating in the study is "*2332 subjects*" (highlighted with orange) while "*age, body mass index, blood pressure, smoking, alcohol drinking, betel nut chewing, physical activity, income, education level, fat%, protein%, carbohydrate% and magnesium*" are the covariates (highlighted with blue). The "*adjusted odds ratios (ORS)*" - effect size mentions (highlighted with pink)- were "*0.77 (0.52-1.13), 0.46 (0.28-0.76) and 0.37 (0.16-0.83)*". No study design mention was observed.

3.2. Method Overview

An overview of the methodology for epidemiological literature mining is shown in Figure 41. In step 1, an epidemiological study design corpus related to a health problem is processed using a rule based approach combined with relevant dictionaries to recognise potential mentions of key epidemiological characteristics through the environment of Minorthird. At this stage, the identification is being performed at the mention level. In the second step, the extracted characteristics are then normalized and classified under the respective UMLS semantic groups and categories with the application of the MetaMap at the document level. Any other detailed information is captured as an attribute (in the cases of study design i.e., "*randomized*", "*non-randomized*", "*serial*", population i.e., age, ethnicity and effect size i.e., confidence interval, effect size type).

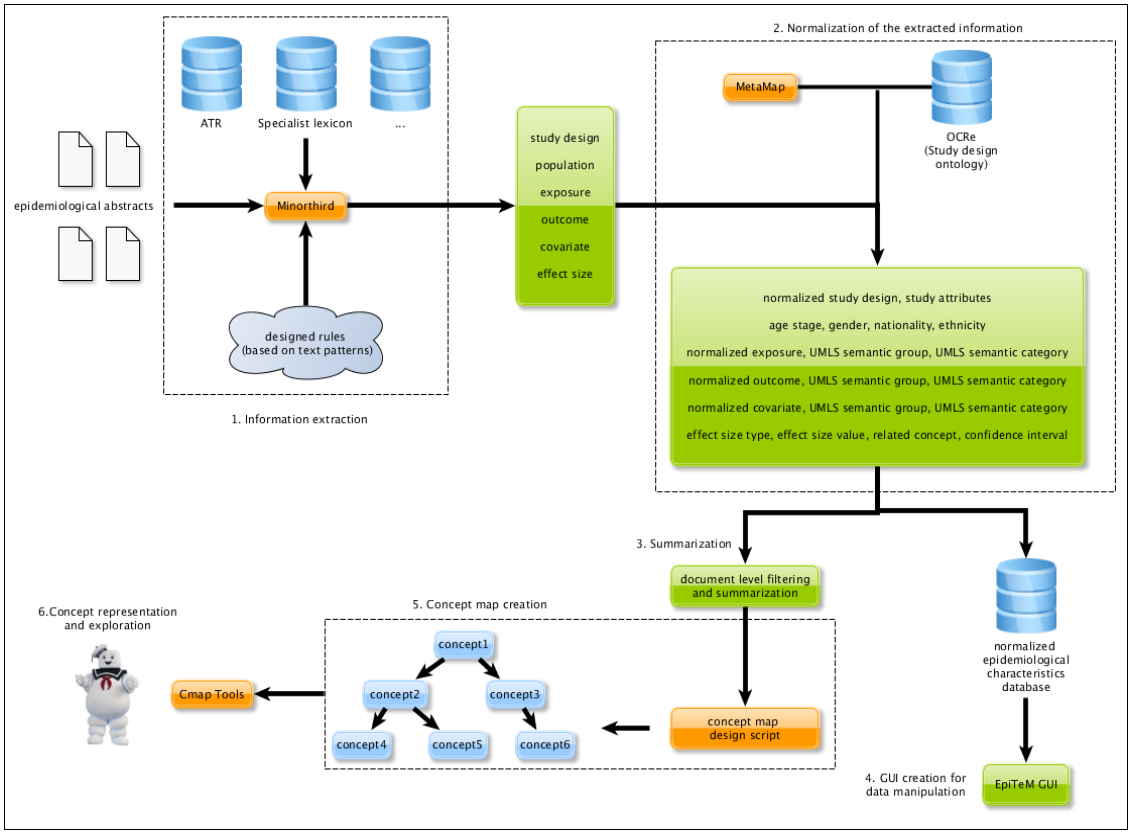


Figure 41: An overview of the proposed methodology. The approach includes identification, normalization and visualization of key characteristics from epidemiological abstracts. Epidemiological text mining is represented by steps 1 (Information extraction), 2 (Normalization of the extracted information) and 3 (Summarization). These steps contain: ATR, a term identification tool; Specialist lexicon, a dictionary that includes various biomedical terms; MinorThird, a toolkit of learning methods for storing, annotating and categorizing text as well as recognizing entities of interest; OCRe a clinical research ontology that describes human studies and their methods; MetaMap, concept mapping tool of entities to the UMLS. Step 4 (GUI creation for data manipulation) refers to the creation and use of a GUI for the manipulation of the normalized data called EpiTeM. Step 5 (Concept map creation) reveals the generation of a concept map from the results of the epidemiological text mining procedure and in step 6 (Concept representation and exploration) the concept map is represented through a concept mapping software that provides the visualization of the extracted and normalized characteristics in the form of a concept map named CmapTools.

Table 21 provides a summarization of the normalization process for each characteristic. The normalized results are then used in step 3 for:

1. the automatic generation of a concept map that will represent the exposure, outcome and covariate concepts along with their respective UMLS classification and their frequency in the epidemiological corpus linked to a health problem.
2. manipulation through a graphical user interface (GUI - EpiTeM) for data navigation observation and exploration.

Figure 42 shows an example of how the identified and normalized characteristics in an epidemiological study abstract can be manipulated through the EpiTeM (see Section 7.3).

Table 21: Summarization of each characteristic's normalization for their attributes. Study design is normalized to the nodes of the OCRe ontology; population is normalized towards its ethnicity, age, nationality and gender, exposure, outcome and covariate mentions are normalized to specific UMLS semantic groups and categories; finally effect size mentions are normalized towards their value, effect size type, related concept and respective confidence interval.

<table>
<tr><th>characteristic</th><th colspan="4">normalized to</th></tr>
<tr><td>study design</td><td colspan="4">OCRe</td></tr>
<tr><td>population</td><td>age</td><td>nationality</td><td>ethnicity</td><td>gender</td></tr>
<tr><td>exposure</td><td colspan="2" rowspan="3">UMLS semantic groups</td><td colspan="2" rowspan="3">UMLS semantic categories</td></tr>
<tr><td>outcome</td></tr>
<tr><td>covariate</td></tr>
<tr><td>effect size</td><td>effect size value</td><td>confidence interval</td><td>related concept</td><td>effect size type</td></tr>
</table>

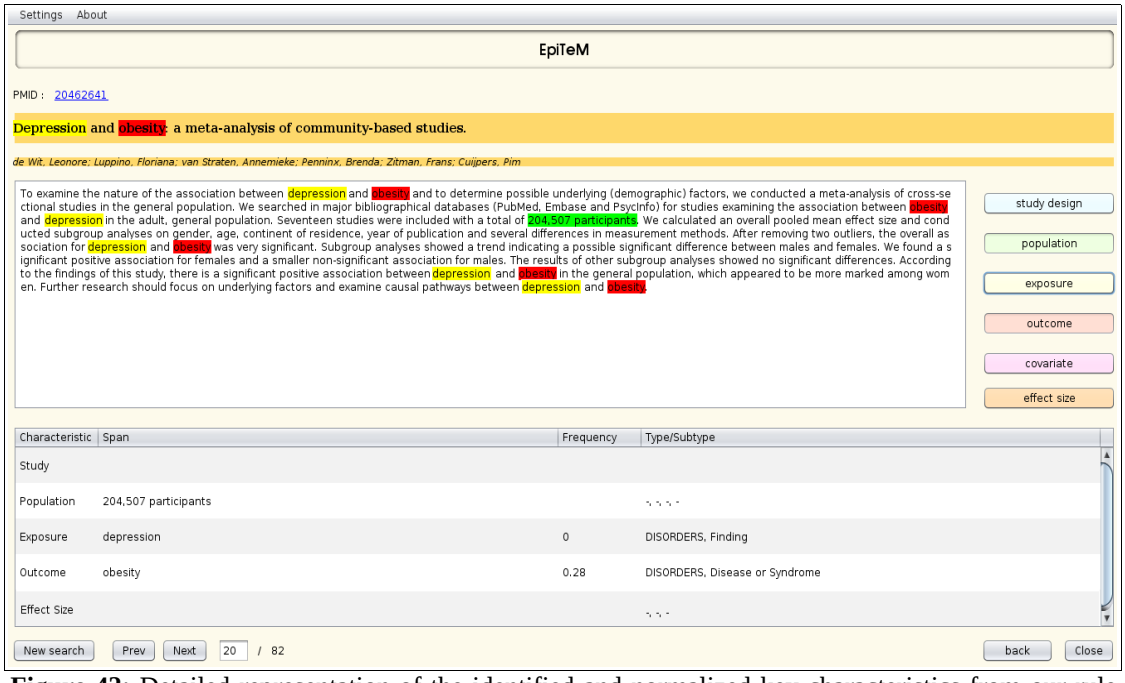


Figure 42: Detailed representation of the identified and normalized key characteristics from our rule based approach through the application of EpiTeM for the PubMed abstract with pmid 20031086.

Overall, the methodology here has the following characteristics:

- **Novel normalization pipeline:** To the best of our knowledge, there is no approach available for the recognition and normalization of key characteristics (e.g., epidemiological study type, population, exposure, outcome, covariate, effect size) from epidemiological literature related to a health problem. The normalization component aims to avoid losing any specific information (e.g., epidemiological study attributes such as "*retrospective*", "*prospective*" or the gender and age of the study population). Concepts belonging to the characteristics of exposure, outcome and covariate are normalized to their canonical term and classified under a UMLS semantic group and a more specific UMLS semantic category; both suggesting and revealing the nature of the concept.
- **Targeted characteristics for extraction:** Previous research work aimed to identify risk factors and exposure terms by applying rule based methods but only to MEDLINE citations and titles of epidemiological articles rather than epidemiological study abstracts. Additionally, a number of efforts in epidemiological text mining focused on the recognition of multiple specific characteristics presented only in clinical trial publications. Our approach includes the identification of six key characteristics that are commonly presented in most epidemiological study types and not only in trials in the form of abstracts related to a health problem (obesity). The data used for epidemiological text mining include both experimental (clinical trials) and observational studies (e.g., case control studies, cohort studies).
- **Large scale extraction:** We aim at performing large scale text mining of epidemiological study abstracts. More specifically, all the related epidemiological study abstracts to a health problem are collected, integrated and processed through the system.
- **Concept map visualization of epidemiological text mining results for representation and exploration:** As opposed to previous work on the generation of a concept map from text for the identification of unique relationships between the recognised concepts, our approach focuses on the visual representation of the identified concepts under their UMLS classification in a semantic group and category. The produced concept map is not aiming for the identification of unique relationships but rather for the observation and exploration of its recognized (and normalized) concepts; an overview of the health problem's concepts that have been studied as exposures, outcomes and covariates along with their respective nature (under the UMLS classification in a specific semantic group and category).

3.3. Corpora for Training, Development and Evaluation

3.3.1. Preliminary Annotation Exercise

A gold standard was created in order to develop and evaluate our approach. A set of 40 epidemiological abstracts were selected from the PubMed results returned by querying "*obesity/epidemiology[mesh]"* as a "*mesh*" descriptor by assigning each abstract an id and then randomly selecting a number of ids through a preudo-random number generator. Obesity was chosen because it has risk factors of various nature (e.g., behavioural, biological), shares underlying links with many diseases (e.g., type 2 diabetes, cancer), affects different population types and contains a wealth of related epidemiological data available for research. The set was manually double-annotated by the author and an external 2nd annotator for all six key characteristics. The annotator was suggested by a clinical professor in Public Health Informatics, due to her epidemiological expertise (BSc in Human Genetics, MSc in Epidemiology) making her ideal for annotating characteristics in epidemiological abstracts related to obesity. Any additional annotators would require more time in order to enable the correct calculation of the inter-annotator agreement. Therefore, due to project time constraints only one annotator was used.

The annotation guidelines required highlighting spans in both the abstract text and its title such as:

1. **study design:** the highlighting of a span (particularly noun phrase) that describes the study design including any detailed information e.g., "*double-blind randomized clinical trial*". If a synonym study span appears (more than once), it should be highlighted. Determiners (if any) are not included.
2. **population:** Spans (noun phrases) regarding the participant population in the study. If spans that contain information regarding age, ethnicity, nationality and gender of the population exist in the abstract but are not part of a single most informative span, they are to be also annotated separately (i.e., in addition).
3. **exposure/outcome:** Any spans (single or multi-word noun phrases) with role as an exposure and as an outcome should be highlighted. For example in these example sentences, the concepts that follow after particular syntactical patterns should be annotated without including any articles or coordinating conjunctions (also called coordinators); "*Onset of post generalized anxiety disorder*$_{\text{outcome}}$ *in Chinese obese individuals*", "*obesity*$_{exposure}$ *is a risk factor for cancer*$_{outcome}$". If variations of the same concept (synonyms) appear, they are to be highlighted.
4. **covariate:** Any spans (single or multi-word noun phrases) that describe concepts for which the studies are adjusting for their results, are annotated as covariates. Articles

should be excluded in the annotation process unless they are directly linked through a conjunction (see example below). For example in the following sentence, "*maternal obesity*" and "*high parity*" should be annotated as covariates; "*The disparity in stillbirth rates between Pacific and European women can be attributed to confounding factors such as maternal obesity and high parity*$_{covariate}$". Any generic mentions for covariate concepts should be excluded. For example in the following sentence, "*panel of cognitive risk factors*" should not be annotated; "*They were statistically interactions between BMI and WHR and incident cognitive impairment and probable dementia with and without adjustment for a panel of cognitive risk factors*".

5. **effect size:** Any spans that state and describe the measures of prevalence, incidence, (adjusted) odds ratio, hazard ratio and relative risk, are annotated as effect sizes including the respective confidence interval (when mentioned) and the related concept (when mentioned). These spans contain numbers suggesting the size of the applied measure and possibly a single or multi-word noun phrase that describes the concept associated with that size. In the case of prevalence and incidence, the related concepts and the respective value of the effect size type should be annotated. Any conjunctions and verbs that are part of the effect size statement can be included while articles should not be considered part of the characteristic of interest e.g., in the following sentence "*the prevalence of obesity was 35%*$_{\text{effect size}}$ *in the male population sample*", the span of interest is "*prevalence of obesity was 35%*".

Following the annotation guidelines, the inter-annotator agreement was calculated by the absolute agreement rate (Kim et al. 2006), P(A). The agreement was focused at the document level since the aim of our method is to provide a summarization of the key epidemiological information presented in each abstract. Disagreements in the length of the annotated spans were expected between annotators due to different considerations of the length for the same span (whether to include acronyms in the annotation of mentions where these were introduced e.g., "*alzeheimer's disease (ad)*"). Therefore, inexact matching was accepted. Table 22 reveals the number of agreements and disagreements in inexact matching at the document and mention level in the golden standard of 40 epidemiological study design abstracts.

Table 22: Statistics for the annotation agreements and disagreements between the author and the 2nd annotator for the 40 epidemiological study abstracts at the document and mention level. TP: true positives, annotations in which both the author and the 2nd annotator agreed. FP: false positives, the number of annotations that the 2nd annotator highlighted in text while the author did not. FN: false negatives, the number of annotations that the author highlighted in text but the 2nd annotator did not. The annotations of the external annotator were used as a guide point to see if the author's annotations were of sufficient quality since more time was spent for the abstract annotations (the author annotated an extra 90 abstracts, 30 of the development set and 60 of the evaluation set).

	document level			**mention level**		
	TP	**FN**	**FP**	**TP**	**FN**	**FP**
study design	37	3	1	37	10	1
population	43	7	2	43	54	2
exposure	62	21	17	65	71	18
outcome	53	17	10	56	128	11
covariate	9	2	16	9	3	17
effect size	41	0	14	41	0	14
total number	244	51	60	250	267	63

A lower agreement at the mention level was expected since :

1. the 2nd annotator has chosen to highlight mainly one span for each characteristic (usually the most informative) while ignored any (synonymous) repeats;
2. the author annotated more than once the same concept, synonymous or not (mainly for the characteristic of population, exposure and outcome) presented in an abstract.

The above reasons led to an agreement of 29.4% at the mention level for the 40 epidemiological abstracts. On the other hand, the inter-annotator agreement at the document level was found to be 61.5% suggesting relatively reliable annotations. However, common sources of disagreement were observed to be:

1. the difference in annotations in the abstract titles;
2. different annotations of exposure and outcome concepts at the document level - either synonymous concepts were selected or spans representing a different meaning;
3. the identification of the most informative population concept.

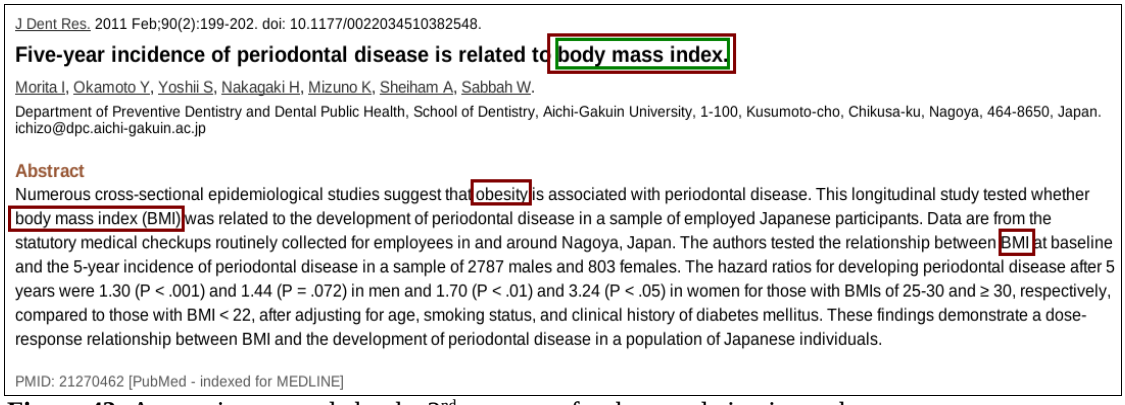
For the recognition of the study design characteristic at the document level, the agreement was relatively good. The only source of disagreement here was the generic mentions of "*epidemiological study*", which were included by the author. For the characteristics of exposure and outcome, the 2nd annotator did not curate any possible mentions in the abstract titles, in contrast to the author. They highlighted concepts of interest in text where their role

was more clear. As it was expected, in a number of cases, the abstract title does not explicitly suggest the role of clinical concepts; i.e., which could be either the exposure or the outcome of a study. For example, the title from a MEDLINE abstract[18]:

"*Depressive symptoms clusters*$_{exposure}$ are differentially associated with *general and visceral obesity*$_{outcome}$"

reports both potential exposure and outcome concepts. The concepts of "depressive symptoms clusters" and "general and visceral obesity" were considered exposure and outcome respectively from a hypothesis generated by the author that the first concept is the variable under examination and the second is the potential consequence of its exposure. However, this could be interpreted vice versa – that the first concept is the outcome and the second is the exposure. Due to the ambiguous relationship in such cases, the 2nd annotator preferred to highlight the respective concepts in text, where their role is more clear.

Additionally, a small number of disagreements were observed in the annotations of exposure and outcome concepts in text. For example, Figure 43 reveals the annotations of the author (highlighted in red) and the 2nd annotator (highlighted in green) for the exposure characteristic in an epidemiological abstract. While the first author identified "*body mass index*", "*obesity*", "*body mass index (bmi)*" and "*bmi*" as potential exposures, the 2nd annotator highlighted only the "*body mass index*".


J Dent Res. 2011 Feb;90(2):199-202. doi: 10.1177/0022034510382548.

Five-year incidence of periodontal disease is related to body mass index.

Morita I, Okamoto Y, Yoshii S, Nakagaki H, Mizuno K, Sheiham A, Sabbah W.

Department of Preventive Dentistry and Dental Public Health, School of Dentistry, Aichi-Gakuin University, 1-100, Kusumoto-cho, Chikusa-ku, Nagoya, 464-8650, Japan. ichizo@dpc.aichi-gakuin.ac.jp

Abstract

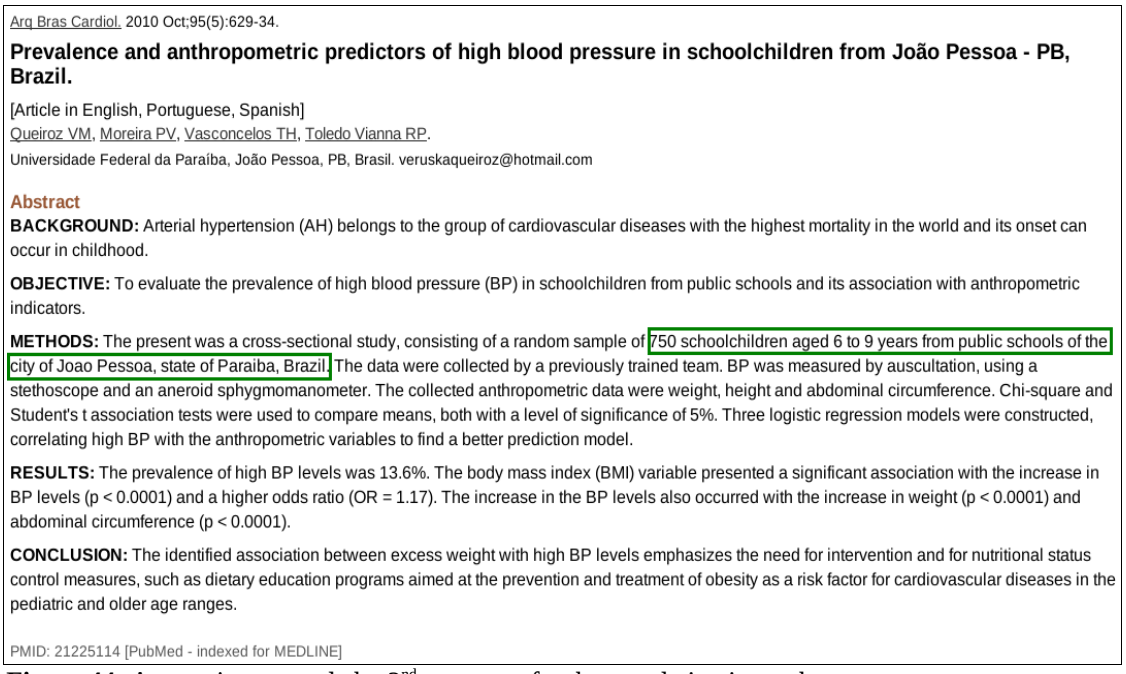
Numerous cross-sectional epidemiological studies suggest that obesity is associated with periodontal disease. This longitudinal study tested whether body mass index (BMI) was related to the development of periodontal disease in a sample of employed Japanese participants. Data are from the statutory medical checkups routinely collected for employees in and around Nagoya, Japan. The authors tested the relationship between BMI at baseline and the 5-year incidence of periodontal disease in a sample of 2787 males and 803 females. The hazard ratios for developing periodontal disease after 5 years were 1.30 (P < .001) and 1.44 (P = .072) in men and 1.70 (P < .01) and 3.24 (P < .05) in women for those with BMIs of 25-30 and ≥ 30, respectively, compared to those with BMI < 22, after adjusting for age, smoking status, and clinical history of diabetes mellitus. These findings demonstrate a dose-response relationship between BMI and the development of periodontal disease in a population of Japanese individuals.

PMID: 21270462 [PubMed - indexed for MEDLINE]

Figure 43: Annotation example by the 2nd annotator for the population in an abstract.

The variations of the population mentions in the abstracts led to some disagreements. One annotator highlighted only the most detailed (or informative) spans for the studied population (Figure 44) while the other additionally curated more general mentions (Figure 45).

18 http://www.ncbi.nlm.nih.gov/pubmed/21226677



Arq Bras Cardiol. 2010 Oct;95(5):629-34.

Prevalence and anthropometric predictors of high blood pressure in schoolchildren from João Pessoa - PB, Brazil.

[Article in English, Portuguese, Spanish]

Queiroz VM, Moreira PV, Vasconcelos TH, Toledo Vianna RP.

Universidade Federal da Paraíba, João Pessoa, PB, Brasil. veruskaqueiroz@hotmail.com

Abstract

BACKGROUND: Arterial hypertension (AH) belongs to the group of cardiovascular diseases with the highest mortality in the world and its onset can occur in childhood.

OBJECTIVE: To evaluate the prevalence of high blood pressure (BP) in schoolchildren from public schools and its association with anthropometric indicators.

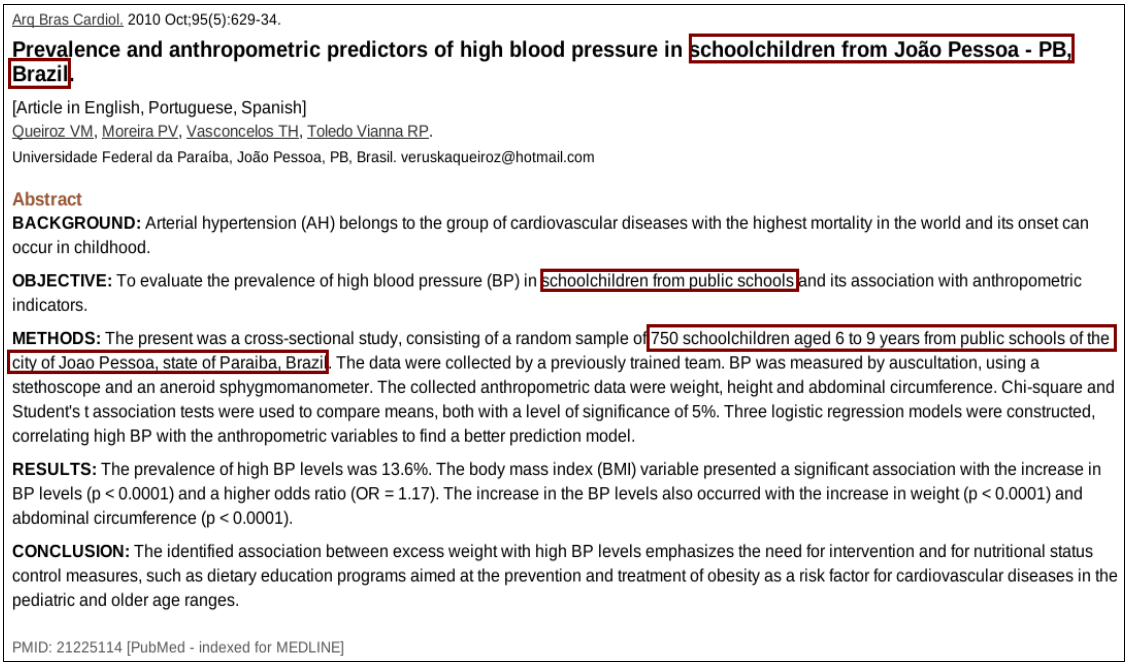
METHODS: The present was a cross-sectional study, consisting of a random sample of 750 schoolchildren aged 6 to 9 years from public schools of the city of Joao Pessoa, state of Paraiba, Brazil. The data were collected by a previously trained team. BP was measured by auscultation, using a stethoscope and an aneroid sphygmomanometer. The collected anthropometric data were weight, height and abdominal circumference. Chi-square and Student's t association tests were used to compare means, both with a level of significance of 5%. Three logistic regression models were constructed, correlating high BP with the anthropometric variables to find a better prediction model.

RESULTS: The prevalence of high BP levels was 13.6%. The body mass index (BMI) variable presented a significant association with the increase in BP levels ($p < 0.0001$) and a higher odds ratio (OR = 1.17). The increase in the BP levels also occurred with the increase in weight ($p < 0.0001$) and abdominal circumference ($p < 0.0001$).

CONCLUSION: The identified association between excess weight with high BP levels emphasizes the need for intervention and for nutritional status control measures, such as dietary education programs aimed at the prevention and treatment of obesity as a risk factor for cardiovascular diseases in the pediatric and older age ranges.

PMID: 21225114 [PubMed - indexed for MEDLINE]

Figure 44: Annotation example by 2nd annotator for the population in an abstract.



Arq Bras Cardiol. 2010 Oct;95(5):629-34.

Prevalence and anthropometric predictors of high blood pressure in schoolchildren from João Pessoa - PB, Brazil.

[Article in English, Portuguese, Spanish]

Queiroz VM, Moreira PV, Vasconcelos TH, Toledo Vianna RP.

Universidade Federal da Paraíba, João Pessoa, PB, Brasil. veruskaqueiroz@hotmail.com

Abstract

BACKGROUND: Arterial hypertension (AH) belongs to the group of cardiovascular diseases with the highest mortality in the world and its onset can occur in childhood.

OBJECTIVE: To evaluate the prevalence of high blood pressure (BP) in schoolchildren from public schools and its association with anthropometric indicators.

METHODS: The present was a cross-sectional study, consisting of a random sample of 750 schoolchildren aged 6 to 9 years from public schools of the city of Joao Pessoa, state of Paraiba, Brazil. The data were collected by a previously trained team. BP was measured by auscultation, using a stethoscope and an aneroid sphygmomanometer. The collected anthropometric data were weight, height and abdominal circumference. Chi-square and Student's t association tests were used to compare means, both with a level of significance of 5%. Three logistic regression models were constructed, correlating high BP with the anthropometric variables to find a better prediction model.

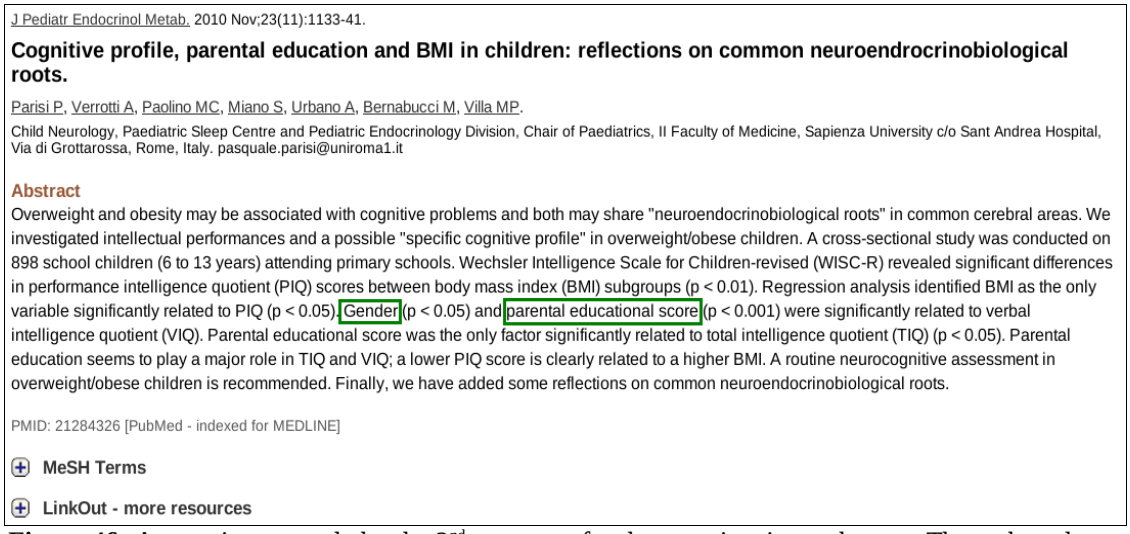
RESULTS: The prevalence of high BP levels was 13.6%. The body mass index (BMI) variable presented a significant association with the increase in BP levels ($p < 0.0001$) and a higher odds ratio (OR = 1.17). The increase in the BP levels also occurred with the increase in weight ($p < 0.0001$) and abdominal circumference ($p < 0.0001$).

CONCLUSION: The identified association between excess weight with high BP levels emphasizes the need for intervention and for nutritional status control measures, such as dietary education programs aimed at the prevention and treatment of obesity as a risk factor for cardiovascular diseases in the pediatric and older age ranges.

PMID: 21225114 [PubMed - indexed for MEDLINE]

Figure 45: Annotation example by the author for the population in an abstract.

The major differences in the annotations of the covariate concepts were observed due to the limited epidemiological expertise from the author's part. The author highlighted only a small number of concepts that could be classified as covariates, most of them explicitly stated in text. However, the 2nd annotator recognised other covariate concepts that required epidemiological expertise in order to be identified. An example regarding this can be seen in Figure 46 in which the author did not highlight any covariate mentions but the 2nd annotator did.



J Pediatr Endocrinol Metab. 2010 Nov;23(11):1133-41.

Cognitive profile, parental education and BMI in children: reflections on common neuroendrocrinobiological roots.

Parisi P, Verrotti A, Paolino MC, Miano S, Urbano A, Bernabucci M, Villa MP.

Child Neurology, Paediatric Sleep Centre and Pediatric Endocrinology Division, Chair of Paediatrics, II Faculty of Medicine, Sapienza University c/o Sant Andrea Hospital, Via di Grottarossa, Rome, Italy. pasquale.parisi@uniroma1.it

Abstract

Overweight and obesity may be associated with cognitive problems and both may share "neuroendocrinobiological roots" in common cerebral areas. We investigated intellectual performances and a possible "specific cognitive profile" in overweight/obese children. A cross-sectional study was conducted on 898 school children (6 to 13 years) attending primary schools. Wechsler Intelligence Scale for Children-revised (WISC-R) revealed significant differences in performance intelligence quotient (PIQ) scores between body mass index (BMI) subgroups ($p < 0.01$). Regression analysis identified BMI as the only variable significantly related to PIQ ($p < 0.05$). Gender ($p < 0.05$) and parental educational score ($p < 0.001$) were significantly related to verbal intelligence quotient (VIQ). Parental educational score was the only factor significantly related to total intelligence quotient (TIQ) ($p < 0.05$). Parental education seems to play a major role in TIQ and VIQ; a lower PIQ score is clearly related to a higher BMI. A routine neurocognitive assessment in overweight/obese children is recommended. Finally, we have added some reflections on common neuroendocrinobiological roots.

PMID: 21284326 [PubMed - indexed for MEDLINE]

MeSH Terms

LinkOut - more resources

Figure 46: Annotation example by the 2nd annotator for the covariate in an abstract. The author, due to limited epidemiological expertise did not annotate these two mentions that were considered as covariates by the 2nd annotator.

For the effect size characteristic, there was no disagreement between the author and the 2nd annotator (no false negatives were detected) due to to the more explicit (numerical) nature of the related concepts and its relatively well-defined mentions in epidemiological text. Although a small number of false positives appeared at the document level, after careful review, they were not related to any of the desirable measures mentioned in the annotation guidelines.

For the above examples and observations, it was concluded that potentially the annotation guidelines were not explicit and detailed enough, in particular for the characteristics of the population, exposure and outcome spans. Despite a relatively reliable agreement at the document level, the guidelines should be more specific to avoid confusion between the selection process of exposure and outcomes while further epidemiological expertise from the author could potentially enhance the agreement for the covariate spans.

As a result, a new, more detailed set of guidelines was followed for further annotation.

3.3.2. Training, Development and Evaluation Sets

The training set used for the identification and normalization of key epidemiological characteristics included the 40 abstracts of the preliminary exercise and a further collection of 20 articles. Since the inter-annotator agreement on the 20 abstracts returned relatively reliable results (80%), the set was deemed appropriate. Since epidemiological abstracts are relatively longer than abstracts of other scientific context, manual annotation by the author required a significant amount of time. Hence, the number of 60 abstracts was selected due to the time constraints inflicted upon the completion of this thesis, as an appropriate balance between time investment and sample size. A further set of 30 epidemiological abstracts related to a health

problem also manually curated by the author were utilized as a development set to optimize the performance of the proposed methodology. The set used for the evaluation of the proposed approach included 60 epidemiological abstracts related to obesity, manually annotated by the author. Table 23 reveals the number of annotations (at the document and mention level) for each of the epidemiological characteristics in the training, development and evaluation sets.

Table 23: Number of annotations by the author for the six characteristics in the training, development and evaluation sets at the document and mention levels.

epidemiological characteristic	training set		development set		evaluation set	
	mention level	document level	mention level	document level	mention level	document level
study design	69	38	15	13	16	13
population	134	99	50	40	55	39
exposure	189	118	77	59	76	56
outcome	258	133	98	66	119	86
covariate	14	13	16	13	21	17
effect size	50	50	55	55	75	75
total number	714	451	311	246	362	286

In chapters 4, 5 and 6 we provide a detailed analysis of the designed and implemented rule based approach for the identification and normalization of key characteristics in epidemiological study abstracts in order to create automatically a concept map.

Chapter 4

Epidemiological Characteristics Extraction

"Everybody gets so much information all day long that they lose their common sense."

Gertrude Stein, 1946

The first task of this project is to design a text mining approach in order to enable the identification of key epidemiological characteristics at the mention level. In this chapter, the information extraction methods are explained:

1. creation of vocabularies for the detection of key characteristics in epidemiological text;
2. identification of biomedical concepts (for the characteristics of exposure, outcome, covariate);
3. key characteristic extraction through the design and implementation of a rule based approach.

4.1. Creation of Vocabularies

A number of semantic classes are identified using custom-made vocabularies that include both unique and synonymous terms in order to detect key characteristics in epidemiological study abstracts. These vocabularies contain mostly generic terms regarding the role of a concept in epidemiological studies and they can be potentially used in other tasks. A total of fourteen vocabularies were created and utilized. Table 24 reveals the vocabularies used along with a brief description and examples.

Table 24: A description of each of the vocabularies in the information extraction step along with examples (continues on the next page).

vocabulary	description	example
epidemiological study design	a collection of epidemiological study designs (e.g., population, epidemiological, primary research, secondary research, experimental, observational, meta-analysis, review) along with (if any) their respective attributes. It contains a total of 60 epidemiological types, identified from the main nodes of the expanded OCRe ontology (Chapter 5). The vocabulary excludes any studies that can have more than one attribute in their design.	"*randomised*", "*epidemiological*", "*non-blinded*", "*cross-sectional*", "*retrospective*", "*serial*"
study synonyms	a group of words (a total of 11) that indicates that a study was conducted.	"*analysis*", "*study*", "*trial*"
population sample noun	a small collection of clue words (a total of 18) that suggests the observed population sample in an epidemiological study.	"*total*", "*sample*", "*group*", "*dataset*"
verb synonym	a compilation of verbs (a total of 54) that indicates the involvement of a population group in the study.	"*studied*", "*comprised*", "*collected*", "*observed*"
population	a cluster of words (a total of 60) that health professionals utilize to generically describe the population of the study.	"*patients*", "*cases*", "*individuals*", "*children*"

Table 24: A description of each of the vocabularies in the information extraction step along with examples.

exposure related clue	a group of words along with their synonyms (a total of 30) indicating that a concept can be an exposure in an epidemiological study.	"*risk factor*", "*risk variant*", "*marker*", "*predictor*"
outcome related clue	a collection of words (a total of 15) suggesting the measurement of an outcome in an epidemiological study.	"*diagnosis*", "*prevalence*", "*incidence*", "*occurrence*"
exposure-outcome association	a compilation of words (a total of 53) that suggest a link or an association between two concepts (usually between the exposure and the outcome concept(s)).	"*association*", "*relation*", "*role*", "*influence*", "*impact*", "*related*", "*associated*"
covariate related clue	a group of words (a total of 24) showcasing that a concept in epidemiological text suggests a covariate.	"*covariable*", "*confounder*", "*covariate*"
effect size	a collection of the available effect size measures (a total of 21) that are used in epidemiological studies to describe the strength of associations among concepts.	"*hazard ratio*", "*adjusted odds ratio*", "*prevalence*"
epidemiological abstract structure	a group of words (a total of 18) referring to various structure sections of an epidemiological study abstract.	"*background*", "*design*", "*research methods*", "*conclusions*
association adverbs	a compilation of words (a total of 13) referring to adverbs that describe the association(s) between concepts (usually between the exposure and the outcome concept(s)).	"*inversely*", "*strongly*", "*similarly*", "*positively*"
adjustment	a cluster of words (a total of 8) that reveal the adjustment of an epidemiological study for a covariate.	"*controlling*", "*adjusting*", "*adjusted*", "*adjustment*"
numbers	an arithmetic vocabulary that includes numbers and words that express numbers.	"*1*", "*twenty two*"
percentage	an arithmetic vocabulary that includes percentages.	"*80%*", "*72.3%*"

4.2. Identification of Biomedical Concepts

For the recognition of biomedical concepts that belong potentially to the characteristics of exposure, outcome and covariate, the Specialist lexicon[19] was applied. Special Lexicon contains a large variety of concepts and commonly occurring English words (more than 415,000), although its focus is in the biomedical domain rather than the clinical one. Additionally, in order to expand the dictionary resources, the related corpus of epidemiological abstracts is processed through the ATR C-value method for the extraction of multi-word candidate concepts (Fratzi et al. 1997). ATR was selected due to its successful application in various text mining problems of biomedical nature showcasing encouraging performance (Nenadic et al. 2004). The returned results contained multi-word concepts based on their document frequency and overall frequency in the corpus. Filtering was applied aiming to remove any concepts of non-biomedical nature, hence improving the resulting dictionary. A common stop-word list was used (created by Fox (1989) and was manually expanded through an empirical validation on the training set.

19 http://lexsrv3.nlm.nih.gov/Specialist/Summary/lexicon.html

4.3. Rules for Epidemiological Characteristics Extraction

A set of text based rules are applied to the corpus with a combination of the dictionaries for each epidemiological characteristic. We used MinorThird (Cohen, 2004), a toolkit of learning methods for storing, annotating and categorizing text as well as recognising entities of interest. It was chosen because it supports rule based approaches while enabling the user to navigate the identified data through a GUI. MinorThird applies tokenization with words or symbols (e.g., "*?*", "*:*", "*;*", etc.) being considered as tokens. The recognised tokens are called spans. A span can be either a single token or multiple ones.

The rules were designed and based on semantic patterns observed in epidemiological text. The semantic patterns make use of two constituent types and are introduced as follows:

1. **frozen lexical expressions** that are used as anchors for specific categories;
2. specific **semantic classes** identified through the vocabularies.

Particularly, the semantic patterns are specific combinations of lexical expressions and semantic classes that indicate the presence of a key characteristic in text. More than one of semantic patterns can exist in one epidemiological study referring either to one characteristic or to many e.g., "*X is associated with Y*", "*after adjusting for confounders such as X, Y, Z*". The frozen lexical expressions can contain particular verbs to prepositions and certain noun phrases and their translation into the rule design includes the usage of regular expressions and the creation of vocabularies.

Through the careful inspection of the training set, common semantic patterns for each of the key characteristics were observed and then used for the design and implementation of the rule based approach. Additionally, after analysing the development set, the rules were expanded to include more similar semantic patterns for each characteristic.

In order to identify a potential span of interest, a semantic pattern in text has to be matched by the related rule and the respective vocabularies. If so, the defined concept will be recognised through the Specialist lexicon and the ATR integrated dictionary resources. Consequently, *candidate mentions* of epidemiological concepts are tagged in text when these are matching any of the concepts in the applied dictionaries.

Table 25 shows the number of rules created for each characteristic and tables 26, 27, 28, 29, 30, and 31 present (detailed) examples of extracted concepts (those being surrounded in square brackets in the rule syntax).

Table 25: Number of rules for the recognition of each characteristic.

epidemiological characteristic	number of rules
study design	16
population	119
exposure	134
outcome	100
covariate	28
effect size	15
total number	412

4.3.1. Study Design

The mentions of study design are relatively straightforward to extract in epidemiological text. In most cases, the mentions of epidemiological study designs in related abstracts are specific enough that can be detected through a usage of domain specialized words that indicate their presence e.g., "*retrospective birth cohort study*". An application of two vocabularies (@st, a(types) –see Table 24 first and second vocabulary respectively) that contain the most common study designs and synonyms of study research (e.g., "*research*", "*trial*", "*study*") can potentially assist in the identification of this characteristic's spans. Some examples of identified study design spans can be seen in Table 26. The rules are following the MinorThird notation[20].

Table 26: Rule examples for the recognition of study design in study abstracts. The rule components in square brackets are the extracted spans that denote the key characteristic.

examples	identified tokens	
Methods: This was a cross-sectional study of 214 overweight/obese ...	**cross-sectional**	**study**
A systematic review of adenotonsillectomy as a risk factor for ...	**systematic**	**review**
Design: prospective cohort study. Setting: ...	**prospective cohort**	**study**
rule	[@st	a(types)]
extraction	@st matches the vocabulary's "*st*" single and multi-word epidemiological study designs e.g., "*cross-sectional*", "*systematic*", "*prospective cohort*" spans in the epidemiological text. a(types) matches the single words that are included in the vocabulary "*types*"; synonyms words for "*study*" e.g., "*survey*", "*analysis*", "*approach*".	

4.3.2. Population

The recognition of population mentions in epidemiological text is challenging since it requires the identification of spans that suggest key attributes such as the participants' age, the sample

20 @X – the multi-word concepts of dictionary X are matched against the text
a(x) – the single word concepts of dictionary X are matched against the text
re('X')- the regular expression X is matched against the text
eq('X') – the token X is matched against the text.

number, their gender as well as their nationality and ethnicity. Hence, the design of the related rules is based on more complex and varied semantic patterns than the ones observed for the characteristic of study design. The number of recognised vocabularies has been increased: they contain clusters of particular verbs and nouns and their combination with regular expressions that cover numerical sequences and prepositions can suggest the participant population. Therefore, the rules can belong to one of the following categories:

1. those built around prepositions;
2. those built around noun phrases that belong to specific dictionaries;
3. those built around noun phrases that belong to specific dictionaries and prepositions;
4. those built around noun phrases that belong to specific dictionaries and verbs.

Detailed examples can be seen below in Table 27 with rules including:

- regular expressions for the matching of prepositions (of, on, in, with);
- various vocabularies ("*total*", "*clusters*", "*sample*" - see Table 24 population sample noun vocabulary) for the identification of semantic classes that suggest the presence of the study population combined with prepositions (of, on, in, with);
- numeric vocabularies ("*stats2*" - see Table 24 numbers vocabulary);
- multi-word biomedical concepts related to the observed participant sample ("*multiple*"); in order to cover potential biomedical concepts that are multi-word terms but are not included in the Special Lexicon or returned by ATR we also include all possible combinations of Specialist Lexicon terms (as a Cartesian product operation). For example, if the following concepts "*eating*" and "*behaviour"* are included in the Specialist Lexicon, with the implementation of the above mentioned rule, an extra term is being generated: "*eating behaviour*". This potentially may increase the chances of detecting more complex concepts in text as well as to detect more than one concepts detected in one semantic pattern (e.g., "***weight gain, alcohol consumption*** *and* ***smoking*** *are associated with obesity*").

Table 27: Rule examples for the recognition of population in study abstracts. The rule components in square brackets are the extracted spans that denote the key characteristic (continuing on the next page).

examples	**identified tokens**			
rule1	a(totals)	re('(of|on|in)')	[@stats2	a(clusters)]
	a(totals) matches the terms contained in the vocabulary "*totals*" that indicate the population participating in the study e.g., "*cohort*", "*sample*". re('(of|on|in)') is a regular expression that matches any of the three "*of*", "*in*" or "*on*". @stats2 matches the multiplied individual arithmetic tokens belonging in the dictionary named "*stats2*". a(clusters) identifies the collection of words in the vocabulary "*clusters*", which are used to describe generically the study population e.g., "*men*", "*individuals*", "*participants*", "*girls*". The presence of "*?*" suggests that this particular part of the rule is not necessary to be matched.			
Sibling study in a prospective cohort of 208,866 men from ...	cohort	of	**208,866**	**men**

Table 27: Rule examples for the recognition of population in study abstracts. The rule components in square brackets are the extracted spans that denote the key characteristic.

rule2	a(verbs)	re('with|in|on')?	[@stats2	a(clusters)	@age?]
	a(verbs) matches the terms of the vocabulary "*verbs*" that includes verbs suggesting the realization of an epidemiological study, e.g., "*conducted*", "*performed*". re('(with|in|on)') is a regular expression that matches any of the three "*with*", "*in*" or "*on*". @stats2 matches the multiplied individual arithmetic tokens belonging in the vocabulary named "*stats2*". a(clusters) identifies the collection of words in the dictionary "*clusters*", which are used to describe generically the study population e.g., "*men*", "*individuals*", "*participants*", "*girls*". @age? Matches the numeric concepts of the vocabulary "*age*", *that* contains multiple tokens that can be a part of an age expression e.g., "*6 to 13 years*" or "*six to thirteen years*" or "*6 to 13 years old*". The presence of "*?*" suggests that this particular part of the rule is not necessary to be matched.				
... study was conducted on 898 schoolchildren (6 to 13 years) ...	conducted	on	**898**	**schoolchildren**	**(6-13 years)**
rule3	@multiple	re('with|in|on')?	[a(clusters)	re('with|without')	@multiple]
	@multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon. re('(with|in|on)') and re('with|without') are regular expression that match any of the respective tokens. a(clusters) identifies the collection of words in the vocabulary "*clusters*", which are used to describe generically the study population e.g., "*men*", "*individuals*", "*participants*", "*girls*". The presence of "*?*" suggests that this particular part of the rule is not necessary to be matched.				
Bone mineral density in patients with type 2 diabetes	Bone mineral density	in	**patients**	**with**	**type 2 diabetes**

4.3.3. Exposures

Exposure concepts are difficult to detect in epidemiological text. The same exposure can have many synonymous forms in the same text so the implementation of a large scale dictionary was considered necessary. Hence, there are six categories of exposure rules based on their design:

1. those built around noun phrases that belong to specific dictionaries or verbs (see Table 24 exposure-related clue and association verbs vocabularies);
2. those built around noun phrases that belong to specific dictionaries or verbs and prepositions;
3. those built around noun phrases that belong to specific dictionaries, prepositions and articles;
4. those built around noun phrases that belong to specific dictionaries and adverbs;
5. those built around verbs and prepositions;
6. those built around verbs, noun phrases that belong to specific dictionaries and prepositions;

Similarly for the reason stated for the characteristic of population, the dictionary "*multiple*" is used to detect (if possible) variations of the biomedical concepts in epidemiological text through the triggering of specific rules. Table 28 shows some detailed rule examples.

Table 28: Rule examples for the recognition of exposure in study abstracts. The rule components in square brackets are the extracted spans that denote the key characteristic.

examples	**identified tokens**						
rule1	a(relations)	eq('of')	[@multiple]		re('in|on|of')		@multiple
	a(relations) recognises the words contained in the vocabulary "*relations*" indicating a relation between two or more concepts, e.g., "*role*", "*association*", "*link*", "*impact*". eq('of') matches the "*of*" token. @multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon. re('(with|in|on)') is a regular expression that matches any of the three prepositions "*with*", "*in*" or "*on*".						
... the pathogenic role of hyperhomocysteinemia in cryptogenic stroke is ...	role	of	**hyperhomocysteinemia**		in		cryptogenic stroke
rule2	a(relations)	eq('between')	[@multiple]		eq('and')		@multiple
	a(relations) recognises the words contained in the vocabulary "*relations*" indicating a relation between two or more concepts, e.g., "*role*", "*association*", "*link*", "*impact*". eq('between') matches the "*between*" token. @multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon. eq('and') matches the "*and*" token.						
... and analyze the association between body mass index and blood pressure in ...	association	between	**body mass index**		and		blood pressure
rule3	[@multiple]	a(be)	a(related)	a(with)	eq('onset')?	eq('of')?	@multiple
	@multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon. a(be) matches the variations of the verb "*be*", e.g., "*is*", "*are*" in the dictionary "*be*". a(related) identifies the words contained in the vocabulary "*related*", which are verbs suggesting a relationship among two or more concepts of interest, e.g., "*related*", "*linked*". a(with) matches two tokens: "*to*" and "*with*" that are included in the vocabulary "*with*". eq('onset')? matches the token "onset". If not, the rule will progress to the next part (eq('of')). eq('of') matches the "*of*" token. The presence of "*?*" suggests that this particular part of the rule is not necessary to be matched.						
Short sleep duration is associated with onset of obesity	**Short sleep duration**	is	associated	with	onset	of	obesity

4.3.4. Outcomes

Similar rule design was followed for the identification of outcome concepts. Outcome mentions are relatively easier to detect since they involve the utilization of common semantic patterns observed in text. The identification of outcomes requires a large dictionary due to the nature of epidemiological text. A concept can have more than one synonym and lexical resources with a variety of terms can potentially increase the possibility of recognising it. Table 29 provides a few detailed examples regarding the rules created for the identification of outcomes.

Table 29: Rule examples for the recognition of outcome in study abstracts. The rule components in square brackets are the extracted spans that denote the key characteristic.

examples	**identified tokens**					
rule1	@factors		eq('of')		[@multiple]	
	@factors matches the terms suggesting an outcome, e.g., "*risk factor*", "*predictor*", "*protective factor*" included in the single and multi-word vocabulary "*factors*". eq('of') matches the "*of*" token. @multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon.					
Cardiovascular and disease related predictors of depression	predictors		of		**depression**	
rule2	@multiple	a(be)	a(adverbs)	a(related)	a(with)	[@multiple]
	@multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon. a(be) matches the variations of the verb "*be*", e.g., "*is*", "*are*" in the vocabulary "*be*". a(adverbs) identified the adverbs used to describe the association between exposure and outcome, e.g., "*inversely*", "*strongly*", in the vocabulary "*adverbs*". a(related) identifies the words contained in the vocabulary "*related*", which are verbs suggesting a relationship among two or more concepts of interest, e.g., "*related*", "*linked*". a(with) matches two tokens: "*to*" and "*with*" that are included in the vocabulary "*with*".					
Conclusions coffee intake is inversely associated with t2dm in Chinese.	coffee intake	is	inversely	associated	with	**t2dm**
rule3	@factors	eq('for')	[@multiple]			
	@factors matches the terms suggesting an outcome, e.g., "*risk factor*", "*predictor*", "*protective factor*" included in the single and multi-word vocabulary "*factors*". eq('for') matches the "*for*" token. @multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon.					
Body mass index is an established risk factor for post-menopausal breast cancer.	risk factor	for	**post-menopausal breast cancer**			

4.3.5. Covariates

The rules created for the identification of covariate mentions in epidemiological study text followed a similar design like those mentioned in the characteristics of exposure and outcome. However, it was relatively straightforward to detect common semantic patterns on text that could suggest the presence of a covariate concept. The rules can be classified under the following categories:

1. those built around noun phrases that belong to specific dictionaries with or without verbs (see Table 24 covariate related clue vocabulary);
2. those built around noun phrases that belong to specific dictionaries and prepositions;
3. those built around noun phrases that belong to specific dictionaries, verbs and adverbs (see Table 24 covariate related clue and adjustment vocabularies);
4. those built around noun phrases that belong to specific dictionaries, idioms with or without adverbs;
5. those built around noun phrases that belong to specific dictionaries and articles with or without prepositions;

6. those built around noun phrases that belong to specific dictionaries, prepositions and either verbs or related idioms;

Table 30 shows some rules for the recognition of covariate concepts in epidemiological text.

Table 30: Rule examples for the recognition of covariate in study abstracts. The rule components in square brackets are the extracted spans that denote the key characteristic.

examples	**identified tokens**				
rule1	a(adj)	eq('for')	[@multiple]		
	a(adj) matches words that suggest standardization of the study regarding covariates e.g., "*adjustment*", "*controlling*", "*standardized*", in the vocabulary of "*adj*". eq('for') matches the "*for*" token. @multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon.				
... after adjusting for **age**, **smoking status**, and **clinical history of diabetes mellitus**.	adjusting	for	**age**, **smoking status**, and **clinical history of diabetes mellitus**.		
rule2	a(adj)	eq('for')	[@multiple]		
	a(adj) matches words that suggest standardization of the study regarding covariates e.g., "*adjustment*", "*controlling*", "*standardized*", in the vocabulary of "*adj*". eq('for') matches the "*for*" token. @multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon.				
Adjusting for **physical activity** attenuated these associations ...	Adjusting	for	**physical activity**		
rule3	eq('including')	[@multiple]		eq('as')	@synonyms
	eq('including') matches the "*including*" token. @multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon. eq('as') matches the "*as*" token. @synonyms identifies the single and multi-word terms of the "*synonyms*" vocabulary that refer to covariates, e.g., "*confounding variables*", "*covariable factors*", "*confounder*".				
... including **visceral adipose tissue (vat)** and **subcutaneous adipose tissue (sat)** as covariates.	including	**visceral adipose tissue (vat)** and **subcutaneous adipose tissue (sat)**		as	covariates

4.3.6. Effect Size

Recognition of effect size mentions relies on mentions of numerics and use of specific vocabularies. Particularly, the rules included the utilization of same numerical dictionaries applied to other characteristics – with the addition of extra ones that focus on different types of numerical data e.g., number percentages ("*perce*" - see Table 24 percentage vocabulary) and the inclusion of vocabularies describing the type of effect size ("*preva*", "*or*" - see Table 24 effect size vocabulary). The covariate rules can be found in the following categories according to their design:

1. those built around noun phrases that belong to the various dictionaries and numerical data;
2. those built around noun phrases that belong to the various dictionaries, prepositions and numerical data;

Table 31 reveals some rule examples for the recognition of covariate concepts in epidemiological text.

Table 31: Rule examples for the recognition of effect size in study abstracts. The rule components in square brackets are the extracted spans that denote the key characteristic.

examples	**identified tokens**			
rule1	@multiple	[a(preva)	a(be)	@perce]
	@multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon. a(preva) recognises the epidemiological terms regarding population proportions and risk measurements, e.g., "*prevalence*" and" *incidence*" that are included in the "*preva*" vocabulary. a(be) matches the variations of the verb "*be*", e.g., "*is*", "*are*" in the vocabulary "*be*". @perce matches the variations of percentages (both arithmetic and lexical) of the vocabulary *"perce"*.			
Hernia **prevalence was 32.4%**	Hernia	**prevalence**	was	**32.4%**
rule2	@multiple	@or		@ci
	@multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon. @or identifies the various effect size measures e.g., "*relative risk*", "*adjusted odds ratio*" that are included in the vocabulary "*or*" along with their respective acronyms as well as their numeric values. @ci matches the various formats contained in the "*ci*" vocabulary that refer to the confidence interval.			
Factors associated with higher risk of gi track leak were **open gastric bypass (adjusted odds ratio [aor], 4.85**) ...	**open gastric bypass**	**adjusted odds ratio [aor]**		**4.85**
rule3	@multiple	@or		@ci
	@multiple matches the single and multi-word biomedical concepts of the "*multiple*" dictionary that belong in the Specialist lexicon. @or identifies the various effect size measures e.g., "*relative risk*", "*adjusted odds ratio*" that are included in the vocabulary "*or*" along with their respective acronyms as well as their numeric values. @ci matches the various formats contained in the "*ci*" vocabulary that refer to the confidence interval.			
... more likely to have **elevated blood pressure (or = 9.05, 95% ci: 1.44, 56.83)**	Elevated blood pressure	**(or = 9.05,**		**95% ci: 1.44, 56.83)**

4.4. Evaluation and Results

In order to evaluate the performance of the rule based approach, precision, recall and F-score were calculated for each of the characteristics at the mention and document level. The returned values of the development set were compared to those of the evaluation set (See Table 32). Table 32 reveals the values of precision, recall and F-score for all six characteristics in the training, development and evaluation set at the document level. Nevertheless, despite focusing at the document level performance, mention level efficiency is still been calculated and reported as Table 33 shows. A true positive (TP) is the exact matching of an identified characteristic span with an annotated concept in text by the author. A false positive (FP) is a span that was not highlighted by the author and was identified as a key characteristic whereas a false negative (FN) is an annotated by the author span of interest that suggests an instance of an epidemiological characteristic in a study abstract, which the system failed to detect.

4.4.1. Document Level Evaluation

The observed precision at the evaluation set ranged from 79.3%-100.0%, while recall ranged from 80.3%-100.0%. The best precision was observed for study design (100.0%) with population having the second best one (97.2%) followed close by covariate (97.0%). The lowest precision was noted for outcome (79.3%). However, despite having a relatively large number of study design mentions in the training set (a total of 38), the subsequent sets of development and evaluation had only 1/3 of those. Therefore the precision value of 100.0% should be taken with caution. Study design, population and covariate saw their precisions increasing (8.4%, 7.2%, and 8.2% respectively) when compared to those of the development set. Effect size was noted to have big increase from 74.6% to 97.5% in precision, the largest in any characteristic. On the other hand, exposure and outcome had a small decrease (8.7% and 4.0%).

Table 32: Values of TP, FP, FN, P, R and F for the training, development and evaluation sets for the extraction process of study design, population, exposure, outcome, covariate and effect size concepts at the document level. Micro averages are calculated across all different document level mentions; macro averages are calculated across different characteristics at the document level.

.	Training set (60 abstracts)						Development set (30 abstracts)						Evaluation set (60 abstracts)					
	TP	**FP**	**FN**	**P(%)**	**R(%)**	**F(%)**	**TP**	**FP**	**FN**	**P(%)**	**R(%)**	**F(%)**	**TP**	**FP**	**FN**	**P(%)**	**R(%)**	**F(%)**
Study design	37	5	1	88.0	97.3	92.5	11	1	2	91.6	84.6	88.0	12	0	1	100.0	92.3	95.9
Population	94	10	5	90.3	94.9	92.6	36	4	4	90.0	90.0	90.0	35	1	4	97.2	89.7	93.3
Exposure	104	21	14	83.2	88.1	85.5	59	4	0	93.6	100.0	96.7	45	8	11	84.9	80.3	82.5
Outcome	125	26	8	82.7	93.9	88.0	65	13	1	83.3	98.4	90.2	73	19	13	79.3	84.8	82.4
Covariate	13	4	0	76.4	100.0	86.6	13	3	0	81.2	100.0	89.6	17	2	0	89.4	100.0	94.4
Effect size	41	5	9	89.1	82.0	85.4	50	17	5	74.6	90.9	81.9	65	2	10	97.0	86.6	91.5
All classes (micro)	414	71	37	85.3	91.7	88.4	234	42	12	84.7	95.1	89.6	247	32	39	88.5	86.3	87.4
All classes (macro)				84.9	92.7	88.4				85.7	93.8	89.5				91.3	88.9	90.0

Covariate returned the best recall (100.0%) and as it was mentioned above, since the number of annotated covariate concepts is limited, this resulting value should be taken with caution. Exposure had the lowest recall (80.3%). With the exception of study design that saw a little increase in recall (7.7%), recall decreases were observed in the rest of the characteristics with population, exposure, outcome and effect size recall having 0.3%, 19.7%, 13.6%, and 4.3% drops respectively (covariate's recall maintained stable – 100.0%) when compared to the values of the development set. It was observed that exposure had the biggest decline in recall (19.7%).

All F-scores were noted to be above 82.0%, with four of them above 91.0%. Study design returned the highest F-score overall with a value of 95.9%. An expected drop in the evaluation set in the F-scores of exposure and outcome mentions was noted with 14.2% and 7.8% respectively. On the contrary, study design, population, covariate and effect size revealed an increase in their F-score values, with effect size having the largest (7.9%).

The micro F-score, precision and recall for all the six epidemiological characteristics is 87.40%, 88.5% and 87.4% respectively. Although at the mention level, the values of precision, recall and F-score tend to decrease, at the document level it is expected to have a (small) increase in the system's performance. Since the focus is not on the detection of every possible (synonymous or not) mention of each characteristic but that one that summarizes mostly the necessary information, this could attribute the increase in the performance of study design, population, covariate and effect size. Additionally, the decrease in the evaluation metrics of exposure and outcome could be also affected by the document level since the identification of a false positive or the neglect to identify a true positive has greater influence on the performance overall due to the decrease of the total number of true positives per document.

4.4.2. Mention Level Evaluation

The values of precision in the evaluation set ranged from 81.2% to 100.0% suggesting reliable results, while recall had a range from 78.9%-100.0%. The best precision belonged to study design with 100.0%. However, due to the limited number of study design spans mentioned in the evaluation set, this value should be taken with caution. Drops were observed in the values of precision for the characteristics of population, exposure and outcome (2.4%, 6.6% and 6.6%) while effect size had the largest increase (22.4%) in comparison with the development set.

Table 33: Values of TP, FP, FN, P, R and F for the training, development and evaluation sets for the extraction process of study design, population, exposure, outcome, covariate and effect size concepts at the mention level.

.	Training set (60 abstracts)						Development set (30 abstracts)						Evaluation set (60 abstracts)					
	TP	**FP**	**FN**	**P(%)**	**R(%)**	**F(%)**	**TP**	**FP**	**FN**	**P(%)**	**R(%)**	**F(%)**	**TP**	**FP**	**FN**	**P(%)**	**R(%)**	**F(%)**
Study design	66	8	3	89.1	95.6	92.3	13	1	2	92.8	86.6	89.6	14	0	2	100.0	87.5	93.3
Population	126	11	8	91.9	94.0	92.9	46	5	4	90.1	92.0	91.0	50	7	5	87.7	90.9	89.2
Exposure	167	23	22	87.8	88.3	881	74	4	3	94.8	96.1	95.4	60	8	16	88.2	78.9	83.3
Outcome	244	32	14	88.4	94.5	91.3	94	13	4	87.8	95.9	91.7	104	24	15	81.2	87.3	84.2
Covariate	14	4	0	77.7	100.0	87.4	16	3	0	84.2	100.0	91.4	21	2	0	91.3	100.0	95.4
Effect size	41	5	9	89.1	82.0	85.4	50	17	5	74.6	90.9	81.9	65	2	10	97.0	86.6	91.5
All classes (micro)	658	83	56	88.7	92.1	90.4	293	43	18	87.2	94.2	90.5	314	43	48	87.9	86.7	87.3
All classes (macro)				87.3	92.4	89.7				87.3	93.5	90.2				90.9	88.5	89.6

Covariate had the best recall (100.0%) - nevertheless, since the evaluation set has only a small number of mentions (21 in total), this value is indicative of the system's performance rather than definitive. Exposure had the lowest one (78.9%). Drops in recall performance were expected with population, exposure, outcome, and effect size showing a decrease of 1.1%, 17.2%, 8.6%, and 4.3% respectively from the development to the evaluation corpus. Only study design revealed an increase in recall from 86.6% to 87.5% although this value should be considered only indicative, while the recall of covariate remained stable (100.0).

All F-scores were observed to be above 83.0% indicating relatively reliable results. Particularly, covariate was noted to have the best F-score (95.4%) and exposure was noted to have the lowest one (83.3%). With the exception of study design, covariate and effect size that saw increases in their F-score (3.7%, 4.0% and 9.9% respectively), a drop was observed in the rest of the characteristics (population – 1.8%, outcome 7.5%) with exposure having the largest one (12.1%) when the returned values of the evaluation set are compared to those of the development set. The micro F-score returned was 87.3% further indicating the reliable performance of the system for the identification of epidemiological characteristics in text.

4.5. Discussion

We discuss the performance of the rule based methodology and analyse the component of extraction for each of the key epidemiological characteristics at the document level. So far, the system demonstrated promising performance on the test dataset with an overall F-score of 87.4%. This suggests that a rule based approach can generate good and reliable results in epidemiological text mining despite the restrained nature of the targeted concepts. Hence, epidemiological professionals and researchers could be assisted in the identification and manipulation process of key epidemiological information from studies related to complex health care problems. The methodology relies on intensive lexical and terminological pre-processing along with rules based on common expressions observed in text for the recognition of key characteristics. The number of rules designed for the identification of key information in epidemiological abstracts with obesity as a case study can be considered relatively high (412), given that they were engineered from small training (and development) datasets (a total of 90 abstracts).

However, the chosen problem, obesity, affects multiple and different population types and includes a variety of determinants along with a number of outcomes of various nature (e.g., anatomical, biological, disease-related, etc). Therefore, it was considered that the saturation for rule development for such generic characteristics for the field of epidemiology should (and could) be achieved quickly due to the existence of certain generic (syntactical) expressions observed in text. On one hand, the number of rules for the characteristics such as study design (16), covariate (28) and effect size (15) were rather small in comparison to others e.g., population (119), indicating the existence of generic expression patterns that can highlight specific concept types. For example, the usage of dictionaries that contain epidemiological study designs can assist in the identification of the respective study design of an abstract. On the other hand, the task of recognising other epidemiological elements (e.g., exposure) through a rule based approach is not an easy task and observed sources of false positives and false negatives are described below for each characteristic.

4.5.1. Study Design

The system achieved the best performance (95.9% F-score) in the recognition of the various types of epidemiological study in abstracts. Due to the limited number of study design mentions in the evaluation set (13), the (overall positive) values of precision, recall and F-score should be taken with caution. Nonetheless, they indicate that a rule based approach combined with lexical resources containing the necessary study designs, is reliable enough for the identification of study types in epidemiological abstracts related to a particular problem. Figure 47 reveals a comparison of the performance of the rule based methodology for the training, development and evaluation set.

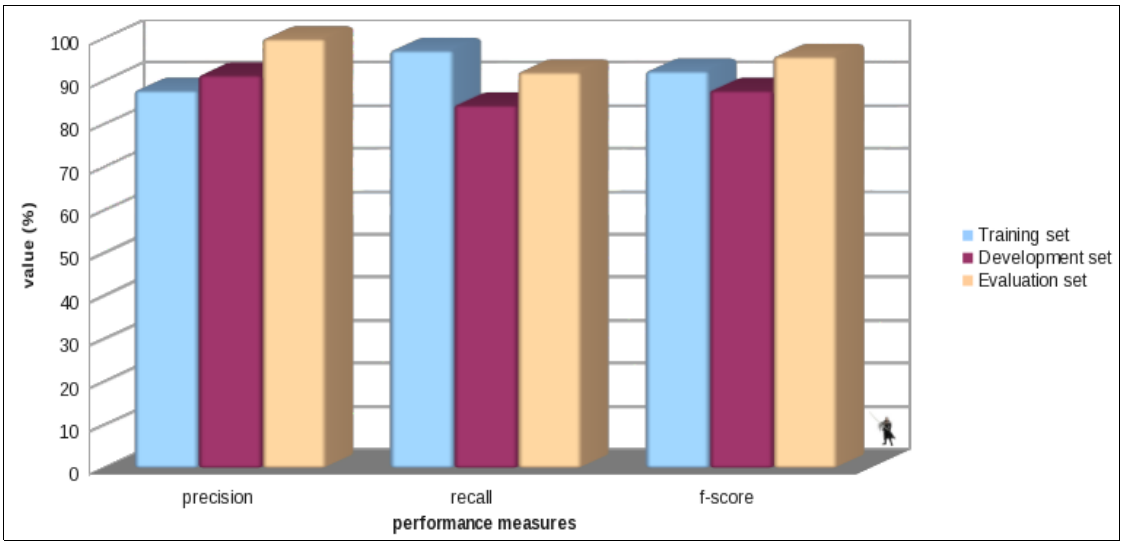


Figure 47: Comparison of the performance of the rule based methodology (precision, recall, F-score) for study design.

False Positives

There were no false positives in the evaluation data set. In all the sets (training, development, evaluation), all the abstracts include only one concept related with the element of study design. However, due to the application of the associated dictionaries, it is possible that in a larger abstract set, false positives could be generated if certain citations report more than one mention of different (or similar) study types, therefore resulting in a confusion as to which span(s) is the actual representative study design(s).

False Negatives

In the evaluation dataset, there was only one false negative. Any study design in epidemiological text without any specific information can pose a problem for their detection. For example:

- "Metabolic and bariatric surgery for obesity: a **review**[21] **[False Negative]**"

No rule based entirely on the recognition of single words that describe a study design (e.g., "*review*") was implemented. Consequently, any study mentions that are described in a similar manner are not and will not be detected by the system. Since certain study designs (e.g., "*review*", etc.) are ambiguous, any incorporation into the current lexical resources as study designs and the rule sets will result to an increase of false positive mentions. Therefore, the system's precision is maximised by avoiding their identification even though the value of recall may be sacrificed in the process. Nevertheless, the recall was increased from the development set from 84.6% to 92.3%.

4.5.2. Population

Most of the recognised population spans contain important demographic elements such as:

- attributes like age and gender;
- environmental and social information (e.g., the geographical region of residence, primary occupation, socioeconomic status, marital status);
- health-related concepts (e.g., the disease the individuals are currently affected from or the therapeutic procedure they have currently undergone).

The performance of the system is promising for the recognition of the studied population samples with an overall F-score of 93.3% (and an improved precision from 90.0% to 97.2% - second best precision in all six characteristics). However, the various ways in which the related population spans appear in text, do not enable the design and implementation of flexible and robust rules. Any attempts aiming to create rules based on generic expression patterns observed in text, resulted in the increase of false negatives on the development set affecting the overall performance of system for the population mentions (and contributing to a (rather insignificant) drop in recall: from 90.0% to 89.7%). Figure 48 reveals a comparison of the performance of the rule based methodology for the training, development and evaluation set.

21 Words highlighted with bold were false positives or false negatives.

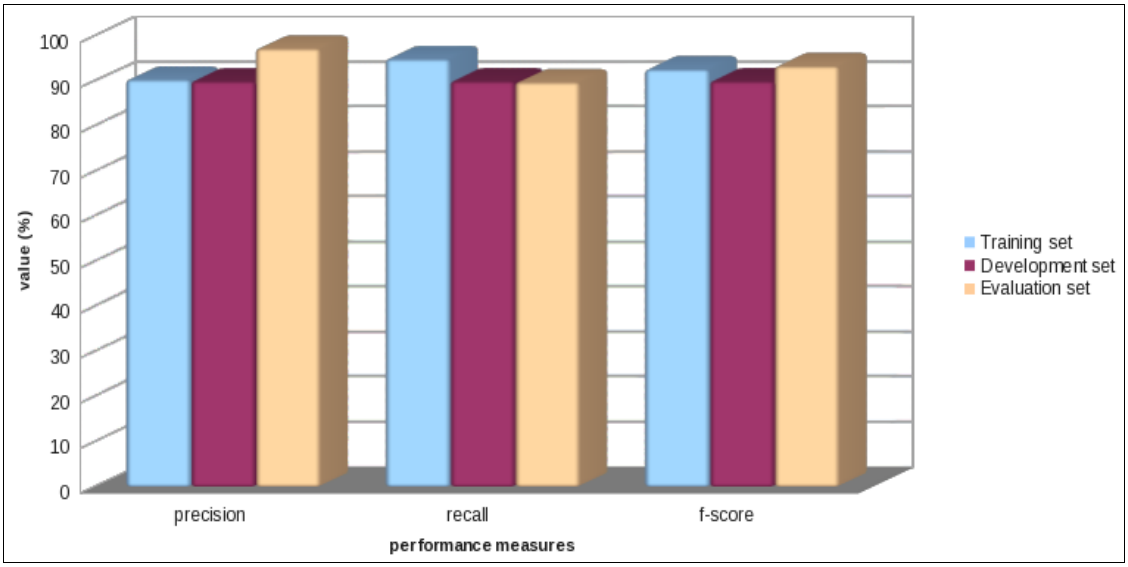


Figure 48: Comparison of the performance of the rule based methodology (precision, recall, F-score) for population.

False Positives

Despite a relatively large increase of 7.2% in precision in comparison to the development set and the second best precision overall in the system, a small number of false positives were noted. Rules that rely on the identification of population samples through generic expressions observed in text that include propositions (e.g., "*among*", "*of*" , "*in*", *etc.*) as well as those with synonym expressions that suggest the existence of participants (e.g., "*in analysis of*", "*among the*", "*in total*", "*involving*", "*studies with*", etc.) contribute to the generation of false positives. More specifically, there are three cases revolving around the falsely identified population spans. The falsely extracted population concept:

1. **is not descriptive enough for the whole population to be extracted;**

In the following example:

- "material and method: the **study population [False Positive]** included 3,715 deliveries ... "

"*study population*" was recognised by the rule based approach as a population span. However, although semantically the span is not wrong (it does refer to the study's population after all), it bares little important information regarding the population sample (or to the attributes of interest such as age, gender, nationality and ethnicity) that was observed in that particular study.

2. **can be completely unrelated with the targeted participant sample;**

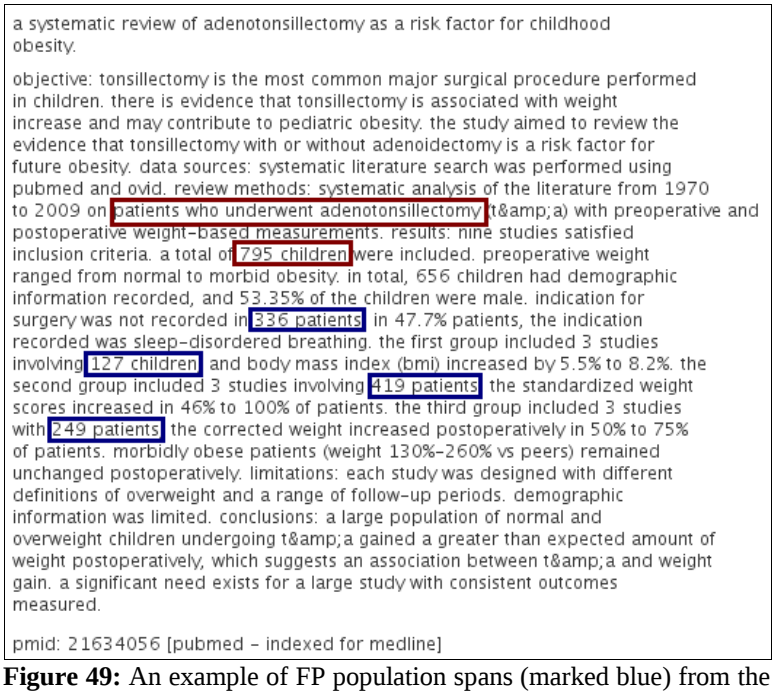
A case was noted in which a span was falsely identified as a population mention in an epidemiological study. In the following sentence:

- "The joint study of Mexican and Jamaican cohorts of early onset type **2 diabetes cases [False Positive]** will be useful ..."

"*2 diabetes cases*" was recognised as a population span due to the presence of the word "*cases*" in the lexical resources that are being used to identify the population sample of a study as well as the existence of a number before that (that could suggest the sample size of the population). The system failed to ignore this case as a mention of "*type 2 diabetes cases*" and extracted wrongfully a population span. It is logical to assume that if any other cases like this exist in the epidemiological literature, the system currently is not able to ignore them due to the ambiguity of certain words existing in the vocabularies. These words could potentially suggest the presence of a population such as "*cases*" that can be used as mentioned above in a different context.

3. **is a subset of the studied population;**

More precisely, certain identified mentions despite correctly considered as a part of the studied population semantically, are not describing the participant sample as a whole. The aim is to be able to detect the entire studied population and in its most descriptive form (that potentially includes the age, gender, nationality, ethnicity). An example can be seen in Figure 49.


a systematic review of adenotonsillectomy as a risk factor for childhood obesity.

objective: tonsillectomy is the most common major surgical procedure performed in children. there is evidence that tonsillectomy is associated with weight increase and may contribute to pediatric obesity. the study aimed to review the evidence that tonsillectomy with or without adenoidectomy is a risk factor for future obesity. data sources: systematic literature search was performed using pubmed and ovid. review methods: systematic analysis of the literature from 1970 to 2009 on patients who underwent adenotonsillectomy (t&a) with preoperative and postoperative weight-based measurements. results: nine studies satisfied inclusion criteria. a total of 795 children were included. preoperative weight ranged from normal to morbid obesity. in total, 656 children had demographic information recorded, and 53.35% of the children were male. indication for surgery was not recorded in 336 patients. in 47.7% patients, the indication recorded was sleep-disordered breathing. the first group included 3 studies involving 127 children and body mass index (bmi) increased by 5.5% to 8.2%. the second group included 3 studies involving 419 patients. the standardized weight scores increased in 46% to 100% of patients. the third group included 3 studies with 249 patients. the corrected weight increased postoperatively in 50% to 75% of patients. morbidly obese patients (weight 130%-260% vs peers) remained unchanged postoperatively. limitations: each study was designed with different definitions of overweight and a range of follow-up periods. demographic information was limited. conclusions: a large population of normal and overweight children undergoing t&a gained a greater than expected amount of weight postoperatively, which suggests an association between t&a and weight gain. a significant need exists for a large study with consistent outcomes measured.

pmid: 21634056 [pubmed - indexed for medline]

Figure 49: An example of FP population spans (marked blue) from the implementation of the rule based system. The marked red spans are the spans that were highlighted by the author and were returned as true positives by the system.

The system highlighted (besides the main descriptive spans), the following: "*656 children*", "*336 patients*", "*127 children*" and "*419 patients*", which are all subsets of the main population sample. Rules that contain propositional phrases or expressions suggesting the total size of the studied population sample, probably require to become stricter in order to enable the accurate capture of patient-related concepts.

False Negatives

The sources for false negatives include:

1. **population spans for which no specific rules were designed;**

Certain population spans followed more specific syntactical expressions that described the participant sample ("*Conclusions:* ***Women with SLE or RA diagnosis [False Negative] ...***") . While currently the relevant rule set was not able to cover such patterns, their incorporation along with other similar variations to the rule design might be able to further improve the performance of the system without alternating the value of precision.

2. **lack of concepts in the respective lexical resources;**

Associated concepts were missing from the dictionaries describing a specific demographic ("*veterans*"), case ("*deliveries*") or concepts that suggested a population sample ("*cohort*", "*diagnoses*") (see Section 4.3.2.). This can be solved by enriching the related dictionaries with a larger variety of respective terms although there is a possibility that precision could be affected. In the following examples, the population spans that the system failed to extract can be seen:

- "... included **3,715 deliveries [False Negative]** in ..."
- "**... 895 veterans who had bariatric surgery [False Negative]** ..."
- "... have built a large cohort of **families with early onset type 2 diabetes [False Negative]** ..."
- "The current and lifetime psychiatric diagnoses of **120 consecutive bariatric patients [False Negative]** ..."

Despite the above observations, the F-score for the population type remained encouraging (93.3%) showcasing that a rule based approach can capture the necessary (and in many cases the most detailed) participant information in epidemiological studies.

4.5.3. Exposures

Exposure concepts are often not explicitly stated (in contrast to outcomes), thus their role may not be clearly presented in text. Their textual (expression) patterns are more complex than those of other epidemiological characteristics mainly because they are interconnected with outcome concepts through words suggesting relationships and associations. Despite a good performance in the recognition of related concepts (F-score, 84.9%), the rule based exposure identification is a challenging task since it is often uncertain if a detected concept is acting as an exposure or an outcome. The recall returned was the lowest from all the six characteristics (80.3%) indicating that a rule based approach no matter how reliable results may returned, there will be still specific cases that will be falsely ignored. Figure 50 reveals a comparison of the performance of the rule based methodology for the training, development and evaluation set. Table 34 provides a summary of the sources noted for false positives and negatives in the evaluation set.

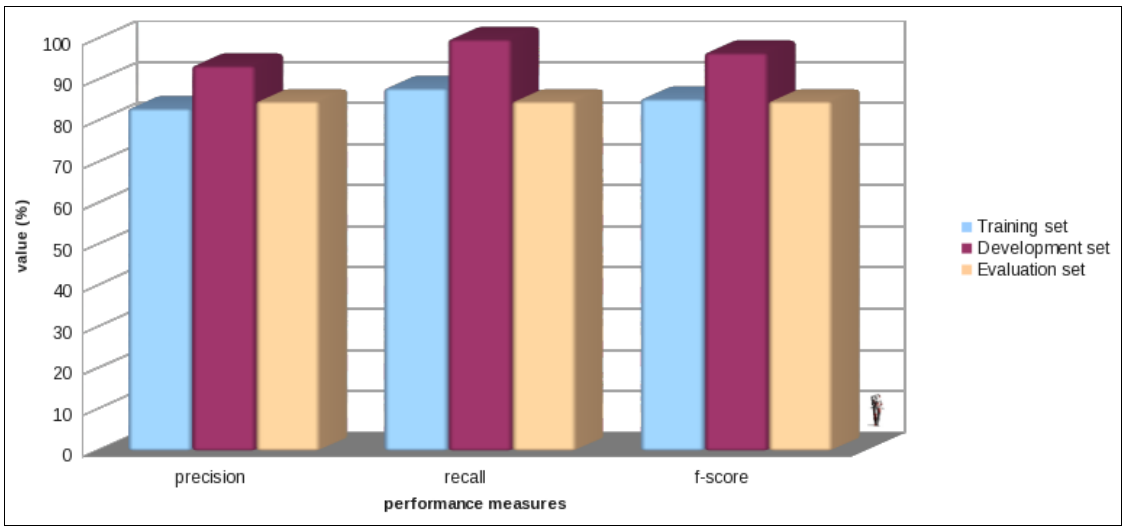


Figure 50: Comparison of the performance of the rule based methodology (precision, recall, F-score) for exposures.

Table 34: Causes for FP and FN exposure mentions with the respective number of concepts and percentage.

	causes	mention number	percentage (%)
false positives	reversed position in text	5	62.5
	generic rule design	3	37.5
total		8	100.0
false negatives	reversed position in text	4	36.0
	lack of implemented rules	4	36.0
	lack of dictionary concepts	3	28.0
total		11	100.0

False Positives

From the reviewing process of false positives, common types of errors were observed (Table 34) that explain the precision drop (8.7%) in the evaluation set. The recognition of false exposure concepts depended heavily on:

1. **the reverse position of the concepts in text;**

   More specifically, generic rules such as:

- "association between <X> and <Y>",
- "<X> associated with <Y>",

generated encouraging results i.e., a number of TPs. However, sometimes they would detect concepts irrelevant to the role of exposure. More specifically, it was concluded after careful consideration that when an expression based on these (and similar) rules associates two concepts, the first concept is typically the exposure, while the latter is the outcome. This looks natural to the researcher's point of view: when a clinical professional is studying the relationship between two concepts, he explores the link between (firstly) an exposure and (secondly) an outcome. To further support this hypothesis, in the training set of 60 epidemiological abstracts, a total of 73 exposure-outcome mentions through "*association*" rules were evaluated. Only 9 had the concepts reversed in text as outcome-exposure (Table 35).

Table 35: Number of exposure/outcome and outcome/exposure mentions in the training set in rules such as "*association between X and Y*" or "*X is associated with Y*".

	exposure - outcome	**outcome - exposure**
total number of mentions	73	9
number of abstracts	28	6

Therefore, due to the assumed hypothesis regarding the concepts' role in a relationship noted in text, the outcome mention would be detected instead of the exposure one despite being semantically positioned vice versa. Cases likes these result in the generation of both false positives and false negatives (in the characteristic of exposure and outcome) since the annotated exposure concept is not identified. Examples of false positives taken from the respective abstracts of the evaluation set can be seen below with the real role of each concept mentioned in brackets:

- "... the association of **survival [False Positive - outcome]** and **bariatric surgery [False Negative - exposure]** for older men is less clear. "

- "... to examine the association of **childhood overweight status [False Positive - outcome]** and **sleep difficulties [False Negative - exposure]** among pre-adolescents."

In the above cases, "*survival*" and "*childhood overweight status*" are false mentions of exposure while "*bariatric surgery*" and "*sleep difficulties*" are the true exposure concepts (false negatives). Despite this phenomenon occurring relatively frequently, it is a challenging task to comprehend the role of each concept and their semantic relationship through a rule based approach only. Therefore, it is more appropriate to take the role of the concepts into consideration within a wider context, e.g., the whole abstract text. The system currently only recognises the exposures at a sentence level.

2. **the trigger of generic rules;**

The design and application of generic rules led to the identification of false positives. Due to the nature of the rules and the large number of (generic) concepts included in the biomedical resources that were utilized, certain mentions were recognised despite not being annotated in text as exposures or having any epidemiological role in the study. An example of a false positive can be seen in the following sentence:

- "... and the association between **race [False Positive]** and gender. "

"*race*" was recognised as an exposure due to the trigger of the generic rule "*association between <X> and <Y>*". However, the particular mention is not an exposure as its role in the study context is different. This is one of the disadvantages of designing and implementing generic rules, due to the possibility of identifying more concepts as false positives, therefore contributing to a drop in precision and an increase in noisy results.

False Negatives

These were the causes for a significant large drop compared to the development set in the recall (19.7%), marking it the largest drop from all characteristics and the lowest recall (80.3%). Sources for false negatives (Table 34) included:

1. **reverse concept position in text;**

As was already discussed above, the reverse position of concepts in text led to the recognition of outcomes incorrectly marked as exposures while the real exposures were not extracted.

2. **lack of implemented rules to represent complex expressions in text;**

Some exposure mentions were not identified in text due to the lack of implemented rules. More specifically, these cases included the design of specific rules in order to

capture the necessary concepts. For reasons of system robustness and to maintain a reliable precision, they were not included in the rule set. False negative examples from the evaluation set abstract can be seen below:

- "**Sarcopenia [False Negative]** was more common among SLE."
- "Only **educational status [False Negative]** had a significant association with exclusive..."

"*Sarcopenia*" and "*educational status*" were not detected by the system due to lack of related implemented rules based on the observed revolving expressions ("*was more common*" and "*had a significant association*").

3. the lack of concepts from the dictionaries;

Finally, the lack of particular concepts from the currently used lexical resources resulted in the generation of false negatives and a decrease in recall (e.g. "*late bedtimes*", "*uncontrollable eating behaviour*", "*pa during leisure time*"-pa, an acronym for physical activity). These concepts are not of biomedical nature but are mostly associated with the cluster of activity and behaviour. The dictionaries used in the system are not focused on the inclusion of such terms and it is logical to assume that if they do, they do not contain a wide range of concepts. Since the targeted exposures in epidemiological studies can be of various nature (e.g., due to the clinical complexity of obesity), the incorporation of dictionaries that include terms related to human individual and social activities/behaviours, could potentially increase the accuracy of the system.

4.5.4. Outcomes

In comparison with the exposure mentions, outcome concepts are better defined through specific expressions observed in epidemiological abstracts. However, due to the implementation of generic rules, the precision returned was 79.3%, which was the lowest value from all the characteristics. Recall was 84.8%, substantially higher than the one of exposure, suggesting the system is more likely to detect the correct outcomes than exposures in epidemiological text. Overall, outcome had the lowest F-score (82.4%), suggesting that despite the easy detection of outcomes in an epidemiological abstract, there is room for false identification of concepts due to the generic rule design. Therefore, a stricter rule design should be applied to order to improve the accuracy of the system and filter out potentially insignificant concepts being detected as outcomes. Upon review of the false positives and negatives, causes of false positives and negatives were detected, identical with those analysed in the characteristic of exposure (Table 36). Figure 51 reveals a comparison of the performance of the rule based methodology for the training, development and evaluation set.

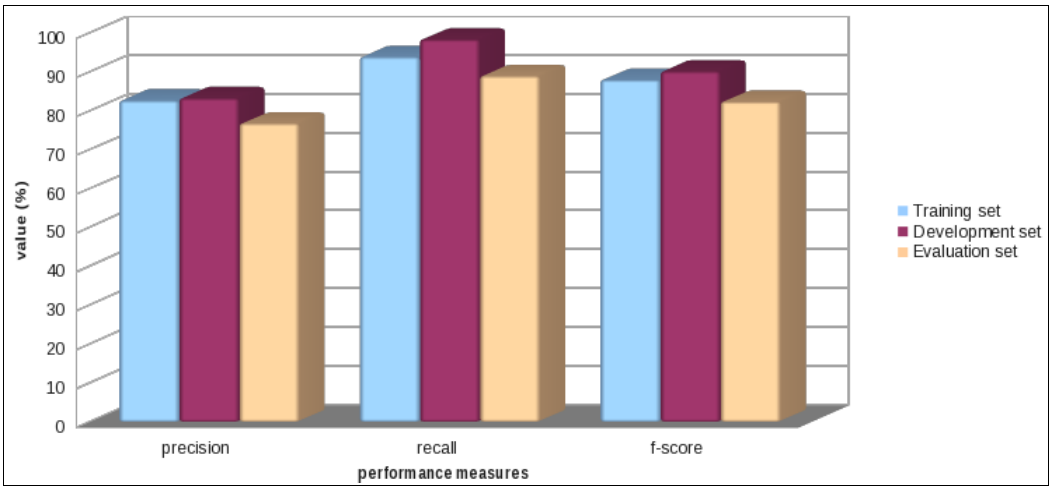


Figure 51: Comparison of performance of the rule based approach (precision, recall, F-score) for outcomes.

Table 36: Causes for FP and FN outcomes with the respective number of concepts and related percentage.

	causes	mention number	percentage (%)
false positives	reversed position in text	5	26.3
	generic rule design	14	73.7
total		19	100.0
false negatives	reversed position in text	4	30.7
	lack of implemented rules	6	46.2
	lack of dictionary concepts	3	23.1
total		13	100.0

False Positives

The rules that were designed and implemented followed generic (syntactical) expressions. Despite a relatively good F-score (82.8%), the outcome characteristic ended with a large number of false positives (19 in total) due to:

1. **generic rule triggering;**

   In the following text taken from epidemiological abstracts of the evaluation set, few examples of false positives can be seen:

- "... a higher prevalence of **sleep difficulties [False Positive]** ..."
- "... relationship of physical activity and **eating behaviour [False Positive]** with obesity and type 2 diabetes mellitus."
- "short sleep is associated with an increased risk for **being or [False Positive]** becoming overweight/obesity..."

Rules such as "*characteristics of <X>*", "*relationship of <Y> and <X>*" and "*<Y> is associated with <X>*" are triggered when the respective expression (pattern) is detected in text. These patterns can be used also in ways that are not indicating potential outcomes but rather describing or stating a concept. Hence, the concepts in the examples above such as "*sleep difficulties*", "*eating behaviour*" and "*being or*" are extracted falsely as outcomes.

2. **reverse concept position in text;**

This is the same problem encountered in the characteristic of exposure (see Section 4.4.3. Exposure, False Positives, False Negatives). The design and implementation of generic rules based on expressions like those reported in Section 4.4.3. resulted into a (small) decrease in the precision (4.0%).

False Negatives

A big drop in recall (13.6%) in comparison to the value from the development set, was noted. A variety of reasons provided can suggest why the system failed to recognise particular mentions of outcomes, the same ones observed in the characteristics of exposure. Sources for the generation of false negatives (Table 36) are listed below along with respective examples of falsely ignored outcomes from abstracts of the evaluation set (The analysis of these error sources is identical to that one of Section 4.4.3. Exposure, False Negatives):

1. **reverse concept position in text;**

- "the association of **survival [False Negative-outcome]** and bariatric surgery **[False Positive-exposure]** ..."
- "... **leptin levels [False Negative-outcome]** were positively correlated with fe(o) levels **[False Positive-exposure]**..."
- "... to examine the association of **childhood overweight status [False Negative-outcome]** and sleep difficulties adjusting **[False Positive-exposure]** for age, gender ..."

The concepts from the above examples "*survival*", "*leptin levels*" and "*childhood overweight status*" were incorrectly identified as exposures in the associations shown above and thus not extracted as outcomes.

2. **lack of implemented rules to represent complex expressions in text;**

Certain outcome concepts were part of complex expressions observed in text that were not covered by the system's rule design:

- "Overweight and obesity could increase risk in **cesarean section**, **pre-eclampsia, dm, pph and severe pph [False Negative]**..."

Due to a very generic nature of particular expressions noted in text, such as:

- "... have a role in **poor asthma control [False Negative]** ..."
- "... the presence of **severe depression [False Negative]**.. "

rules were not designed and implemented. These expressions clearly showcase outcome concepts within the context of a study, however, due to their generic nature, they result in an increasing number of false positives. While recall may be improved slightly, there is a risk for the precision to drop significantly. Thus, it was decided to find an acceptable balance between the two metrics.

3. **the lack of concepts from the respective lexical resources ;**

A minority of outcome concepts were not included in the related dictionaries and as a result their recognition was not possible by the system (e.g. "*excessive gestational weight gain*", "*bone mineral density (bmd) of femoral neck", "overfat")*. We have noticed that these include very specific biomedical concepts ("*excessive gestational weight gain*", "*overfat*") that are not covered by the current dictionaries as well as the existence of acronyms in the concept ("*bone mineral density (bdm) of femoral neck"*). With potential future improvements in the outcome rule design, it can be said that a generic rule based approach can generate encouraging and intriguing results.

4.5.5. Covariates

The implementation of rules based on associated with term (e.g., "*confounding factors*", "*covariates*", etc) and verb/noun dictionaries indicating a covariate (e.g., "*adjusting*", *controlling*", "*adjustment*", etc) along with respective prepositions (e.g., "*for*", "*of*", etc) revealed reliable results with the F-score (94.4%) and the best recall (100.0%). Covariates had the fewest number of mentions in all three sets from all the characteristics, hence any conclusion regarding the system's performance is at most indicative: due to the limited number of covariate mentions (17) in comparison to those of other epidemiological characteristics (e.g., outcome - 96, effect size - 75), the values of precision, recall and F-score should be taken with caution. Particularly, the steady recall that returns 100.0% value in both the development and the evaluation set. Since the number of the observed false positives is small, it is difficult to draw any terminal conclusions regarding the accuracy of the information extraction of covariates in epidemiological abstracts. However, these results could provide an initial indication that covariate mentions could be detected relatively easy without any noise

recognized. Figure 52 reveals a comparison of the performance of the rule based methodology for the training, development and evaluation set.

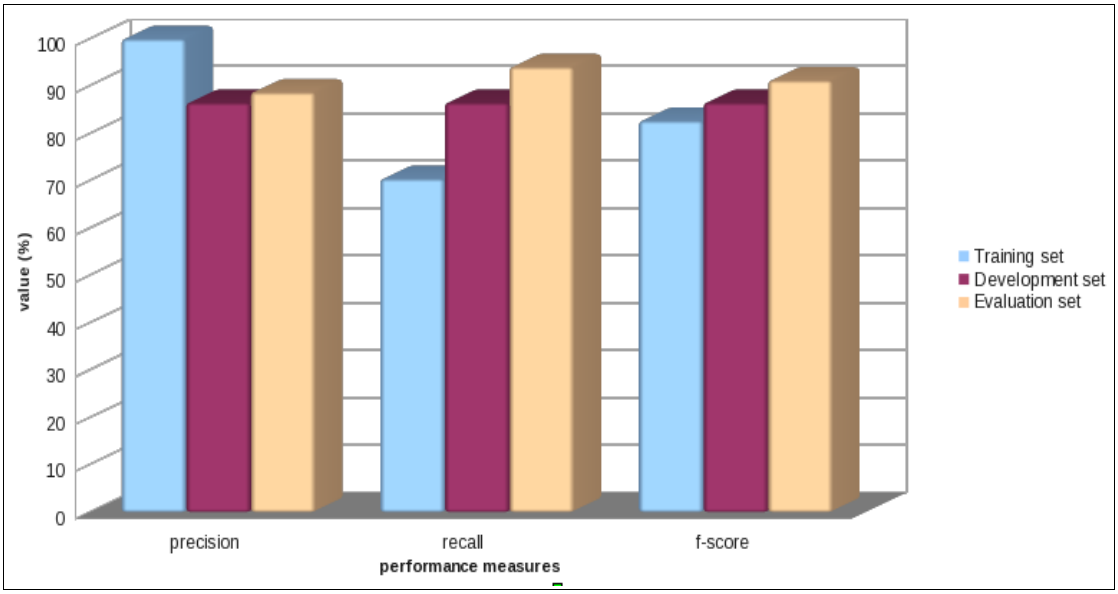


Figure 52: Comparison of performance of the rule based methodology (precision, recall, F-score) for covariates.

False Positives

Only two false positives were generated from the usage of rules based on dictionaries containing words associated with covariates. In the following examples of false positives:

- "... after adjustment for **potential confounders [False Positive]**."
- "... with adjustment for **potential confounders [False Positive]** (age, education ..."

"*potential confounders*" was recognised as a covariate mention. The utilization of generic rules in combination with large dictionaries may lead to the recognition of undesirable concepts. In the second example despite the recognition of related mentions from the system (age, education, etc.), "*potential confounders*" is not adding any information in the summarization of the covariate characteristic relevant to a study, therefore its extraction is not desirable.

False Negatives

No false negatives were returned. The stability of the covariate's recall from the development set to the evaluation set could be attributed to a limited number of covariate mentions in the abstract dataset. Despite a promising recall, perhaps in a larger dataset, the system may fail to recognise any covariates that are associated with unique and different expressions that the current rule design is based on.

4.5.6. Effect Size

The identification of effect size mentions returned encouraging results (97.0% overall precision). The rules designed to recognise the related (and mostly arithmetic in nature) mentions, were based on the combination of numerical and syntactical expressions observed in text. Effect size mentions incorporate various numerical characters, symbols (e.g., "&", "%"), parenthesis, quotes or brackets. Therefore, the rule design focused on the detection of these symbols and the numeric characters along with dictionaries that include specific words suggesting the presence of effect size mentions e.g., "*odds ratio*", "*relative risk*", "*confidence interval*", etc. A relatively high recall (86.6%) revealed that this approach returns encouraging results with only a small number of mentions being ignored by the system (10). Figure 53 reveals a comparison of the performance of the rule based methodology for the training, development and evaluation set.

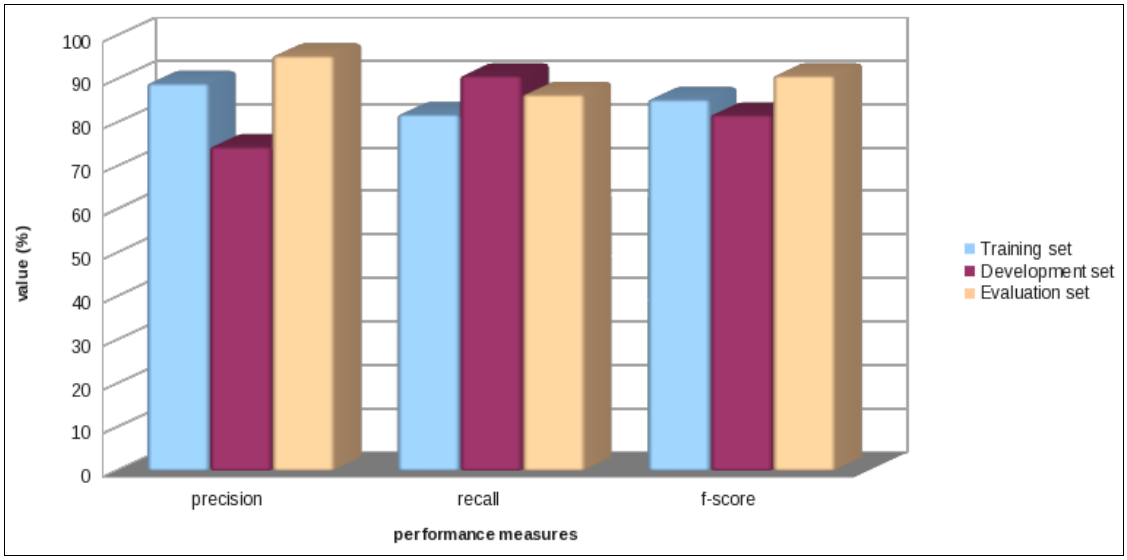


Figure 53: Comparison of performance of the rule based methodology (precision, recall, F-score) for effect size.

False Positives

Since the precision was high (97.0%), the number of detected false positives was limited. There were only two false positives. These "*mentions*" were incorrectly classified as effect size concepts due to the presence of numeric characters. False positives of effect size mentions can be seen below:

- "... despite having **normal or overweight bmi. Pmid: 21649469 [false positive]**"
- "tc and ldl-c were positively correlated with c**d4 [false positive]**+ cell count (r=0.13 ..."

Since some rules are based on combinations of dictionaries that include a variety of biomedical concepts (as well as related acronyms) with (present or non present) symbols e.g., ":" suggesting the mention of a numeric value (representing the effect size), it is possible that other similar expressions like the ones listed above will be extracted. This consequently contributes to the generation of false positives. Despite this, the system's performance can be characterised as reliable and precise for the identification of effect size mentions along with their corresponding attributes (e.g., confidence interval).

False Negatives

Upon reviewing a relatively small number of false negatives (a total of 10), it was observed that these mentions were following expression formats not covered by the current rule design. More specifically, in the following examples:

- "... increased risks of overweight/obesity at the age of 4 years **(odds ratio (95% confidence interval): 15.01 (9.63, 23.38))** **[False Negative]** ..."
- "... bmi statistically significantly increased by **2.8% (95% confidence interval: 1.5% to 4.1%**; **[False Negative]** p<0.001) ..."
- "... the relative risk (rr) for postmenopausal breast cancer is around **1.5 [False Negative]** for overweight women and **> 2 [False Negative]** for obese women ..."

The highlighted mentions were not detected due to their format. The rules did not cover the existence of certain symbols (such as parenthesis in the first example) before the confidence interval span as it had not been observed before (in the training and development sets); in the second example, a value as percentage was not included for the identification of effect size measures with existing confidence interval; in the third case, a semantic pattern that follows the format of "*the relative risk (rr)* **[effect size type]** *for ..... is ... 1.5* **[effect size value]**" was not encountered previously. Despite not being difficult to detect in text (particularly if the effect size mentions are fully described with type, value, confidence interval and related concept), any expression (like the ones mentioned above) regarding the description of effect size mentions that was not observed in the training and development sets, was excluded from the rule design, therefore, contributing to a recall decrease.

Despite observing a relatively large number of false negatives, the recall of the system was 86.6%, which suggests reliable performance. If both precision and recall values are being taken into consideration, the system is accurate enough to recognize efficiently effect size mentions in epidemiological text. Identified cases which the system failed to detect, could be easily done in the future by incorporating these observed expression patterns to the rule design, thus potentially contribute to the performance improvement.

4.6. Summary

A text mining approach was designed in order to identify key epidemiological characteristics at the mention level from study design abstracts. Custom-made vocabularies that include terms for the identification of semantic classes that will assist in the detection of key characteristics in study abstracts were created and biomedical concepts were identified in text through the application of the Specialist lexicon and the ATR C-value method.

We extracted the key characteristics through a rule based approach. A set of text based rules (through MinorThird) was applied to the training set with a combination of the dictionaries. The rules were designed and based on semantic patterns that are combinations of frozen lexical expressions (that are used as anchors for specific categories) and specific semantic classes (that have been identified through the vocabularies) that indicate the presence of a key characteristic in text. The translation of these lexical expressions into the rule design includes the usage of regular expressions and the creation of vocabularies. By inspecting the training set, common semantic patterns for each of the key characteristics were observed and used for the design and implementation of the rule based method. The rules were expanded to include more similar semantic patterns for each characteristic after analysing the development set.

Precision, recall and F-score were calculated at the mention and document level in order to evaluate the performance of our methodology. At the document level the precision ranged from 79.3%-100.0%, while recall ranged from 80.3%-100.0%. The best precision was observed for study design (100.0%). The lowest precision was noted for outcome (79.3%). Covariate returned the best recall (100.0%), while exposure had the lowest one (80.3%). All F-scores were noted to be above 82.0% with study design returning the highest value (95.9%). At the mention level, precision ranged from 81.2% to 100.0% suggesting reliable results, while recall had a range from 78.9%-100.0%. The best precision belonged to study design with 100.0%. Covariate had the best recall (100.0%) - and exposure had the lowest one (78.9%). All F-scores were observed to be above 83.0% with covariate having the best F-score (95.4%) and exposure the lowest one (83.3%) suggesting relatively reliable results.

The system demonstrated promising performance with an F-score of 87.4% indicating that a rule based approach can generate good results in epidemiological text mining despite the restrained nature of the targeted concepts. The number of rules designed for the identification of key information in epidemiological abstracts (related to obesity) can be considered relatively high, given the small training (and development) datasets. However, obesity includes various determinants along with a number of outcomes, hence it was expected that the saturation for rule development for such generic epidemiological characteristics should (and could) be achieved quickly due to the existence of certain generic (syntactical) expressions.

The performance of our method in the study design characteristic suggests that a rule based approach combined with lexical resources, is reliable enough for the identification of study types. However, it is possible that in a larger abstract set, false positives could be generated if certain citations report more than one mention of different (or similar) study types resulting in a confusion as to which span(s) is the actual representative study design(s). Since certain study designs (e.g., "*review*") are ambiguous, any incorporation into the current dictionaries as study types and the rule sets will result to an increase of false positive mentions. The various ways in which the population spans appear in text, do not enable the design and implementation of flexible and robust rules. Any attempts aiming to create rules based on generic expression patterns (that include propositions e.g., "*among*", "*of*"), contributed to an increase of false negatives on the development set affecting the system's performance.

Despite a good performance in the recognition of exposures (F-score, 84.9%), it is often uncertain if a detected concept is acting as an exposure or an outcome. Common sources of errors for both exposure and outcome concepts were observed (e.g., reverse positions of the concepts in text, trigger of generic rules, lack of dictionary concepts). It is more appropriate to take the role of the concepts into consideration within a wider context e.g., the whole abstract text. On the other hand, outcome had the lowest F-score (82.4%), indicating that despite its easy detection in epidemiological abstracts, there is room for false identification of concepts due to the generic rule design.

Covariates had the fewest number of mentions, hence any conclusion regarding the system's performance is at most indicative. The number of the observed errors is small, thus any terminal conclusions regarding the accuracy of the covariate recognition in epidemiological abstracts are not entirely reliable. The system's performance can be characterized as reliable for the identification of effect size mentions. Since some rules are based on combinations of dictionaries that include a variety of biomedical concepts with symbols e.g., ":" suggesting the mention of a numeric value (representing the effect size), it is possible that a number of errors could occur with the system identifying wrong spans and ignoring those who contain acronyms. Despite not being difficult to detect in text, any expression regarding the description of effect size mentions that was not observed in the training and development sets, was excluded from the rule design, therefore, contributing to a decrease of recall.

Chapter 5

Epidemiological Characteristics Normalization

"It is a very sad thing that nowadays there is so little useless information"

Oscar Wilde, 1894

After the identification of the related characteristic concepts, a normalization procedure is required since it will enable the recognition of descriptive attributes (such as the biomedical classes a concept can belong to) that can assist in the understanding of epidemiological information related to a health problem. Additionally, it is easier to cluster the data under their (if any) particular attributes e.g., age for population mentions or a specific attribute for study design and perform statistical analysis that could reveal important patterns about a health care problem.

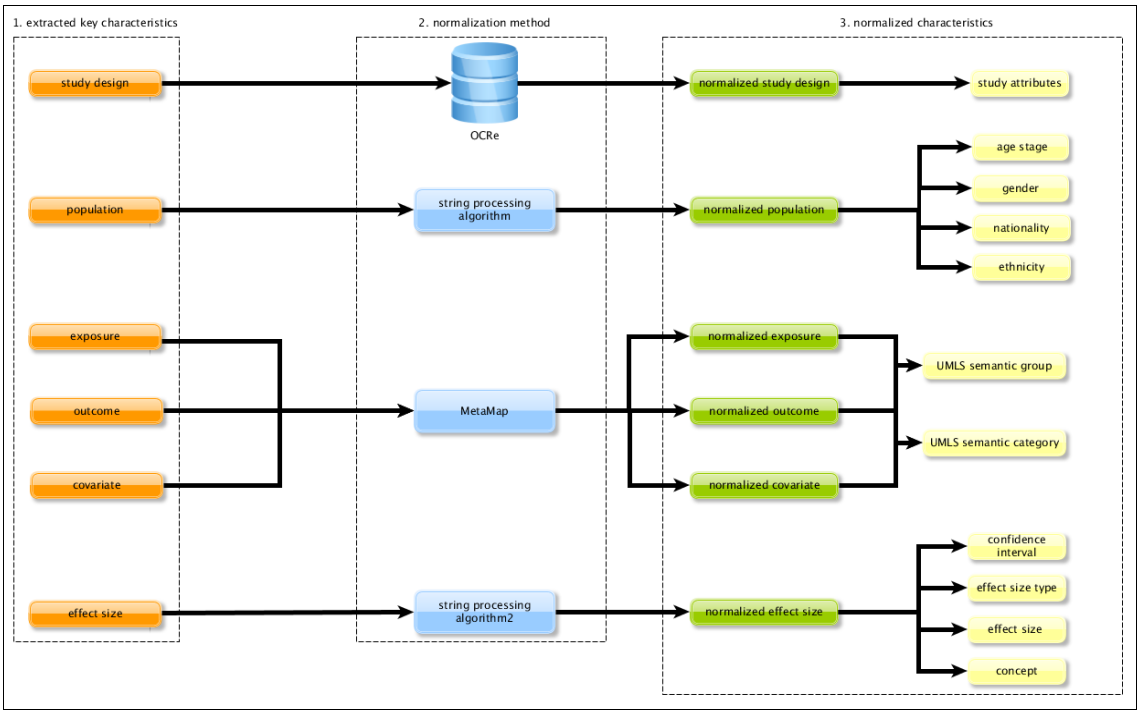


Figure 54: Overview of the normalization approach. In the second step, the extracted mentions of the six epidemiological characteristics are normalized through the use of a study design ontology named OCRe (study design), the application of string processing algorithms (population, effect size) and the implementation of the MetaMap tool. In step 3, study design mentions have been normalized in their study attributes; population spans have been normalized into the attributes of age, gender, nationality, and ethnicity; exposure, outcome and covariate mentions have been classified under a UMLS semantic group and category; effect size spans have been normalized into the attributes of effect size type, effect size value, effect size related concept and the respective confidence interval.

To be more specific, the extracted mentions of epidemiological characteristics in MEDLINE abstracts are normalised and either classified into specific semantic classes (any spans related to exposures, outcomes and covariates) or their descriptive attributes are detected through the

use of external resources and tools such as a study design ontology, MetaMap and text processing algorithms. Figure 54 reveals an overview of the normalizing method applied to the generated results of the information extraction approach. Before the normalization of each of the identified characteristic mentions takes place, an approach was followed that allowed the elimination of similar or identical spans (with the exception of effect size mentions). More specifically, it was hypothesized that the lengthiest span is the most informative in each characteristic. Therefore the normalization process is applied to the mentions that have been considered unique by using a string comparison module (see Section 5.1.) between the lengthiest span and the rest of other mentions for each characteristic. Spans that were similar to the longest one are ignored; those that were no similar were used for normalization. The identified effect size mentions are excluded from this procedure as epidemiological abstracts report the same effect size only once.

5.1. Normalization of Study Design

An adapted version of a clinical research ontology (Ontology of Clinical Research, OCRe, Tu et al. 2009) is used for the normalization of epidemiological study designs. OCRe is an ontology designed to describe human studies and their methods. OCRe's study design branch was expanded to include experimental research, observational types such as "*correlational*" and "*descriptive*" and secondary research ones e.g., "*meta-analyses*". Each identified study design is to be mapped to one of the nodes of this ontology (23 nodes in total, Figure 55).

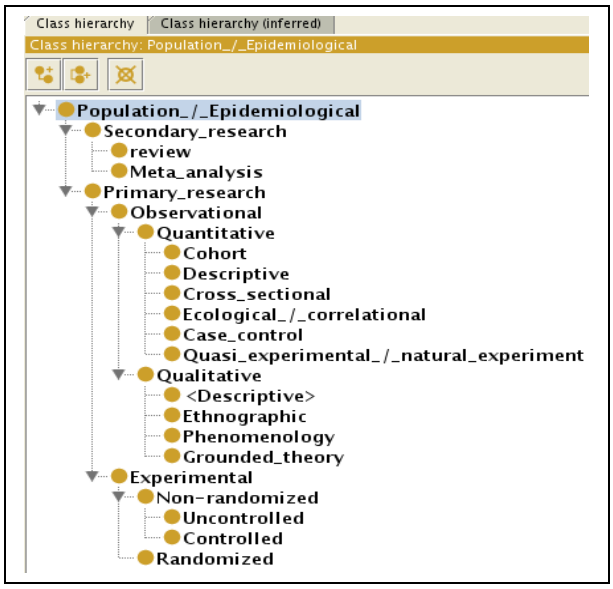


Figure 55: Expanded ontology of epidemiological and clinical study design. The ontology is used for the normalisation of extracted study designs from epidemiological text (visualization through Protege[22]).

22 http://protege.stanford.edu/

In order to match the identified study design mentions to the ontology, we used a string comparison module (Tresoldi, 2009). The method is based on previous work by Yang et al. (2001). More specifically, the module compares two strings and estimates the similarity between them through edit distance. Edit distance takes into account inserting, deleting or substituting single characters although it excludes the repetition of substrings. The more edit operations performed to alter the input string in order to match the one that is being compared to, the lower the similarity/match score is (out of 100%).

Since the aim is the representation of key epidemiological characteristics in each abstract, it was assumed that the longer the mention is, the more information it will include about the study design that was conducted in the abstract, and therefore more comprehensive/detailed normalization could be performed. Through the application of the string comparison module (mentioned above), the extracted study design mention is compared to those that form a part of the expanded ontology. The match that returns the highest score is then chosen as the normalized version of the input study design. No threshold was used in the similarity score. Following that, the normalized span (if necessary) is classified into higher level nodes of the ontology with any additional information stored as attributes. More detailed information (where available) that was captured is considered as a study attribute. For example, each of the mentions "*prospective cohort study*" and "*retrospective cohort study*" are mapped to the node of "*cohort study*", but with different additional attributes that provide more specific categorisation of the study type (in this case it is "*prospective*" or "*retrospective*"). Table 37 shows the attributes that each study type possess including their potential values. Table 38 shows some examples of identified study design mentions and their normalized versions.

Table 37: Specific attributes assigned to various study types with potential values (continuing on the next page).

epidemiological study design	study attribute	values
experimental research (trials)	clinical	yes, no
	blinded	non blinded, double-blind, triple blind
	treatment response	placebo-controlled, equivalence, non-inferiority
	arm	crossover, two arm, multi-arm
	type of randomization	cluster, individually
	sponsored	publicly,industrially
	phase	2,3
meta-analysis study	synthesis of evidence	quantitative, qualitative
review	type of review	systematic, literature
case control study	case control type	nested case, propensity matched

Table 37: Specific attributes assigned to various study types with potential values.

quasi-experimental natural experiment	quasi experimental type	before after, interrupted time series, natural experiment
	natural experiment	yes, no
cohort study	time attribute	retrospective, prospective
	cohort type	birth, case, patient
	occupational	yes, no
cross-sectional	serial	yes, no

Table 38: Examples of extracted study designs with their matched OCRe study design, respective score produced from the String Comparison Module, the assignment of a key node from the expanded ontology and the attribute values. The "-" stands for unknown value in the respective attribute.

extracted study design mention	matched study design (OCRe+)	score (%)	main OCRe node	attribute value(s)
case-control study	case control study	76.0	case control study	-
longitudinal epidemiological study	epidemiological study	76.0	epidemiological study	-
nested case-control analysis	nested case control study	52.0	case control study	case control nested
cross-sectional study	cross-sectional study	100.0	cross-sectional study	-
literature review	literature review	100.0	review	literature
prospective cohort study	prospective cohort study	100.0	cohort study	prospective
randomized double-blind placebo-controlled trial	randomised double-blinded placebo-controlled trial	62.0	randomised trial	double blind
				placebo-controlled
cross-sectional population-based sample	cross-sectional study	50.0	cross-sectional study	-
randomized clinical trial	randomised clinical trial	74.0	randomised trail	clinical
retrospective cohort study	retrospective cohort study	100.0	cohort study	retrospective

5.2. Normalization of Population

The identified population spans were normalized according to specific attributes that seem to be prevalent in their description:

1. **age:** The human age was divided into four stages:
   a) **juvenile** (0 - 19 years old)
   b) **early adulthood** (20 - 39 years old)
   c) **middle adulthood** (40 - 59 years old)
   d) **late adulthood** (60 + years old).
   If participants span more than one class of adulthood e.g., "*adults between 25 - 54 years old*" (early and middle adulthood), then the normalized age is considered to be "*adulthood*". In cases where the population age includes juvenile and any stage of adulthood, it was decided that no specific class should be assigned in order to focus mostly on these four stages.
2. **gender:** Populations can be classified as female, male or mixed when mentions of gender are explicitly mentioned.

3. **nationality:** A list of nationalities was obtained from the following web pages (http://www.peoplegroups.org/, http://en.wikipedia.org/wiki/Lists_of_people_by_nationality). The compiled dictionary included 229 nationalities (available in the Appendix).
4. **ethnicity**: A list of the most common ethnicities was obtained from the UK Office of National Statistics (http://www.ons.gov.uk/ons/index.html) and the United States Census (http://www.census.gov/) with a total of 26 ethnicities (available in the Appendix).

Before the normalization of the population spans, as it was mentioned above, the lengthiest span was chosen in each abstract. Nationality, ethnicity and gender are being detected from the use of respective dictionaries while the age is recognised from the applied regular expressions due to its relatively structured format when mentioned. If any of the attributes were not detected in the chosen longest span, then a script was applied to the other identified mentions. If they contain any information regarding these four attributes, then the information is extracted. Mentions that included ethnicity that can be partly considered as a nationality (e.g., "*Chinese American men*", "*Chinese British children*"), were labelled as problematic. Ethnicity was not recognised since the related terms were lacking from the respective dictionary while the substrings were identified as the nationality of the population span. Examples of normalized population mentions can be seen in Table 39 below.

Table 39: Examples of normalized populations into their respective attributes. The table includes age, gender, nationality and ethnicity. The "-" stands for unknown value in the respective attribute.

extracted population	age	gender	nationality	ethnicity
567 rural native Hawaiian men	-	male	-	native Hawaiian
909 women aged 45-54 years	middle adulthood	female	-	-
African American patients	-	-	American	African American
Canadian children aged 7-13 years	juvenile	-	Canadian	
5002 subjects aged 70 years	late adulthood	-	-	-
Chinese women aged between 41 and 63 years	adulthood	female	Chinese	Chinese
Indian men 30 to 50 years	adulthood	male	Indian	Indian
Hispanic children aged 6 to 13 years (n = 1017)	juvenile	-	-	Hispanic
25-34-year-old males	early adulthood	male	-	-
Thai adults aged 35 years	early adulthood	-	Thai	-

5.3. Normalization of Exposures, Outcomes and Covariates

In order to perform normalization to the necessary concepts that highlight information regarding exposures, outcomes and covariates at the document level for each abstract, a similar approach to the one applied to the characteristics of study design and population was followed. Spans that appeared more than once in the same abstract were filtered out through the application of the string comparison module. Any span with a similarity score above 40.0% is removed. Any span with a similarity score under a certain threshold is considered a different concept and is incorporated into the document level representation of the abstract's necessary information. Table 40 reveals some examples where each selected span is compared to other mentions of each abstract in order to eliminate those who posses a similarity above the threshold. In the experiment reported here, we used 40.0% for the threshold.

Table 40: Each selected span (before normalization) is compared to other potentially meaningful spans of each abstract in order to remove those with a similarity score above 40.0%.

abstract	selected span	other potentially meaningful spans	similarity score	removed?
1	leptin levels	leptin	63.1	yes
	leptin levels	ethnicity	22.2	no
2	obesity	obesity	100.0	yes
		self-esteem loss	21.2	no
3	socioeconomic status	low-self esteem	21.3	no
4	social class	obesity	23.5	no
	education	obesity	25	no
	smoking	obesity	24.7	no
5	HDL cholesterol concentration	HDL concentration	68.1	yes

The normalization of exposures, outcomes and covariates is essentially normalization of biomedical concepts and we applied the state-of-the-art software for this procedure, MetaMap (see Chapter 2, Section 2.1.4) (Aronson et al. 2010). Each concept is classified in one of 135 UMLS semantic categories and then clustered into one of the 15 higher-level UMLS semantic groups (McCray et al. 2001) (Table 41). The UMLS classification of concepts into semantic groups and categories was chosen because it is often used in Medical Informatics with a standardized biomedical vocabulary that includes a variety of synonymous terms.

Table 41: UMLS semantic groups and their respective categories.

UMLS semantic group	UMLS semantic category
activities/behaviours	activity, behaviour, daily or recreational activity, event, governmental or regulatory activity, individual behaviour, machine activity, occupational activity, social behaviour
anatomy	anatomical structure, body location, body part organ or organ component, body space or junction, body substance, body system, cell, cell component, embryonic structure, fully formed anatomical structure, tissue
chemicals/drugs	amino acid, peptide, or protein, antiobiotic, biologically active substance, biological or dental material, carbohydrate, chemical, chemical viewed functionally, chemical viewed structurally, clinical drug, eicosanoid, element ion or isotope, enzyme, hormone, hazardous or poisonous substance, immunologic factor, Indicator, reagent or diagnostic aid, Inorganic chemical, lipid, neuroactive substance or biogenic amine, nucleic acid nucleotide or nucleotide, organosphosphorus compound, pharmacologic substance, receptor, organic chemical, organic, chemical, pharmacologic substance, steroid, vitamin
concept/ideas	classification, conceptual entity, functional concept, group attribute, idea or concept, intellectual product, language, qualitative concept, quantitative concept, regulation or law, spatial concept, temporal concept
devices	drug delivery device, medical device, research device
disorders	acquired abnormality, anatomical abnormality, anatomical abnormality disease or syndrome, cell or molecular dysfunction, congenital abnormality, disease or syndrome, experimental model of disease, finding, injury or poisoning, mental process, mental behavioural dysfunction, neoplastic process, pathologic function, sign or symptom,
genes/molecular	amino acid, sequence, carbohydrate sequence, gene or genome, molecular sequence, nucleotide sequence
geographic areas	geographic area
living beings	age group, amphibian, animal, archaeon, bacterium, eukaryote, family group, fish, fungus, group, human, mammal, patient or disabled group, plant, organism, population group, professional or occupational group, reptile, vertebrate, virus
objects	entity, food, manufactured object, physical object, substance
occupations	biomedical occupation or discipline, occupation or discipline
organizations	health care related organization, organization, professional society, self-help or relief organization
phenomena	biologic function, environmental effect on humans, human caused phenomenon or process, laboratory or test result, natural phenomenon or process, phenomenon or process
physiology	cell function, clinical attribute, genetic function, mental process, molecular function, organ or tissue function, organism attribute, organism function, physiologic function
procedures	diagnostic procedure, educational activity, health care activity, laboratory procedure, molecular biology research technique, research activity, therapeutic or preventive procedure

The greatest challenge in the normalization of exposures, outcomes and covariates is the resolution of ambiguity issues when two or more UMLS Metathesaurus concepts share a common synonym, therefore the World Sense Disambiguation (WSD) option in MetaMap was used. MetaMap attempts to disambiguate among the concepts that have scored equally well in matching input text. It selects the concept that has the most likely semantic type for a given context in which the ambiguity rises.

An example can be seen in figures 56 and 57 below. The term "*cancer*" was used as an input term. Figure 56 shows that there are three mappings – the term "*cancer"* is considered to belong to the "*Eukaryote*", "*Neoplastic* P*rocess*" or "*Finding*" UMLS semantic categories with equal scores. With the activation of the WSD option (Figure 57), it was observed that from the three mappings, the second one was selected by MetaMap ("*Neoplastic process*") in this context.

```
Phrase: "cancer"
Meta Candidates (Total=3; Excluded=0; Pruned=0; Remaining=3)
  1000   Cancer (Malignant Neoplasms) [Neoplastic Process]
  1000   Cancer (Cancer Genus) [Eukaryote]
  1000   Cancer (Primary malignant neoplasm) [Finding]
Meta Mapping (1000):
  1000   Cancer (Cancer Genus) [Eukaryote]
Meta Mapping (1000):
  1000   Cancer (Malignant Neoplasms) [Neoplastic Process]
Meta Mapping (1000):
  1000   Cancer (Primary malignant neoplasm) [Finding]
```

Figure 56: Returned UMLS mapping of the "*cancer*" term. "*Cancer*" can be mapped to three different concepts with three different UMLS semantic categories.

```
Phrase: "cancer"
Meta Candidates (Total=3; Excluded=0; Pruned=0; Remaining=3)
  1000   Cancer (Malignant Neoplasms) [Neoplastic Process]
  1000   Cancer (Cancer Genus) [Eukaryote]
  1000   Cancer (Primary malignant neoplasm) [Finding]
Meta Mapping (1000):
  1000   Cancer (Malignant Neoplasms) [Neoplastic Process]
```

Figure 57: Returned UMLS mapping of the "*cancer*" term with WSD on. This results to the return of one mapping result.

Examples of exposure, outcome and covariate concept normalization along with their respective UMLS classification into semantic groups and categories can be seen in Table 42.

Table 42: Examples of normalized exposures, outcomes and covariates. Their respective normalized version along with their UMLS semantic group and category is being shown.

extracted mention	UMLS normalized version	UMLS semantic category	UMLS semantic type
low-density lipoprotein particle size	low-density lipoprotein particle	cell component	anatomy
weight loss	weight loss	finding	disorders
mortality	mortality	quantitative concept	concepts/ideas
child-feeding practices	feeding practices	therapeutic or preventive procedure	procedures
cardiovascular disease	cardiovascular disease	disease or syndrome	disorders
breast cancer recurrence	breast cancer recurrence	neoplastic process	disorders
life insurance	life insurance	idea or concept	concepts/ideas
interleukin-6	interleukin 6	amino acid, peptide, or protein,Immunologic factor, pharmacologic substance	chemicals/drugs
age	age	organism attribute	physiology
tobacco consumption	tobacco consumption	individual behavior	activities/behaviours

5.4. Normalization of Effect Size

Since it is not likely for the same abstract to report the same effect size more than once, the process of eliminating any spans that are similar or identical was not followed. For effect size mentions, the normalization was performed through an automatic text processing algorithm based on regular expressions that focused on their individual attributes (e.g., confidence interval, effect size type, etc.). Since effect size spans follow a relatively structured format comprised from different types of data, hence making their processing relatively straightforward in order to detect the respective attributes. Usually, the extracted effect size concepts contain the following attributes (not necessarily in that order):

1. the **effect size measure type:** This includes *adjusted odds ratio*, *odds ratio*, *hazard ratio*, *relative risk*, *prevalence*, *incidence*, *adjusted relative risk* and *adjusted hazard ratio*;
2. the respective (numeric) **value** of the effect size measure usually in the form of a percentage;
3. **confidence interval:** an observational (numeric) interval that indicates the reliability of an estimate;
4. the **concept** that the effect size is linked to (either as an exposure or as an outcome).

Regular expressions were implemented in order to match the multiple patterns in which effect size values and confidence intervals are being mentioned. However, certain effect size mentions may not follow a typical written format with sometimes extensive use of symbols e.g., ";", "(", ")", around the related concept, and the reporting of the confidence internal. Table 43 shows examples of identified effect size mentions in epidemiological abstracts and their respective normalized versions. Due to the nature of this characteristic, effect size spans might have more than one different mentions in each epidemiological study.

Table 43: Examples of normalized effect size mentions. The respective effect size value, type, confidence interval and related concept is shown (if available) for each mention.. The "-" stands for unknown value in the respective attribute.

extracted mention	type	value	confidence interval	concept
...underweight subjects had a higher risk of **death (HR 1.64; 95% CI 1.11-2.40)**, and conversely, **overweight (HR 0.69; 95% CI 0.49-0.96)** or **obese (HR 0.61; 95% CI 0.43-0.88)** subjects showed a lower risk of post-ICH death.	hazard ratio	1.64	95% [1.11-2.40]	death
	hazard ratio	0.69	95% [0.49-0.96]	overweight
	hazard ratio	0.61	95% [0.43-0.88]	obese
There were also trends toward increased risk for wound dehiscence with **hypothyroidism (odds ratio, 4.3; p = 0.06) and Ehlers-Danlos syndrome (odds ratio, 18.7; p = 0.05)**.	odds ratio	4.3	-	hypothyroidism
	odds ratio	18.7	-	Ehlers-Danlos syndrome
The **prevalence of diabetes was 9.02%**.	prevalence	9.02%	-	diabetes
A 5-unit increase in body mass index was associated with an 35% increased risk of **knee osteoarthritis (RR: 1.35; 95%CI: 1.21, 1.51)**.	relative risk	1.35	95% [1.21-1.51]	knee osteoarthritis
After adjustment for gender, ethnicity, overweight, and age, the relative risk of hypertension was significant for **gender (relative risk: 1.50; confidence interval: 1.15, 1.95) and overweight (relative risk: 3.26; confidence interval: 2.50, 4.24).**	relative risk	1.5	[1.15-1.95]	gender
	relative risk	3.26	[2.5-4.24]	overweight
In ordinal logistic regression, obesity was associated with current **depression (odds ratio [OR]= 1.86, 95% confidence interval [CI]: 1.25 to 2.78) and anxiety (OR = 1.58, 95% CI: 1.12 to 2.22).**	odds ratio	1.86	95% [1.25-2.78]	current depression
	odds ratio	1.58	95% [1.12-2.22]	anxiety
Regression analysis demonstrated that **4G homozygosity (OR = 0.176), hypertension (OR = 6.288), and body mass index (OR = 1.325)** were independent predictors of stroke.	odds ratio	0.176	-	4G homozygosity
	odds ratio	6.288	-	hypertension
	odds ratio	1.325	-	body mass index
Moreover, when the model was adjusted for age, male subjects carriers of the 27Glu allele had a significant ten-fold higher risk of **abdominal obesity (OR = 10.31; 95 % CI: 1.4-76.8)**...	odds ratio	10.31	95 % [1.4-76.8]	abdominal obesity
...with a 27% increased risk for **diabetes (hazard ratio 1.27 [95% CI 1.02-1.57])**.	hazard ratio	1.27	95% [1.02-1.57]	diabetes
The **incidence of morbid obesity was 0.6%** in 5,824 women	incidence	0.6%	-	morbid obesity

5.5. Evaluation and Results

In this Section, the results returned from the application of the normalization approach to the evaluation set are shown. The evaluation results are not reported for the exposure, outcome and covariate characteristics since the performance of MetaMap has been extensively reported with an overall F-score of 81.3% for the normalization of biomedical concepts, suggesting reliable results (Aronson, 2001; Denny et al. 2003; Aronson et al. 2010).

At the document level, spans of study design, population and effect size mentions are considered to have been normalized correctly only if all their reported attributes have been successfully recognised. For example, in the population span "*boys aged from 13-15 years old*", the attribute gender should be identified as "*male*" while the age should be "*juvenile"*. If one of these attributes is not normalized, then the population span has not been normalized correctly. In addition, if a span does not have any attributes mentioned and the system does not return any falsely recognised ones, then the span's normalization does not count as incorrect. At the attribute level, we consider all attributes separately i.e., independently. If one span has been normalized for one attribute but in another the normalization process has failed, the normalization procedure is considered correct for the attribute that was recognized. For example, in the population span, "*boys aged from 13-15 years old*", if gender was identified ("*boys*") but the age was not ("*13-15 years old*" suggesting the juvenile age), then the population has been normalized correctly for the gender attribute but not for the age.

Due to the relatively small number of concepts in certain characteristics e.g., 13 study design and 24 population mentions (each one is the chosen lengthiest span of each abstract), a further random sample of 100 MEDLINE epidemiological abstracts was selected from a larger corpus for each characteristic in order to perform a more meaningful evaluation of the normalization method when needed (when a characteristic had a limited number of mentions in the evaluation set) to further validate the normalization methods. Tables 44 and 45 reveal the accuracy of the normalization text processing algorithms for the characteristics of study design, population, and effect size at the attribute level while, Table 46 shows the accuracy at the document level.

5.5.1. Attribute Level Evaluation

Tables 44 and 45 show the accuracy of the normalization process for study design, population and effect size mentions at the attribute level in general and in detail respectively. Particularly, in Table 44 the number of correctly normalized spans can overlap for each attribute. For example, in the following span "*Lebanese adolescents*", the span have been correctly normalized for both its age ("*adolescents*") and its nationality ("*Lebanese*"), therefore, it would be considered as an accurately normalized span twice, once for its age and another for

its nationality. Same approach has been applied for the study design and effect size attributes. Figures 58 and 59 illustrate the accuracy for the population and effect size attributes.

Table 44: Accuracy of the normalization for study design, population and effect size concepts from both the evaluation set and the random sample at the attribute level. The accuracy results are being shown in spans normalized (in total) for their attributes and in spans that contained no attribute.

evaluation method		**study design**		**population**		**effect size**	
		attribute	no attribute	attribute	no attribute	attribute	no attribute
evaluation set	correct	3	10	14	9	343	0
	incorrect	0	0	1	0	12	0
	accuracy	100.0%	100.0%	93.3%	100.0%	96.6%	-
random evaluation sample	correct	33	66	63	28	551	0
	incorrect	1	0	9	0	18	0
	accuracy	97.0%	100.0%	87.5.0%	100.0%	96.8%	-

Table 45: Accuracy of the normalization for population and effect size concepts from both the evaluation set and the random sample at level of each attribute specifically. The accuracy results are being shown in spans normalized for each of their attributes and in spans that contained no attribute.

evaluation method		**population**					**effect size**				
		attribute				no attribute	attribute				no attribute
		gender	age	nationality	ethnicity		type	value	concept	CI	
evaluation set	correct	10	8	4	0	9	75	111	76	81	0
	incorrect	0	0	1	0	0	2	2	4	4	0
	accuracy	100.0%	100.0%	80.0%	-	100.0%	97.4%	98.2%	95.0%	95.2%	-
random evaluation sample	correct	36	39	15	4	28	107	198	100	146	11
	incorrect	3	0	4	2	0	0	0	8	10	0
	accuracy	92.3%	100.0%	78.9%	66.6%	100.0%	100.0%	100.0%	92.5%	93.5%	100.0%

The evaluation of the normalization method at the attribute level revealed encouraging results. The returned accuracies were noted to be above 93.3% in the evaluation set for all the characteristics. More specifically, study design was noted to have 100.0% accuracy in both mentions that had been normalized to their attributes and in those who were not since they had no specific information in their spans. Nevertheless, in the random evaluation sample of 100 abstracts, only one study design was not normalized correctly to its respective attribute (see Section 5.6.1.), resulting to an accuracy of 97.0%. Additionally, any mention that did not carry specific information regarding study design attributes, returned an accuracy of 100.0%. This indicates that the normalization method followed for the characteristic of study design is reliable.

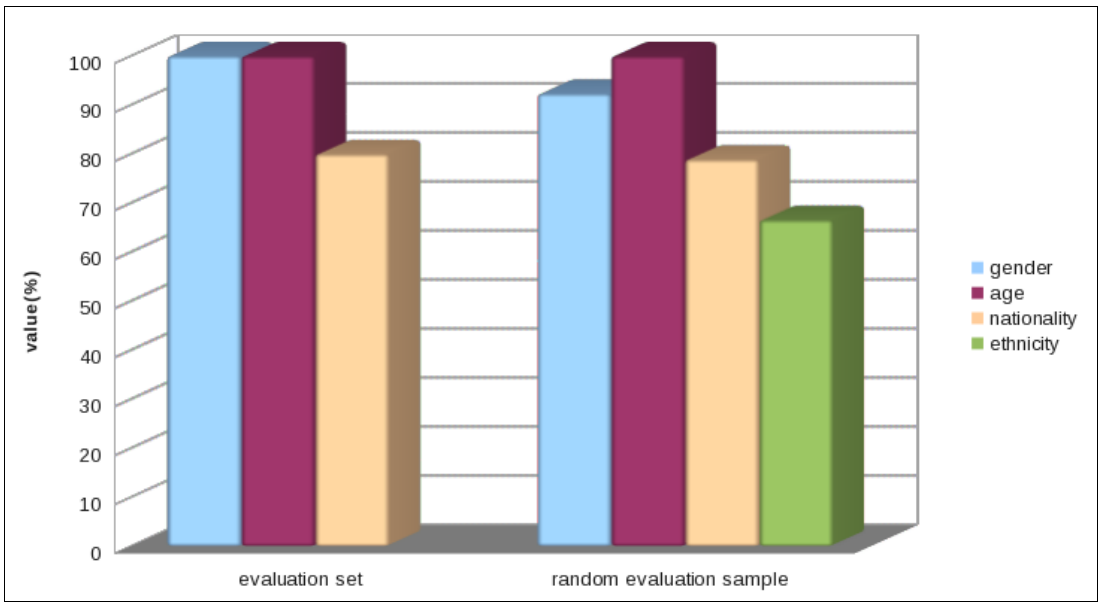


Figure 58: Accuracy results for the attributes of population in the evaluation set and the random evaluation sample. There is no value for ethnicity in the evaluation set as there was no span to perform normalization.

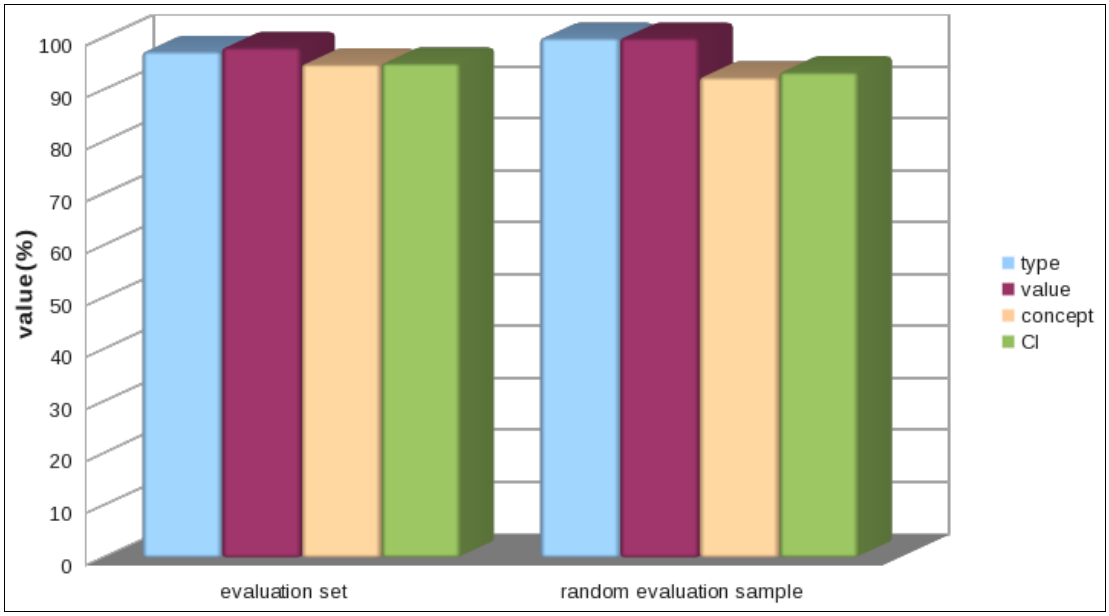


Figure 59: Accuracy results for the attributes of effect size in the evaluation set and the random evaluation sample.

It was observed that in the characteristic of population, for the evaluation set, a total of 14 population attributes were recognised with only one being normalized incorrectly. This returned an accuracy of 93.3%. However, it was noted that in the random test sample, the accuracy dropped 5.8% (87.5%). Table 45 reveals that the best accuracy for the normalization of population spans remained in the attribute of age with a value of 100.0%. The lowest accuracy in an attribute was noted in that one of ethnicity and nationality with values of 66.6%

and 78.6% respectively. However, the accuracy of ethnicity should be taken with caution since only 6 total mentions were present in the random sample and the dictionary resources used for the normalization are not still comprehensive. Additionally, the same conclusion could be applied to a lesser degree for nationality.

For the characteristic of effect size, the accuracy at the attribute level overall was 96.6% in the evaluation set. A small increase was observed in the random set in comparison with the evaluation corpus with the accuracy being 96.8% (+0.2%). More specifically, it was noted that all values at the attribute level in both sets were above 92.5% (Table 45), indicating good normalization results. Drops were noticed in confidence interval and related with the effect size concept from the evaluation set to the random set with 2.5% and 1.7% respectively. Best attribute accuracy in the evaluation set was that of the effect size value with 98.2%. On the random set, it was observed that the best accuracy was that one of the effect size type and value with 100.0%. From Table 45, it can be inferred that it is relatively straightforward to recognise the effect size type and value since both returned 100.0% accuracies in a larger set than that one of the evaluation. However, it appears that the normalizing script requires more flexibility in order to detect confidence interval and the related concept (See Section 5.6.3).

5.5.2. Document Level Evaluation

Table 46 reveals the accuracy results of the evaluation process at the document level for study design, population and effect size mentions. It should be noted that if an effect size mention contains more than one effect size spans, these are considered independent from each other in the normalization process (e.g., in the following effect size mention "*the risk for overweight (or = 0.89) and abdominal obesity (or = 0.91)*", there are two effect size spans "*overweight (or=0.86)*" and "*abdominal obesity (or=0.91)*"). This explains the increased number of effect size spans in both the evaluation set and random sample when compared with those of study design and population.

Table 46: Accuracy of the normalization for study design, population and effect size concepts from both the evaluation set and the random sample at the document level.

evaluation method		**study design**	**population**	**effect size**
evaluation set	correct	13	23	113
	incorrect	0	1	5
	accuracy	100.0%	95.8%	95.7%
random evaluation sample	correct	99	91	232
	incorrect	1	9	11
	accuracy	99.0%	91.0%	95.4%

It was observed that overall the accuracy at the document level was above 91.0% in both the evaluation set and the random sample of abstracts. The highest accuracy was noted for study design (100.0%) in the evaluation set. However, the study design accuracy should be taken with caution due to a limited number of study design concepts presented in the set (a total of 13). Still, it was observed that in a random sample of 100 study design mentions, the accuracy dropped only 1.0% suggesting reliable normalization of the study design mentions. For the characteristic of population, it was observed that only one span was normalized incorrectly in the evaluation set (95.8%). In a larger sample of epidemiological text, the normalization accuracy at the document level for study participants decreased with a relatively significant drop of 4.8% (91.0%), still maintaining good accuracy performance. In the effect size, normalizing concepts into their (if existing) related attributes such as confidence interval, effect size value, type and associated concept was 95.7% accurate in the evaluation corpus. However, a small decrease (0.3%) was observed in the larger sample of abstracts with an overall accuracy 95.4%.

By comparing the results from both the evaluation and random sample, it was noted that the values returned in the larger set are more accurate in their description of the normalization method performance due to the random selection of mentions and their increased number. Hence, despite a good performance in the limited number of concepts present in the evaluation set, the accuracy values of the random sample suggest a normalization that efficiently detects the related attributes of study design, population and effect size.

5.6. Discussion

From the normalization methods that were implemented against the recognised mentions of epidemiological characteristics, the accuracy in both the evaluation set and the random sample from the large corpus was above 90.0%. Despite the promising results, certain error types that could affect the system's performance were noted in the normalization process of study design, population and effect size mentions.

5.6.1. Study Design

During the normalization of study mentions from the test set, it was observed that no span was normalized incorrectly. However, few mentions of epidemiological study design were identified in the evaluation set (13 in total) and they were relatively simple (e.g., "*cross-sectional study*", "*cohort study*", etc.) without any related attributes. It is possible that the application of the string comparison module could have generated incorrect normalized versions if it was tested in more complex epidemiological study designs. Therefore, the performance of the system in the evaluation set should be taken with caution.

However, more supportive evidence about the system's good performance can be found from the application of the normalization procedure in a random sample of 100 study design mentions selected from the large corpus with 99.0% accuracy. From this random sample, it was observed that only one was normalized wrongly:

- "*randomized controlled trial*" normalized into the "*non-randomised controlled trial"* node

"*Randomized controlled trial*" has not been integrated as a node in the expanded study design branch of OCRe. Therefore, with the implementation of that particular module, the lexically closest node is the "*non-randomized controlled trial*". This suggests that perhaps any other mentions that report "*randomized controlled trials*" have been normalized to as "*non-randomized controlled trials*". The lack of specific nodes in the study design ontology may have lead to other wrong normalizations that were not detected from the random sample. While the current performance is an indicator of reliable normalization for various study designs and their associated attributes, by assimilating more study designs could actually improve the accuracy of the normalization and add another layer of information to the extracted data.

5.6.2. Population

In the evaluation set, only one population mention was not normalized correctly. Particularly, in the span "*132 Kirghiz patients with insulin resistance*", the "*Kirghiz*" nationality was not recognized since it was not included in the associated lexical resources. The total mentions for population were relatively few and despite an overall promising accuracy, in a larger set, the system may produce more errors. In the random population sample of 100 mentions selected from the large corpus, the system returned an accuracy of 91.0% (see Table 46). This reveals a more plausible performance and in combination with the one returned from the evaluation set, the normalization process for population is relatively good. In the 100 mentions, 9 were incorrectly normalized. Two types of errors were noted:

1. **lack of associated nationalities from the lexical resources;**

More specifically, in the identified populations spans:

- "***US** population*"
- "***US** children***"**
- "***US** women*"

surprisingly, “US” was not included as a nationality in the lexical resources (instead “*American*” is present). An argument could be made that the current nationality dictionaries may lack the variations that can describe an individual's nationality due to the formality of the resources that were obtained from. For example, people that are from Sweden can be called “*Swedish*” or “*Swedes*”. The dictionaries contained official descriptions of nationalities but cases like those above may not be included as the related terms could be unofficial, informal or new that require time for their establishment e.g., “*US*”.

2. **lack of related ethnicities from the lexical resources;**

A similar observation was noted for ethnicities. While not as common as the lack of nationalities (since it has less mentions), the description of various ethnic backgrounds has been a more complicated one in order to incorporate all the possible ones in a dictionary due to the nature of the attribute. For example in the following identified population mentions:

- “*sixty-seven **Chinese American** children*”
- “*white women / **black** women*”

“*Chinese American*” can be considered an ethnicity. However, this specific ethnic background was not present in the official USA census that described the multiple ethnic categories and groups. A common interpretation could be the term is referring to the nationality of the population. However, this could lead to the generation of false positives and false negatives respectively as both “*nationalities*” can be used to express either the “*Chinese*” or the “*American*” nationality. Furthermore, it is logical to assume that other rarer ethnic backgrounds would not be recognised and normalized by the system since they are not incorporated currently in the lexical resources.

Additionally, in the second case, it is noted that “*black*” was used as the typical term to describe the ethnicity of the population sample since “*white*” was not included in the related dictionary. Depending from which perspective this issue is seen, the normalization of that span can be considered wrong. A possible solution to this problem could be the implementation of a method that mixes the nationalities and ethnicities in order to produce compound terms such as “*Chinese American*” or “*Greek American*”. These terms can then be used in the respective dictionary and potentially improve the normalization accuracy of the system.

From the analysis of the mis-normalizations at the population's attributes level, in the random sample that the most difficult attribute to normalize population spans for is the ethnicity

(50.0% accuracy) and then the nationality (78.9%). This further enhances the remark that larger dictionaries with more associated ethnicities and nationalities could potentially improve the accuracy of the system to detect population spans. However, these accuracy values should be taken with caution due to the limited number of nationality and ethnicity mentions in a sample of 100 population samples (19 and 6 respectively).

5.6.3. Effect Size

The results generated in the normalization of effect size spans in the evaluation set revealed encouraging performance with an accuracy of 95.7%. A limited number of effect size spans were normalized wrongly for their attributes (a total of 5). The examples taken from the evaluation set are shown below:

- "meat-fat dietary pattern was positively associated with **obesity (odds ratio for high tertile vs low tertile intake=2.78 [95% confidence interval: 1.43 to 5.42**]"
- "**low high-density lipoprotein cholesterol (hr: 1.88, 95% ci: 1.29- to -2.77)**"

These effect size mentions were not normalized for their attributes since the implemented regular expressions did not cover their syntactical expressions. In the first case, the incorporation of the "*odds ratio for high tertile vs low tertile intake*" after the related concept and before the reporting of the effect size and its respective confidence interval did not match the format in which the implemented regular expression was based on "<*concept*> <*effect size type*> <*effect size*> <*confidence interval*>". This resulted in the non-normalization of this span for its attributes. In the second example, the inclusion of extra symbols such as "-" in the reporting of confidence interval has not been encountered before, therefore this lead in a failure to recognise the respective value of confidence interval.

Additionally, in a random sample of 100 mentions from the large corpus, a drop in accuracy was noted by 0.3% indicating reliable performance overall. Since the number of effect size mentions was smaller in the test set, and from the larger sample of 100. The performance was 95.4%, this demonstrates that the normalizing method is robust enough to generate good results. This is further supported with the fact that the recognised effect size spans can be normalized easily with regular expressions (see Section 5.4.). The incorrectly normalized spans from the population sample can be classified into two categories:

1. **spans containing acronyms;**

In the following example:

- "... **high total cholesterol (tc) (or: 1.26, 95% ci: 1.03-1.54)** and **high triglycerides (tg) (or: 1.38, 95% ci: 1.16-1.64)** ..."

the regular expressions implemented in the normalizing algorithm aiming to recognise any repeated patterns in an effect size mention:

- <concept> <effect size type> <effect size> <confidence interval>

failed to detect the related attributes. More specifically, due to the presence of acronyms (e.g., "*(tc)*", "*(tg)*") in the effect size concept, the regular expressions were not able to detect the above mentioned example. Despite only two incorrect normalizations, it is possible that in a much larger set, spans with identified concepts containing parenthesis/brackets will not be normalized to their concepts but only to the effect size value and respective confidence interval (if presented).

2. **lacking the words from dictionaries that suggest the presence of an attribute;**

A relatively small number of identified effect size spans were incorrectly normalized at the document level to their attributes due to the presence of brackets and parenthesis. In the following example taken from the large sample of effect size spans:

- "increased risk of **gastric cancer [odds ratio (or)=1.22; 95% confidence intervals (cis)=1.06-1.41]"**

the normalization procedure was performed incorrectly since the confidence interval was not recognised. More specifically, "*confidence intervals (cis)*" was not included in the respective regular expression since it has not been encountered before. Therefore, the system ignored the values of confidence interval in this effect size span. It can be assumed that perhaps in a larger sample, more cases similar like this one mentioned above could be ignored by the system. The incorporation of cases like this into the design of the normalizing regular expressions could improve the performance of the system.

By reviewing the accuracy at the attribute level in the random sample of effect size mentions, the best one belonged to the effect size type and value (100.0% both). Nevertheless, the lowest performance belonged to the effect size concept (92.5%) due to the inclusion of acronyms and other epidemiologically related concepts. However, these cases can be addressed with the incorporation of larger vocabularies that will include acronyms as well as more flexible regular expression rules to allow identification of the related with the effect size concepts.

5.7. Comparison with Other Approaches

"A scientist's aim in a discussion with his colleagues is not to persuade, but to clarify."
Leo Szilard, 1961

Here we summarize the differences between the rule based system developed in this thesis and other related approaches.

1. The methodology includes recognition of key epidemiological information from observational study designs (e.g., "*cohort study*", "*cross-sectional study"*, etc.) as well as from any type of secondary research ("*meta-analysis study*", "*literature review")* rather than focusing solely on experimental study types (e.g., "*clinical trials*"). Consequently, a significantly more diverse literature space is being addressed like case-control, cross-sectional and cohort study types rather that relying only in randomized clinical trials. Clinical trials are subject to strict regulations and are reported in highly standardised ways, hence making it a less challenging task to discover generic concepts of interest such as exposures or outcomes.
2. The extraction of key epidemiological characteristics presented in the majority of epidemiological study designs (both observation and experimental) is a much more different procedure than those already applied to clinical trial text. The aim is to enable the capture of key information observed in all types of epidemiological study rather than focusing in the extraction of specific elements from clinical trial text e.g., "*secondary outcome*", "*eligibility criteria*", "*duration of the treatment*", etc.
3. The approach includes the normalization of the extracted key characteristics into the UMLS semantic classes and for their (if present) attributes in order to identify more accurate information. The normalized characteristics aim to contribute to the aggregation and easy exploration of epidemiological data through EpiTeM and the automatically generated concept map (Chapter 7).

The above differences make it difficult to contrast the generated results with those of other studies of similar research so direct comparison cannot be done. However, an indirect inspection of the methods and their performance can be made with studies focused on the identification of clinical trial attributes. The range of the precision and recall can be compared to those aiming to extract:

- similar or same epidemiological information such as exposures (Xu et al. 2010) and risk factors (Fizman et al. 2007);
- any type of epidemiological information from clinical trial text (Hara et al. 2007;

Hansen et al. 2008; de Bruijin et al. 2008; Chung, 2009a; Chung, 2009b; Kiritchenko et al. 2010; Luo et al. 2011; Luo et al. 2012).

More specifically, the system's (micro) precision was 88.5% (see Section 4.4.1.) for key epidemiological characteristics indicating good performance overall across the sets of data. It also suggests reliable performance in comparison to studies aiming to identify similar epidemiological information. Studies conducted by Fizman et al. (2007) and Xu et al. (2010) reported 67.0% and 61.0% precision in 350 MEDLINE sentences and 450 epidemiological article titles for the identification of risk factors and exposures respectively, while their returned recall was 83.0% and 69.% respectively (in contrast with the system's more promising recall of 86.3%). Figure 60 shows the (macro) performance of the system for all the characteristics in comparison to the performance of Xu et al. (2010) and Fizman et al. (2007). However, they aimed to recognise the targeted information from MEDLINE citations and titles of epidemiological articles rather than epidemiological abstracts.

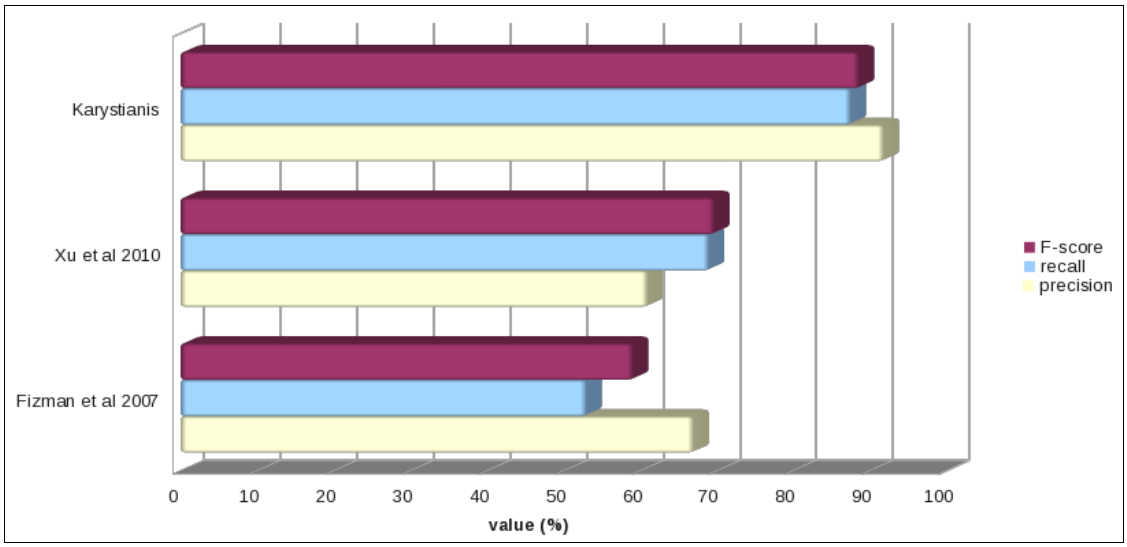


Figure 60: Comparison of the system's (micro) performance with studies aiming to recognize key epidemiological information from MEDLINE citations and epidemiological article titles.

Furthermore, studies that focused on the identification of various elements from clinical trials revealed more competitive (and detailed) performance (Figure 61). Research focused on the identification of the participant population in a clinical trial yielded encouraging values of 79.8% F-score (Hara et al. 2007) and 84.0% F-score (Hansen et al. 2008). Additionally, any identification of various elements e.g., "*coordinating instructions*", "*outcome measures*", "*interventions*", etc, presented in clinical trials returned a large range of precision, recall and F-score values (65.0%-93.0%, 75.0%-91.0% and 69.6-91.9% respectively). Different recognition element methodologies (machine learning techniques, etc) are applied for different types of

epidemiological studies (clinical trials) for the identification of more specific key features mainly presented only in these types, explaining the multiple values that precision, recall and F-score returned. The variations in performance for the extraction procedures for RCT characteristics, suggests that clinical trial text is better structured and less challenging than epidemiological study design abstracts.

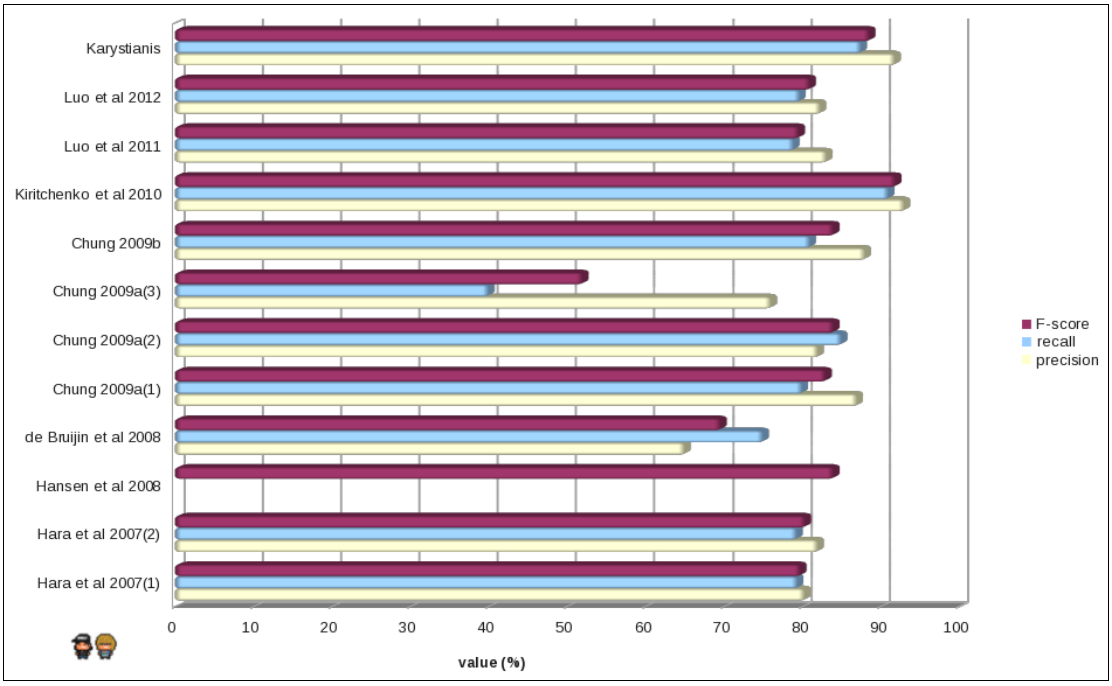


Figure 61: Comparison of the system's performance with studies aiming to recognize key elements from clinical trial text.

A direct comparison of previous efforts in epidemiological data cannot be done with our system since various datasets that are different than those we have used, have been selected as well as for different identification tasks with specific targeted characteristics rather than key ones that can be found in most of types of epidemiological research.

The mining and continuous integration of text-mined results is paramount for uncovering the complexity and enhancing the clinical understanding of the most common chronic diseases in society such as obesity or depression. Despite focusing on obesity mainly for the purpose of evaluation (and as a case study), this rule based approach applied for the recognition of key epidemiological characteristics related to a particular health problem, is generic. As indicated earlier, articles in epidemiology are relatively structured, with their related vocabulary being stable. Epidemiological abstracts are deliberately written to improve the study design as well as the collaboration between respective scientists internationally. Despite that, certain domains will still have certain characteristics (for example exposure models in air pollution studies), in

which the same approach should be feasible for the extraction and normalisation of key epidemiological features.

In addition, most of the previous work in the automatic generation of a concept map from unstructured text was performed in non-clinical/epidemiological data such as articles from the e-Learning domain, on-line texts and unspecified unstructured textual resources. The method we developed and implemented, focused on the automatic creation of a concept map from the epidemiological text mining results generated by our rule based approach for the recognition and normalization of key characteristics in epidemiological study abstracts related to a health problem. Previous efforts (e.g., Leximancer – Watson et al. 2005) automatically create concept maps from unstructured informally written text (e.g., patient records and clinical cases) rather than relatively structured text such as epidemiological study abstracts. Their aim was the identification of the information that should be sorted in clinical databases rather than the representation of concepts researched in epidemiological studies associated to a health problem. The automatically produced concept map can reveal:

1. the exploration of its concepts through observation and cluster detection;
2. the generation of new research hypothesis by making new associations.

Most other tools have been designed and developed with particular fields in mind such as the e-Learning domain (Chen et al. 2008) or in different languages (Zubrinic, 2011) rather than epidemiological text. However, this rule based extraction method is generic enough and could potentially be applied to other domains of the biomedical field and other types of semi-structured data. Additionally, these CMM tools are focusing more on the identification of the concepts and their respective relationships (Villanon et al. 2008; de la Villa et al. 2012). Since our aim is to represent the epidemiological text mining results in a coherent and easily readable form that could show an overview of a health problem as well as could promote potential exploration, it was not required to recognise the relationships between the concepts. Due to the application of the method in a large scale text mining procedure, the amount of generated data is large enough, so currently if any tool could be potentially applied to visualize the following results, an extremely large concept map would be created, making it difficult to navigate and manipulate it.

5.8. Limitations and Challenges

The design and implementation process of the rule based system for the identification, normalization and visualization of key characteristics from epidemiological study abstracts came with certain challenges.

1. **Type of data**: Since the aim was the recognition of key epidemiological elements, all types of studies were included (both experimental and observational with the extra addition of secondary research such as meta-analysis studies and literature reviews). Clinical trials contain more specific context than observational studies with various syntactical expressions indicating the existence of unique elements (e.g., "*route of treatment*", "*start date of enrollment*", etc.), therefore it is less of a challenge to detect and annotate key epidemiological information. However, observational studies and secondary research have a more descriptive nature and the identification of characteristics such as exposures, outcomes or covariates through a rule based approach is not a straightforward task. Certain exposures and outcomes are being described in text through lengthy expressions or complex sentence structures. Despite using a relatively generic rule based approach to enable the capture of most characteristics mentions, there were a number of cases in which the system was not able to the identification of epidemiological spans. Based on these observations, any rules created for the detection of epidemiological characteristics in text might require a stricter design, although a balance should be pursued between the robustness of the system in order to avoid falling to the trap of over-fitting.
2. **Multiple study designs in one abstract**: It was assumed that most of epidemiological abstracts mention the design that was followed for the given study only once. However, in few cases, the system detected more than one epidemiological study types. At this point, it is assumed that the lengthiest span is the one that captures the most detailed information regarding the type of study. Any other spans are then considered synonymous (due to the summarization of the key epidemiological information at the document level). As a result, a wrong span may have been selected as it could be referencing another conducted study mentioned in the same document (as it was observed to be the case). Consequently, in cases like this, it is difficult to understand which of the multiple spans a) is referring to the abstract's study design and b) if more than one indeed do, which one describes most accurately the study design(s). A potential solution to this is to take into consideration the abstract's structure. Most epidemiological abstracts are semi-structured indicating their method under headings such as "*method*", "*design*", etc. Designing rules that are incorporating the headings in their expression might decrease the extraction of multiple study designs in one citation.
3. **Size of the population sample:** In some cases, the system highlighted more than one population spans regarding the observed population sample in one abstract, resulting in the increase of false positives. These spans are not semantically wrong as they are subsets of the population sample, but they bear no significant information regarding

the population's attributes. A set of stricter rules regarding the identification of the population samples could potentially assist in the decrease of the number of undesirable spans.

4. **Confusion between exposure and outcome:** As it was mentioned in Section 4.4.3, in certain situations there is confusion between concepts that act as exposures and those that are considered outcomes. Exposure concepts are not explicitly stated in text and their role could be easily misinterpreted as an outcome rather than a studied determinant (and vice versa). Additionally, the presence of more than one exposure in one study makes their identification and differentiation from the outcomes a challenging task. These two characteristics share a studied cause and effect association in most studies. Therefore, a rule based methodology based on syntactical expressions may struggle to understand a concept's role in order to identify it as an exposure or an outcome respectively. This leads to incorrect classification of concepts between the characteristics of exposure and outcome. In order to resolve this issue, we hypothesised that in these syntactical associations, the first mentioned concept is the exposure and the second is the outcome. While in the majority of these associations this was true, there are still a number of cases in which the outcome was mentioned first. Potentially, the implementation of stricter rules or a machine learning approach with features related to either the exposure or the outcome could address this challenge.
5. **No covariate recognition:** As it was indicated in Chapter 4, covariate mentions are relatively straightforward to be recognised with a rule based approach in epidemiological text. Despite the limited number of false negatives in the evaluation of covariate concepts, it is possible that covariate concepts may still be present in an epidemiological study without being explicitly mentioned. Their role as a covariate could be inferred through reading and comprehending the entire conducted research study. Consequently, the detection of covariates may be a bit more challenging than it was initially thought and perhaps the application of machine learning techniques could capture the necessary information.
6. **Lexical resources:** The utilization of the current lexical resources did not include a number of concepts for most of the characteristics, thus larger dictionaries are probably required. More precisely, any concepts of different nature than biomedical, may be studied as determinants and outcomes, or being mentioned as covariates (e.g., "*high school environmental activity*"). These may no be detected by ATR or included in the SPECIALIST lexicon. Since the lexicon is a thesaurus of (mostly) biomedical terms, it is likely that it is missing concepts of i.e., social aspects, drug names, etc leading to the increasing number of false negatives in some cases. Therefore, the

incorporation of other dictionaries e.g., a list of the medications from the DrugBank[23] database, could help in the identification of concepts that act as exposures, outcomes and covariates.

7. **Existence of synonyms:** It is often the case that the same concept is mentioned several times using different synonyms. For example "*carcinoma*", "*cancer*", and "*neoplastic process*" could all be extracted as outcomes in an epidemiological study as there is little lexical similarity currently. The system treat cases like these as individual outcomes rather than one. The generated results could contain a variety of synonyms that could affect to various degrees the statistical results presented in Chapters 4, 5 and 7. Consequently, the representation of epidemiological text mining results in a concept map that includes multiple synonyms under the same characteristic, contributes to its (relatively large) size and makes it more difficult to explore. A search and identify procedure for respective synonyms of each concept could potentially solve this problem. For the above example, if any of these three terms have been extracted from text and are considered synonymous, then probably, then only one will be kept in order to represent the document.

5.9. Summary

A normalization procedure is applied to the identified mentions of the characteristics at the document level in order to enable the recognition of descriptive attributes that can assist in the understanding of epidemiological information. In order to eliminate similar or identical spans, it was hypothesized that the lengthiest span is the most informative. The normalization process is applied to the mentions that have been considered unique by using a string comparison module between the lengthiest span and the rest of other mentions for each characteristic. Spans similar to the longest one are ignored while those that are not, are used for normalization.

For the normalization of the study design, an adapted version of OCRe is used. Through the application of the string comparison module, the extracted study design mention is compared to those that form a part of the expanded ontology and the match with the highest score is chosen as the normalized version of the input design. Any additional information that is captured is being stored as study attributes. The identified population spans were normalized according to (if existing) specific attributes (age, gender, nationality, ethnicity). Nationality, ethnicity and gender are being detected from the use of respective dictionaries while age is recognised from applied regular expressions due to its relatively structured format when reported. If any of the attributes are not detected in the chosen longest span, then a script is

23 http://www.drugbank.ca/

applied to the other identified mentions in case they contain any related to these attributes information.

The normalization of exposures, outcomes and covariates is essentially normalization of biomedical concepts and since biomedical classification is required MetaMap was used. Each concept is classified into a UMLS semantic category and a higher-level UMLS semantic group. For effect size mentions, an automatic text processing algorithm based on regular expressions that focused on their individual attributes (effect size measure type, the respective value of the effect size, confidence interval, related concept) was applied.

At the document level, spans of study design, population and effect size mentions have been normalized correctly only if all their existing attributes have been successfully recognised. If one span does not have any attributes to be normalized for the span's normalization does not count as incorrect. At the attribute level, if one span has been normalized for one attribute but in another the normalization process has failed, the normalization procedure is considered correct for the attribute that was recognized. Due to the relatively small number of concepts in the study design and population characteristic in the evaluation set, a further random sample of 100 MEDLINE abstracts was selected from a larger corpus for each characteristic in order to perform a more meaningful evaluation of the method.

At the attribute level, the returned accuracies were above 93.3% in the evaluation set for all the characteristics therefore suggesting reliable results. However, in the random evaluation sample, a decrease was observed in the performance of the study design and the population (97.0% and 87.5% respectively) with a small increase in the accuracy of effect size (from 96.6% to 96.8%). Age had a steady accuracy of 100.0% in both the evaluation set and the random sample while the lowest accuracy belonged to ethnicity (66.6%) although this value should be taken with caution since few mentions were present in the random sample and the dictionary resources used for the normalization are not still comprehensive. At the document level, the accuracy was above 91.0% in both corpora and despite study design having the highest accuracy (100.0%) the value should be taken with caution due to a limited number of study design concepts. Despite a good performance in the limited number of concepts present in the evaluation set, the accuracy values of the random sample suggest a normalization process that efficiently detects the related attributes of study design, population and effect size.

The lack of specific nodes in the study design ontology may lead to wrong normalizations that are not detected from the random sample (e.g., the mapping of the "*randomized controlled trial*" span to the "*non-randomized controlled trial*" node). While the current performance is an indicator of reliable normalization for various study designs and their associated attributes, by assimilating more study designs could actually improve the accuracy of the normalization

and add another layer of information to the extracted data. Lack of associated nationalities and ethnicities from the lexical resources generated incorrect normalizations of population spans with rarer ethnic backgrounds being unable to be recognised by the system. In addition, certain effect size mentions were not normalized for their attributes since the implemented regular expressions did not cover their syntactical format with spans containing acronyms and the dictionaries lacking the words that suggest the presence of an effect size attribute.

Chapter 6

Automatic Construction of Concept Maps from Epidemiological Text Mining

"Visualize this thing that you want, see it, feel it, believe in it. Make your mental blue print, and begin to build"
Robert Collier, 1978

One of the objectives of this thesis is to represent the results of our epidemiological text mining method in the form of a concept map. Epidemiologists interested in synthesising and manipulating evidence about a particular outcome (such as "*clinical depression"*) would likely consider any possible determinants linked to that outcome, such as various environmental and behavioural factors e.g., "*smoking*", "*alcohol intake*", "*physical activity*", and even potential covariates such as "*age*", "*gender*" and "*occupation*". Therefore, it would be useful for the system to enable the recognition of different types of exposures, outcomes and covariates such as "*diseases*" or "*activities*/*behaviours*" in a high level or in a low level e.g., "*findings*", "*therapeutic or preventive procedure*". Consequently, the user has the option to enhance his understanding and complete his knowledge of related factors in a particular study by emphasizing connections between the background of the exposure, the resulting outcomes and the influential covariates. Additionally, through the normalization of the effect size mentions, health professionals can understand the strength of the association between exposures and outcomes and perform a more detailed search that could suggest otherwise for the same concepts but with different effect size types.

The same principle can be applied for the normalization of study design and population. However, most of the attributes to which the associated spans have been normalized for are indicating more specialized information that is more of interest in epidemiologists rather than to other types of health professionals. To be more specific, any user who might developed an interest in the examination of various study designs and their respective population sample, can have the opportunity to associate the type of study designs applied to the population samples under which targeted determinants and outcomes for investigation. Users could make new relationships by observing e.g., which ethnicities can be susceptible in the development of a disease, or which nationalities have most influential covariates, which gender can have a role in disorder progression and if it can influence the outcome of a invasive or non-invasive procedure and at what age stage most of outcomes are being noted.

Therefore, in this stage, the results from the implementation of the epidemiological text mining system for the characteristics of exposure, outcome and covariate are used as an input for the automatic construction of a concept map. Study design, population and effect size were not included since they are not the focus of any concept map exploration due to their nature and the limited information they bare. The aim of the map is to shed light onto the various concepts related to obesity as exposures, outcomes or covariates as well as to enable users to make associations between the represented concepts. The concept map represents the normalized exposures, outcomes and covariates along with their respective UMLS classification under a semantic group and a semantic category.

6.1. Concept Map Building Method

We have developed a method for the automatic generation of a concept map that takes as an input the normalized epidemiological text mining results and a number as the threshold level for the mentions of concepts under which they will be displayed in the map.

The concept map was designed as a structured hierarchy with the root node displaying the health care problem of interest (in this case obesity). This is followed by three primary nodes, each one representing an epidemiological characteristic (exposure, outcome, covariate). Each primary node includes the UMLS semantic groups and categories and, at the lowest level of information, the normalized concepts. Each concept is characterised by the number of total mentions in the entire corpus under its assigned specific classification (UMLS group and category). Labelled relationships are utilized to associate the concepts with the related categories and groups:

- the root node is linked with the characteristic nodes through a "*has*" relationship;
- the epidemiological characteristic nodes (exposure, outcome, covariate) are linked through a "*can be*" relationship with the UMLS semantic groups;
- the UMLS semantic groups are linked through an "*is*" relationship with the respective UMLS semantic categories;
- the UMLS semantic categories are linked through an "*is*" relationship with the respective concept.

Any specific relationships among the concepts are not represented (e.g., "*smoking contributes to the onset of obesity*"). The main objective is to enable the conversion of input information into a concept map file, which is an XML based format used by a concept mapping software IHMC CmapTools[24]. We have chosen CmapTools to visualize the results of the epidemiological text mining approach due to:

24 http://cmap.ihmc.us/download/

1. Easy navigation. CmapTools offers the opportunity to import files representing concept maps allowing easy manipulation of their content and friendly and easy to understand visualization.
2. Ability to allow export of its files to other tools. Users can choose to export their concept maps into various formats such as XML and CXL[25] (concept mapping extensible language based on XML aiming to describe the content of the concept maps) or as LifeMaps and image files or even web pages.
3. It is free to use.

A threshold level (regarding the total number of mentions at the document level for the concepts) is being introduced in order to enable variations of concept display. This threshold can be changed according to the user's requirement.

6.2. Results

In order to be able to display and inspect the automatically generated concept map, the threshold level of document frequency was chosen to be 2 so not all of the normalized concepts are being displayed. Figure 62 reveals the concept map visualized through Cmap for the normalized concepts of the exposure, outcome and covariate characteristics with a threshold of 2 from the evaluation set. The generated concept map can be used to explore exposures, outcomes and covariates of interest from epidemiological studies. Epidemiologists and other clinical professionals are able to manipulate and inspect large amounts of knowledge more efficiently through the overview of a domain's information that the concept map can present. Associations between rare concepts can be inferred and point out potential new areas and concepts of future research regarding a certain health care problem.

In addition, concept maps can be used as a backbone for literature exploration through the observation of concepts that have been assigned into one of the three epidemiological characteristics. For example, by inspecting Figure 62, we witness that "*metabolic syndrome*" has been represented as an outcome under the UMLS semantic group classification of disorder and "*alcohol intake*" as a covariate under the UMLS semantic group classification of activities/behaviours. These criteria (metabolic syndrome as an outcome, alcohol intake as a covariate) could be potentially used as variables of interest while searching related epidemiological literature.

In Chapter 7 (Section 7.8.), a specific concept map for obesity from the related epidemiological literature has been built automatically through the application of our approach and it represents its associated exposures, outcome and covariates.

25 http://cmap.ihmc.us/xml/CXL.html

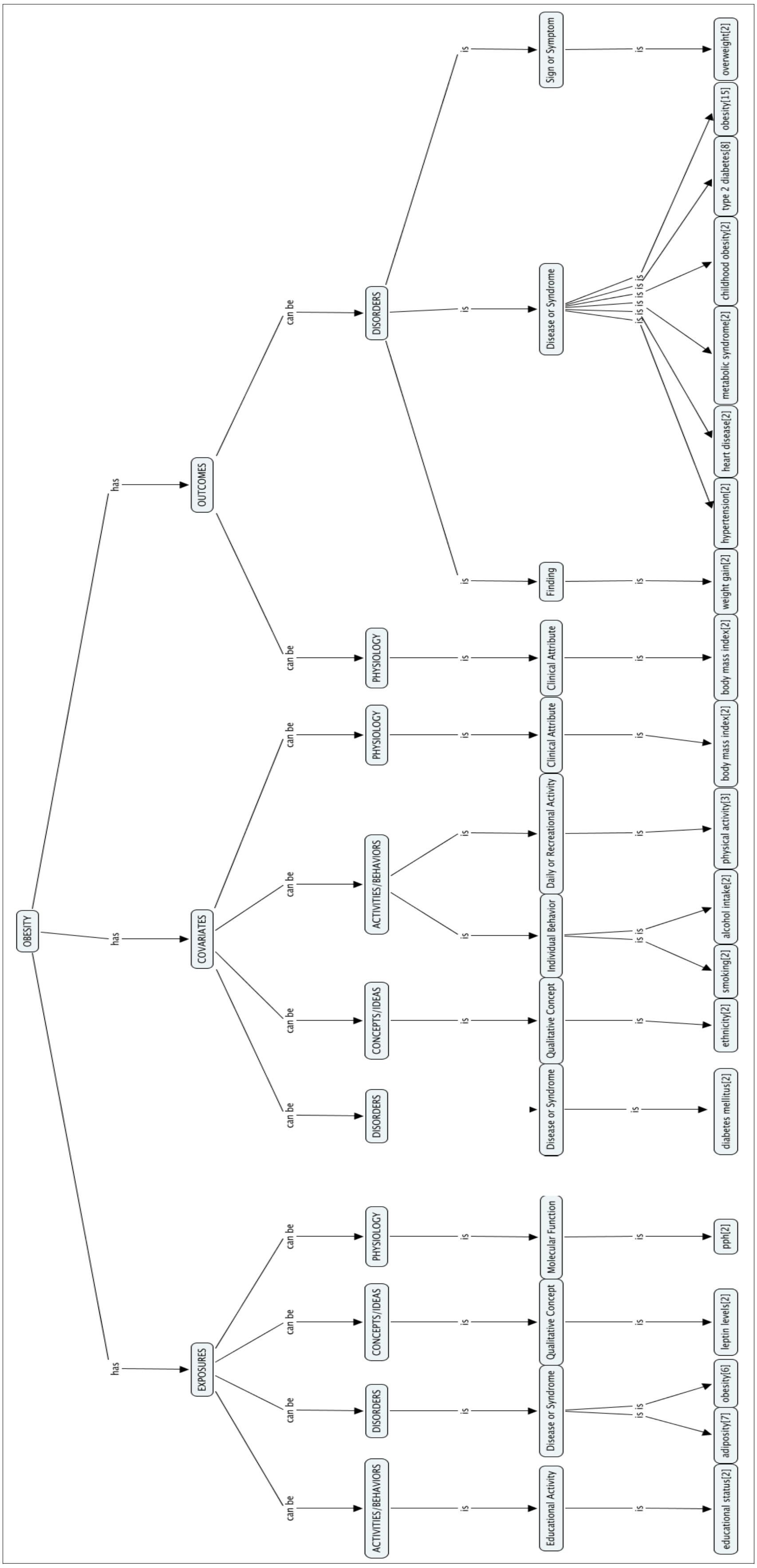


Figure 62: An example of the generated concept map that represents the normalized exposures, outcomes and covariates. The threshold level is 2 and their respective UMLS classification through Cmap.

Chapter 7

Extraction of Key Characteristics from Epidemiological Literature on Obesity: a Case Study

This Chapter describes the application of the proposed approach to the recognition and normalization of key characteristics from a large scale corpus of epidemiological study abstracts related to a health problem. Obesity was chosen as a case study because of its complexity and public health importance. We aimed to generate automatically a concept map that will represent exposures, outcomes and covariates identified and normalized from epidemiological literature related to obesity.

7.1. Obesity as a Major Health Problem

"*The people in power have created an obesity epidemic.*"
Robert Atkins, 1996

Obesity was included as a disorder in the International Classification of Diseases (ICD) in 1948. At the dawn of the 21st century, there have been growing concerns about obesity as it has emerged in recent decades as one of the most important and global health problems with its rates increasing at an alarming rate and reaching epidemic proportions (Wang et al. 2008a; Duncan et al. 2010; Nguyen et al. 2010; Henry et al. 2011; Vucenik et al. 2012; WHO, 2013). It is considered by many experts a growing challenge in the developing world for adults, adolescents and even children (Calle et al. 2004; Hossain et al. 2007). Although obesity was initially the "*epidemic*" described in developed countries (particularly in the United States), its incidence and prevalence is rising sharply in the past 20 years in low/middle income nations that have adopted a Westernised lifestyle with regions in the Middle East, South Asia, Japan and China seemingly having the biggest problem so far (Hossain et al. 2007; Ford et al. 2008; Low et al. 2009; Henry et al. 2011) (Figure 63). Obesity is a complex disease with underlying relationships with various other disorders and it should be treated as a part of a disease cluster (e.g., metabolic syndrome) rather than an isolated condition.

According to WHO, obesity is defined as abnormal or excessive fat accumulation that presents a risk of death or may impair health (Canoy, 2010; WHO, 2013). Fat is the principal energy storage form metabolised from the human body in order to meet its energy needs (Ogden et al. 2007). The degree of obesity can be calculated from the measurements of the individual's height and weight resulting in the individual's body mass index (BMI) (the person's weight in kilograms divided by the square of its height in metres (kg/m^2)) (CDC, 2012). BMI originating

from Quetelet's "*average man*", is widely used in clinical settings and population studies as a measure for obesity since it is strongly correlated with the total body fat content (Pi-Sunyer, 2002). So far, BMI has provided the most useful population level measure of obesity because it is the same for both sexes and for all adult ages (WHO, 2013). A BMI value of 30 or more suggests an obese person, while an individual with a BMI equal to or more than 25 is considered overweight (Ogden et al. 2007; Vucenik et al. 2012). Despite its strong correlation with body fat percentage and the risks to develop obesity-related diseases i.e., cardiovascular diseases and cancer, BMI has been characterized as imprecise due to the conflation of fat mass in body weight (Buchan et al. 2007; Ogden et al. 2007).

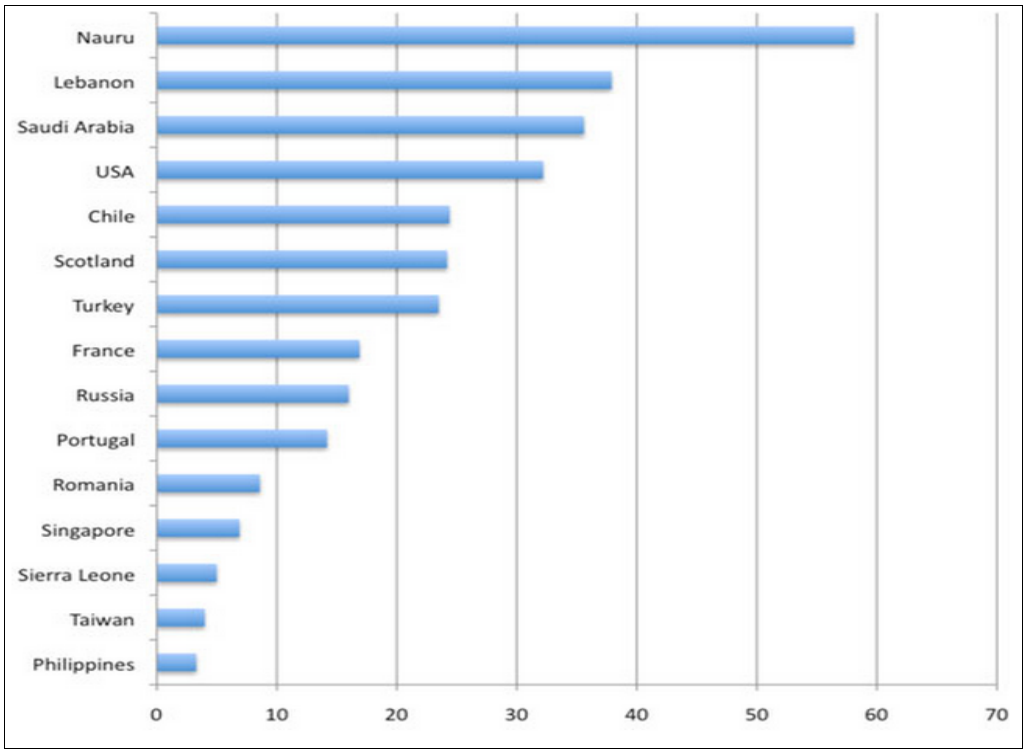


Figure 63: Global prevalence of obesity (in percentage) from the International Association for the studies of Obesity with data collected between 1996-2005 (WCRF 2010).

Its worldwide prevalence has been double since the early 80's and has seen since then a steady increase (Wang et al. 2008a, WHO, 2013). The World Health Organization (WHO) estimated in 2005 that 1.6 billion people were overweight and at least 400 million were obese worldwide. By 2015, 2.3 billion adults will be overweight and at least 700 million will be obese (Malik et al. 2010; Ryan et al. 2011; WHO, 2013). In the US, which is the main (developed) affected country, the prevalence of obesity between the years 1980 and 2004 saw an increase of 18% and currently more than 78 million adults and 12.5 million children and adolescents are obese (Ogden et al. 2007; Ogden et al. 2012). Obesity figures in the UK trembled over the last quarter of century with a rise of 9.5% in men and 6.8% in women (Buchan et al. 2007; Duncan et al. 2010).

WHO has predicted that in the 21st century the economical, clinical and social impact of obesity will surpass that one of other major health problems such as malnutrition and infectious diseases (Hossain et al. 2007; Henry et al. 2011). In the US, the provided health care services related to obesity were estimated to be a 10% of the total health care cost per year (Malik et al. 2010). It has been predicted that by 2048 all American adults would become overweight or obese while the clinical costs for obesity will be accounting for 16-18% of the total US health care costs (Wang et al. 2008a). Childhood obesity is also posing an epidemic threat with more than 40 million children under the age of five worldwide classified as overweight in 2011, while in Europe 20% of children and adolescents were overweight with one third of these classified as obese (Monasta et al. 2010, WHO, 2013). In the young ages, more than 155 million children worldwide are presently either obese or overweight. The prevalence of childhood obesity in UK has increased from 1.5% to 6.3% within a time period of almost 20 years (Duncan et al. 2010). Due to the multi-dimensional clinical nature of obesity, the number of articles and published research related to obesity have been vastly increased per year. These can be found in clinical databases such as National Health Service Centre for Reviews and Dissemination (NHS CRD), MEDLINE and the Cochrane Library on CD-ROM (Low et al. 2009).

7.1.1. Complex Obesity Risk Factors

Obesity follows the technological progression that characterized the end of the 20th century and the beginning of the new millennium. It represents a major and constantly growing has challenge for the technological society (Grundy, 2004). The rapid spread of obesity across the world affected people of all ages, locations, genders and ethnicities. This strongly indicates that both behavioral and environmental determinants are behind its epidemic proportions with only a small percentage explained by genetics (Monasta et al. 2010; Nguyen et al. 2010). The fundamental cause of obesity is the energy imbalance between the calories consumed and expended (CDC, 2012; Demark-Wahnefried et al. 2012; WHO, 2013). A chronic positive energy balance leads to and contributes to weight gain (Henry et al. 2011). However, the control of the obesity epidemic requires understanding the complex relevance of factors that link energy intake and expenditure (Trasande et al. 2010). Obesity may be the consequence of constant dietary and physical activity pattern alterations that result from environmental and social changes (Wang et al. 2008a; CDC, 2012; Gumpta et al. 2012). Therefore, obesity can have a variety of direct risk factors such as lack of physical activity and consumption of sugar-sweetened beverages and food as well as indirect contributors like sleep disturbance (Monasta et al. 2010; Henry et al. 2011).

BMI reflects only body weight, thus any healthy factors such as increased skeletal muscle may not be captured (Ogden et al. 2007). An individual's body weight should be considered as a reflection of a genetic factors, physiologic status, individual behaviors, environmental and social influences cluster, hence making the identification of health risk a difficult and complicated task (Ogden et al. 2007). More research is required in order to understand how obesity leads to early death (Franks et al. 2010). The most common risk factors for obesity related mortality are:

1. **Sedentary lifestyle:** Although genes may determine how individuals process energy, lifestyle and environmental determinants dominate obesogenesis (Eckel et al. 1998; Kahn et al. 2006; Trasande et al. 2010). The prevalence of excessive weight in a population is driven by the energy intake along with a sedentary lifestyle that follows a western approach and includes unhealthy nutrition, increase caloric availability, fat consumption, physical inactivity, smoking, alcohol and soft drink consumption (Alberti et al. 2004; Kahn et al. 2006; Buchan et al. 2007). A global shift of the human diet towards the consumption of low in sugar, vitamins and minerals but rich in fat eating habits and of soft drinks or sugar sweetened beverages contributes to obesity and can increase the risk for diabetes and cardiovascular diseases (cvds) (Alberti et al. 2004; Malik et al. 2010). The widespread availability of electronic devices in households as well as the tobacco use are considered among the environmental determinants of weight and overweight resulting to their association with weight increase (Ogden et al. 2007).
2. **Physical inactivity:** Inactivity is one of the main contributors to obesity and overweight (Alberti et al. 2004; Buchan et al. 2007). There is a trend towards decreased physical activity that has been associated with weight gain (Buchan et al. 2007). The workplace has become largely sedentary due to labour saving technologies and increased mechanization procedures, and modes of transport are changing along with the increase of the urbanization (Pi-Sunyer, 2002; WHO, 2013). Consequently, the amount of physical activity that was part of the everyday life is greatly reduced and risk of obesity and its complications is increased (Davy et al. 2004; WHO, 2013).
3. **Gender, age and socio-economic status:** Several other factors are associated with obesity although it is not clear why or how. Sex, age, race, and socio-economic status have an impact on weight gain which is more likely to happen in women, elders, minority races and individuals of low socio-economic status (Pi-Sunyer, 2002). Mortality from obesity is higher for women than men but the rates are rising steadily for the male population, showing that sex differences may not exist in recent trends in the obesity (Duncan et al. 2010). In England in 2004, 22.7% of the total population

was obese with higher prevalence in some ethnic groups such as Irish, Black Caribbean (Pi-Sunyer, 2002; Buchan et al. 2007). Figure 64 shows the prevalence of obesity among adults with age from 20 years old and race/ethnicity in the US for a 44 year period (1960-2004). In the developed countries, obesity tends to be more common in individuals with low levels of education or occupational status. On the other hand, in the developing nations, higher weight could be linked with wealth and higher socio-economical status, suggesting a positive association between body size and status in both sexes (Ogden et al. 2007).

		Race-ethnicity[a]							
		Total		Non-Hispanic white		Non-Hispanic black		Mexican American	
Sex	Age, y[b]	%	SE	%	SE	%	SE	%	SE
All	20+	31.0	.7	29.8	.8	41.4	1.1	34.2	1.4
	20–39	26.8	.8	24.4	1.0	38.1	1.9	29.4	1.9
	40–59	34.8	1.2	34.4	1.4	43.6	1.5	39.5	2.1
	60–79	35.2	1.0	35.0	1.1	45.9	2.3	36.2	1.8
	80+	17.3	1.3	17.4	1.4	25.9	4.7	11.9	4.7
Men	20+	28.8	.7	29.2	.9	30.0	1.5	28.9	1.4
	20–39	24.7	1.1	24.3	1.5	27.3	2.1	26.8	2.5
	40–59	32.2	1.3	32.8	1.6	32.5	1.8	31.5	2.5
	60–79	33.1	1.3	34.7	1.5	32.5	3.3	30.0	2.4
	80+	14.0	1.4	14.5	1.6	12.9	6.4	10.4	4.4
Women	20+	33.2	1.0	30.5	1.1	50.6	1.6	39.8	2.0
	20–39	29.0	1.2	24.5	1.6	47.7	2.4	32.9	2.9
	40–59	37.4	1.6	35.9	1.9	53.1	2.2	47.9	2.7
	60–79	36.9	1.4	35.3	1.6	55.4	3.7	41.5	2.5
	80+	19.2	1.8	19.1	2.0	31.1	5.5	13.1	6.3

NOTE. Obesity, BMI ≥ 30; National Health and Nutrition Examination Survey 1999–2004.
SE, standard error.
[a]No significant race/ethnic differences among men. Among women, non-Hispanic black > Mexican American (20–39 y and 60–79 y), non-Hispanic white < Mexican American (40–59 y), non-Hispanic black > non-Hispanic white (20–39 y, 40–59 y, and 60–79 y). Significance at $P < .05$ with Bonferroni correction.
[b]Among both men and women, significant age differences 40–59 years > 20–39 years and 60–79 years > 80+ years. Significance at $P < .05$ with Bonferroni correction, without regard to race/ethnicity.

Figure 64: Prevalence of obesity among adults with age from 20 years old and race/ethnicity in the US in the years 1960-2004. SE stands for standard error (Ogden et al. 2007).

4. **Genetics:** Not all members of a population are at risk of obesity and its consequences although some individuals will have a genetic predisposition more than others (Sperrin et al. 2013). The individual effect of genes in obesity though is quite modest and genetic predisposition to obesity might be the result of a combination of genes (Marinou et al. 2009). There is a chance though for this particular genotype to be expressed only under specific environmental conditions which include a sedentary lifestyle (Pi-Sunyer, 2002; Marinou et al. 2009). In the Western countries, many people are being exposed to obesogenic conditions, which include increased consumption of energy dense foods, limited energy expenditure and lack of physical activity. Consequently, the population percentage expressing that genotype is rising, leading to the current increase of obesity and pointing out the importance of the interaction between the environment and the genetics (Ogden et al. 2007) (Figure 65). However, genetic factors are unable to explain comprehensively the current prevalence of obesity in countries such as USA, UK and many others (Ogden et al. 2007).

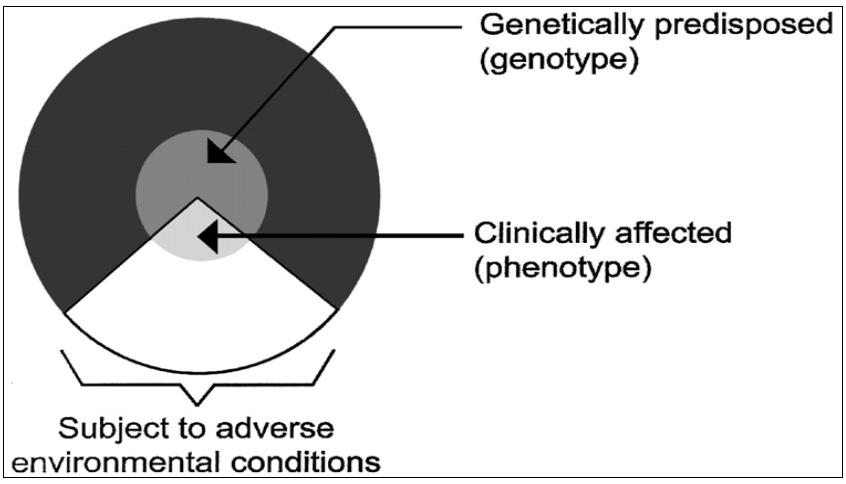


Figure 65: Interaction of genetics and environment (Pi-Sunyer, 2002).

7.1.2. Complications of Obesity

The hazards of obesity were first noted by the ancient Greeks. Hippocrates, the "*father of medicine*" was the first to describe the dangers of excessive food consumption and lack for individual physical activity (Demanrk-Wahnefried et al. 2012). Obesity has diverse consequences: biological (e.g., diabetes, various cancers, cvds), psychological (e.g., depression, low self-esteem) and social (e.g., prejudice, discrimination), and for society (e.g., economic loss) (Wang et al. 2008a; Puhl et al. 2009; Stone et al. 2012). However, obesity is often simply as an issue of body weight that can be dealt with by weight loss while it should be considered as a multi-dimensional disorder (Ogden et al. 2007). Obesity is a factor for numerous chronic diseases such as cardiovascular diseases (i.e., coronary artery disease, stroke), amplifying for diabetes and leading to premature death (Calle et al. 2004; Grundy, 2004; de Koning et al. 2007; Canoy, 2008; Ford et al. 2008; Wolin et al. 2008; Duncan et al. 2010; Malik et al. 2010; Ogden et al. 2012). The mechanisms that link obesity, related diseases and their outcomes, however, are poorly understood (Brown et al. 2000; Onyike et al. 2003; Low et al. 2009; Whitlock et al. 2009; Canoy, 2010; Ogden et al. 2010; Ryan et al. 2011; CDC, 2012; Vucenik et al. 2012):

- **Cardiovascular diseases:** Obesity is a risk factor for cardivoascular diseases such as heart failure, coronary heart disease (CHD), stroke, all accelerated by diabetes and hypertension which are themselves causes by obesity. It is associated with reduced overall survival (Eckel et al. 1998; Van Gaal et al. 2006; Lavie et al. 2008). There is evidence that obesity along with glucose tolerance and hypertension in childhood has a strong connection with increased rates in premature death from cvd, suggesting early origins of obesity related disease (Franks et al. 2010).

- **Type 2 diabetes:** The rapid rise in of type 2 diabetes to a global epidemic has mirrored that of obesity, since the early 70's and is considered to be a direct result of the obesity "*epidemic*" (Calle et al. 2004; WHO, 2013). The term "*diabesity*", proposed by Sims in 1970 reflects the close relation between these two disorders (Marinou et al. 2009). Obesity is currently the most important predictor of diabetes and its association is perhaps stronger than any other comorbidity due to the existence of clearer biological causal pathways (Kahn et al. 2006; Lorenzo et al. 2007). Most type 2 diabetics are overweight (Marinou et al. 2009). Furthermore, more than 197 million individuals worldwide have impaired glucose tolerance due to obesity, and this number is set to rise to 420 million by 2025 (Hossain et al. 2007).
- **Hypertension:** Obesity causes hypertension directly in all populations (Brown et al. 2000; Mertens et al. 2000; Faith et al. 2002; Davy et al. 2004; Lavie et al. 2008). Almost a third of hypertensive patients are obese. Hypertension incidence is five times more common in overweight people than those of normal weight (Mertens et al. 2000; Grundy, 2004; Hossain et al. 2007; Henry et al. 2011). The risk of hypertension increases in line with the BMI and can be observed also in the non-obese range of affected individuals (Davy et al. 2004). Research has demonstrated that the prevalence of hypertension is increasing due to the obesity epidemic and analysis from large cohorts suggest that obesity-related hypertension poses a significant risk to morbidity and mortality (Brown et al. 2000; Flynn et al. 2011; Henry et al. 2011).
- **Depression:** Obesity shares a complicated relationship with depression; it may lead to depression or be one of its consequences, thus causal and reverse causal association co-exist in populations (Onyike et al. 2003; Atlantis et al. 2008; Patten et al. 2009; Duncan et al. 2010; Faith et al. 2011). Multiple meta-analyses of epidemiological studies have highlighted for the co-occurrence of obesity and depression (Onyike et al. 2003; de Wit et al. 2010; Luppino et al. 2010; Faith et al. 2011). Despite this, more research is required to clarify this bi-directional relationship, and it is unknown if gender, age, race and socioeconomic status covary (Onyike et al. 2003; Faith et al. 2011).
- **Cancer:** Obesity has been established as a risk factor for several types of cancer (Adami et al. 2003; Larsson et al. 2007; Renehan et al. 2008; Vucenik et al. 2012). Results from epidemiological and meta-analytical studies since the early 70's demonstrated that obesity contributes to the increase incidence of cancer (Calle et al. 2003b; Calle et al. 2004; Vucenik et al. 2012). Although, the risk for cancer development has not been fully characterized in obese and overweight individuals, obesity has been linked to cancer mortality and evidence continues to suggest

connections with cancers in particular sites such as thyroid, liver and prostate (Calle et al. 2003a; Calle et al. 2003b; Calle et al. 2004; Renehan et al. 2008; Wolin et al. 2008; Duncan et al. 2010; Demark-Wahnefried et al. 2012; Vucenik et al. 2012). Nevertheless, more research is required in order to explore the link between obesity and cancer as well as the role of gender and cancer site in their association (Larsson et al. 2007; Wolin et al. 2008).

- **All-cause mortality:** WHO reports that obesity is the fifth leading risk for global mortality (2.8 million adult deaths each year) (WHO, 2013). Strong evidence indicates that high BMI is related to an increase in overall and cause-specific mortality (Adams et al. 2006; Marinou et al. 2009; Whitlock et al. 2009; Nguyen et al. 2010; Vucenik et al. 2012). Despite recent increases in life expectancy, the global prevalence of obesity may reverse this trend as the mortality rates seem to rise every decade (Franks et al. 2010). In the UK, it was proposed that obesity is responsible for 7.0% of morbidity and mortality with a significant impact on the National Health System (NHS) (Duncan et al. 2010). The number of deaths per year attributable to obesity in UK and in the US were roughly 30,000 and 300,000 respectively indicating that obesity is set to replace smoking as the main preventable cause of illness and premature death (Marinou et al. 2009; Jia et al. 2010). Figure 66 shows the ten most common conditions that have been associated with obesity on death certificates in England (and the Oxford region) in the time period of 1995-2006.

ICD code	Condition	England (N[a])	Oxford (N[b])
Obesity as underlying cause			
428, 150	Heart failure	531	58
415, 126	Pulmonary embolism	469	69
250, E10–E14	Diabetes mellitus	227	21
485, 486, J12–J18	Pneumonia unspecified	221	35
799, R09, R64, R99	Other ill defined and unknown causes of morbidity and mortality	215	16
414, 125	Chronic ischaemic heart disease	196	34
451, 180	Phlebitis and thrombophlebitis	177	21
401–404, I10–I13	Essential hypertension	167	18
490–492, 496, J44	COPD	155	11
429, I51	Complications and ill defined descriptions of heart disease	136	24
Obesity as contributing cause			
414, I25	Chronic ischaemic heart disease	1086	106
410, I21	Acute myocardial infarction	704	146
490–492, 496, J44	COPD	503	43
250, E10–E14	Diabetes mellitus	359	32
485, 486, J12–J18	Pneumonia unspecified	334	31
415, I26	Pulmonary embolism	281	29
451, I80	Phlebitis and thrombophiebitis	279	26
401–404, I10–I13	Hypertensive heart disease	264	30
428, I50	Heart failure	167	19
129, I51	Complications and ill-defined descriptions of heart disease	118	17

a: The total number of deaths in England where obesity was given as the underlying cause was 2152; there was a total of 6364 deaths where obesity was given as a contributing cause
b: The total number of deaths in the Oxford region where obesity was given as the underling cause was 267; there was a total of 760 deaths where obesity was given as the underlying cause

Figure 66: The ten most common related to obesity medical conditions on death certificates in England 1995-2006 and in the Oxford region 1979-2006 (Duncan et al. 2010).

The above complications (and more) have been the focus of a number of research efforts. It is expected to for these diseases to appear in the automatically generated concept map either as

exposures or outcomes when we apply our rule based approach in study design abstracts related to obesity. However, it is not certain how focused on these concepts epidemiological studies related to obesity are.

7.2. Mining obesity-included Epidemiological Literature

We will apply our rule based method in epidemiological literature related to obesity. More specifically, a large scale epidemiological corpus related to obesity will be retrieved from MEDLINE. The extraction and normalization method of key epidemiological characteristics will be performed and the generated results will produce automatically a concept map that will represent the exposures, outcomes and covariates of epidemiological study design abstracts related to obesity for further exploration and learning.

7.2.1. Information Retrieval

Information retrieval is the first step in the identification method of key characteristics in epidemiological studies (see Chapter 3). A corpus of MEDLINE articles was obtained from PubMed through "*obesity/epidemiology*" as a search term, with "*[mesh]*" as a specific descriptor that aimed to incorporate epidemiological related abstracts that mentioned the term obesity.

Articles were restricted to English. The corpus consisted from 23,690 epidemiological MEDLINE abstracts. A number of returned MEDLINE citations (4,502) did not contain any abstract text besides the title and the author list. These citations were excluded, therefore resulting in the total number of 19,188 abstracts.

7.2.2. Information Extraction at Mention and Document Level

The rule based approach for the identification of key characteristics from epidemiological abstract text was applied on the collected corpus of MEDLINE abstracts. It is interesting that from 838 citations no characteristics were identified. In order to evaluate if any key characteristics were present and the rule based approach failed to recognise them, 35 MEDLINE abstracts were randomly selected. It was observed that no important information was mentioned in text for a variety of reasons:

- **Article nature:** these abstracts were articles referring to epidemiological context without mentioning any specific information regarding any of the six characteristics. Therefore, the style and format were not based on epidemiological research.
- **No key epidemiological information:** no concept belonging to any of the six epidemiological characteristics was present. More specifically, there were no spans for study design, population, covariate and effect size. However, any mentions of potential

exposures and outcomes failed to trigger any relevant rules since there were not based on (epidemiological) semantic patterns in text.

From the remaining 18,350 abstracts, the number of relevant articles have been growing per year in the health sciences literature, particularly after the new millennium with 2,382 of those in 2011 (Figure 67).

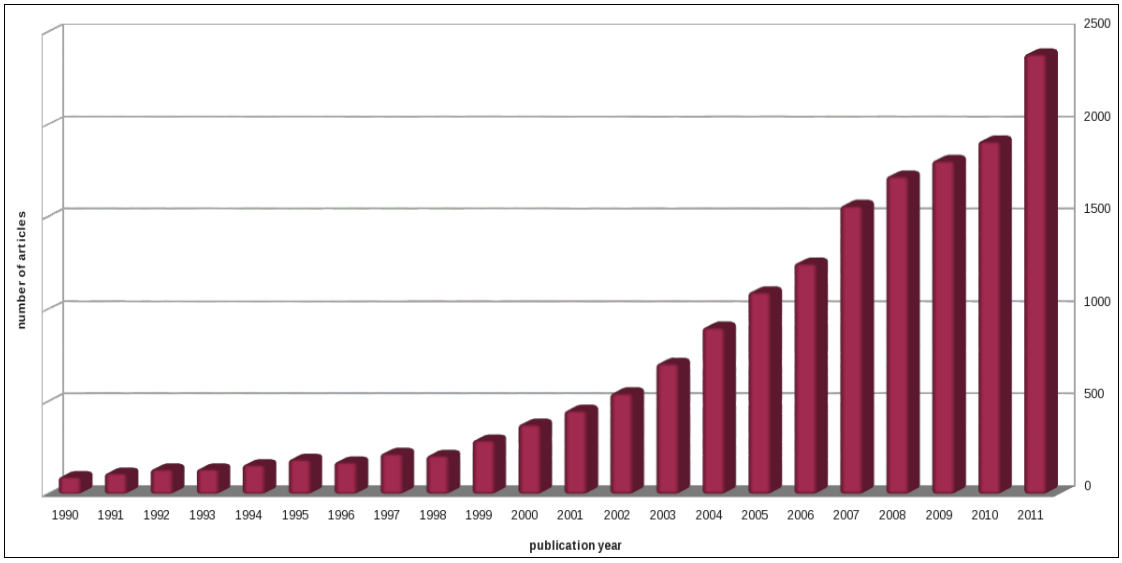


Figure 67: Number of published epidemiological articles related to obesity for the period 1990-2011. Due to the limited number of epidemiological articles for the years 1965-1989, it was decided to show only the numbers for the time period 1990-2011 in order to present clearly the escalating chronically numbers of epidemiological study abstracts related to obesity.

Identified Concepts at Mention Level

The method extracted a total of 140,619 mentions of the targeted epidemiological characteristics at the mention level: 7,697 (5.0%) study design, 33,531 (24.0.%) population, 27,667 (20.0%) exposure, 56,523 (40.0%) outcome, 5,500 (4.0%) covariates and 9,701 (7.0%) effect size concepts respectively (Figure 68). Outcome has the highest number of spans detected in the corpus at the mention level (40.0%), double that of the exposure ones (20.0%).

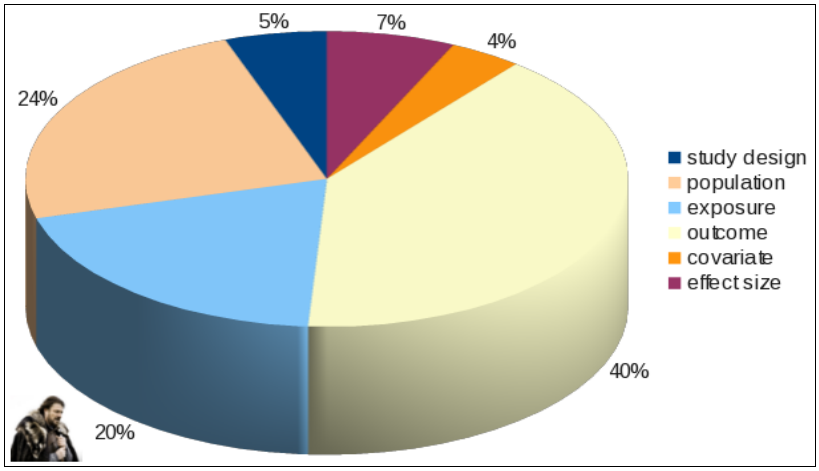


Figure 68: Extraction results at the mention level for each epidemiological characteristic from a corpus of 19,188 MEDLINE abstracts.

Identified Concepts at Document Level

The method extracted a total of 98,649 mentions of the targeted epidemiological characteristics at the document level: 6,060 (6.0%) study design, 13,537 (14.0%) population, 23,518 (24.0%) exposure, 40,333 (41.0%) outcome, 5,500 (6.0%) covariates and 9,701 (10.0%) effect size concepts respectively (Figure 69). Outcome has again the highest number of mentions extracted in the corpus (41.0%), almost double that of the exposure ones (24.0%).

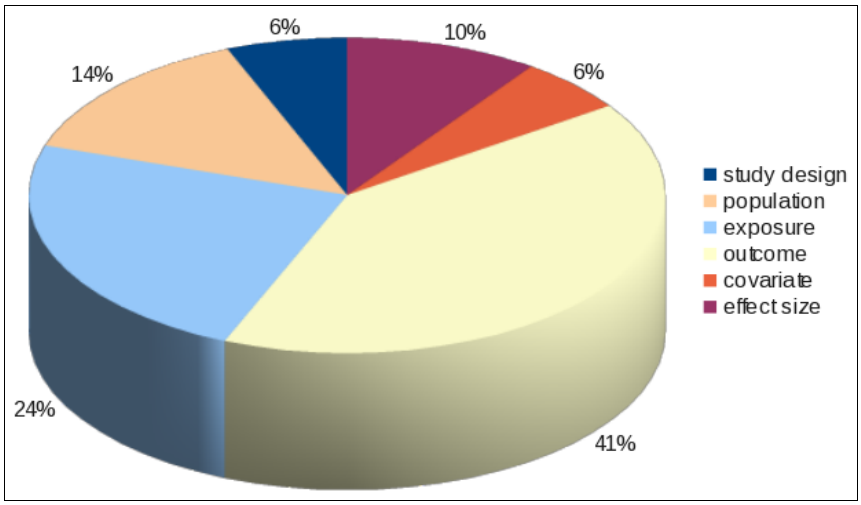


Figure 69: Extraction results at the document level for each epidemiological characteristic from a corpus of 19,188 MEDLINE abstracts.

Table 47 shows in more detail the most frequent study types in obesity epidemiology before mapped to the study design ontology (at the document level).

Table 47: Most frequent study designs extracted from 19,188 epidemiological abstracts at the document level.

epidemiological study design	mentions (%)	epidemiological study design	mentions (%)	epidemiological study design	mentions (%)
cross-sectional study	1,912 (31.5)	nested case control study	31 (0.5)	case crossover study	5 (0.0)
prospective cohort study	586 (9.6)	case cohort study	30 (0.5)	case only study	4 (0.0)
longitudinal study	465 (7.6)	randomised clinical trial	28 (0.4)	self-controlled case series study	3 (0.0)
review	430 (7.0)	multicenter cohort study	27 (0.4)	primary research	3 (0.0)
retrospective cohort study	400 (6.6)	ecological study	24 (0.4)	occupational study	3 (0.0)
population study	345 (5.6)	hospital based study	21 (0.3)	grounded theory study	3 (0.0)
case control study	270 (4.4)	ethnographic study	21 (0.3)	case report study	3 (0.0)
cohort study	229 (3.7)	quantitative study	19 (0.3)	randomised crossover trial	2 (0.0)
observational study	191 (3.0)	randomised double-blinded placebo controlled trial	18 (0.3)	case finding study	2 (0.0)
systematic review	179 (2.9)	serial cross-sectional study	12 (0.1)	randomised non-blinded trial	1 (0.0)
epidemiological study	176 (2.9)	correlational study	12 (0.1)	randomised multi-arm trial	1 (0.0)
birth cohort study	113 (1.8)	cluster randomised trial	12 (0.1)	quantitative descriptive study	1 (0.0)
non-randomised controlled trial	109 (1.7)	quasi-experimental study	11 (0.1)	qualitative evidence synthesis	1 (0.0)
qualitative descriptive study	95 (1.5)	case comparison study	11 (0.1)	phenomenology study	1 (0.0)
literature review	69 (1.1)	experimental study	9 (0.1)	phase 2 trial	1 (0.0)
qualitative study	49 (0.8)	randomised double-blinded trial	7 (0.1)	lifecourse study	1 (0.0)
meta analysis	45 (0.8)	case-base study	6 (0.0)	historical cohort study	1 (0.0)
randomised trial	39 (0.5)	case series study	6 (0.0)	before after study	1 (0.0)

"*Cross-sectional*" was the most conducted epidemiological study design (1,912 abstracts, 31.5%) while experimental types (trials) were expected to be in small numbers (a total of 227 abstracts, 3.7%) since obesity is mostly studied through observational research rather than experimental. "*Prospective cohort study*" is found at the second place with 586 total mentions (9.6%) indicating that most population under epidemiological investigation have to be studied at present in order to potentially comprehend the nature of various exposures that have been linked in obesity and to collect high quality data for further analysis. The cohort type of epidemiological study design was noted to be second if its various subtypes are to be collated (prospective, retrospective, birth, case, historical, occupational, longitudinal, hospital based, etc.) with 1,876 mentions in total (30.9%) (Figure 70).

"*Reviews*" are in the fourth place (430 abstracts, 7%) by analyzing the various components of the complex health care problem that obesity denotes by integrating together various sources of observational research. Additionally, abstracts focusing on research related to obesity reported generic study types for their designs such as "*population/epidemiological*" (521 abstracts, 8.5%). There were 341 (5.6%) total mentions of case control study design and its respective subtypes (Figure 71). Quantitative studies had higher mention number (4,225 abstracts, 69.7%) than qualitative studies (169 abstracts, 2.7%– Figure 72), both belonging in the observational study design branch.

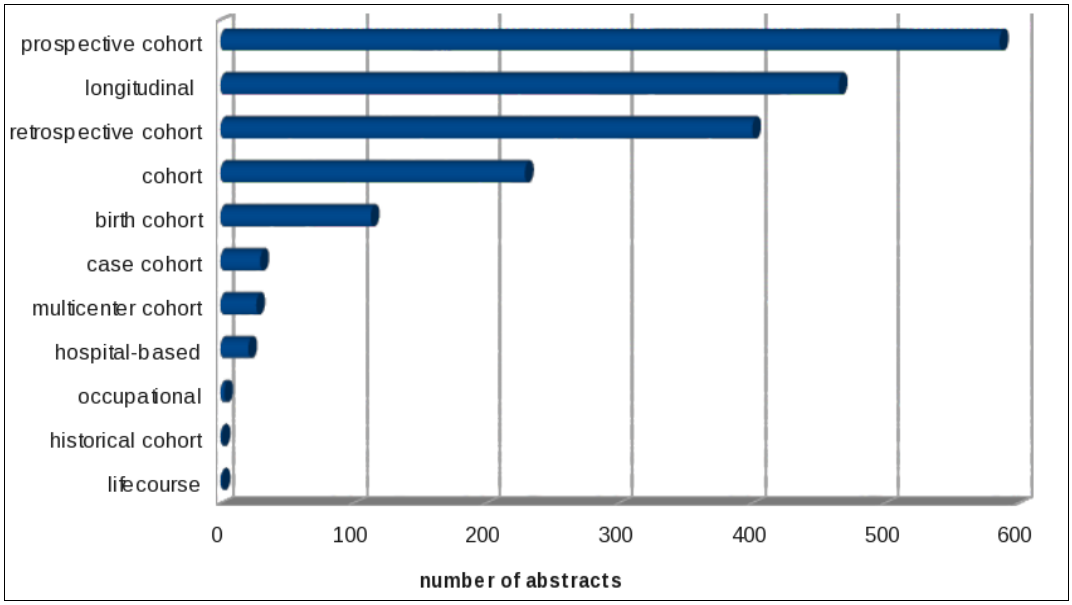


Figure 70: Number of mentions for un-normalized cohort related study designs in a corpus of 19,188 MEDLINE abstracts.

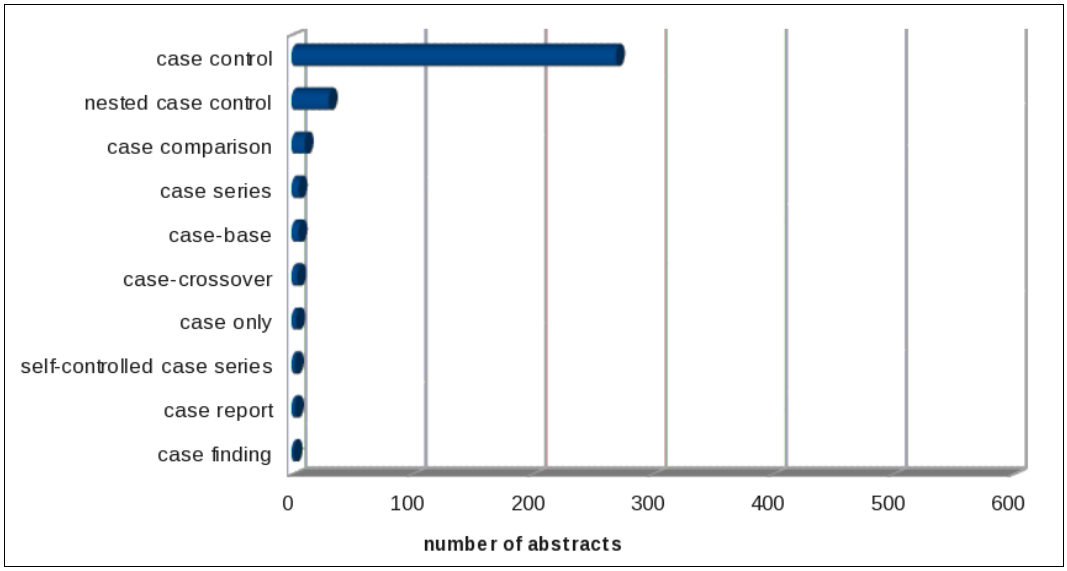


Figure 71: Number of mentions for un-normalized case control related study designs in a corpus of 19,188 MEDLINE abstracts.

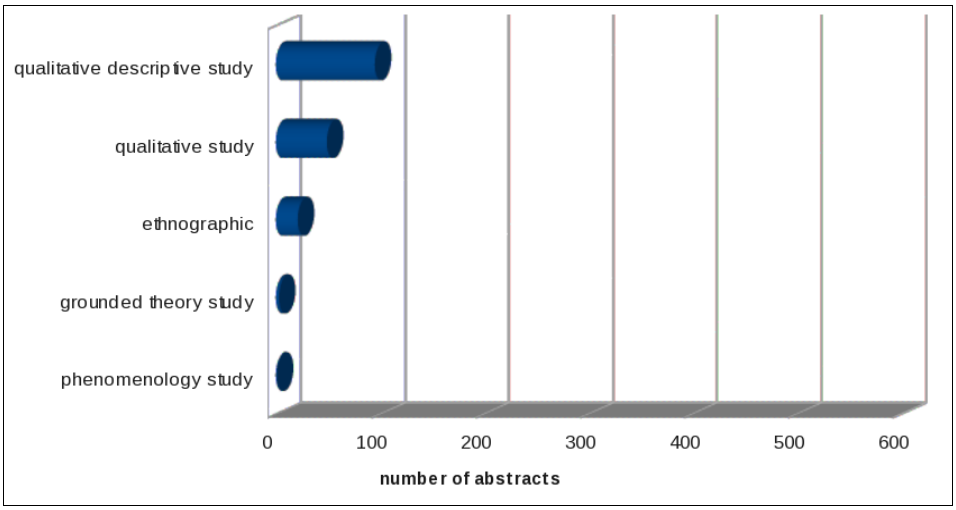


Figure 72: Number of mentions for qualitative related study designs in a corpus of 19,188 MEDLINE abstracts.

Tables 48, 49 and 50 present the top forty most frequent exposure, outcome and covariate concepts at the document level, while figures 73, 74, and 75 reveal the top one hundred most frequent exposures, outcomes and covariates that were extracted from the corpus in the form of word clouds. Obesity was the most mentioned concept in the exposure and outcome characteristics (2,450 and 5,220 respectively). On the other hand, "*obesity*" was spotted in covariates in the 11th place (58 mentions) while the top concept in the covariate group was revealed to be "*age*" (1,066) with almost twice more mentions than the next one ("*gender*", 631). Since obesity has been used as a case study for our methodology, it is understandable why concepts representing it and its subsequent measures e.g., "*body mass index*", "*waist circumference*", "*overweight*", "*obesity*", etc have been extracted from multiple documents either as exposures or outcomes.

Table 48: Top 40 most frequent exposures in 19,188 epidemiological abstracts related to obesity.

most common exposures	frequency
obesity	2,450
body mass index	1,351
overweight	531
age	394
waist circumference	291
physical activity	289
hypertension	256
metabolic syndrome	240
body weight	218
type 2 diabetes	206
gender	193
smoking	186
abdominal obesity	135
insulin resistance	128
mortality	117
adiposity	116
weight gain	108
diet	98
childhood obesity	92
weight loss	89
waist to hip ratio	82
education	79
childhood	79
socioeconomic status	75
ethnicity	75
depression	70
central obesity	69
pregnancy	67
race	66
blood pressure	66
overweight/obesity	59
cvd risk factors	59
height	55
morbidity	54
leptin	52
birth weight	49
asthma	49
bariatric surgery	48
physical inactivity	47
family history	45

Table 49: Top 40 most frequent outcomes in 19,188 epidemiological abstracts related to obesity.

most common outcomes	frequency
obesity	5,220
overweight	2,058
type 2 diabetes	1379
body mass index	1,084
hypertension	728
cardiovascular disease	712
metabolic syndrome	659
mortality	460
insulin resistance	297
childhood obesity	289
coronary heart disease	260
death	250
health	225
waist circumference	211
abdominal obesity	209
smoking	194
physical activity	193
weight gain	181
morbidity	180
cvd risk factors	175
weight	162
adiposity	161
overweight/obesity	155
asthma	127
blood pressure	122
dyslipidemia	116
body weight	110
stroke	101
central obesity	98
depression	95
weight loss	94
underweight	91
chronic diseases	91
hypercholesterolemia	88
cancer	86
survival	85
cardiovascular risk	85
atherosclerosis	81
coronary artery disease	78
inflammation	68

Table 50: Top 40 most frequent covariates in 19,188 epidemiological abstracts related to obesity.

most common covariates	frequency
age	1,066
gender	631
body mass index	346
smoking	260
education	160
race	117
physical activity	108
alcohol consumption	83
ethnicity	70
type 2 diabetes	67
race/ethnicity	60
obesity	58
waist circumference	53
income	43
hypertension	42
socioeconomic status	39
height	36
marital status	33
demographics	32
parity	27
smoking status	25
energy intake	25
lifestyle	22
educational level	20
birth weight	20
weight	17
maternal age	17
family history	17
exercise	16
depression	15
total energy intake	14
region	13
insulin resistance	13
occupation	12
family income	12
blood pressure	12
adiposity	11
social class	10
gestational age	10
area	10

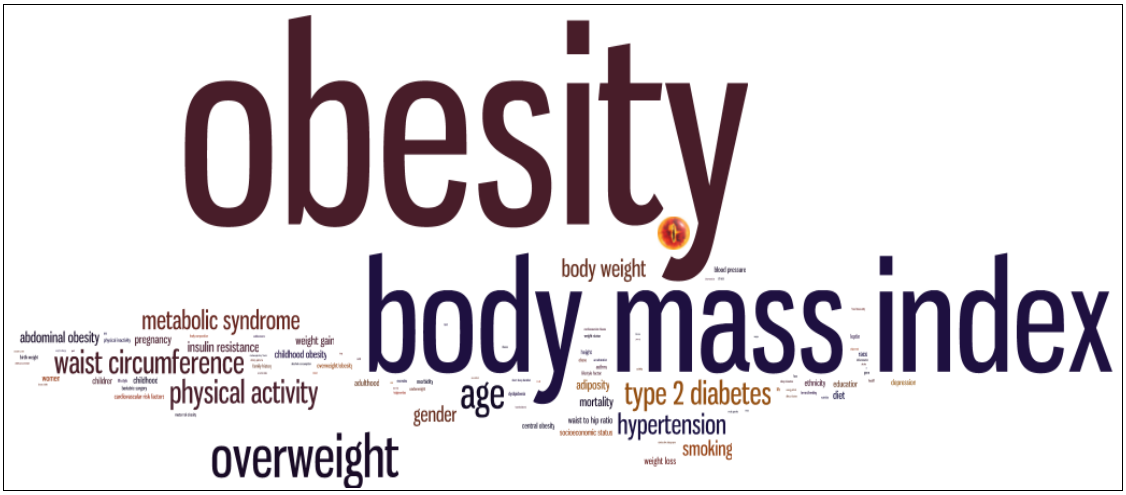


Figure 73: Word cloud for the top 100 most frequent exposures with the highest number of mentions from 19,188 epidemiological abstracts related to obesity.

Figure 74: Word cloud for the top 100 most frequent outcomes with the highest number of mentions from 19,188 epidemiological abstracts related to obesity.

Figure 75: Word cloud for the top 100 most frequent covariates with the highest number of mentions from 19,188 epidemiological abstracts related to obesity.

By removing any concepts that represent directly obesity (e.g., "*obesity*", "*waist circumference*", "*body mass index*") in the characteristics of exposure and outcome, a clearer picture regarding the potentially associated-with-any-way-to-obesity concepts can be revealed as it is shown by figures 76 and 77 and tables 51 and 52. More specifically, for the exposures (Figure 76), it can be observed that most epidemiological studies focused on the research of organism attributes ("age", "height", etc), individual behaviour concepts ("smoking", "physical activity", etc) and various disorders that have been linked to obesity such as diabetes, hypertension and metabolic syndrome. Additionally, the majority of outcomes are disorder-related e.g., "*hypertension*", "*type 2 diabetes*", "*asthma*", "*cardiovascular disease*", "*chronic diseases*", "*stroke*", etc (Figure 77). This further enhances the fact that obesity is a complex disease that shares underlying relationships with various other disorders and conditions and when treated, it should be taken into consideration as a part of a disease cluster rather than an isolated concept.

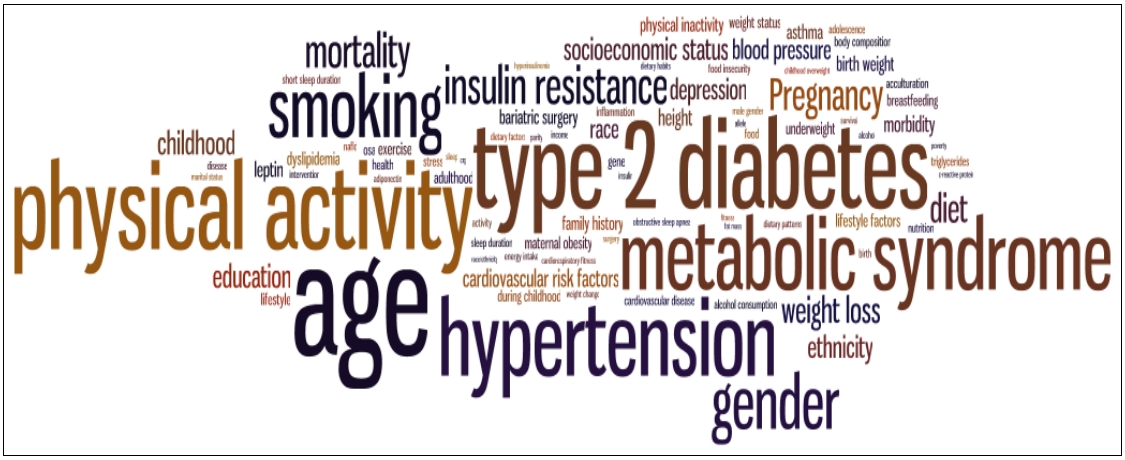


Figure 76: A word cloud for the top 100 most frequent identified exposures after filtering out concepts representing or being directly linked to obesity.

Figure 77: A word cloud for the top 100 most frequent identified outcomes after filtering out concepts representing or being directly linked to obesity.

Table 51: Top forty most frequent exposures after filtering out concepts representing or being directly linked to obesity.

most common exposures	frequency
age	394
type 2 diabetes	295
physical activity	289
hypertension	256
metabolic syndrome	240
smoking	208
gender	183
insulin resistance	128
mortality	117
pregnancy	109
diet	98
weight loss	89
education	79
childhood	79
socioeconomic status	75
ethnicity	75
depression	70
race	66
blood pressure	66
cardiovascular risk factors	59
height	55
morbidity	54
leptin	52
birth weight	49
asthma	49
bariatric surgery	48
physical inactivity	47
family history	45
dyslipidemia	43
lifestyle factors	42
exercise	41
underweight	40
maternal obesity	40
weight status	39
lifestyle	39
adulthood	39
breastfeeding	37
osa	35
stress	33
body composition	32

Table 52: Top forty most frequent outcomes after filtering out concepts representing or being directly linked to obesity.

most common outcomes	frequency
type 2 diabetes	1279
hypertension	770
metabolic syndrome	659
cardiovascular disease	612
mortality	460
insulin resistance	297
coronary heart disease	260
death	250
cardiovascular risk factors	241
blood pressure	190
morbidity	180
chronic diseases	146
asthma	127
dyslipidemia	116
stroke	101
depression	95
weight loss	94
underweight	91
hypercholesterolemia	88
cacncer	86
survival	85
cardiovascular risk	85
disease	84
atherosclerosis	81
coronary artery disease	78
all-cause mortality	75
inflammation	68
osa	65
chronic kidney disease	53
diet	52
nafld	51
heart disease	51
body composition	51
pregnancy	49
physical inactivity	49
triglycerides	48
sleep	47
education	47
myocardial infarction	46
breast cancer	46

The population mentions in epidemiological abstracts at this step are simply identified and due to their variable mentions in text, any meaningful figure could not be produced regarding common attributes often used to describe them such as ethnicity, gender, age and nationality (see Section 7.4.). A similar comment applies to the effect size spans. Due to their numerical nature, a graph representing the identified effect size metrics, the utilized confidence interval, the related concepts and their associated values could not be produced until the application of the normalization process (see Section 7.4.).

7.2.3. Analysis of Extracted Results at Document Level

Considering the promising performance of the system in the evaluation set (see Chapter 4, Section 4.4.1.), a further analysis of the information extraction and normalization results from the large scale corpus related to obesity revealed some interesting issues. Certain characteristics (study design, covariates) had a limited number of concepts identified in the corpus while others contained a much higher number of mentions than others (e.g., outcome with a total of 40,333 concept mentions at the document level). These findings are further discussed and analysed below.

Study Design

Despite having a relatively good recall in the study design characteristic (in both development and evaluation sets, 84.6% and 92.3% respectively), epidemiological study designs were identified only in 33% of the corpus (6,060). In order to evaluate the recall of the system, 50 articles from the rest of the corpus that did not have their study design identified, were randomly selected and upon review (Table 53), four possible explanations were noted:

(a) **No mention of study design**: articles contained epidemiological context, but no specific epidemiological study had been conducted (and thus there was no need to specify study design);
(b) **Limited epidemiological description**: articles contained summarized epidemiological information but without reporting a conducted study and its findings;
(c) **Irrelevant study designs**: studies (e.g., comparative studies, surveys, pilot studies, follow-up studies, reports) that were not targeted for identification;
(d) **False negatives:** study designs that incorrectly were not recognised.

Table 53: Causes for the non-identification of study designs in a random sample from the corpus.

cause	number of abstracts	percentage
no mention of study design	25	50.0%
limited epidemiological information	14	28.0%
undesirable study designs	8	16.0%
false negatives	3	6.0%
total	50	100.0%

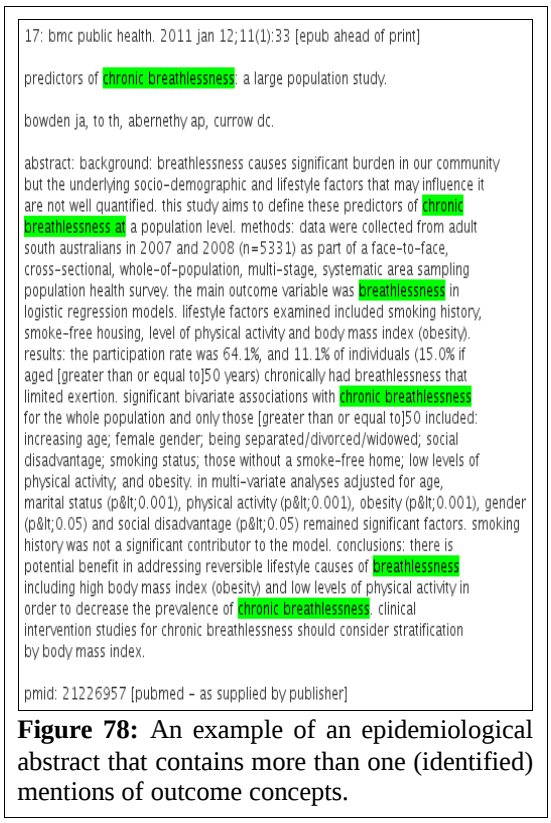
The above conclusions indicate that the system performs relatively well in the identification of study designs in epidemiological abstract text. Analogically if we project Table 53 to the entire corpus, it seems that from 19,188 abstracts, 9,594 articles do not mention their associated epidemiological study design while 5,372 are articles of epidemiological nature. 3,070 articles are probably abstracts that contain irrelevant study designs that currently are not of interest but could be incorporated in the future. Only in 3 abstracts the system failed to identify the related study design span. Although this initially could be attributed to the generic rule designs associated with that particular characteristic, it was observed that designs such as "*review*" and "*meta-analysis*" were considered false negatives because they were not included in the lexical resources as single word concepts (see Section 4.5.1.). This aimed to avoid any potential increase in false positives as certain words (e.g., "*review*", "*meta-analysis*") associated with study designs, are of ambiguous nature and can be used differently in text besides to describe an epidemiology study. This is suggesting that perhaps the system is missing a total of 1,151 study designs (of probably secondary research nature) that are constituted from a single word. Additionally, it is highly likely that the generic rule based approach may have missed certain study mentions due to complex syntactical expressions although the examination of the random sample has not suggested such conclusion.

Outcomes

The number of outcome concepts was almost half of the identified mentions in the whole corpus for all the associated characteristics (41.0%). According to the results in the evaluation set, the precision was 79.3%, suggesting reliable performance in the recognition of outcomes at the document level. However, most of the epidemiological studies:

- ✗ include more than one outcome of interest;
- ✗ have sometimes a concept that is mentioned through a synonymous term or phrase;
- ✗ contain various expressions that trigger the associated rules.

The above remarks indicate that multiple outcomes can be identified in one abstract (including synonyms). Therefore, these observations can potentially explain the relatively large number of outcome mentions in comparison to those of other characteristics. Examples of these cases can be seen in figures 78 and 79.



17: bmc public health. 2011 jan 12;11(1):33 [epub ahead of print]

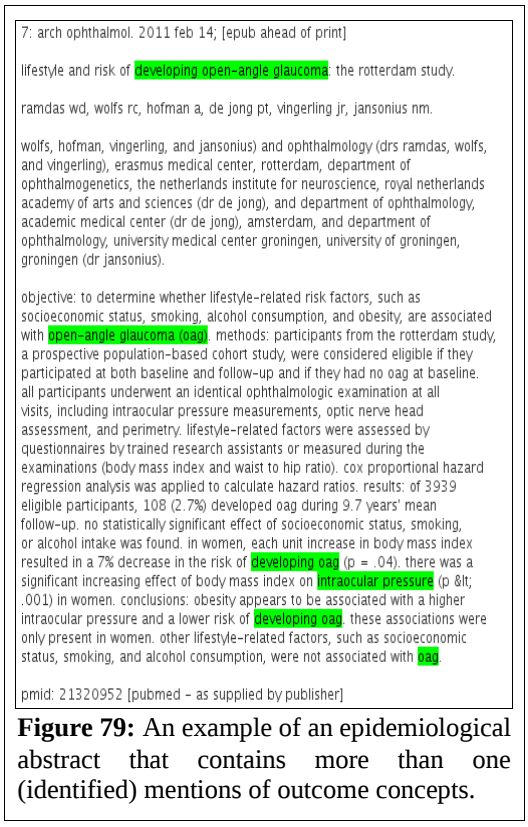
predictors of chronic breathlessness: a large population study.

bowden ja, to th, abernethy ap, currow dc.

abstract: background: breathlessness causes significant burden in our community but the underlying socio-demographic and lifestyle factors that may influence it are not well quantified. this study aims to define these predictors of chronic breathlessness at a population level. methods: data were collected from adult south australians in 2007 and 2008 (n=5331) as part of a face-to-face, cross-sectional, whole-of-population, multi-stage, systematic area sampling population health survey. the main outcome variable was breathlessness in logistic regression models. lifestyle factors examined included smoking history, smoke-free housing, level of physical activity and body mass index (obesity). results: the participation rate was 64.1%, and 11.1% of individuals (15.0% if aged [greater than or equal to]50 years) chronically had breathlessness that limited exertion. significant bivariate associations with chronic breathlessness for the whole population and only those [greater than or equal to]50 included: increasing age; female gender; being separated/divorced/widowed; social disadvantage; smoking status; those without a smoke-free home; low levels of physical activity; and obesity. in multi-variate analyses adjusted for age, marital status (p<0.001), physical activity (p<0.001), obesity (p<0.001), gender (p<0.05) and social disadvantage (p<0.05) remained significant factors. smoking history was not a significant contributor to the model. conclusions: there is potential benefit in addressing reversible lifestyle causes of breathlessness including high body mass index (obesity) and low levels of physical activity in order to decrease the prevalence of chronic breathlessness. clinical intervention studies for chronic breathlessness should consider stratification by body mass index.

pmid: 21226957 [pubmed - as supplied by publisher]

Figure 78: An example of an epidemiological abstract that contains more than one (identified) mentions of outcome concepts.



7: arch ophthalmol. 2011 feb 14; [epub ahead of print]

lifestyle and risk of developing open-angle glaucoma: the rotterdam study.

ramdas wd, wolfs rc, hofman a, de jong pt, vingerling jr, jansonius nm.

wolfs, hofman, vingerling, and jansonius) and ophthalmology (drs ramdas, wolfs, and vingerling), erasmus medical center, rotterdam, department of ophthalmogenetics, the netherlands institute for neuroscience, royal netherlands academy of arts and sciences (dr de jong), and department of ophthalmology, academic medical center (dr de jong), amsterdam, and department of ophthalmology, university medical center groningen, university of groningen, groningen (dr jansonius).

objective: to determine whether lifestyle-related risk factors, such as socioeconomic status, smoking, alcohol consumption, and obesity, are associated with open-angle glaucoma (oag). methods: participants from the rotterdam study, a prospective population-based cohort study, were considered eligible if they participated at both baseline and follow-up and if they had no oag at baseline. all participants underwent an identical ophthalmologic examination at all visits, including intraocular pressure measurements, optic nerve head assessment, and perimetry. lifestyle-related factors were assessed by questionnaires by trained research assistants or measured during the examinations (body mass index and waist to hip ratio). cox proportional hazard regression analysis was applied to calculate hazard ratios. results: of 3939 eligible participants, 108 (2.7%) developed oag during 9.7 years' mean follow-up. no statistically significant effect of socioeconomic status, smoking, or alcohol intake was found. in women, each unit increase in body mass index resulted in a 7% decrease in the risk of developing oag (p = .04). there was a significant increasing effect of body mass index on intraocular pressure (p < .001) in women. conclusions: obesity appears to be associated with a higher intraocular pressure and a lower risk of developing oag. these associations were only present in women. other lifestyle-related factors, such as socioeconomic status, smoking, and alcohol consumption, were not associated with oag.

pmid: 21320952 [pubmed - as supplied by publisher]

Figure 79: An example of an epidemiological abstract that contains more than one (identified) mentions of outcome concepts.

Covariates

From 19,188 abstracts, only 5,500 confounding factors were recognised. A random sample of 50 abstracts in which no covariate concept was identified, was investigated and no abstracts contained any covariates. Most abstracts neglect to report any potential covariates of interest or they mentioned generic patterns (e.g., "*after adjustement for confounding factors*", "*after controlling for covariates*") without specifying the respective concepts. This resulted in the system's inability to detect any potential concepts that could be linked to the covariate characteristic. This probably is occurring because the text is a summarization of an epidemiological study and it focuses in the presentation of the implemented method and generated deductions. It is most likely that in full text, covariates are well defined with more detail. This conclusion could indicate the importance of covariate concepts being included in an epidemiological study abstract. Covariate is a key characteristic in a study design that enables clinical professionals to understand the various concepts that left unexamined and can/could potentially influence the outcome of a particular (epidemiological) hypothesis.

7.2.4. Normalization Results at Document Level

After the identification of 98,649 mentions of key characteristics at the document level from a large scale epidemiological abstract corpus related to obesity, the normalization method was applied. The returned results are presented below for each key characteristic.

Study Design

At the document level, the 6,060 identified study designs were normalized through the extended branch of the OCRe ontology. The results can be seen in Figure 78 below. The most common study design is the observational sub-type "*cross-sectional*" study (1,940 mentions, 32% of total study design mentions) with "*cohort*" and "*review*" being the second and third most commonly mentioned types respectively (1,876, 31% and 678, 11% respectively). Obesity is not a rare disease and therefore, it seems appropriate that most epidemiological research is following the cross-sectional design due to its relatively low cost and its applicability in large scale data. Since the concept of obesity has been associated with a number of diseases, the observation of a population sample in a defined time point could reveal co-existing morbidities and their respective prevalence. "*Randomised*" and "*non-randomised*" trials are in 7th and 8th place with 109 (1.7%) mentions each. It should be noted that any randomized trial mentions are normalized into the OCRe node of "*non-randomized controlled*", suggesting that no trials with attributes besides "*non-controlled*" were identified from the corpus (see Section 5.6.1.). The study designs with the fewest mentions are "*phenomenology*" and "*quantitative descriptive*" with only one mention each. This is not surprising as obesity is considered a complex disease with multiple potential risk factors and health consequences. Therefore conducting epidemiological designs that:

1. may study obesity in its own right (phenomenology) rather than determine its possible causes;
2. aim to discover the nature of obesity and its involved concepts by discovering relationships through quantitative information (quantitative descriptive);

might not produce any intriguing results.

Overall, the observational study type is the most prevalent study design related to obesity accounting for 83.0% of the total number of identified designs. Its number (4,586, 75.6%), is almost 20 times more than experimental study types (227, 3.7%), whereas among the secondary research, the number of reviews is 14 times bigger than the total of meta-analysis studies (678 – 11.1% and 46 – 0.7% mentions respectively) (Figure 81). Additionally, due to the nature of the related exposures and outcomes, the chances of conducting experimental research are limited since any trials can violate ethical boundaries or can be expensive. In a

lower level of the epidemiological study design hierarchy, from the identified observational studies, 218 (5.0%) mentions were qualitative study designs while 4,244 (70.0%) mentions belonged to the quantitative cluster (Figure 82).

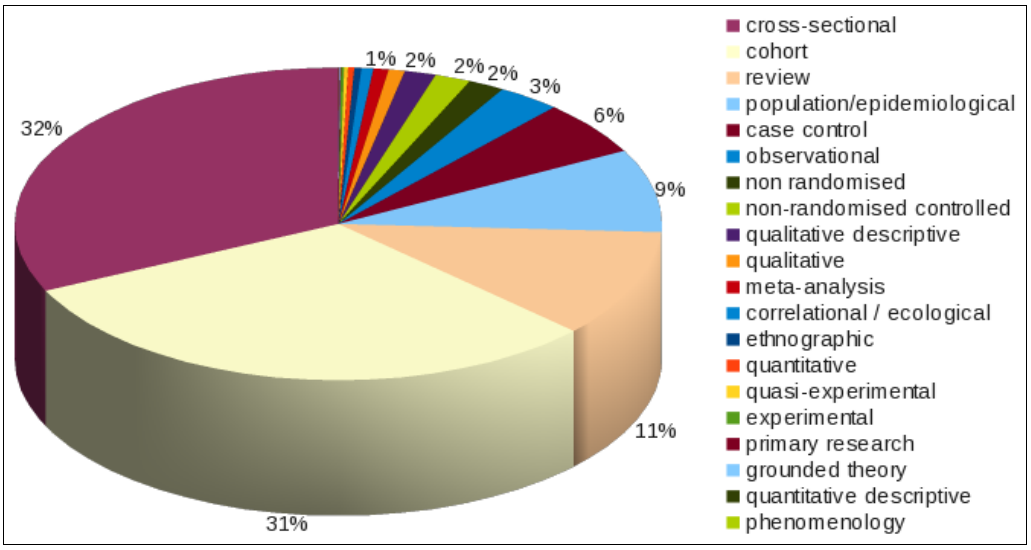


Figure 80: Distribution of the normalized study designs at the document level from the epidemiological corpus related to obesity of 19,188 MEDLINE abstracts.

This reveals that the vast majority of the epidemiological observational studies related to obesity are of quantitative nature. Obesity has a large variety of exposures and outcomes and their influence and generation respectively are noticeable through time rather than instantly. Hence the prevalence of this study design cluster was not unexpected since it focuses to a group of people (cases) observed through time in order to reach specific conclusions regarding potential obesity variables.

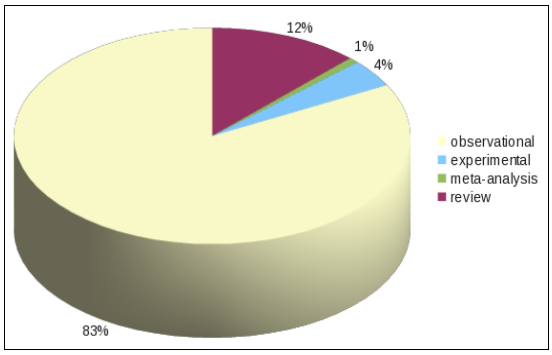


Figure 81: Distribution of the main four epidemiological study designs at the document level

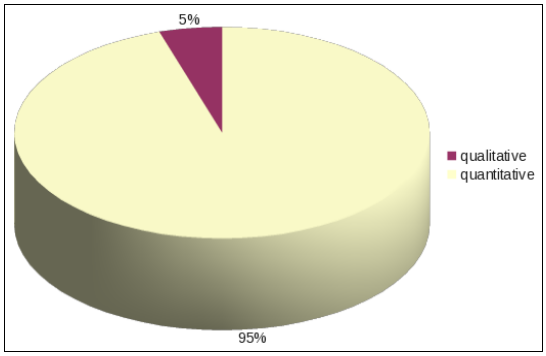


Figure 82: Distribution of the two observational study designs; qualitative and quantitative at the document level.

In the lowest level of the normalized study design hierarchy (observational study > quantitative study, qualitative study), "*cross-sectional*" is the prevalent type of conducted study (44.0%) while "*cohort*" and "*case control*" follow with 42.0% and 8.0% respectively. Qualitative lower level study designs have few mentions (less than 2.0%) e.g., "*grounded theory*", "*quantitative descriptive*" and "*phenomenology*" while from experimental study designs, only "*non-randomised controlled*" was detected with 2.0%. Figure 83 displays the various epidemiological study designs belonging to the quantitative type.

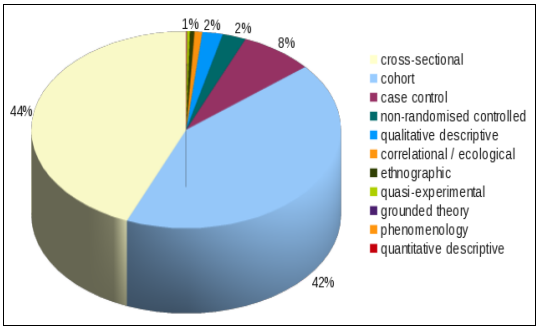


Figure 83: Distribution of the normalized quantitative study designs at the lowest level of the epidemiological study design ontology at the document level.

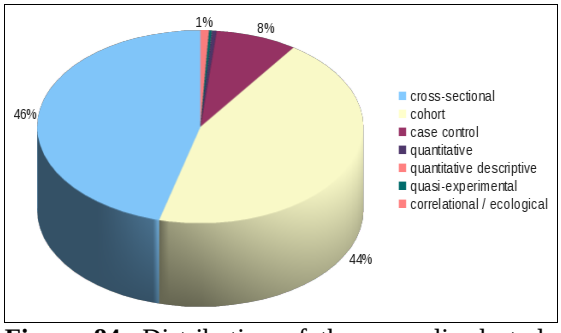


Figure 84: Distribution of the normalized study designs at the lowest level of the epidemiological study design ontology at the document level.

Population

From the 13,537 normalized population mentions at the document level, 5,521 (40.7%) have been normalized for their age, 5,309 (39.2%) for their gender, 2,137 (15.7%) for their nationality and 689 (5.0%) for their ethnicity. A total of 377 (2.7%) abstract populations had both their ethnicity and nationality recognised. Table 54 shows these results in more detail.

Table 54: Number of population spans normalized for attributes at the document level. "-" suggests that there is no distinct class assigned in the attribute but a large range of potential values.

attribute	value	number of spans	(%)
age	juvenile	3,532	26.0
	early adulthood	161	1.1
	middle adulthood	174	1.2
	late adulthood	129	0.9
	adulthood	1,497	11.0
gender	male	1,386	10.2
	female	2,436	17.9
	male/female	1,484	10.9
	unknown	8,231	60.8
nationality	-	2,137 (107 distinct nationalities)	15.7
ethnicity	-	689 (14 distinct ethnicities)	5.0
nationality and ethnicity	-	377	2.7

Age: A total of 5,493 (40.0%) population mentions were normalized for the attribute of age. More than half (65.0%) were referring to juveniles (from infancy to adolescence, Figure 85). Adulthood's three stages had few normalized mentions in the identified population spans with middle adulthood being the most prevalent with 3.0% (174) followed by the early and late stages (3.0% - 161, 2.0% - 129 respectively), although there was no major difference between their mentions. It was observed that adulthood as a generic age – containing participants with age ranging within all three adulthood phases – was the second most common class 27.0% (1,497) suggesting that most population samples include patients from all ages .

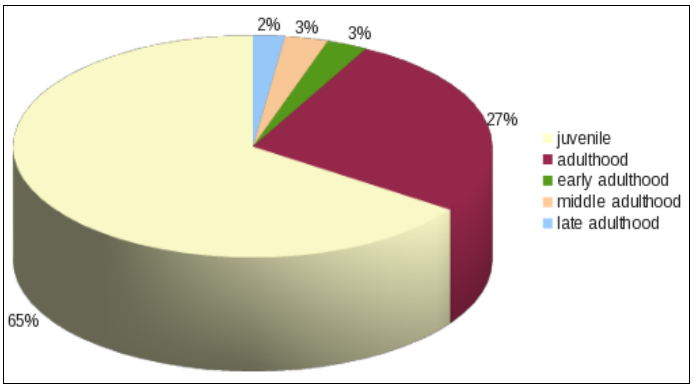


Figure 85: Distribution of population age at document level.

Gender: Figure 86 shows that only 40.0% of studies revealed the gender of the studied subject group. It could be assumed that study abstracts neglecting to report the gender of the participant sample will likely contain both female and male subjects. The number of studies performed on female populations (18.0%) was observed to be almost double than those focusing on male subjects alone (10.0%), while the number of studies that explicitly mentioned both sexes was slightly above than that (11.0%).

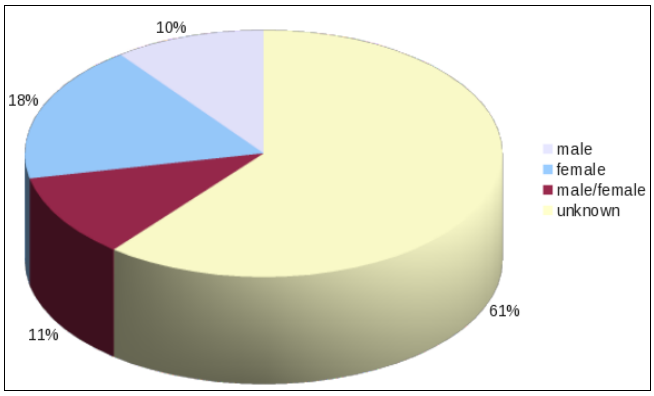


Figure 86: Distribution of male/female population at the document level.

Nationality: 107 nationalities in total were recognised in the obesity related epidemiological corpus. Europe had the largest number of studied nationalities (37), followed by Asia (28) and Africa (18) (Figure 87). Oceania had population samples from only 6 countries indicating it as the continent with the least studied populations. Figure 88 reveals the top twenty nationalities of the normalized identified populations in epidemiological study abstracts related to obesity. Most studies were focused on subject samples of Chinese nationality (229 mentions). Europe was represented with 10 countries (Danish, Dutch, Finnish, French, German, Greek, Italian, Spanish, Swedish, Turkish,) the Americas with 4 (American, Brazilian, Canadian and Mexican), 5 from Asia (Chinese, Japanese, Indian, Iranian, Korean), and only 1 from Oceania (Australian). Surprisingly, despite high rates of obesity in the UK, British populations were identified in 36 epidemiological studies. Figure 89 shows the top 10 countries that were identified in the corpus in a world map.

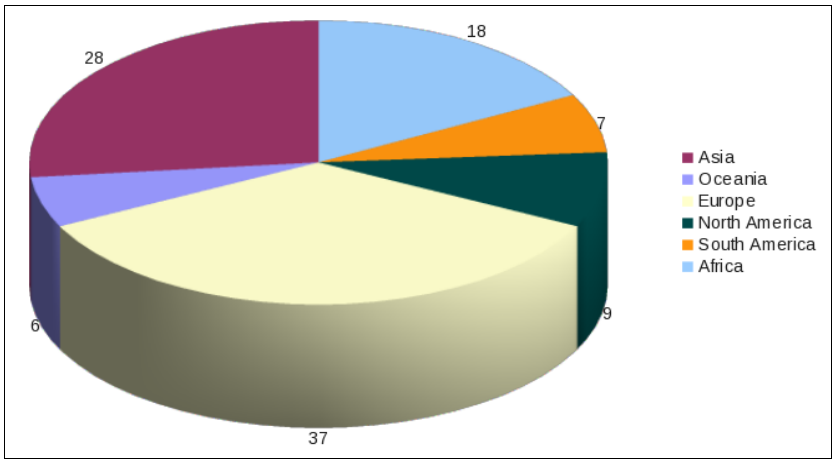


Figure 87: Number of nationalities in epidemiological abstracts related to obesity, grouped by five continents.

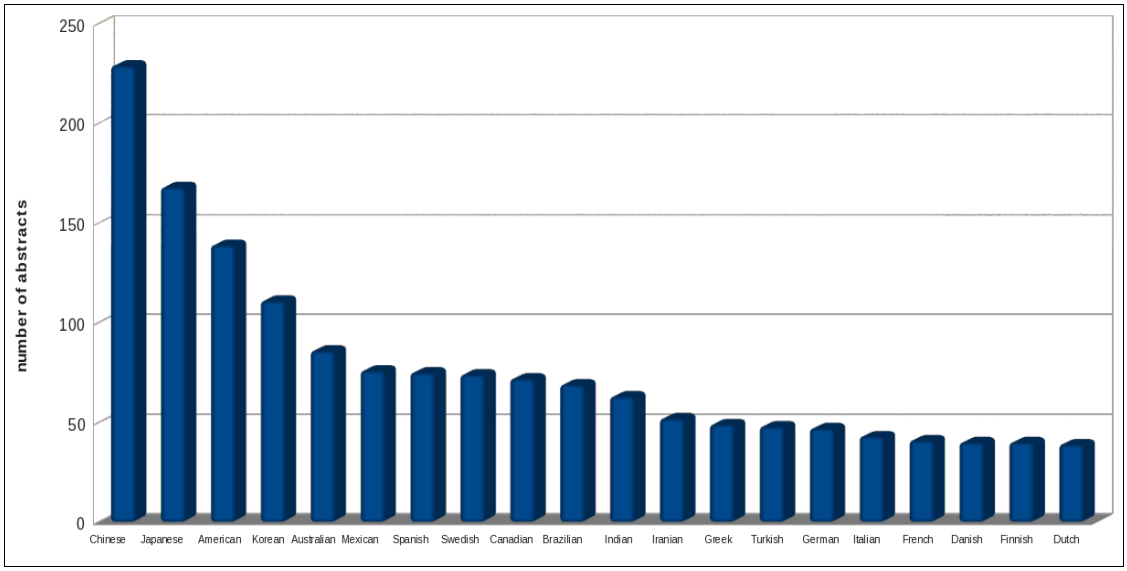


Figure 88: Top 20 identified nationalities from normalized populations at the document level.

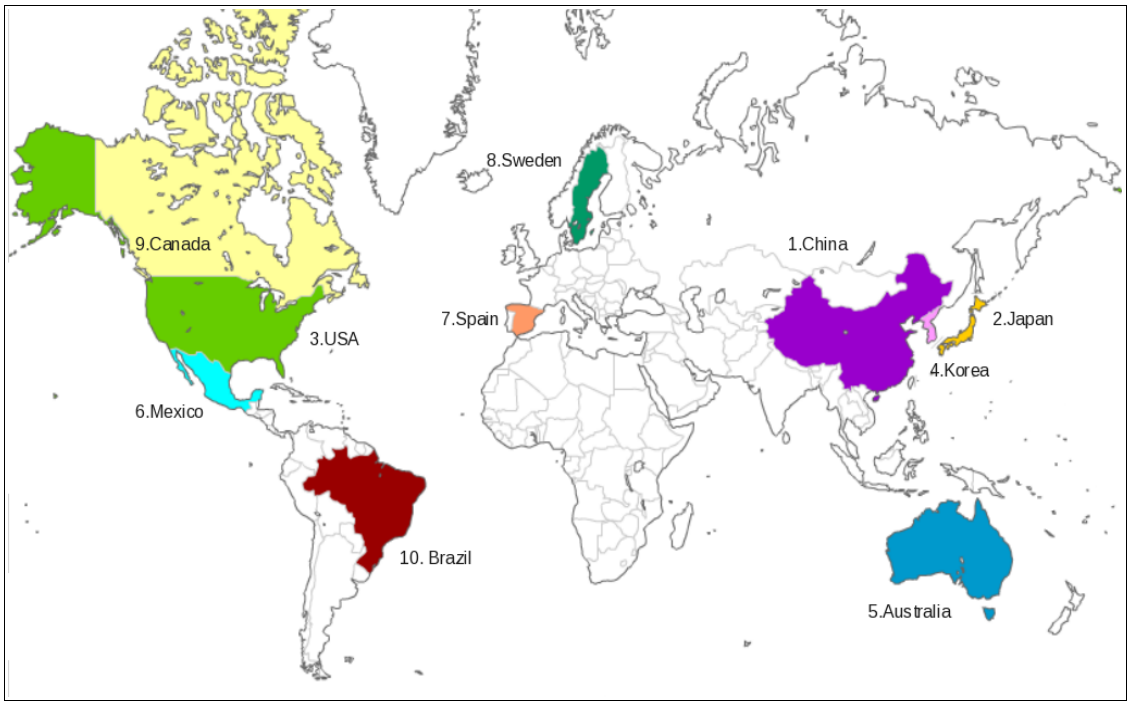


Figure 89: Top ten nationalities from normalized population mentions at the document level (generated through P&P World Map tool [http://english.freeman.jp]).

Ethnicity: From the entire corpus, only a total of 11 different ethnicities were identified (see Figure 90). Chinese was the primary ethnicity (a total of 226 mentions at the document level), while Native American and Hawaiian were the least prevalent ones with three and two mentions respectively. The top five ethnicities were identified in more than 50 abstracts while the rest showcased low rates with a frequency below 10 articles. Nevertheless, these results though should be taken with caution due to low rate of mentions of ethnicity in the corpus. 377 documents had both the ethnicity and nationality of their population identified. In these cases the ethnicity of the studied population:

- was a part of the respective nationality, e.g., "73 *African American men*" with "*African American*" identified as the ethnicity and "*American*" as the nationality;
- or was considered the same with the respective nationality from the applied lexical resources e.g., "*200 Chinese children*" where "*Chinese*" is a part in both the related ethnicity and nationality dictionaries.

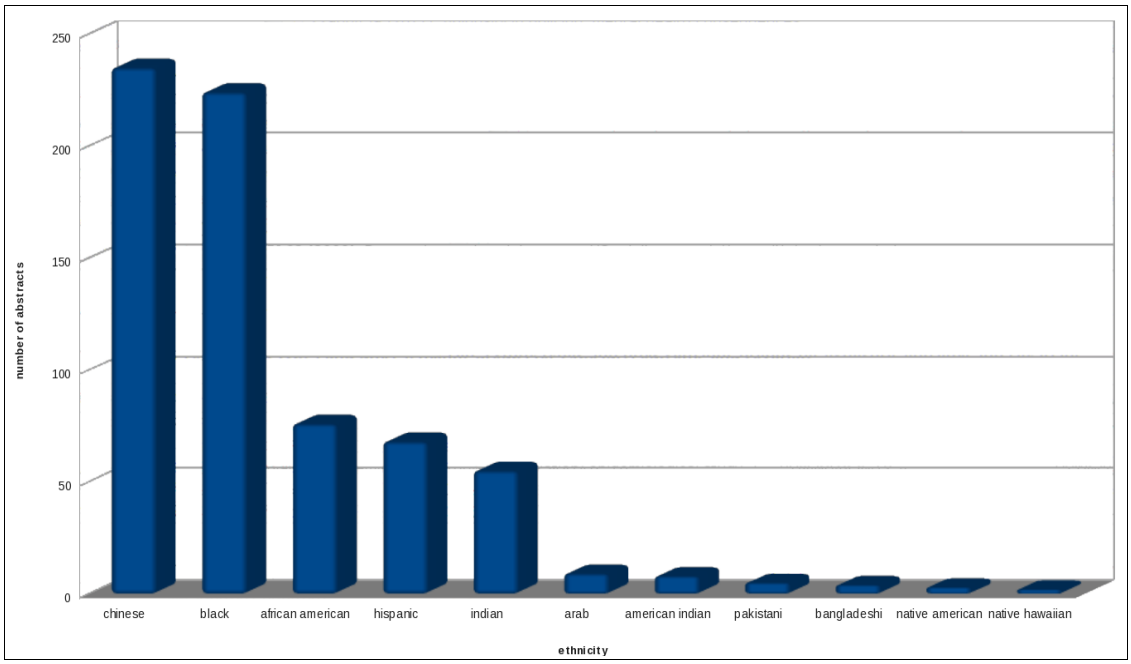


Figure 90: The eleven recognized ethnicities from the normalized populations at the document level.

Exposures

The most prevalent UMLS semantic group is "*disorders*" (8,700, 36.9%), followed by "*concepts/ideas*" (4,635, 19.7%) (Table 55). In only 16 cases, we were not able to classify exposure mentions. The top six most frequent semantic groups have more than 1,000 exposure mentions whereas the bottom four have below 100. In 23,518 exposure mentions identified and normalized in epidemiological abstract text, 7,072 (30.0%) are unique, suggesting that each concept is being repeated almost in 3.3 documents in the corpus. It was observed that the highest number of unique concepts belong to the group of "*concepts/ideas*" (2,257, 31.9%) with "*disorders*" ranking in second place (1,620, 22.9%). The smaller the frequency of exposure mentions in text, the higher the number of unique concepts in certain UMLS groups e.g., "*organizations*", "*devices*", "*geographic areas*", "*phenomena*", etc.

Figure 91 shows that among the ten UMLS semantic categories, four are part of the "*concepts*/*ideas*" group (functional concept, idea or concept, qualitative concept, quantitative concept), three are part of the "*disorders*" group (disorder or syndrome, finding, sign or symptom), two belong to "*physiology*" group (clinical attribute, organism attribute), one to "*procedure*" (therapeutic or preventive procedure). "*Disease or syndrome*" is the prevalent UMLS semantic category for the classification of exposure mentions with more than double the concept numbers from the second one ("*finding*").

Table 55: Number of (unique) exposures for each UMLS semantic group.

UMLS group	number of concepts	%	number of unique concepts	%
disorders	8,700	36.9	1,620	22.9
concepts/ideas	4,635	19.7	2,257	31.9
physiology	3,969	16.8	615	8.6
procedures	1,611	6.8	716	10.1
activities/behaviors	1,285	5.4	351	4.9
living beings	1,030	4.3	355	5.0
chemicals/drugs	857	3.6	410	5.7
objects	368	1.5	143	2.0
genes/molecular	344	1.4	182	2.5
anatomy	252	1.0	144	2.0
phenomena	180	0.7	109	1.5
geographic areas	145	0.6	79	1.1
occupations	73	0.3	39	0.5
devices	30	0.01	25	0.3
organizations	21	0.0	20	0.2
other	16	0.0	7	0.0
total number	23,518	100.0	7,072	100.0

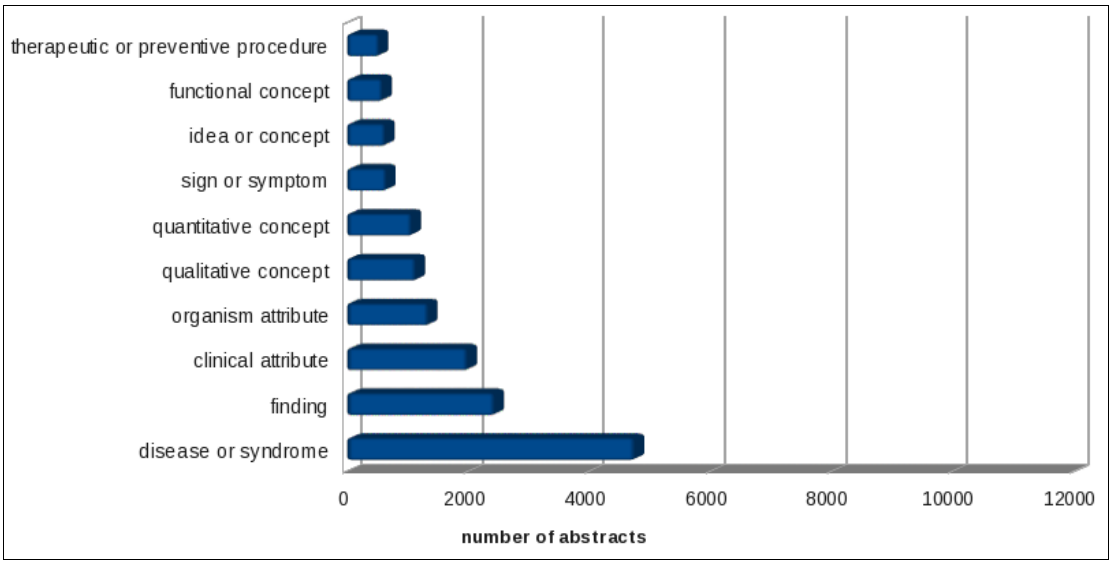


Figure 91: Top ten UMLS semantic categories of normalized exposures.

Outcomes

"*Disorders*" is the semantic group with the largest number of outcome mentions (21,809, 54.0%). The number of disorder outcomes is almost 3 times more that one of the second group, "*concepts/ideas*" (7,277, 18.0%) (Table 56). This is not surprising as the majority of the most frequent concepts were disease related (see Table 49 and Figure 74). "*Devices*" and concept

mentions that are unclassified under the UMLS mapping procedure have the lowest number of outcome concept mentions with a frequency of 28 and 16 respectively. In a similar pattern observed in the exposure characteristic, the six semantic groups have more than 1,000 outcome mentions whereas the bottom three have below 100 ("*organizations*", "*devices*", "*other*").

In 40,333 total outcome mentions, only 9,301 (23.0%) are unique, indicating that each concept is on average, observed in 4.3 documents in the corpus. Outcome concepts are repeated almost double than the exposure ones. The highest number of unique concepts belong to the group of "*concepts*/*ideas*" (2,831, 30.4%) with "*disorders*" ranking again in second place (2,768, 29.7%). Each concept clustered under the disorder group probably appears more frequently in the corpus as an outcome rather than being examined as an exposure with each one being repeated relatively in 8 articles. Almost in every UMLS group, the number of unique concepts is at least half of the total mentions group, suggesting that outcome mentions (no matter what their nature) are reported in text at least twice (with the exception of those placed under "*genes*/*molecular*", "*phenomena*", "*organization*" and "*devices*" groups). This indicates that most epidemiological studies include a single exposure concept and multiple outcomes.

In the top ten UMLS semantic categories with the highest document frequency (Figure 92), the same categories observed in the exposure characteristic can be seen here with only one difference: the replacement of the "*therapeutic or preventive procedure*" category with the "*pathology function*" from the "*physiology*" group. The top three categories belonged to the disorder group suggesting a relationship between outcomes and disease concepts. Additionally, "*disease or syndrome*" appear almost three times the number of concept that the next category did ("*finding*").

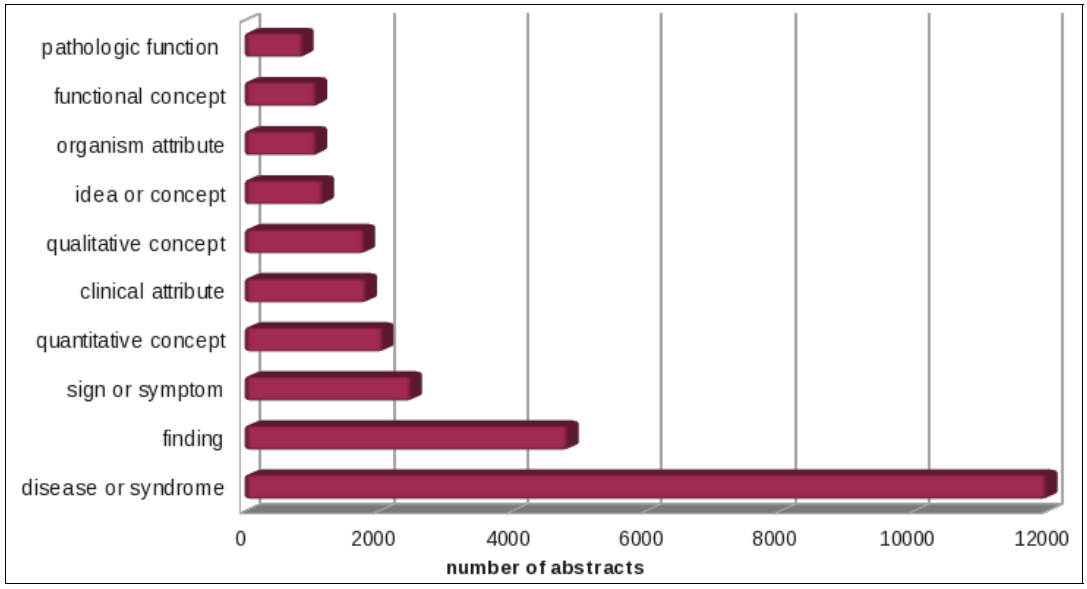


Figure 92: Top ten UMLS semantic categories of normalized outcomes.

Table 56: Number of (unique) outcomes for each UMLS semantic group.

UMLS group	number of concepts	%	number of unique concepts	%
disorders	21,809	54.0	2,768	29.7
concepts/ideas	7,277	18.0	2,831	30.4
physiology	3,810	9.4	773	8.3
procedures	1,697	4.2	678	7.2
living beings	1,616	4.0	540	5.8
activities/behaviors	1,413	3.5	442	4.7
chemicals/drugs	990	2.4	425	4.5
anatomy	577	1.4	283	3.0
objects	314	0.7	110	1.1
genes/molecular	265	0.6	146	1.5
phenomena	250	0.6	131	1.4
geographic areas	137	0.3	63	0.6
occupations	102	0.2	51	0.5
organizations	36	0.0	33	0.3
devices	28	0.0	20	0.2
other	16	0.0	7	0.0
total number	40,337	100.0	9,301	100.0

Covariates

From a total of 5,500 normalized covariates, "*physiology*" is the group with the most mentions (2,381, 43.2%); "*concepts/ideas*" and "*disorders*" are following with 1,044 (18.9%) and 783 (14.2%) mentions respectively. "*Organizations*" and "*devices*" are the groups with the fewest mention numbers (4 and 1 respectively) while the bottom three ones have less than 10 mentions overall. 1,234 (22.4%) out of 5,500 mentions are unique concepts with an overall average repetition in 4.4 documents in the entire corpus. There were 138 unique "*physiology*" concepts indicating that each one could be presented in 17 corpus documents, more than any other concept belonging to any other semantic group. Additionally, in every semantic group, it is observed that each distinct concept is repeated in at least 2 documents, with the exceptions of those belonging to the groups of "*objects*", "*phenomena*","*genes/molecular*", "*anatomy*", "*organizations*" and "*devices*". The characteristic of covariate has the smallest number of mentions in contrast to that one of the exposure and outcome. This suggests that most epidemiological studies do not mention any specific covariates in contrast to a single exposure and multiple outcomes in one abstract.

Table 57: Number of (unique) covariates for each UMLS semantic group.

UMLS group	number of concepts	%	number of unique concepts	%
physiology	2,381	43.2	138	11.1
concepts/ideas	1,044	18.9	454	36.7
disorders	783	14.2	249	20.1
activities/behaviors	591	10.7	84	6.8
living beings	232	4.2	67	5.4
procedures	184	3.3	84	6.8
chemicals/drugs	112	2.0	64	5.1
geographic areas	41	0.7	18	1.4
occupations	34	0.6	8	6.4
objects	29	0.5	18	1.4
phenomena	26	0.4	18	1.4
genes/molecular	17	0.3	13	1.0
anatomy	17	0.3	13	1.0
other	4	0.0	2	0.1
organizations	4	0.0	3	0.2
devices	1	0.0	1	0.0
total number	5,500	100.0	1,234	100.0

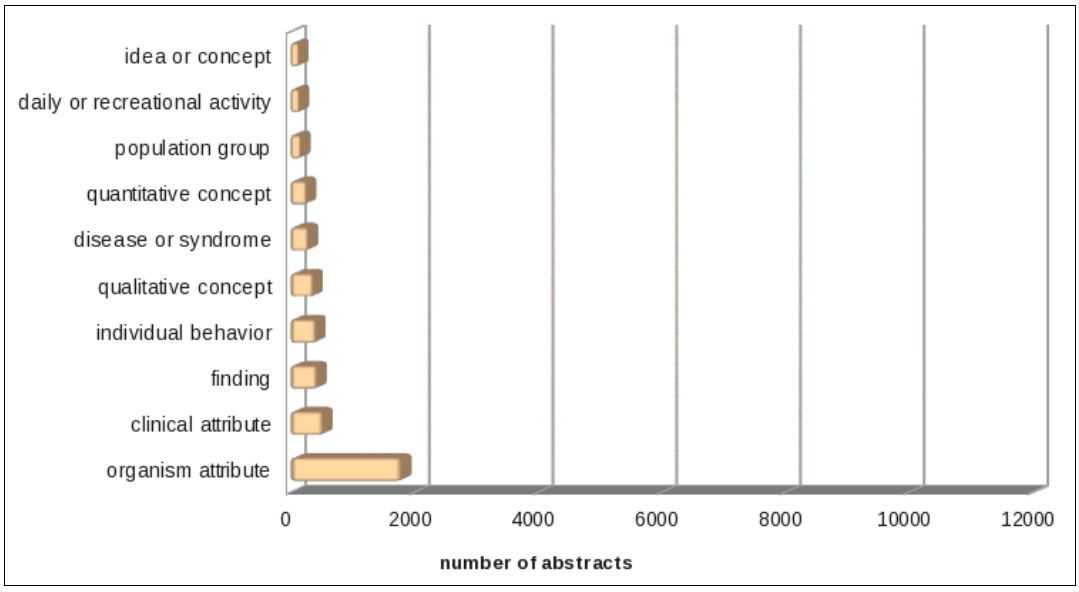


Figure 93: Top ten UMLS semantic categories of normalized covariates.

In Figure 93, the UMLS semantic category "*organism attribute*" had the highest number of covariates with "*clinical attribute*" following with almost 1/3 mentions. With the addition of "*daily recreational activity*", "*population group*" and "*individual behaviour*", the rest of the categories are the same observed in the exposure and outcome ("*idea or concept*", "*quantitative concept*", "*qualitative concept*", "*clinical attribute*", "*organism attribute*", "*functional concept*", "*disease or syndrome*") .

Effect size

From a total of 9,701 normalized effect size mentions, more than half (6,421, 66.1%) report confidence interval (Table 58). Half of the mentions (5,261, 54.2%) contain a related concept with the effect size while half of the effect size spans (5,474, 56.4%) have been normalized to their type. The most prevalent measure of effect size was odds ratio (3,213 mentions more than half of the recognized effect size spans) followed by prevalence (1,011 mentions, 10.4%) (Figure 94).

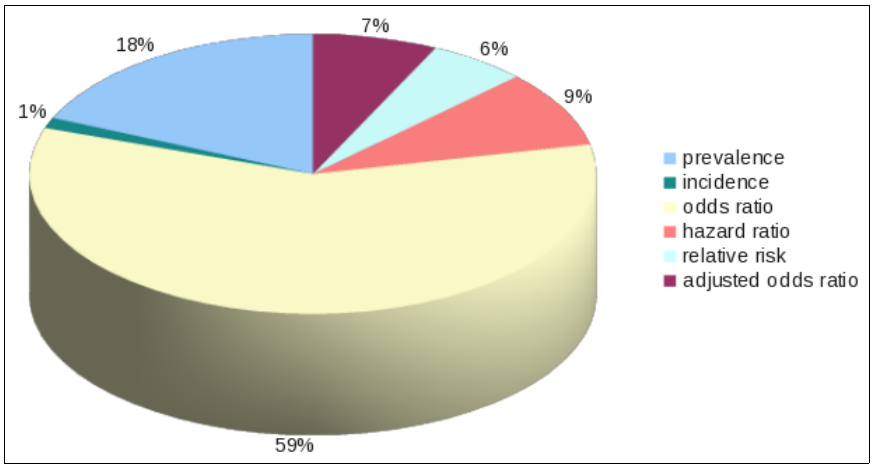


Figure 94: Distribution of the various effect size types.

This is not surprising since the most frequent study design is the "*cross-sectional*" ("*Cross-sectional*" study designs are showing the odds ratios from the studied prevalence of related diseases in a population sample). The metrics with the lowest frequency are those of relative risk and incidence with 308 (3.1%) and 67 (6.9%) mentions respectively in total. 4,227 (44.0%) effect sizes remained without a recognized effect size type.

Table 58: Number of normalized effect sizes. The tables includes effect sizes mentions with value, confidence interval, related concept and known type.

effect size	**number of mentions**					
with value	9,701					
with CI	6,421					
with concept	5,261					
with type	5,474					
	incidence	relative risk	adjusted odds ratio	hazard ratio	prevalence	odds ratio
	67	308	390	485	1,011	3,213

7.2.5. Temporal Analysis of Identified Exposure, Outcome and Covariate Concepts

Through automated database queries, the year of publication for each concept mention detected in an abstract was obtained. Hence, the document level frequency of each of the identified concepts for the characteristics of exposure, outcome and covariate was calculated. The period between the years 1965 and 1989 is excluded due to the limited information obtained for almost all of the identified concepts. The document level frequencies for the top five most frequent concepts for each of the above mentioned characteristics are represented in frequency diagrams (one for overall and five diagrams for five separate time periods, 1990-1995, 1996-2000, 2001-2005, 2006-2010, 2011-present). However, these diagrams exclude any concepts that represent obesity/overweight (e.g., "*obesity*", "*overweight*", "*body mass index*", etc) in the characteristics of exposure and outcome since our aim was to mainly enable the identification of other (potentially underlying) concepts related to or co-existing with obesity such as various co-morbidities or possible health outcomes. Certain interesting patterns regarding the nature of the identified concepts are noted and discussed below.

Exposures

Obesity and body mass index were the top exposure concepts mentioned in epidemiological studies. This suggests that there is a great deal of conducted research involving the role of obesity as a risk factor for the onset of other various disorders such as depression, diabetes or cardiovascular ones e.g., stroke, coronary heart disease. However, if we are to excluded concepts that represent obesity or those that are directly linked to it as measures (e.g., "*waist circumference*"), a clearer picture regarding the possibly associated concepts with obesity can be revealed.

Figure 95 displays the document level frequency for the top five most frequent concepts (age, physical activity, hypertension, metabolic syndrome, diabetes) identified as exposures for the time period of 1990-2011. Generally, through the years, the mentions of exposure are increasing, particularly post 2000. All five concepts displayed an increase after 2006, reaching their highest frequency during the period 2007-2008. A small decline occurring in the following years was observed. From the top five concepts, it was noted that three are disease related ("*hypertension*", "*metabolic syndrome*" , "*diabetes*"), while two are associated with the population clinical attributes and individual behaviour ("*age*", "*physical activity*"). Diseases/disorders such as "*hypertension*", "*metabolic syndrome"* and "*diabetes*" can be posing as potential risk factors to obesity or are related in a potential underlying clinical relationship. More precisely, after the year 2000, it is indicated that research has focused on the inspection of these specific concepts in epidemiological studies with the number of related ones increasing towards the present day.

This suggests that clinical professionals have become more aware of the (missing?) link between obesity and various other disorders, an indication that is being covered currently by the cluster of diseases that are part of metabolic syndrome. After careful inspection of the top forty most frequent exposures (see Table 95), it was revealed that not many concepts were associated with diseases (if so, their document frequency was observed to be low) suggesting that either it is not clear to researchers the relationship they are sharing with obesity or research has yet to focus on their links yet.

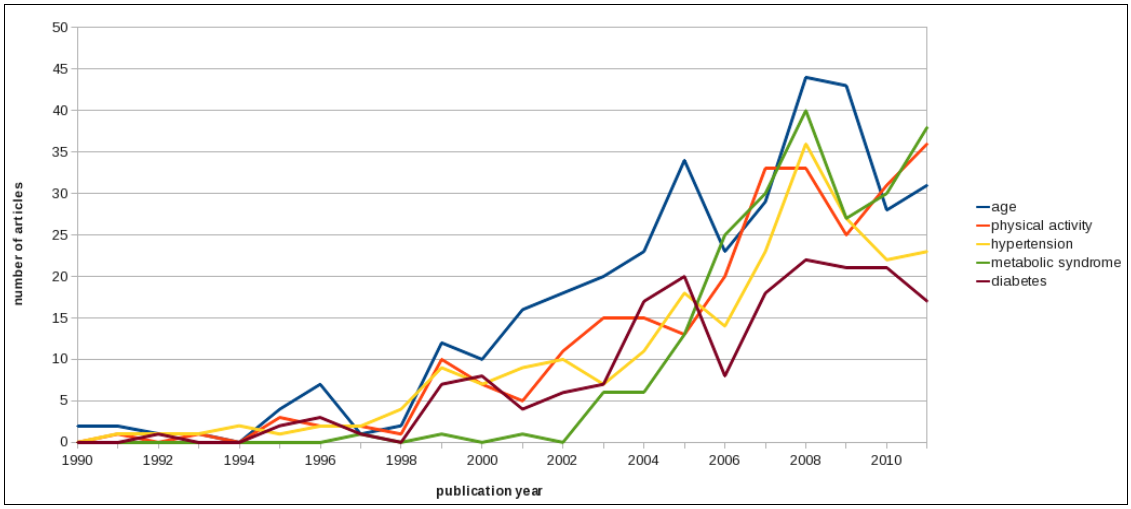
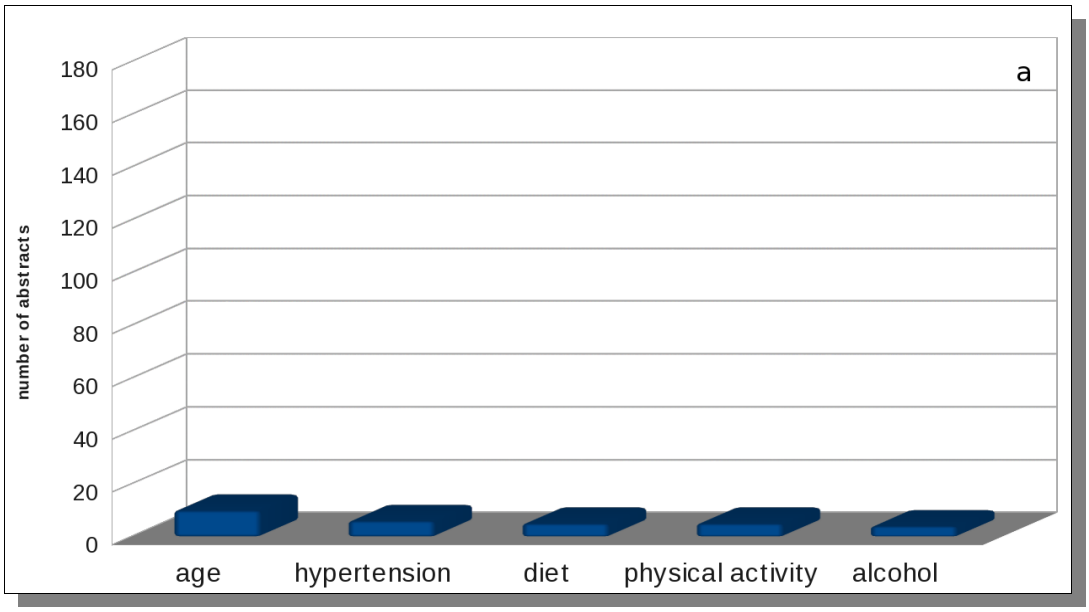


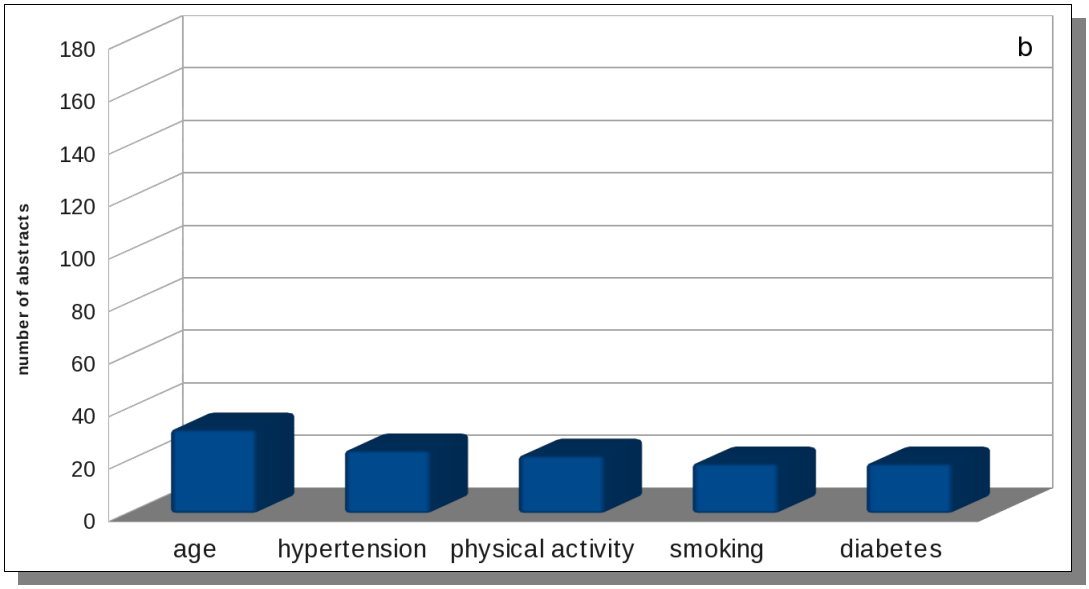
Figure 95: Document frequency according to the publication year of the top five most frequent exposures.

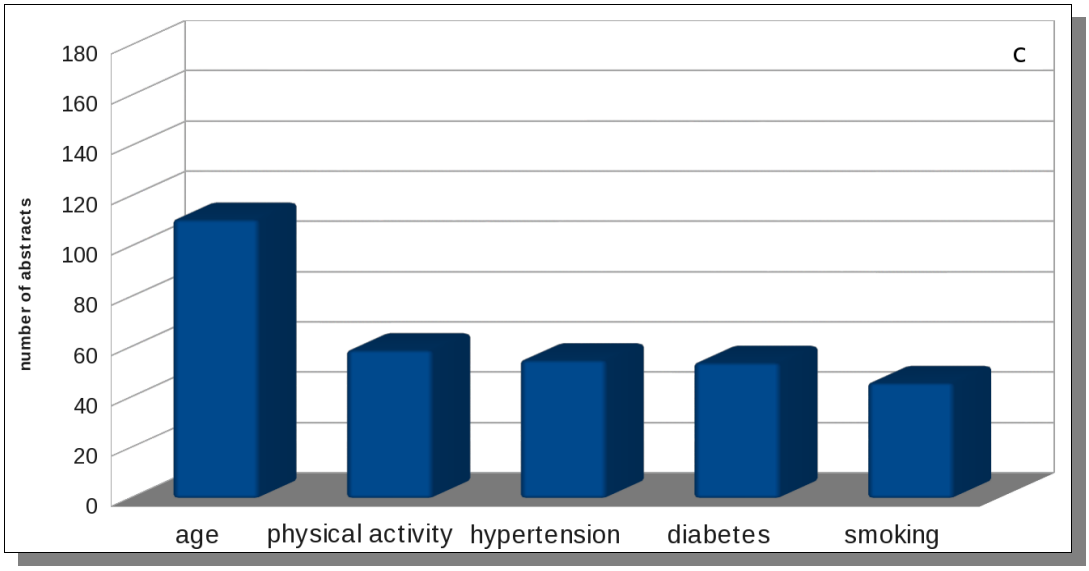
"*Age*" and "*physical activity*" were observed to have the highest mentions among specific exposure concepts overall. In addition to these two, it appears that most of the exposures recognized belonged to the semantic groups of "*physiology*" ("*gender*", "*weight*", "*sex*", etc) and "*activity*/*behaviour*" ("*smoking*", "*diet*", "*education*", etc). Since not many exposures were disease-related, this indicates that with the exception of hypertension, metabolic syndrome and diabetes, diseases such as depression, asthma, cardiovascular disease and non-alcoholic fatty liver disease are receiving the same attention of organism attributes (e.g., age). Surprisingly, most studies are focused on the inspection of the lifestyle and the (unique) physiology of the studied population (e.g., "*gender*", "*education*", "*physical activity*") and how it affects its health status in order to comprehend their hypothesis and reach a reliable conclusion.

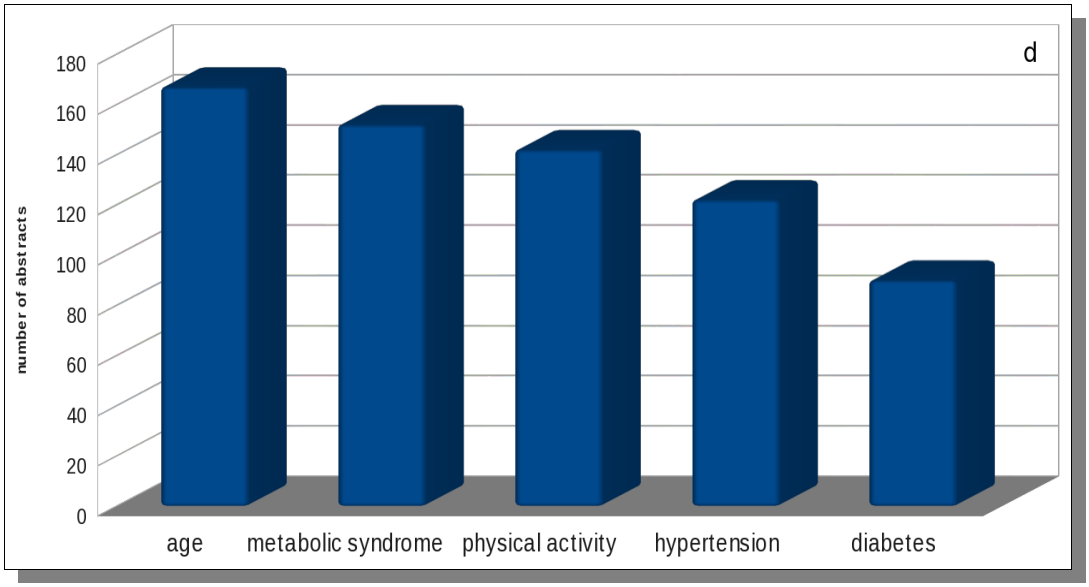
Figures 96a-e show the top five most frequent exposures non-directly related to obesity (from Table 51) through the periods 1990-1995, 1996-2000, 2001-2005, 2006-2010 and 2011. Certain concepts appear as exposures in epidemiological studies in all periods: "*age*", "*hypertension*" and "*physical activity*". Besides these three concepts, other entries observed in the top five of frequent exposures included "*diet*" and "*alcohol*" (1990-1995), "*smoking*" and "*diabetes*" (figures 96a and d) and "*metabolic syndrome*" and "*gender*" (figures 98c and e). More specifically, "*age*" was the most frequent exposure from 1990 till 2010 with increasing number of documents, while in the year 2011, it was moved to the second place, after "metabolic syndrome". "*Metabolic syndrome*" seems to have an increase in document mentions as an exposure between the years 2006-present day. This could explain the disappearance of "*diabetes*" and the decreased number of mentions for" *hypertension*" as both are disorders that are included in the definition of the metabolic syndrome and the shift of research is towards a cluster of diseases rather than inspecting each one separately.

Additionally, it was observed that there is a tendency in the majority of research to study organism attributes such as "*age*" and "*gender*" of population samples as well as various related disorders with obesity rather than social and individual behaviour concepts. This indicates the decreasing interest in relatively straightforward (to understand their influence in the obesity problem) concepts such as "*diet*" and "*smoking*" and the augmenting focus towards related co-morbidities and underlying risk factors.


a
number of abstracts
180
160
140
120
100
80
60
40
20
0
age
hypertension
diet
physical activity
alcohol


b
number of abstracts
180
160
140
120
100
80
60
40
20
0
age
hypertension
physical activity
smoking
diabetes


c
number of abstracts
180
160
140
120
100
80
60
40
20
0
age
physical activity
hypertension
diabetes
smoking



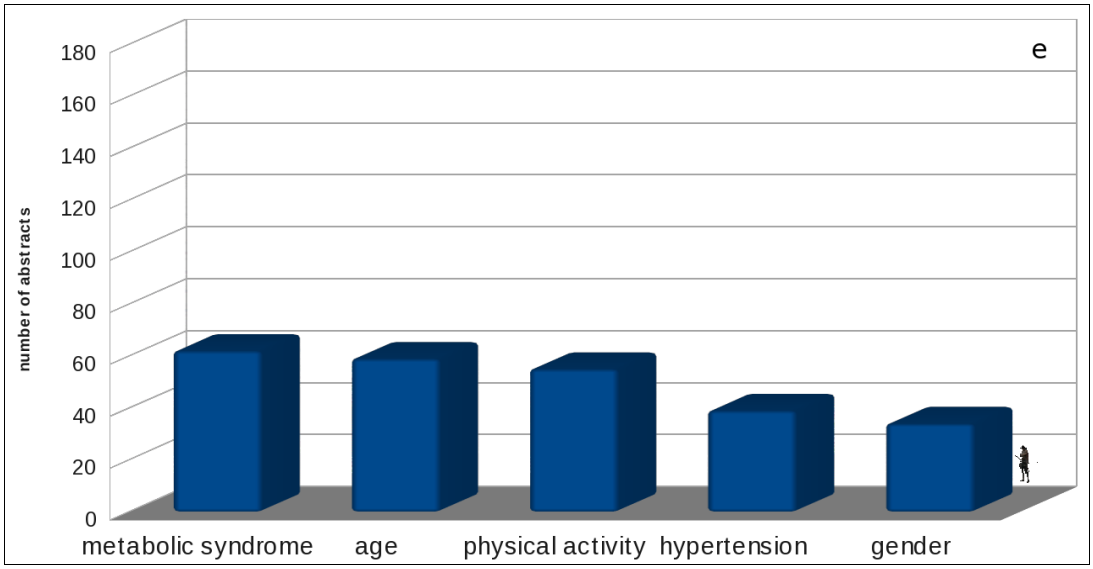


Figure 96a-e: The top five most frequent exposures at the document level. (a) 1990-1995 (b) 1996-2000 (c) 2001-2005 (d) 2006-2010 (e) 2011.

Outcomes

Figure 97 displays the document level frequency for the top five most frequent concepts recognized as outcomes for the time period of 1990-2011. Similar observations to the characteristic of exposure are noted. There is a steady increase in document mentions till the years 2004-2006 (hypertension, diabetes), where a sudden drop occurs, only to lead to a higher augmentation of mentions. Similarly to the temporal analysis of recognized exposures, concepts such as "*obesity*" and "*overweight*" or directly linked to it (e.g., "*body mass index*") were excluded in order to enable a more detailed analysis of potential outcomes studied in association to obesity.

The top five most frequent concepts identified as outcomes are diseases/disorders ("*hypertension*", "*diabetes*", "*metabolic syndrome*", "*cardiovascular disease*"), whereas one concept was of quantitative nature regarding the influence of obesity to the health status of an affected individual ("*mortality*"). This indicates that studies around obesity aim to understand health consequences on affected populations and other potential clinical outcomes that can be attributed to obesity through existing co-morbidities. During the time frame 1998-present, it seems that scientific research has developed an apparent interest into comprehending the complex relationship between obesity and various other disorders as the document frequencies seem to increase towards the present day with almost three times the mentions from early 90's.

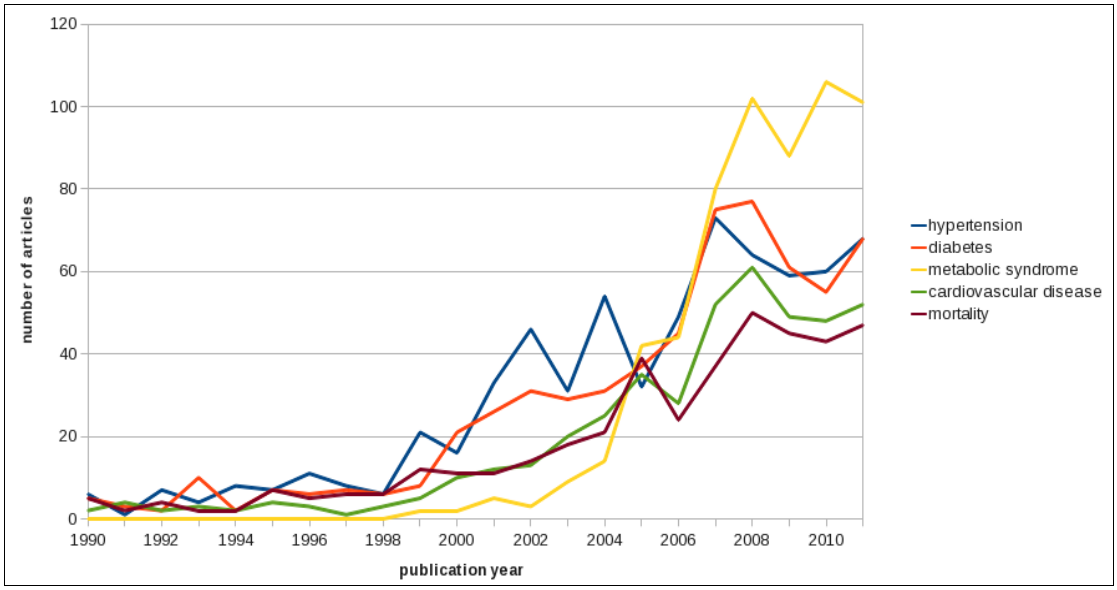
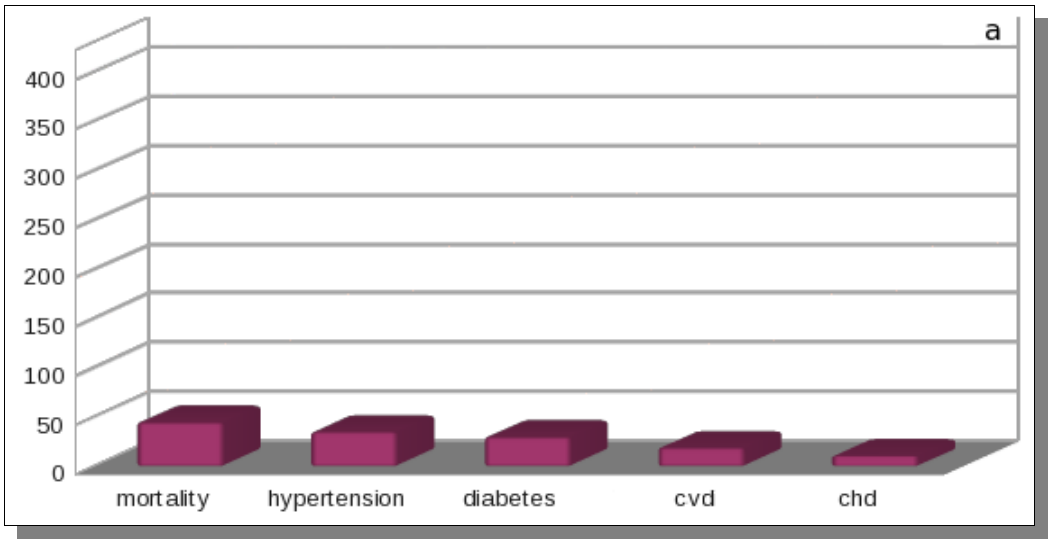
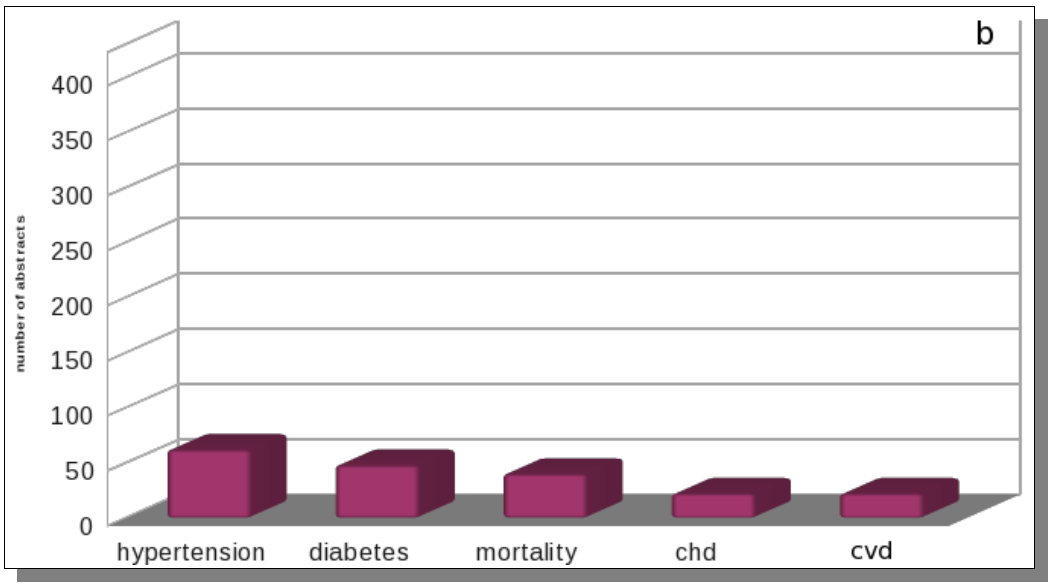


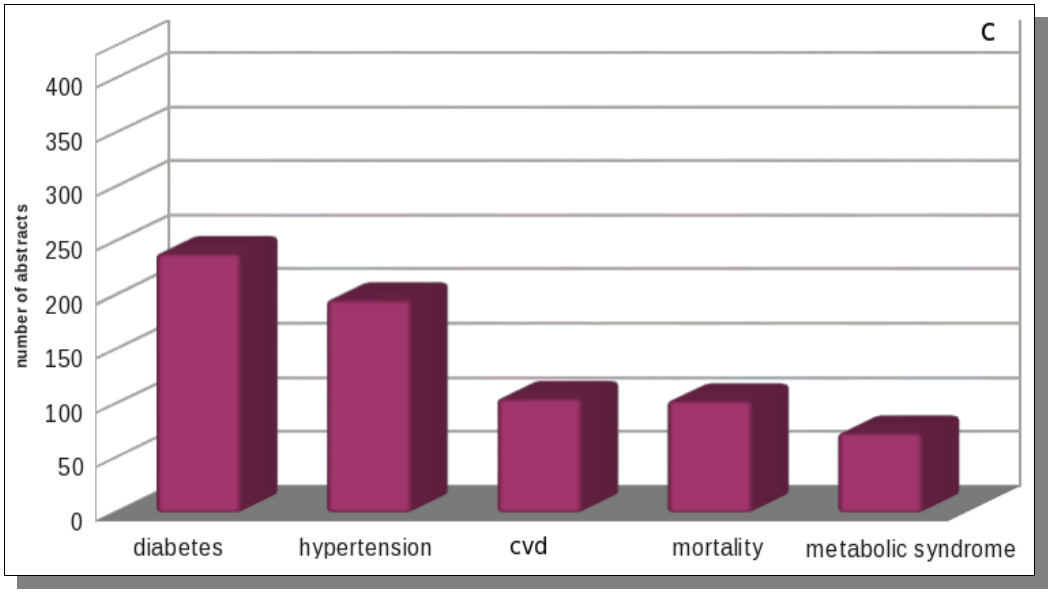
Figure 97: Document frequency according to the publication year of the top five most frequent outcomes.

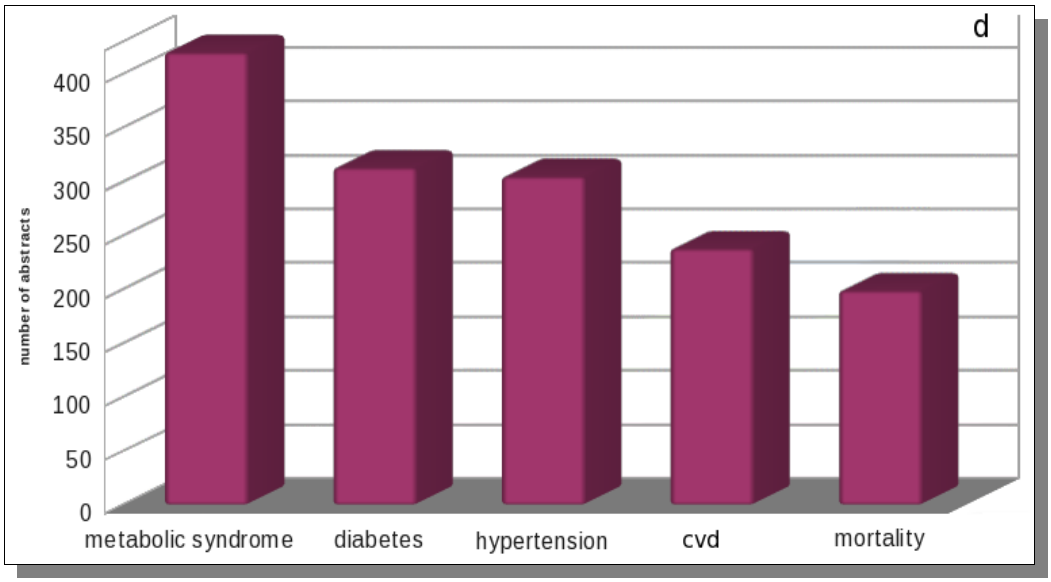
Figures 98a-e show the top five identified outcomes for different time spans between 1990-2011. It was observed that four out of five outcomes were the same through the years: "*diabetes*", "*hypertension*", "*mortality*" and "*cardiovascular diseases*". More specifically, for the periods 1990-1995, 1996-2000 and 2001-2005, the most frequent concepts were "*mortality*", "*hypertension*" and "*diabetes*" respectively. From 2006-2011 (figures 98d-e), it was observed that "*metabolic syndrome*" was the top most frequent concept. Since "*metabolic syndrome*" is a cluster of diseases that contain "*hypertension*" and "*diabetes*", this could explain its rise in mentions in various epidemiological abstracts and respectively the decrease of mentions for the outcomes of "*diabetes*", "*hypertension*" and "*cardiovascular diseases*".

Additionally, the inclusion of "*metabolic syndrome*" in the most frequent outcomes revealed the exclusion of "*coronary heart disease*". The above two conclusions suggest that epidemiological studies have demonstrated a pattern of studying obesity and its co-related morbidities through the inspection of the metabolic syndrome rather than investigating each disorder separately. "*Mortality*" was noted to have a constant drop in epidemiological study mentions as an outcome. This could suggest that clinical professionals are focusing more on the understanding of the complex nature of obesity since its association with mortality could have potential underlying risk factors (posing as diseases and co-existing with obesity in individuals) that have not been fully explored yet. A closer look at the top forty most frequent mentioned outcomes reveals that most of the identified concepts are associated to a disease or disorder e.g., "*type 2 diabetes*", "*cvd*", "*asthma*", "*stroke*", etc. This indicates that obesity has been studied as a risk factor for the onset of not only well-known and studied diseases but others of less popularity (e.g., "*hypercholesterolemia*", "*atherosclerosis*").









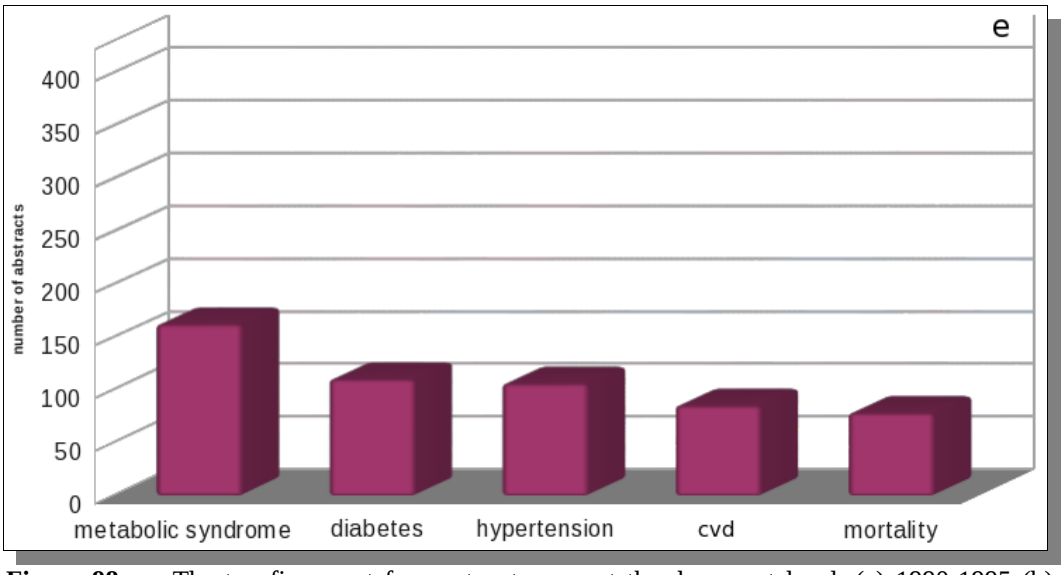


Figure 98a-e: The top five most frequent outcomes at the document level. (a) 1990-1995 (b) 1996-2000 (c) 2001-2005 (d) 2006-2010 (e) 2011.

Covariates

Figure 99 displays the document level frequency for the top five concepts most frequent recognized as covariates for the time period of 1990-2011. There are a few mentions of covariate concepts during the period 1990-1998. However, from the year 1998 and onwards, an increase in the concept document mentions has occurred with their number increasing drastically. Obesity attracted a significant amount of research attention during the 90's due to its evolution into a pandemic health problem, which potentially explains this augmentation of covariate mentions due to the focus of epidemiological interest on it and the isolation of related concepts for study.

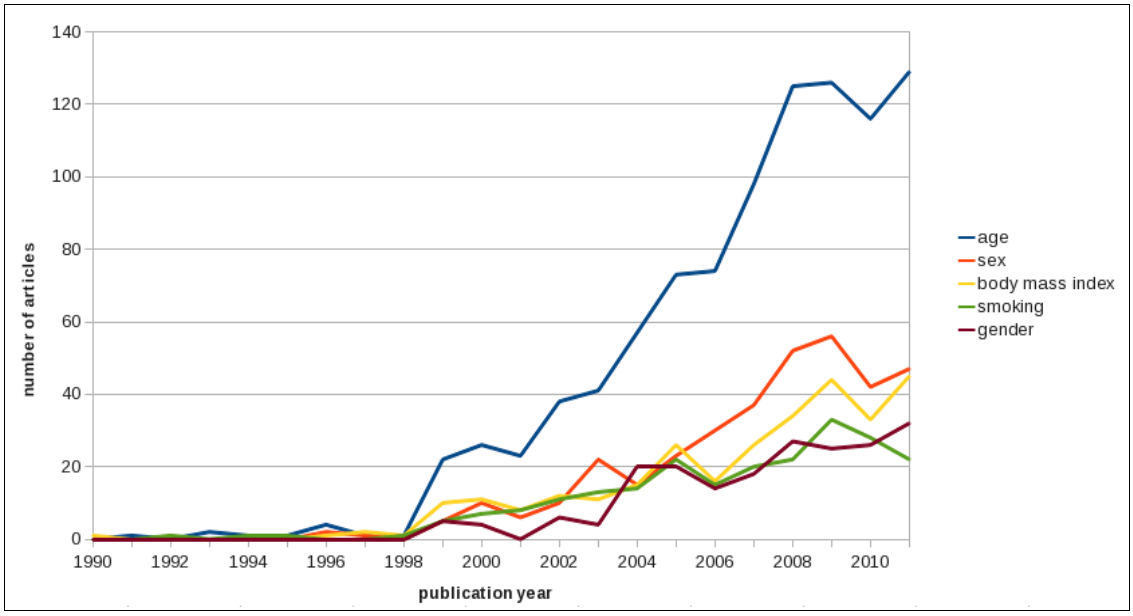


Figure 99: Document frequency according to the publication year of the top five most frequent covariates.

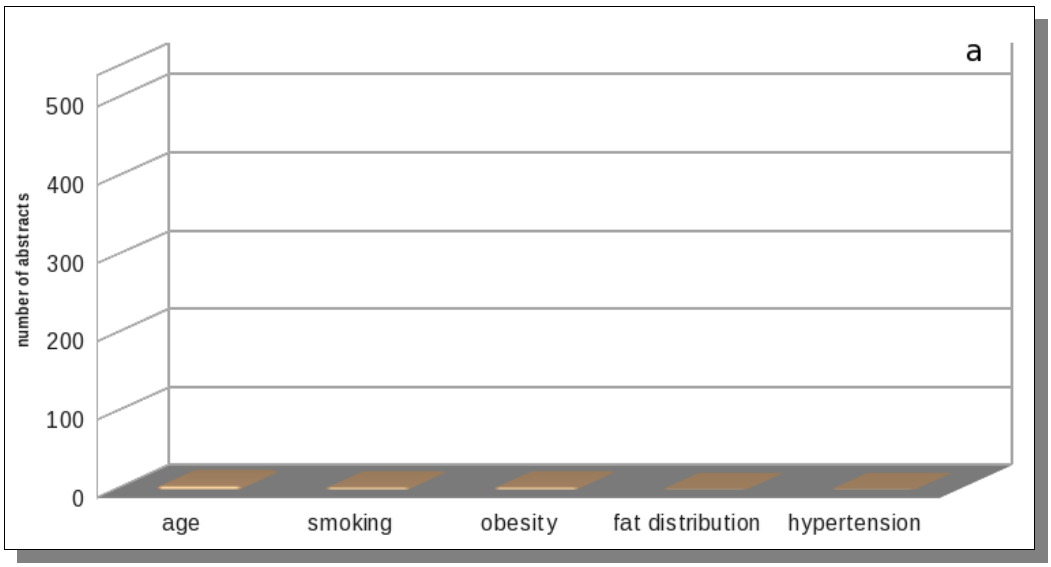
Figures 100a-e reveal the most frequent (top five) identified covariates for different time spans between 1990-2011. "*Age*" was the most mentioned covariate in epidemiological studies for all the years with "*gender*" being in the second place. This is not surprising as the relationship between obesity and its associated co-morbidities has not been clear enough. Therefore, it is most logical for observational studies to adjust their results for "*age*" in order to focus on other influential concepts such as co-existing disorders. Furthermore, from 1996 till 2011, the top four most mentioned covariate concepts remain the same and in the same places: "*age*", "*sex*", "*body mass index*" and "*smoking*". This suggests that most studies are aware of potential (but not easy to understand) influence that any of these concepts could have in their examined relationship between a particular exposure with an outcome.

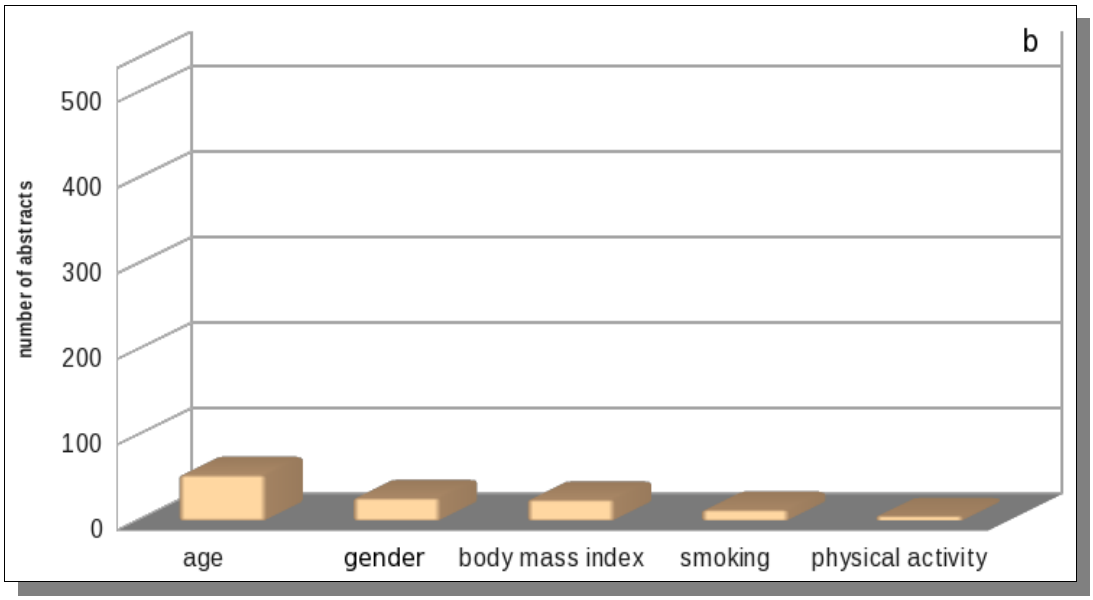
"E*ducation*" appeared steadily in the 5th position since the dawn of the new millennium (figures 100c-e) - in contrast to its appearance as the the 5th most frequent covariate in Figure

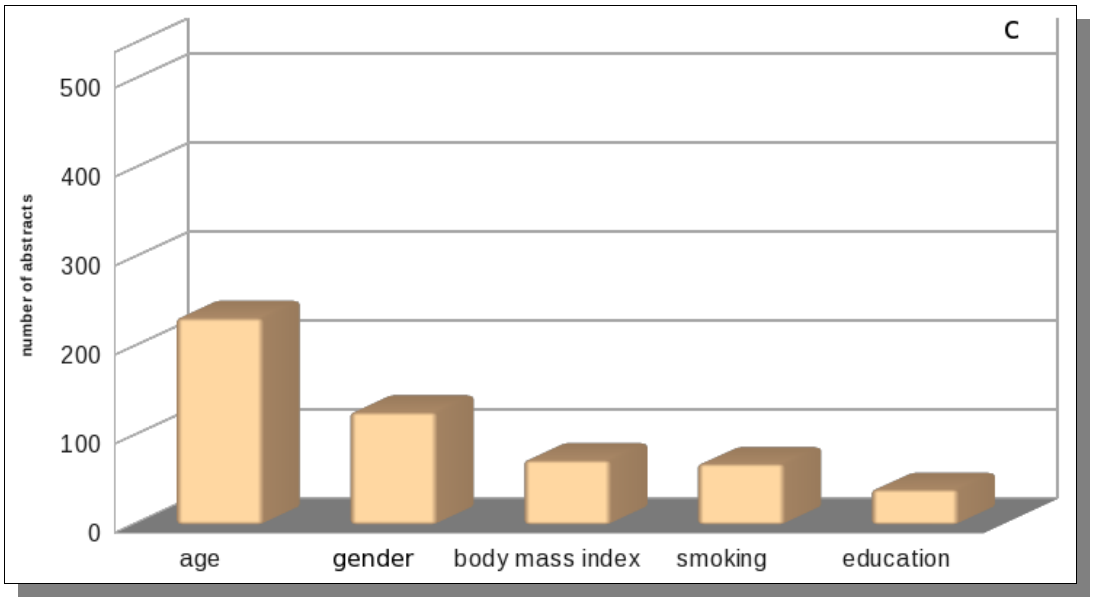
99. The level of education was not considered a part of the obesogenic risk factors till the obesity epidemic reached developing countries at the of 20th century. It is widely considered that poor education status is linked to higher body mass index in developed countries, while in the developing nations it has been mostly related to lower BMI. Therefore, until this association becomes more stable and clear, it is only logical for most studies to consider the level of education a covariate during their inspection of obesity related concepts.

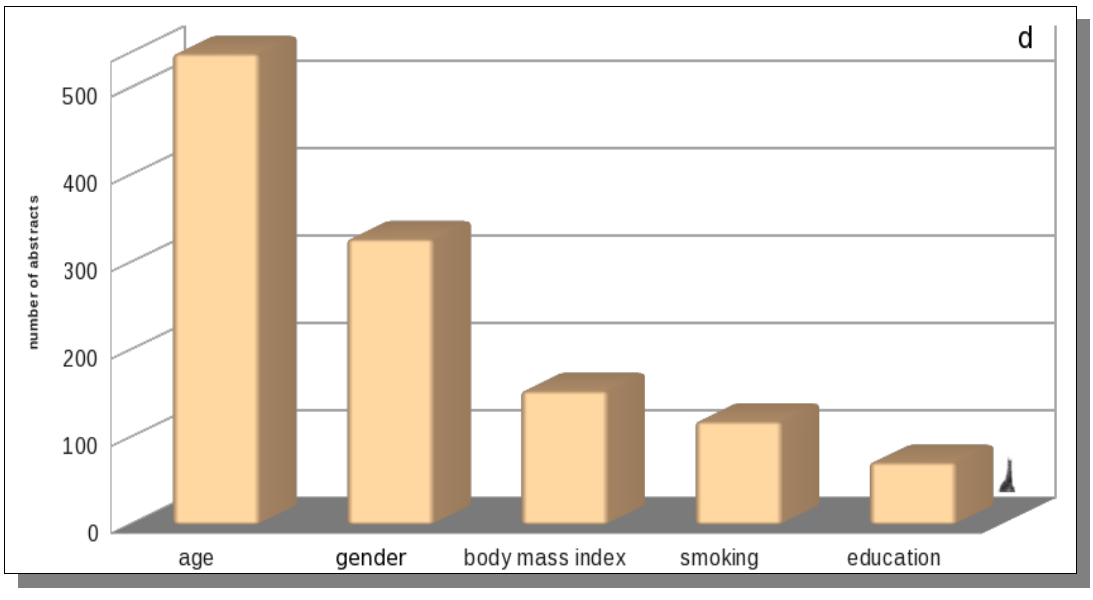
The role of gender in the obesity onset has been widely reported in the respective epidemiological literature but findings have not been consistent (along with their respective metrics - for the measure of obesity in females, waist to hip ratio is preferred, while for males body mass index is chosen). The differences between the sexes for the obesity issue are more complex than just the utilization of appropriate measures e.g., "*BMI*", "*WC*", "*WHR*". Since most epidemiological studies (mainly observational ones) require time to set up their hypothesis and to reach specific conclusions after careful inspection, gender adjustment seems to be the most appropriate action in order to reach a (relatively) stable conclusion regarding a targeted population and a health care problem.

A more closer inspection at the top forty most frequent covariates can reveal that a large number of concepts appeared to be of activity/behavioural nature (e.g., "*alcohol consumption*", "*physical activity*", "*exercise*", "*diet*"). This indicates that more researchers look at other important population features besides demographical information such as the amount of physical activity and its potential effect on the outcome of obesity. However, since it is very difficult to record and measure these type of concepts, more and more epidemiological studies will consider these concepts as covariates.




b
number of abstracts
500
400
300
200
100
0
age
gender
body mass index
smoking
physical activity


c
number of abstracts
500
400
300
200
100
0
age
gender
body mass index
smoking
education


d
number of abstracts
500
400
300
200
100
0
age
gender
body mass index
smoking
education

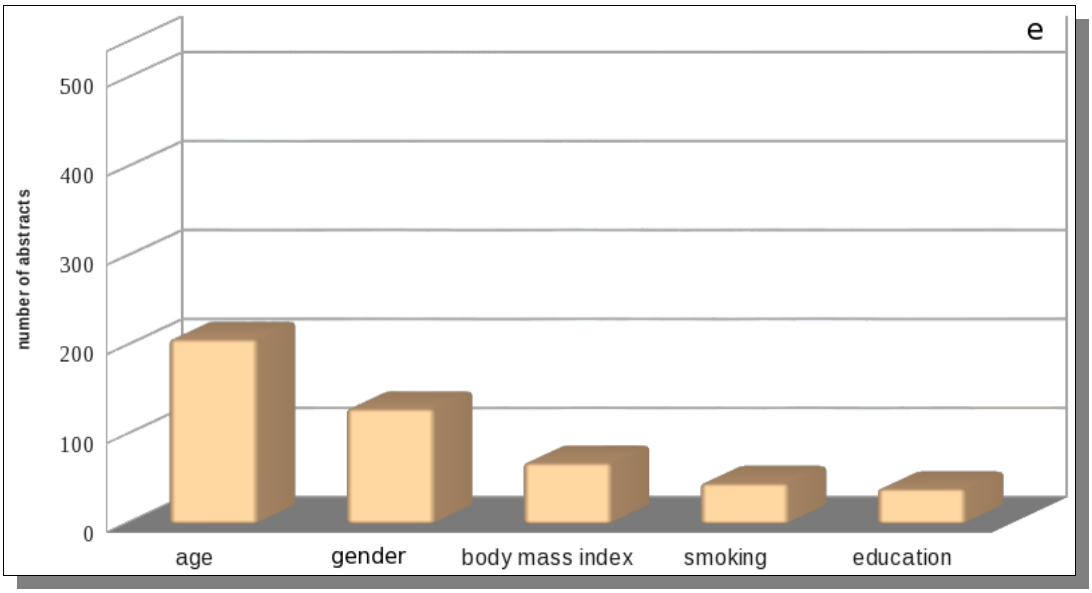


Figure 100a-e: The top five most frequent covariates at the document level. (a) 1990-1995 (b) 1996-2000 (c) 2001-2005 (d) 2006-2010 (e) 2011.

7.2.6. Pairing Identified Characteristics

From the normalized results, we studied the most common pairs of concepts between certain key characteristics. More specifically, we inspected the most frequent pairs of exposures-outcomes, exposures-covariates and outcomes-covariates in order to:

1. detect any specific association between obesity related concepts;
2. identify which concepts are classified as covariates under which common exposures and outcomes.

The inspection of common pairs could potentially reveal the current focus and trends of the research community regarding epidemiological studies related to obesity.

Exposure-Outcome Pairs

Table 59 reveals the most common pairs of exposures and outcomes that have been mentioned in the corpus. From the above Table, certain pairs have been established through long time research. It is now know that the lack of physical activity as well as the progression into an older physical state are likely to lead to the onset of obesity. Besides these relatively straightforward and examined associations, obesity has been studied as an outcome with a variety of exposures that include organism attributes such as "*body mass index*" and "*age*", individuals behaviours such as "*physical activity*" and disorders like "*diabetes*" and "*hypertension*". This conclusion is not surprising since all these exposures have been linked to the onset of obesity. Additionally, it can be seen that obesity (and its substitute concept body

mass index) has been examined as a risk factor (exposure) for a number of diseases such as hypertension, diabetes, cardiovascular diseases, asthma and metabolic syndrome as well as it has been linked to death and mortality. This strengthens the hypothesis that obesity is potentially related to a cluster of diseases and shares with them a complex relationship (hypertension, diabetes). The inclusion of metabolic syndrome as an outcome the contains most of these disorders only enhances this conclusion.

Table 59: The top 15 most frequent exposure-outcome pairs with their respective numbers of documents that they appear in.

exposure	outcome	number of documents
body mass index	obesity	280
obesity	hypertension	133
obesity	mortality	123
obesity	type 2 diabetes	110
age	obesity	106
body mass index	overweight	99
body mass index	mortality	88
physical activity	obesity	82
hypertension	obesity	81
obesity	cardiovascular disease	72
obesity	asthma	69
body mass index	hypertension	67
obesity	death	64
type 2 diabetes	obesity	64
obesity	metabolic syndrome	61

Exposure-Covariate Pairs

Table 60 reveals the most common pairs of exposures and covariates that have been studied in the corpus. "*Age*" was the most frequent covariate in the pairs exposure-covariate. A large number of studies have been considering "*age*" a covariate when they researched as exposures the concepts of obesity and its substitutes (body mass index, waist circumference and overweight), metabolic syndrome, hypertension and physical activity. It seems "*age*" can have an unknown effect to the onset of obesity-associated disorders, therefore contributing to the complexity of the relationships that obesity shares with other diseases. Additionally, other common covariates were noted to be "*gender*", "*smoking*", "*body mass index*" and "*education*". With the exception of "*age*", all the above concepts were considered as covariates in the studies researching obesity as an exposure. These conclusions further strengthen the

links between the complex disease of obesity with these variables that can potentially affect to an unknown effect its relationship with health-related outcomes.

Table 60: The top 15 most frequent exposure-covariate pairs with their respective numbers of documents that they appear in.

exposure	covariate	number of documents
obesity	age	190
body mass index	age	187
obesity	gender	110
body mass index	gender	96
body mass index	smoking	60
obesity	smoking	56
waist circumference	age	48
obesity	body mass index	44
overweight	age	37
metabolic syndrome	age	37
hypertension	age	36
obesity	education	33
physical activity	age	30
body mass index	education	29
body mass index	race	26

Outcome-Covariate Pairs

Table 61 reveals the most common pairs of exposures and covariates that have been studied in the corpus. Similar observations from the inspection of the common exposure-covariate, can be seen in the pairs of outcome-covariate. "*Age*" was the most frequent covariate that studies are adjusting their results for when they examine as outcomes a variety of obesity associated concepts: obesity, body mass index, overweight, hypertension, diabetes, death, cardiovascular disease, metabolic syndrome and insulin resistance. Most of the age related outcomes seem to be disorders (including obesity and overweight) suggesting that "*age*" can have a potentially unknown effect to their onset. Additionally, other covariates were noted to be "*gender*", "*body mass index*", "*smoking*" and "*education*". Their respective outcomes were associated with obesity with the exception of "*gender*" that further included "*body mass index*" and "*overweight*". This indicates that all the above are considered as important variables that can influence the onset and progression of obesity in a population sample.

Table 61: The top 15 most frequent outcome-covariate pairs with their respective numbers of documents that they appear in.

outcome	covariate	number of documents
obesity	age	305
obesity	gender	190
body mass index	age	125
overweight	age	101
metabolic syndrome	age	66
obesity	body mass index	66
obesity	smoking	64
diabetes	age	62
hypertension	age	61
obesity	education	46
cardiovascular disease	age	40
body mass index	gender	39
insulin resistance	age	34
overweight	gender	31
death	age	30

7.2.7. Obesity Concept Map

The generated concept map (see Chapter 6) represents a total of 69,351 concepts: 25,518 exposures, 40,333 outcomes and 5,500 covariates. For obvious size reasons, the threshold level (document frequency) was chosen to be 150 so not all of the normalized concepts were included. The following screenshots (figures 101-103) reveal parts of the automatically generated concept map for each characteristic. The concept map has been slightly edited in order to fit as an image and to showcase the different semantic clusters. For example, in Figure 101, no concepts related to outcomes or covariates can be observed, focusing only to the representation of the exposures. The same approach was utilized in figures 102 and 103 for outcomes and covariates respectively. This concept map was used as a backbone to provide a literature exploration system which is used to follow the extracted links and relationships between exposures, outcomes and covariates (see next section).

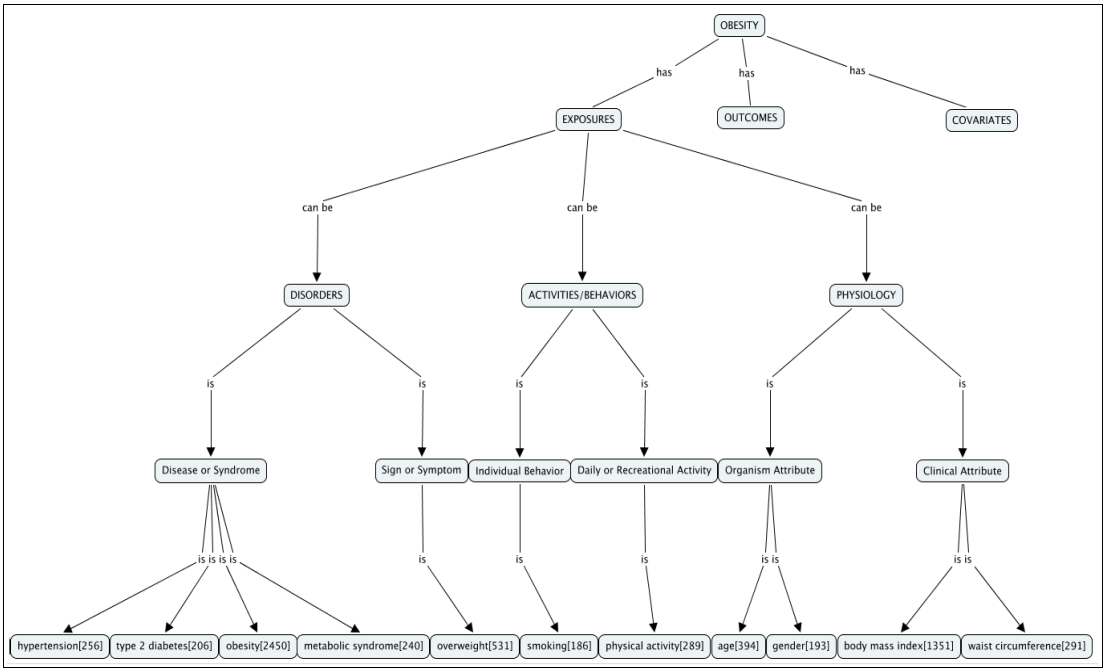


Figure 101: Part of the automatically generated concept map that represents normalized (with a minimum document frequency of 150) exposures along with their respective UMLS classification.

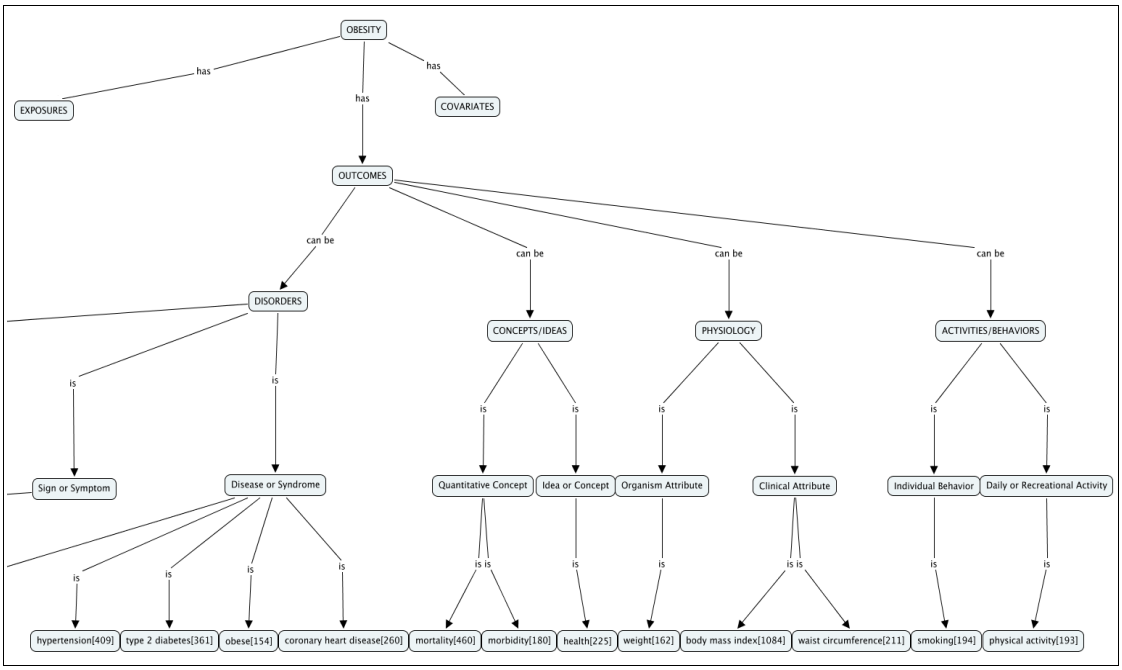


Figure 102: Part of the automatically generated concept map that represents normalized (with a minimum document frequency of 150) outcomes along with their respective UMLS classification.

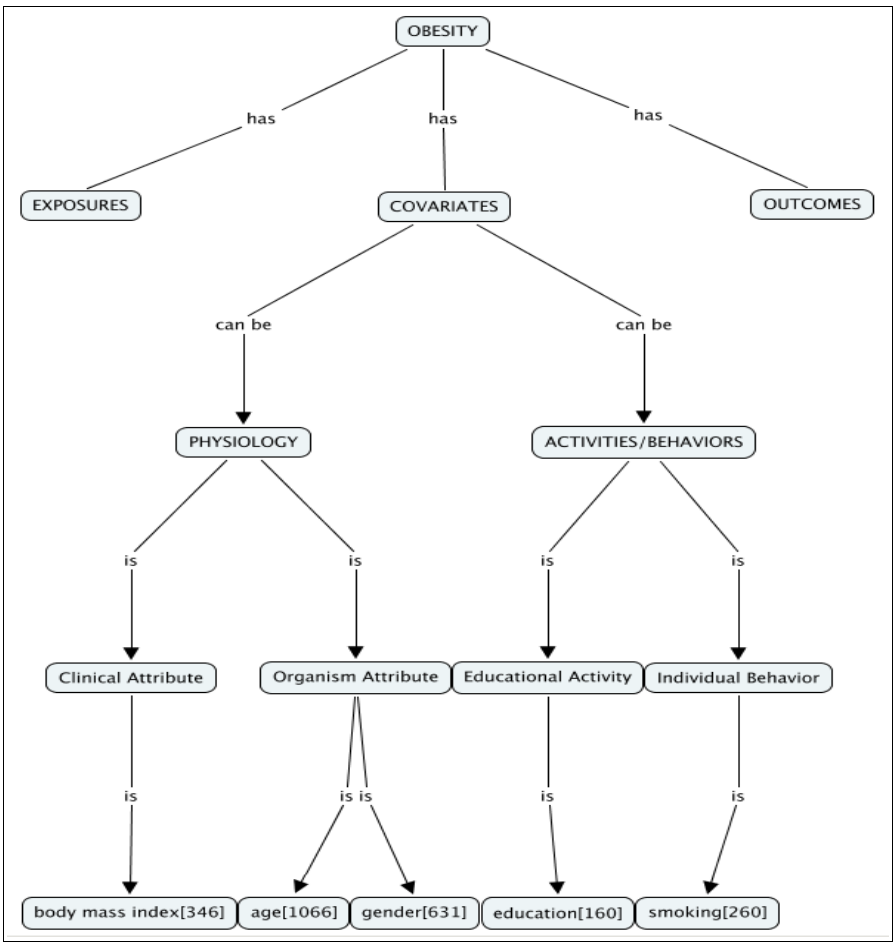


Figure 103: Part of the automatically generated concept map that represents normalized (with a minimum document frequency of 150) covariates along with their respective UMLS classification.

7.3. EpiTeM – Exploration of Epidemiological Literature

EpiTeM (Epidemiological Text Miner) was designed by the author in order to enable the navigation and manipulation of epidemiological text mining results. As a demonstrator project, the system uses obesity-related data as a case although EpiTeM is a generic framework aiming to represented any data identified and normalized from the field of epidemiology related to a health care problem. Due to the overwhelming number of characteristic mentions in the corpus, EpiTeM offers the ability to manipulate easily these large amounts of data, enabling the user to search for a particular characteristic or term and collect the necessary subset of information from a vast corpus of epidemiological research. EpiTeM was implemented in Java with Swing, and can be executed as a stand alone application. The extracted and normalized mentions of the targeted characteristics are stored in a MySql database that includes the publication year and the PMID of each abstract. Figure 104 shows the initial interface of EpiTeM that contains a welcome message and available search options for concepts belonging to all the epidemiological characteristics. The following screenshots (figures 104, 105 and 106)

display how EpiTeM works along with a brief description of its functions for an example with the terms "*obesity*" and "*depression*" as an exposure and as an outcome respectively.

Figure 104: Welcome screen of the EpiTeM system. EpiTem is an interface that enables the user to manipulate the identified and normalized key characteristics from a MEDLINE corpus.

EpiTeM enables the user to input specific concepts of interest as exposure, outcome and covariate, particular epidemiological study designs, population description according to the normalized classes of its attributes (gender, age, nationality, ethnicity) or effect size types. If the user desires to perform a search of multiple input variables, then the intersection of results is returned. EpiTeM manipulates all the abstract information that is being stored in the MySql database and returns those epidemiological studies that match the input criteria (through strict string matching). The information displayed is the total number of abstracts matching the input criteria, their respective PMID, year of publication and their title. Additionally, EpiTeM presents the top five study design types for this subset of abstracts, along with their respective frequency, and the top five (in terms of mentions) UMLS semantic groups for the exposure, outcome and covariate characteristics (Figure 105).

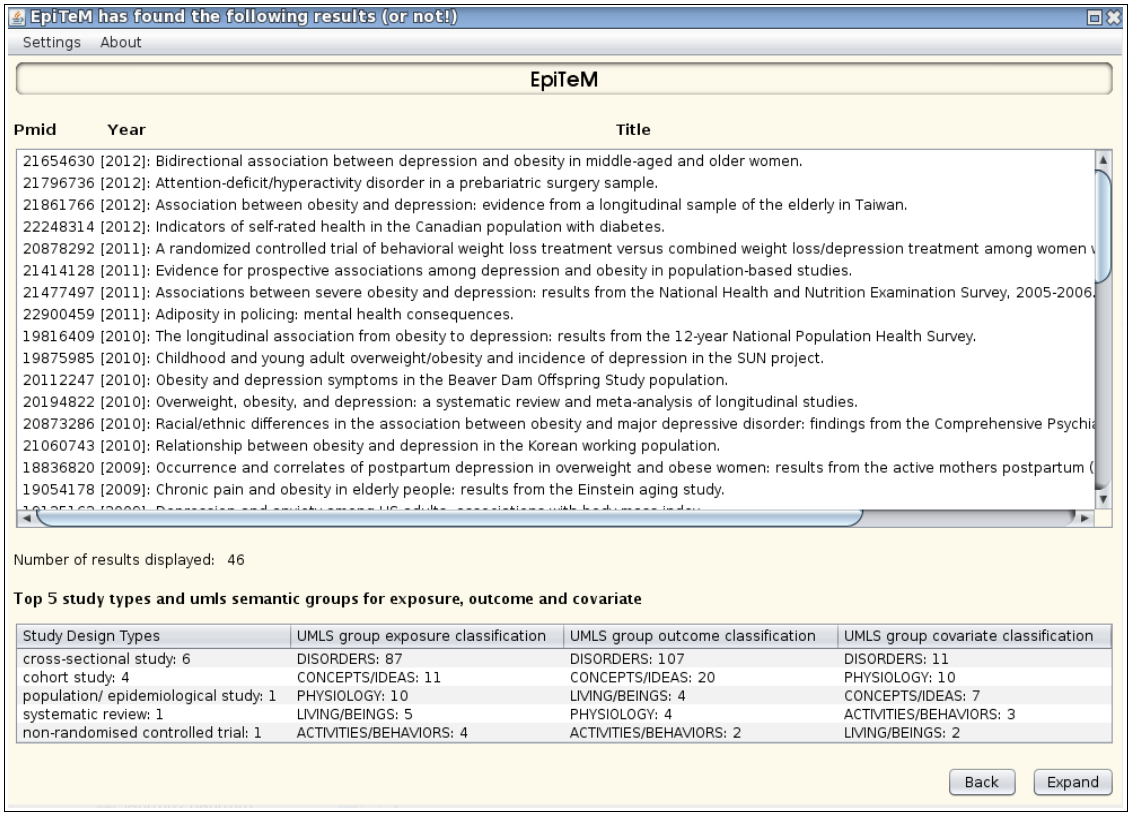


Figure 105: EpiTeM results. After searching the literature for "*obesity*" as exposure and "*depression*" as outcome. Each study matches both of those search characteristics.

Each title leads to the expanded results window that contains the extracted and normalized epidemiological information for the chosen study, including a link to the related PubMed webpage. The extracted mentions for each characteristic (if existing) can be highlighted in text in their original form. Their normalized versions are displayed in a table that contains four columns:

1. **Characteristic:** the type of the recognized epidemiological characteristic;
2. **Span:** the identified and normalized span assigned to the to type of the characteristic;
3. **Frequency:** the respective document frequency of each concept in the corpus (in how many abstracts each concept appears).
4. **Type/Subtype:** the study attributes for the study design, population characteristics such as gender, age, nationality and ethnicity, the UMLS semantic group and category classification for the exposures, outcomes and covariates and the effect size types, values, confidence interval and relate concept. EpiTeM assigns a "-" in the Type/Subtype column when no values have been recognized for the various attributes of the epidemiological characteristics.

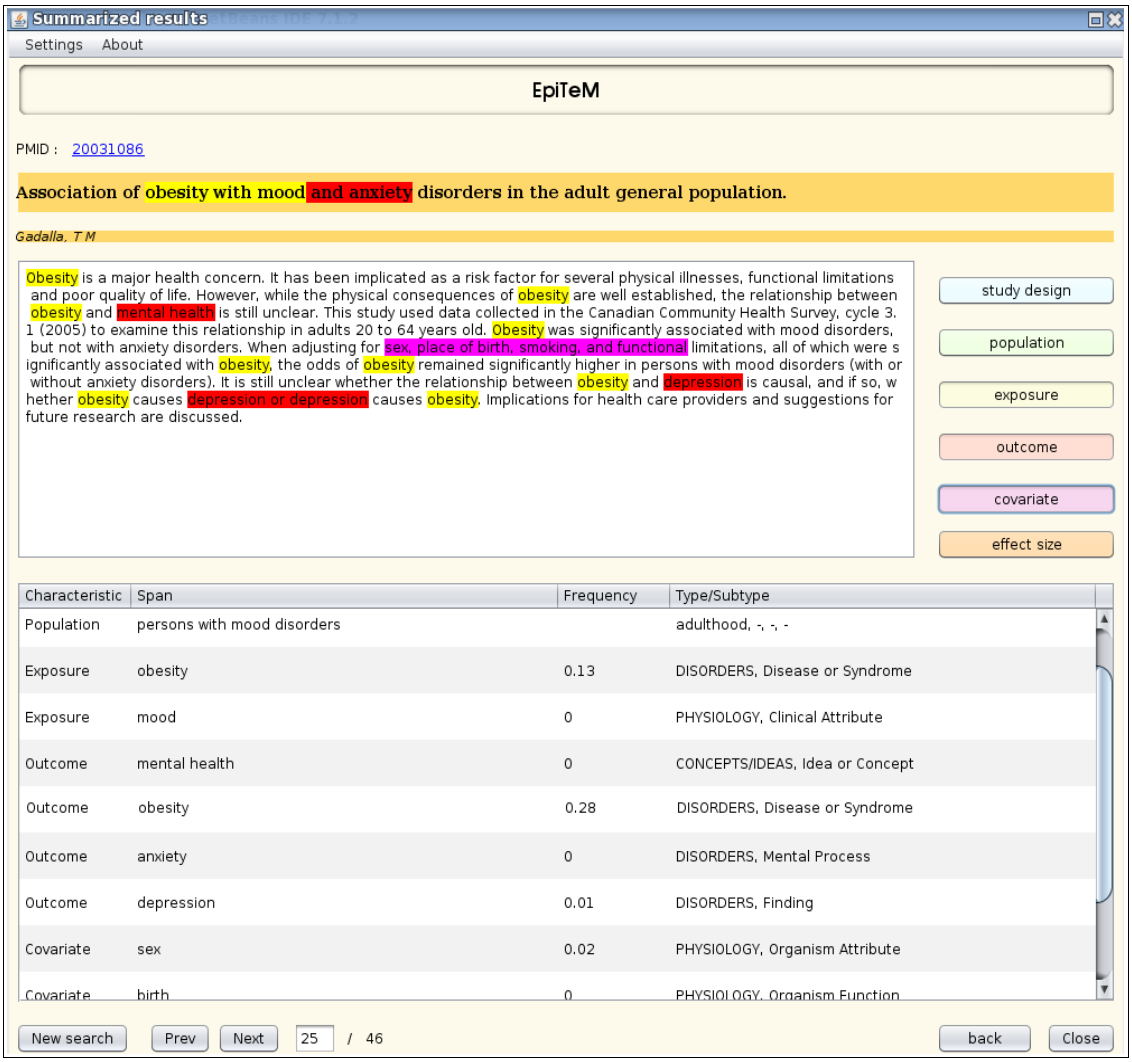


Figure 106: Detailed representation of the epidemiological information extracted from the PubMed abstract with pmid 20031086. The identified characteristic mentions are highlighted in the abstract text while their normalized and classified versions are displayed below.

EpiTeM can be populated with any extracted data for any epidemiological literature subset. EpiTeM aims to facilitate epidemiological data exploration as it enables clinical researchers to search for multiple criteria in a variety of epidemiological studies such as:

1. a variety of study designs with specific exposures e.g., "*cross-sectional studies*" with "*hypertension*" as an exposure (Figure 107);
2. studies with particular exposures and covariates e.g., "*cohort study*" with "*age*" as covariate (Figure 108);
3. studies that investigate specific outcomes under the observation of a targeted population e.g., "*cross-sectional studies*" with "*obesity*" as an outcome and the population sample being of "*Chinese*" nationality (Figure 109).
4. studies that inspect specific exposures and have displayed their effect size e.g., "*cross-sectional*" studies with "*hypertension*" as exposure and its reported prevalence (Figure 110).

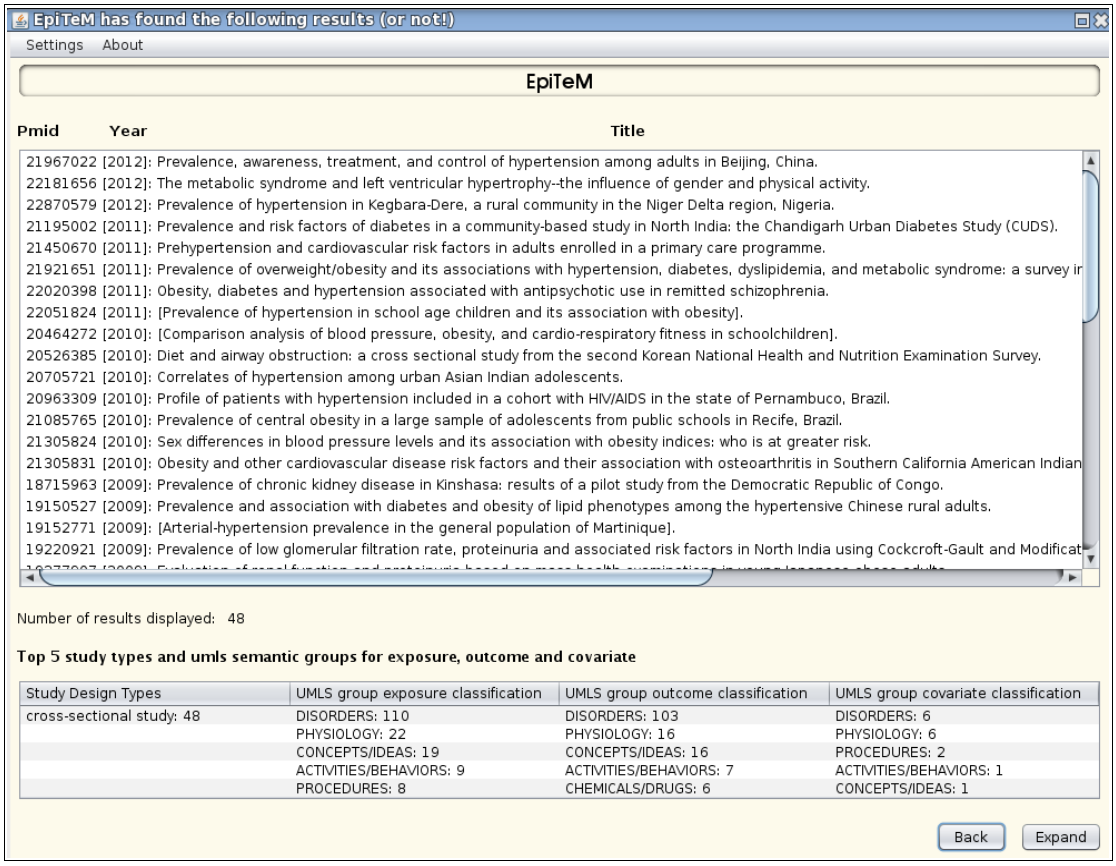


Figure 107: An example of epidemiological data exploration through EpiTeM for "*cross-sectional study*" as study design and "*hypertension*" as an exposure variable.

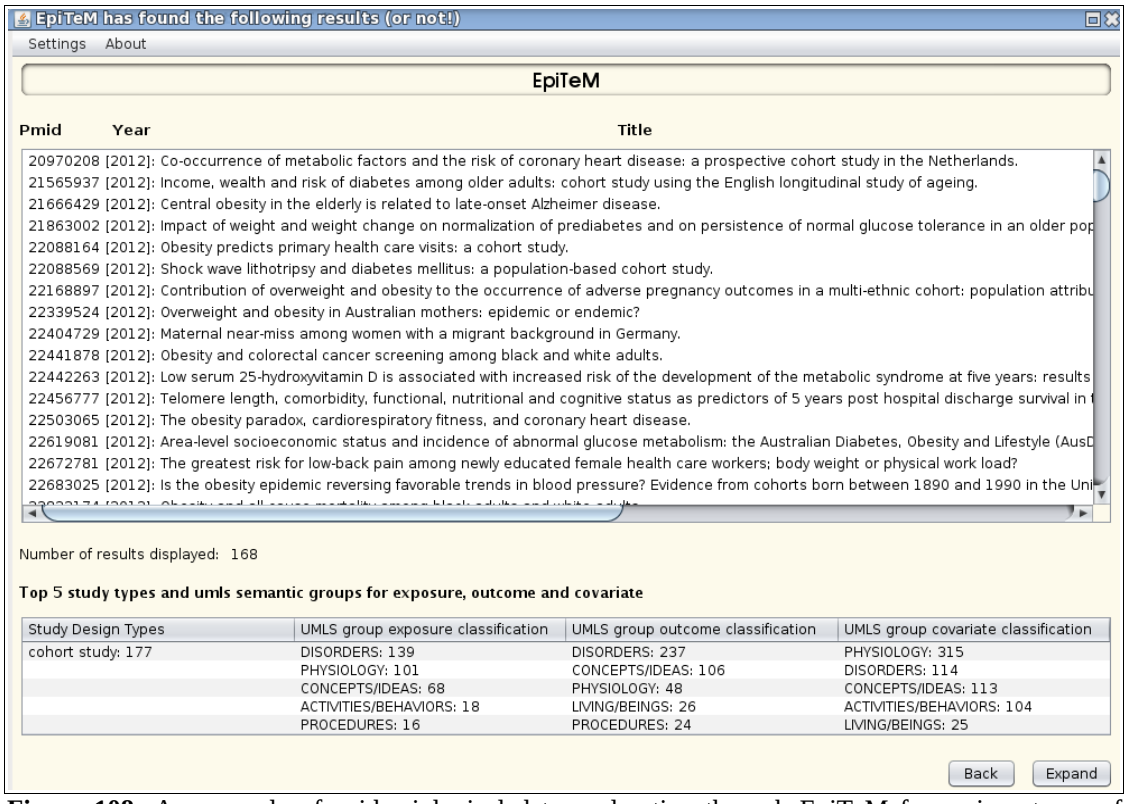


Figure 108: An example of epidemiological data exploration through EpiTeM for various types of cohort studies with "*obesity*" as an outcome for "*Chinese*" population samples.

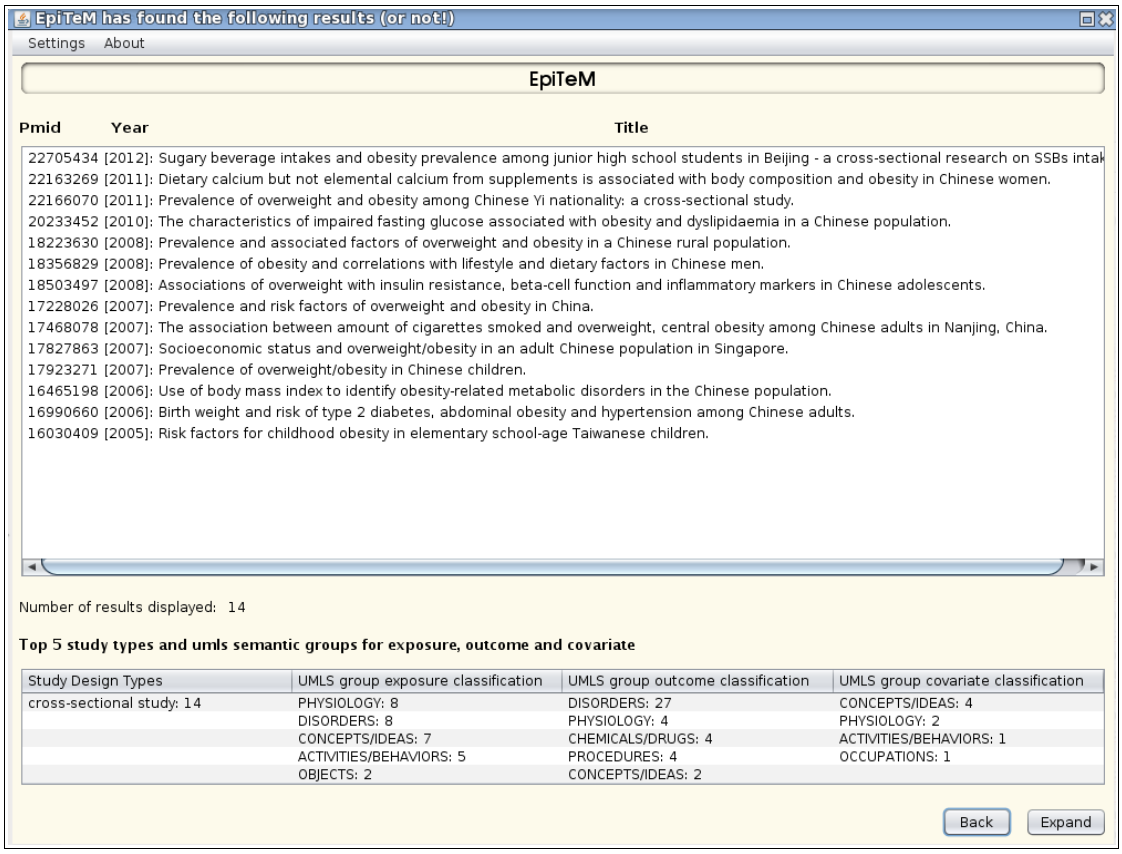


Figure 109: An example of epidemiological data exploration through EpiTeM for various types of cohort studies with "*age*" as a covariate.

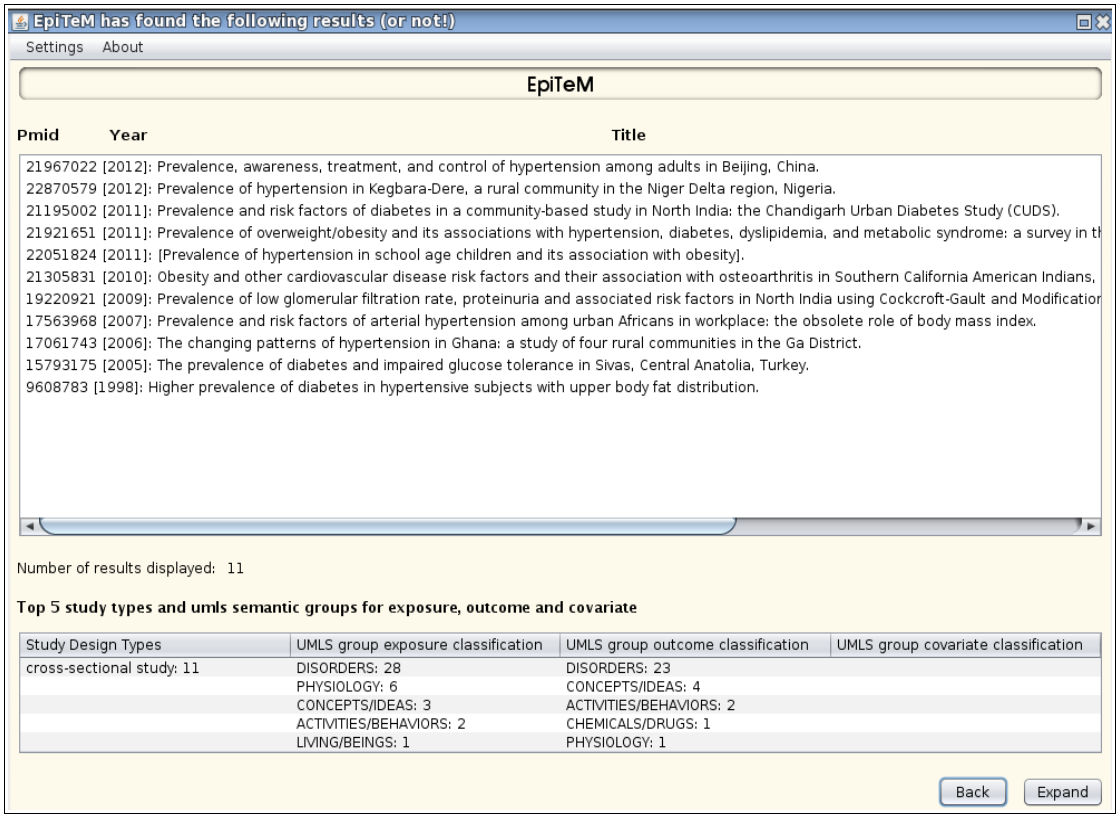


Figure 110: An example of epidemiological data explration through EpiTeM for "*cross-sectional study*" as study design, "hypertension" as an exposure and reported effect size as "*prevalence*".

7.4. Summary

The system described in chapters 4-6 was applied to a large scale corpus of epidemiological abstracts related to obesity as a case study (19,188). At the document level, a total of 98,649 spans were recognized: 6,060 study designs, 13,537 population mentions, 23,518 exposures, 40,333 outcomes, 5,500 covariates and 9,701 effect size concepts.

The most common study design is the observational sub-type "*cross-sectional*" study with a total of 1,940 mentions. Since obesity is a not a rare disease, it seems appropriate for most epidemiological research to follow the cross-sectional design due to its relatively low cost and applicability in large scale data. "O*bservational*" is the most prevalent study design (4,586 documents) in comparison with the experimental that has 227 citations since obesity can affect various populations from all geographical areas and its onset as well progression can be investigated through time and observation of population samples.

For the characteristic of population, 5,521 have been normalized for their age, 5,309 for their gender, 2,137 for their nationality and 689 for their ethnicity. A total of 107 and 11 distinct nationalities and ethnicities are detected respectively. "*Chinese*" is the most studied nationality and ethnicity in the corpus while the continent with the most mentioned nationalities is Europe (a total of 37 nationalities). 377 documents have both the ethnicity and nationality of their population identified, cases in which either the ethnicity is a part of the recognized nationality e.g., "*African-American*" or the ethnicity is the same with the nationality e.g., "*Chinese*".

The most frequently mentioned exposure and outcome concept is "*obesity*". However, since obesity has been used as a case study to illustrate the methodology, it is understandable why concepts representing it and its subsequent measures e.g., "*body mass index*", "*waist circumference*", "*overweight*", "*obesity*", etc have been extracted many times from multiple documents either as exposures or outcomes. The removal of these concepts from the returned identified results of the exposure and outcome characteristics, can reveal a clearer picture regarding the potentially associated-with-any-way-to-obesity concepts. Most epidemiological studies are focusing on research of organism attributes, individual behaviour concepts and various disorders as exposures with the most frequent being "*age*". The most prevalent UMLS semantic group for the characteristic exposure is "*disorders*" while the most frequently observed UMLS semantic category is "*disease or syndrome*". The number of unique exposure concepts is 7,072.

The majority of identified outcomes in the epidemiological literature related to obesity are disorder-related e.g., "*hypertension*", "*type 2 diabetes*", which further highlights that obesity is a complex disease sharing underlying relationships with various other disorders and when

treated, it should be taken into consideration as a part of a disease cluster rather than an isolated concept. Similarly to exposures, for the characteristic of outcome, "*disorders*" is the UMLS semantic group with the largest number of mentions with "*disease or syndrome*" being the top UMLS semantic category. A total of 9,301 unique outcome concepts are observed. The total number of outcomes is almost double than the number of exposures indicating that in each abstract there is potentially one exposure concept and multiple outcome ones.

The most frequent covariate for which studies are adjusting their results for is "*age*". The overall number of covariates is way smaller in comparison to those of exposure and outcomes. This indicates that most studies tend to mention a single exposure in relation to multiple outcomes with a limited number of specific covariate spans. "*Physiology*" is the UMLS semantic group with the most covariate mentions while the semantic category "*organism attribute*" has the highest number of covariates. 1,234 is the number of unique covariates.

Effect size mentions has 6,421 spans with recognized confidence interval, 5,261 with the identified related concept while a total of 5,474 mentions have the respective effect size type detected. Since the most frequent study design is the "*cross-sectional*" type, it is not surprising that odds ratio is the prevalent effect size measure with a total of 3,213 mentions.

Temporal analysis in the characteristics of exposure, outcome and covariate was performed. The period between the years 1965 and 1989 is excluded due to the limited information obtained for almost all of the identified concepts. The exposure of "*metabolic syndrome*" was observed to have an increase in document mentions towards the later years (2006-2011), explaining the drop in mentions of concepts such as "*diabetes*" and "*hypertension*" and suggesting the shift of epidemiological study to the investigation of disease clusters rather than each one separately. The top five most frequent concepts detected as outcomes are diseases/disorders ("*hypertension*", "*diabetes*", "*metabolic syndrome*", "*cardiovascular disease*") or related to mortality ("*mortality*"). This indicates that studies around obesity aim to understand health consequences on affected populations and other potential clinical outcomes that can be attributed to obesity through existing co-morbidities. "M*etabolic syndrome*" is the top most frequent concept in the latter years (2006-2011) suggesting that epidemiological research is inspecting the nature of obesity and its co-related morbidities through the examination of the metabolic syndrome. As for covariate concepts, the most mentioned concepts through the whole period are "*age*", "*gender*", "*body mass index*" and "*smoking*". "*Education*" appears steadily in the 5th position since the dawn of the new millennium revealing that until its role to the problem of obesity becomes more stable and clear, it is logical for most studies to consider the level of education a covariate during their inspection of obesity related concepts.

Additionally, the most common pairs of concepts between certain key characteristics are studied in order to potentially reveal any underlying associations between obesity related concepts as well as shown a cluster of concepts that can have a potential effect in the relationships of exposure-outcome that epidemiological studies adjust their results for.

- **Exposure-outcome:** obesity has been studied as an outcome with a variety of exposures that include organism attributes such as "*body mass index*", individuals behaviours such as "*physical activity*" and disorders like "*diabetes type 2"*. However, the majority of exposure-outcome pairs reveal that obesity (and its substitute concepts e.g., body mass index) has been examined as a risk factor (exposure) for a number of diseases related to metabolic syndrome.
- **Exposure-covariate:** "*age*" is the most frequent covariate in the pairs of exposure-covariate since it has an unknown effect to the onset of obesity-associated disorders, therefore contributing to the complexity of the relationships that obesity shares with other diseases. "G*ender*", "*smoking*", "*body mass index*" and "*education*" are considered as covariates in the studies researching obesity (and its variants – overweight, body mass index, etc) as an exposure, further strengthening the links between the complex disease of obesity with these variables that can potentially affect to an unknown effect its relationship with health-related outcomes.
- **Outcome-covariate:** "a*ge*" is the most frequent covariate that studies are adjusting their results for when they examine as outcomes a variety of obesity associated concepts. Most of the age related outcomes seems to be disorders suggesting that "*age*" can have an potentially unknown effect towards their onset and progression in a population sample.

Finally, using the generated concept map as a backbone, a user interface was designed and implemented in order to explore the results from the application of the epidemiological text mining method. EpiTeM offers the ability to navigate and manipulate easily this large amount of data, enabling the user to search for a particular characteristic or term and collect the necessary subset of information from a vast corpus of epidemiological research.

Chapter 8

Conclusions and Future Work

"I think and think for months and years.
Ninety-nine times, the conclusion is false. The
hundredth time I am right."
Albert Einstein, 1947

Epidemiological studies contain information that could improve the understanding of the concept complexity of a health problem and are important sources for evidence based medicine. However, epidemiologists experience difficulties in recognising key characteristics in related research due to the growing amount (and abundance) of published literature they have to explore. In addition, the biomedical domain is known for its complex language with the use of multiple acronyms, abbreviations and synonyms. Consequently, the task of identifying key epidemiological characteristics becomes burdensome and time consuming.

The main aim of this project was to develop a methodology for the extraction of key characteristics from epidemiological literature (both observational and experimental studies) in order to explore concepts related to a health care problem, such as obesity. The approach aims to decrease the required time and effort for an epidemiologist or clinical professional to navigate the large amounts of literature for the identification of potential concepts of interest as well as contributing to future review and meta-analysis studies. Text mining techniques can assist epidemiologists to identify important pieces of information to detect and integrate key concepts for further research and exploration. On the other hand, concept maps, as a knowledge visualization form focused around a specific topic or theme, can enhance the user's understanding of complex concepts while new associations can be easily made and observed through links. Therefore, the application of concept mapping can represent extracted epidemiological information that summarizes a health care problem in a simple and coherent form, in particular for medical education.

The main hypothesis of this work is that a systematic analysis of knowledge related to a given health care problem through epidemiological text mining can enable a generic framework for the design of associated concept maps through the reduction of the required time for the inspection of targeted information. The case study of obesity was used to highlight the viability of the proposed approach, which could be applied to other problems in the field of epidemiology due to the implementation of rules based on generic and common syntactical patterns observed in epidemiological study text.

Previous work (Hara et al. 2007; Hansen et al. 2008; Kiritchenko et al. 2010; Luo et al. 2012) has mainly focused on the recognition of key elements present in clinical trial text (particularly

RCTs) rather than key epidemiological characteristics (e.g., exposure, outcome, covariate) from different types of epidemiological study, or aimed to extract specific elements for a health care problem from articles (Fizman et al. 2007). Even when studies aimed to extract key characteristic concepts, the research was performed on titles of epidemiological articles rather than study abstracts that potentially can contain more detailed information (Xu et al. 2010). These efforts have not applied any normalization to targeted characteristics thus preventing any large scale aggregation of epidemiological information. Additionally, various attempts have been conducted for the automatic generation of a concept map (concept map mining) to represent unique relationships among the concepts (Oliveira et al. 2001; Watson et al. 2005). These efforts used non-epidemiological data of various formats (articles, manuals, educative texts) and created concept maps were used mostly for educational purposes, rather than for research exploration.

8.1. Thesis Contributions

In this thesis, a generic rule based approach was designed, implemented and evaluated for the recognition of six key epidemiological characteristics (study design, population, exposure, outcome, covariate, effect size) in all types of epidemiological study from MEDLINE abstracts that are related to a given health care problem. Various existing lexical resources and results produced from the application of automatic term recognition were combined with the rule sets to identify the characteristic mentions in text. The extracted spans of study design, population and effect size concepts were normalized for their attributes using a set of rules, while identified exposure, outcome and covariate concepts were normalized to respective UMLS semantic group and category. The normalized mentions of exposure, outcome and covariate concepts were then represented in the form of a concept map for concept and relationship exploration, learning and validation of epidemiological knowledge.

The overall contributions of this thesis are:

1. A rule based approach combined with already existing and newly created lexical resources that enables the identification of key characteristics from epidemiological study abstracts related to a specific health problem.
2. A methodology for the normalization of extracted key characteristic mentions to their respective attributes:
    - study design has multiple attributes according to each particular study type performed (e.g., retrospective cohort study);
    - population has age, nationality, ethnicity and gender (e.g., 250 Chinese students aged 20 years old);

- effect size has effect size value, confidence interval, related concept and effect size type (e.g., odds ratio 2.13).

3. An existing branch of an ontology of clinical research (OCRe) was expanded in order to include all the necessary epidemiological study designs and was used for the normalization of identified study designs mentions.
4. The creation of methods that enable the visualization, exploration and validation of normalized epidemiological key information. A GUI was designed and implemented in order to browse and navigate normalized characteristic mentions from epidemiological abstracts related to a health problem. A methodology for the automatic construction of a concept map that represents identified and normalized mentions of key epidemiological characteristics (exposure, outcome and covariate) was developed.
5. These contributions have been developed as generic approaches, but have been evaluated using the case study of obesity to demonstrate how the developed method can be used to understand its complexity. Obesity is considered one of the most important health problems of the 21st century, with its rates increasing at an alarming rate worldwide. Due to its complex nature, a number of studies have studied its underlying risk factors and outcomes. The methodology was applied to a large dataset (19,188 MEDLINE epidemiological abstracts) and the extracted data was analysed.

Each of the contributions is explained in more detail below.

8.1.1. Identification of Key Characteristics from Epidemiological Study Abstracts

A generic rule based approach was designed and implemented for the identification of key characteristics in MEDLINE epidemiological study abstracts. In order to recognize biomedical concepts (single and multi-word), lexical resources and a total of 412 rules were designed for the identification of:

- **study design:** We used specific dictionaries that contain various epidemiological study types since these spans are represented a relatively small set of terms.
- **population:** A variety of rules based on propositions and verbs along with a controlled set of noun phrases observed in text indicating the presence of a population sample, were utilized.
- **exposure and outcome:** Most exposures and outcomes are biomedical concepts linked together with a particular set of expressions that suggest an association between these concepts. Therefore, rules were designed based on noun phrases and a cluster of verbs, adverbs and prepositions that indicate a relationship among two concepts

combined with the Specialist lexicon along with the results of the automatic term recognition to capture any type of phenomena that are mentioned.

- **covariate:** Rules were created according to a common set of noun phrases, verbs and idioms that express the adjusting of epidemiological studies' results for certain concepts. These were combined with the Specialist lexicon along with the results of the automatic term recognition.
- **effect size:** Since effect size spans are based on regular patterns of expressions (effect size related concept, type, value, confidence interval) that suggest the presence of an effect size mention in text, rules based on a limited number of noun phrases (that indicate the effect size type and confidence interval), numbers (effect size and confidence interval value) and prepositions were created.

For the evaluation of the rule based method, a golden standard corpus of 40 articles was created for development and testing. The evaluation process revealed precision between 79.3%-100.0%, with recall between 80.3%-100.0%. Study design extraction had the best precision (100.0%), and outcome had the lowest precision (79.3%). Covariate mining returned the best recall (100.0%) while exposure had the lowest recall (80.3%). The above values suggested reliable results in the detection of key characteristic mentions in epidemiological abstracts.

8.1.2. Normalization of the Extracted Key Characteristic Mentions

The identified concepts of key characteristics are normalized into their attributes, semantic groups and categories at the document level. More specifically for the normalization of:

- study design characteristic, an adapted version of the clinical research ontology (OCRe) was used. OCRe was chosen because it provided the default designs in which experimental and observational research could be performed. The recognised study designs were mapped to OCRe nodes (through a string comparison), while any detailed information is considered an attribute;
- population characteristic, where a cluster of regular expressions and dictionaries were used to recognize specific attributes (age, gender, nationality, ethnicity).
- exposure, outcome and covariate characteristics, MetaMap was used to normalize and classify to the related UMLS semantic groups and categories the identified concepts. Biomedical classification is required in order to assist health professionals when it comes to the application of an '*epidemiological sieve*', which could help them decide if they are going to include abstracts for a more detailed inspection.
- effect size characteristic, regular expressions were applied for the identification of

attributes that mostly describe the related concepts (effect size value, effect size type, confidence interval and related concept).

The accuracy overall was above 91.0% in both the evaluation set and the random sample at the document level, suggesting reliable normalizations. The evaluation of the normalization procedure (MetaMap) of the exposure, outcome and covariate mentions has already been presented separately and, thus, it was not included in the thesis.

8.1.3. Automatic Construction of a Concept Map and EpiTeM

A method was designed for the automatic creation of a concept map from the normalized exposure, outcome and covariate mentions. Through an automatic process, a concept map is generated by CmapTools. The concept map is structured and separated into three main clusters, each one representing one type of characteristic's mentions along with their respective UMLS semantic groups and categories. The map can be used for relationship exploration among the presented concepts as well as epidemiological knowledge validation, and is used as a backbone in EpiTeM.

EpiTeM was designed in order to enable the navigation and exploration of epidemiological text mining results in studies related to a health problem. It enables the user to search for a specific characteristic or term, and collect the necessary subset of documents from a corpus of epidemiological research through the interaction of a MySQL database that contains the extracted and normalized concepts. It displays the recognized information (highlighted in text) as well as the normalized mentions of each characteristic for each abstract.

8.1.4. Obesity as Case Study

The rule based method was applied to a large scale corpus consisting of 19,188 epidemiological abstracts returned by the "*obesity/epidemiology[mesh]"* query. A total of 98,649 concepts at the document level were identified.

The normalized results revealed that "*cross-sectional*" design was the epidemiological study type with the most mentions (32.0%), while 40.5% population spans have been normalized in their age, 39.2% in their gender, 15.7% in their nationality and 5.0% in their ethnicity. The most prevalent UMLS semantic group in the characteristics of exposure and outcome was "*disorders*", while the UMLS semantic category with the most concepts was "*disease or syndrome*". Surprisingly, the prevalent UMLS semantic group for covariates was "*physiology*" and the respective category was "*organism attribute*". Almost two thirds of effect size mentions include confidence interval, whereas 54.2% contain a related concept. More than half

of all effect size mentions have been normalized in their type with the most prevalent effect size being odds ratio (33.1%).

Despite having a relatively good recall in the study design characteristic, only in 33% of the corpus were designs identified. A review of 50 randomly selected articles that did not include the characteristic revealed that there was no reporting of the study design. The number of outcome mentions was almost half of the total identified mentions in the corpus, suggesting that most studies may contain more than one outcome of interest or various synonyms of a particular concept. Similarly to study design, only in 29% of the corpus, covariate mentions were recognized. The inspection of 50 randomly selected abstracts with no identified covariate mentions, concluded that no related concepts were (indeed) present in text suggesting that most likely covariates are reported in full text rather than in the abstract.

Through a temporal analysis of the top five most frequent exposures, outcomes and covariates, it was observed that diseases/disorders have been studied as risk factors to obesity or are potentially related in an underlying (and probably not fully explored) clinical relationship. Four of five outcomes were disease related, indicating a strong relationship between the obesity epidemic and multiple disorders that may exist. Most of the covariate involved concepts are attributes of the targeted for epidemiological study population. The mentions of exposure, outcome and covariate seemed to be increasing towards the new century indicating the research attention that obesity has received in order to understand its complex associations with a variety of concepts. The inspection of the most common concept pairs between exposures, outcomes and covariates was done aiming to reveal the focus of the research community for obesity related studies. Obesity has been studied as an outcome with a variety of exposures that include organism attributes, individuals behaviours and disorders. The majority of exposure-outcome pairs reveal that obesity (and its substitute concepts) has been examined as a risk factor (exposure) for a number of diseases related to metabolic syndrome. For both exposure-covariate and outcome-covariate pairs, the majority of studies were adjusting their results for demographic attributes or individual activities with "*obesity-age*" being the most frequent one.

8.2. Limitations, Challenges and Future Work

Suggestions for future work regarding the limitations and challenges encountered are discussed below:

1. The system is currently used on abstracts only. A next step can be the application of this methodology to full-text epidemiological documents, and its performance could be

compared to that one returned when used on study abstracts. It would be interesting to explore whether this approach would improve the overall accuracy of the extraction procedure or whether it is already accurate enough to perform characteristic extraction from abstracts only. It would be interesting to explore if document level performance can be improved through the identification of mentions of epidemiological characteristics in full text, or if the use of full text would introduce more noise (e.g., mentions (comparisons to) of other epidemiological studies).

2. The methodology proposed in this thesis does not take into consideration the structure of abstract for the identification of epidemiological characteristics. The role of structure e.g., the existence of headings (e.g., "*Introduction*", "*methods*", etc.) could be an important feature, while the current methodology is relying on rules that detect mentions in sentences, it could be of interest to identify key characteristics in abstracts while taking into consideration the position in which they were detected. This approach could assist in the detailed capture of certain characteristic mentions (e.g., "*methods*" for study design, "*participants*" for population, etc.), potentially differentiate the confusion between exposure and outcome and improve the overall accuracy of the current system.
3. An exploration of the semantic relationships between exposures and outcomes could be conducted in titles of epidemiological abstracts. Most titles of epidemiological citations mention exposure and outcome of interest. However, it is not clear which concept is the studied variable. Studying the patterns could help understand and potentially differentiate these two characteristics from each other. Through a semi-automatic epidemiological workbench, the time and effort required to cluster the concepts in the relevant characteristic will be more efficient.
4. The methodology could be used to produce automatically concept maps from epidemiological literature related to a particular health care problem. These generated concept maps could be used to update existing structured information available. Any observed differences between these two would contribute to the understanding of the disorder's complexity and may point out existing knowledge gaps between the scientific literature and the constructed maps from a group of clinical professionals.
5. Currently there is not a "flexible" approach for the of population attributes. The recognition of more specific demographic information regarding the study populations could assist to obtain a clearer picture of a disease. More specifically, besides the identification of attributes such as age, gender, nationality and ethnicity, the recognition of the geographical areas in which the population samples are residing or the related occupations, could potentially reveal interesting aspects of a disease. Identifying urban or not areas that potentially might prevent or contribute to the onset

of a disease can assist in the recognition of underlying risk factors for a health care problem. The combination with detected occupations and recognized ethnicities could shown what groups of individuals are susceptible to a disease or not as well as providing statistical information that could be used for future meta-analysis and reviews.

Overall, this dissertation presents a step towards epidemiological text mining and concept map generation by providing a rule based approach that identifies six key characteristics from epidemiological studies of various types along with additional attributes such as their UMLS semantic classification and by automatically representing these results in the form of a concept map for data manipulation and further exploration.

Appendix 1

List of Nationalities

List of nationalities taken from the following web pages **http://www.people.groups.org/** and **http://en.wikipedia.org/wiki/Lists_of_people_by_nationality**

Algerian	Chadian	Herzegovinian	Micronesian	Seychellois
American	Chilean	Honduran	Moldovan	Sierra Leonean
Andorran	Chinese	Hungarian	Monacan	Sikhs
Angolan	Colombian	I-Kiribati	Mongolian	Singaporean
Antiguans	Comoran (Comorian)	Icelander	Moroccan	Slovak
Argentine	Congolese	Indian	Mosotho	Slovakian
Argentinean	Costa Rican	Indonesian	Motswana	Slovene (Slovenian)
Armenian	Croatian	Iranian	Mozambican	Solomon Islander
Aromanian	Cuban	Iraqi	Namibian	Somali
Aruban	Cypriot	Irish	Nauruan	South African
Australian	Czech	Israeli	Nepalese	South Korean
Austrian	Dane	Italian	Netherlander	Spanish
Azerbaijani	Danish	Ivorian	New Zealander	Sri Lankan
Azeri	Djibouti	Jamaican	Ni-Vanuatu	Sudanese
Bahamian	Dominican (dominicans)	Japanese	Nicaraguan	Surinamer (surinamese)
Bahraini	Dutch	Jordanian	Nigerian (nigerien)	Swazi
Baltic German	Dutchman	Kawaiti	North Korean	Swede (swedish)
Baltic Russian	Dutchwoman	Kazakh	Northern Irish	Swiss
Bangladeshi	East timorese	Kazakhstani	Norwegian	Syrian
Barbadian	Ecuadorean (Ecuadorian)	Kenyan	Omani	Taiwanese
Barbudans	Egyptian	Kittian and Nevisian	Pakistani	Tajik
Batswana	Emirian	Korean	Palauan	Tanzanian
Belarusian	Equatorial Guinean	Kosovo Albanian	Palestinian	Thai
Belgian	Eritrean	Kuwaiti	Panamanian	Tibetan
Belizean	Estonian	Kyrgyz	Papua New Guinean	Tobagonian
Beninese	Ethiopian	Lao	Paraguayan	Togolese
Bermudian	Fijian	Laotian	Persianivoirian	Tongan
Bhutanese	Filipino	Latvian	Peruvian	Trinidadian
Boer	Finnish	Lebanese	Pole	Tunisian
Bolivian	Finnish-Swedish	Liberian	Polish	Turk (turkish)
Bosnian	French	Libyan	Portuguese	Turkish Cypriot
Brazilian	Gabonese	Liechtensteiner	Puerto Rican	Tuvaluan
Breton	Gambian	Lithuanian	Qatari	Ugandan
British	Georgian	Luxembourger	Quebecer	Ukrainian
British Virgin Islander	German	Macedonian	Romanian	Uruguayan
Bruneian	Ghanaian	Malagasy	Russian	Uzbekistani
Bulgarian	Gibraltar	Malawian	Rwandan	Vanuatuan
Burkinabe	Greek	Malaysian	Saint Lucian (St Lucian)	Venezuelan
Burmese	Grenadian	Maldivan (maldivian)	Salvadoran	Vietnamese
Burundian	Guatemalan	Malian	Samoan	Welsh
Cambodian	Guianese	Maltese	San Marinese	Yemenite (Yenemi)
Cameroonian	Guinea-Bissau National	Manx	Sao Tomean	Zambian
Canadian	Guinea-Bissauan	Marshallese	Saudi	Zimbabwean
Cape Verdean	Guinean	Mauritanian	Scottish	

Appendix 2

List of Ethnicities

List of the most common ethnicities of the UK Office of National Statistics **(http://www.ons.gov.uk/ons/index.html)** and the United States Census and the United States Census **(http://www.census.gov/).**

African American
Alaska Native
American Indian
Arab
Arab American
Asian British
Bangladeshi
Black
Black African
Black Caribbean
Chinese
Hawaiian Native
Hispanic
Indian
Native American
Oceanic American
Other Asian
Other Black
Other Mixed
Other White
Pacific Islander
Pakistani
White British
White Irish
White and Black African
White and Black Caribbean

References

1. Aarts S, Vos R, van Boxtel MP, Verhey FRJ, Metsemakers JF, van den Akker M. Exploring medical data to generate new hypotheses: an introduction to data and text mining techniques in epidemiology. Multimorbidity in general practice: Adverse health effects and innovative research strategies (2012).

2. Abacha AB and Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach. Journal of biomedical semantics 2011b, 2(Suppl 5):S4 [http://www.jbiomedsem.com/content/2/S5/S4].

3. Adami H, Trichopoulos D. Obesity and Mortality from Cancer. New England of Medicine, 348:1623-1624, April 24 2003, Number 17.

4. Adams FK, Schatzkin A, Harris BT, Kipnis V, Mouw T, Ballard-Barbash R, Hollenbeck A, Leitzmann FM. Overweight, Obesity, and Mortality in a Large Prospective Cohort of Persons 50 to 71 Years Old. Journal of New England of Medicine. 2006 Aug 24;355(8):763-78. Epub 2006 Aug 22.

5. Agirre E, Edmonds PG. Word sense disambiguation: Algorithms and applications. Vol. 33. Springer Science+ Business Media, 2006.

6. Alberti G, Zimmet P, Shaw J, Bloomgarden Z, Kaufman F, Silink M. Type 2 diabetes in the young: The evolving epidemic. Diabetes Care, 27(7):1798–1811, July 2004.

7. Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends in Biotechnology, 24(12):571-579. 2006.

8. Andritsos P. Data clustering techniques. Mar. 13 [Online]. Available: [http://citeseer.ist.psu.edu/607632.html;http://www.cs.toronto.edu/periklis/./pubs/depth.ps.gz], 2002.

9. Apte C, Damerau F, Weiss SM. Text mining with decision trees and decision rules. Workshop on learning from text and the Web, Conference on automated learning and discovery. 1998.

10. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Mashuichi H, Ohe K. TEXT2TABLE: medical text summarization system based on named entity recognition and modality identification. InProceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pp. 185-192. Association for Computational Linguistics, 2009.

11. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, Ohe K. Extraction of adverse drug effects from clinical records. Studies in health technology and informatics, 160:739, 2010.

12. Aronson AR. MetaMap evaluation. Bethesda MD: National Library of Medicine, 2001. [http://skr.nlm.nih.gov/papers/references/mm.evaluation.pdf].

13. Aronson AR, Lang FM. An Overview of MetaMap: Historical Perspective and Recent Advances. J Am Med Inform Assoc. 2010 May-Jun; 17(3) : 229-36.

14. Ausubel DP, Novak JD, Hanesian H. Educational psychology: A cognitive view. (1968).

15. Atlantis E, Baker M. Obesity effects on depression:systematic review of epidemiological studies. Int J Obes (Lond). 2008 Jun;32(6):881-91. Epub 2008 Apr 15.

16. Azfal H. A literature-based framework for semantic descriptions of E-science resources. Thesis. University of Manchester. 2009.

17. Bai SM, Chen SM. Automatically constructing concept maps based on fuzzy rules for adapting learning systems. Expert Systems with Applications 35, no. 1 (2008): 41-49.

18. Baker LJ, Olsen WL, Sørensen IAT. Childhood Body-Mass Index and the Risk of Coronary Heart Disease in Adulthood. Journal of New England of Medicine. Dec 6, 2007. Vol. 357:2329-2337, number 23.

19. Baraldi A, Blonda P. A Survey of Fuzzy Clustering Algorithms for Pattern Recognition—Part I. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: Cybernetics, Vol. 29, No. 6, December. 1999.

20. Barrett MA, Humblet O, Hiatt RA, Adler NE. Big Data and Disease Prevention: From Quantified Self to Quantified Communities. Big Data 1, no. 3 (2013): 168-175.

21. Bashyam V, Divita G, Bennett DB, Browne AC, Taira RK. A normalized lexical lookup approach to identifying UMLS concepts in free text. Studies in health technology and informatics 129, no. Pt 1 (2007): 545.

22. Binder AB. Methodological issues in the meta-analysis of observational studies: a discussion. SSC-JSM 2010.

23. Blake C. A text mining approach to enable detection of candidate risk factors. Medinfo (2004).

24. Brandt P, Voorrips L, Hertz-Picciotto I, Shuker D, Boeing H, Speijers G, Guittard C, Kleiner J, Knowles M, Wolk A, Goldbohm A. The contribution of epidemiology. Food and Chemical Toxicology 40 (2002) 387-424.

25. Brown CD, Higgins M, Donato KA, Rohde FC, Garrison R, Obarzanek E, Ernst ND, Horan M. Body mass index and the prevalence of hypertension and dyslipidemia. Obesity research Vol.8 No. 9 December 2000.

26. Buchan I, Canoy D. Challenges in obesity epidemiology. Obesity Reviews; 8(suppl 1):1–11, 2007.

27. Burke JG, O'Campo P, Peak GL, Gielen AC, McDonnell KA, Trochim WMK. An introduction to concept mapping as a participatory public health research method. Qualitative health research15, no. 10 (2005): 1392-1410.

28. Calderon LR. Measuring risks in humans: the promise and practice of epidemiology. Food and Chemical Toxicology 38 (2000) S59-S63.

29. Calle EE, Rodriguez C, Walker-Thurmond K, Thun MJ. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of US adults. The New England journal of medicine, 348(17):1625–1638, April 2003a.

30. Calle EE, Rodriguez C, Walker-Thurmond K, Thun MJ. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. Adults. N Engl J Med 348;17, April 24, 2003b.

31. Calle EE, Kaaks R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. Nature reviews, Cancer. Vol. 4, August 2004.

32. Cañas AJ, Carvalho M. Concept maps and AI: An unlikely marriage? Proceedings of SBIE 2004: Simpósio brasileiro de informática educativa. Manaus, Brasil: SBC.

33. Canas JA, Carff R, Hill G, Carvalho M, Arguedas M, Eskridge CT, Lott J, Carvajal R. Concept maps: Integrating knowledge and information visualization. Pages 205–219. 2005.

34. Cano C, Blanco A, Peshkin L. Automated identification of diagnosis and co-morbidity in clinical records. Methods Inf Med, 2009.

35. Canoy D. Distribution of body fat and risk of coronary heart disease in men and women. Curr Opin Cardiol. 2008 Nov;23(6):591-8.

36. Canoy D. Coronary heart disease and body fat distribution. Curr Atheroscler Rep. 2010 Mar;12(2):125-33.

37. Centre for Disease, Control and Prevention (CDC). Obesity, [http://www.cdc.gov/vitalsigns/AdultObesity/], 2012.

38. Chapman WW, Cohen KB. Current issues in biomedical text mining and natural language processing. Journal of Biomedical Informatics, 42(5):757–759, October. 2009.

39. Chapple M. E/R model, [http://databases.about.com/cs/specificproducts/g/er.htm], 2010.

40. Chen PP. The entity-relationship model: Toward a unified view of data. ACM Transactions on Database Systems, 1:9–36. 1976.

41. Chen PP. English sentence structure and entity-relationship diagrams. Information Sciences, 29(2-3):127–149, 1983.

42. Chen PP. Entity-relationship modeling: historical events, future trends, and lessons learned. Pages 296–310. 2002.

43. Chen NiS, Wei CW, Chen HJ. Mining e-learning domain concept map from academic articles. Computers & Education 50, no. 3 (2008): 1009-1021.

44. Chen SM, Bai SM. Using data mining techniques to automatically construct concept maps for adaptive learning systems. Expert Systems with Applications 37, no. 6 (2010): 4496-4503.

45. Chiang JH, Lin JW, Yang CW. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). J Am Med Inform Assoc. 2010 May-Jun;17(3):245-52.

46. Chowdhury MF, Lavelli A. Disease Mention Recognition with Specific Features. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 83-90.

47. Chung YG. Sentence Retrieval for Abstracts of Randomized Controlled Trials. BMC Medical Informatics and Decision Making 2009a, 9:10, doi:10.1186/1472-6947-9-10.

48. Chung YG. Towards identifying Intervention Arms in RCTs: Extracting Coordinating Constructions. J Biomed Inform 2009b, 42(5):790-800.

49. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model, J Biomed Inform 42 (5) (2009), pp. 937–949.

50. Cohen WW. MinorThird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data, http://github.com/TeamCohen/MinorThird/, 2004.

51. Cohen MA, Hersh RW. A survey of current work in biomedical text mining. Brief Bioinform, 6(1):57–71, January. 2005.

52. Cohen, KB, Johnson H, Verspoor K, Roeder C, Hunter L. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC bioinformatics 11, no. 1 (2010): 492.

53. Cukier K. The Economist: A special report on managing information. Feb 27, [http://www.economist.com/], 2010.

54. Dabbagh N. Concept mapping as a mindtool for critical thinking. Journal of Computing in Teaching Education. 17(2),16-23.

55. Davy KP, Hall JE. Obesity and hypertension. Am J Physiol Regul Integr Comp Physiol 286: R803-R813, 2004.

56. De Bruijn B, Carini S, Kiritchenko S, Martin J, Sim I. Automated Information Extraction of Key Trial Design Elements from Clinical Trial Reports. AMIA Annual Symposium 2008, 141-145.

57. De Bruijin B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc 2011;18:557-562. doi:10.1136/amiajnl-2011-000150.

58. de la Villa M, Aparicio F, Mana MJ, de Buenage M. A learning support tool with clinical cases based on concept maps and medical entity recognition. Proceedings of the 2012 ACM international conference on Intelligent User Interfaces. ACM, 2012.

59. de Koning L, Merchant AT, Pogue J, Anand SS: Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: meta-regression analysis of prospective studies. Eur Heart J 2007, 28(7):850-856.

60. de Wit L, Luppino F, van Straten A, Pennix B, Zitman F, Cuijpers P. Depression and obesity: a meta-analysis of community-based studies. Psychiatry Research 178 (2010) 203-2335.

61. Deleger L, Grouin C, Zweigenbaum P. Extracting Medical Information from Narrative Patient Records : the Case of Medication-related Information. J Am Med Inform Assoc , 17 (5), 555–558. 2010.

62. Demark-Wahnefried W, Platz EA, Ligibel Ja, Blair CK, Courneya KS, Meyerhardt JA, Ganz PA, Rock CL, Schmitz KH, Wadden T, Philip EJ, Wolfe B, Gapstur SM, Ballard-Barbash R, McTiernan A, Minasian L, Nebeling L, Goodwin PJ. The role of obesity in cancer survival and recurrence. Cancer Epidemiol Biomarkers Prev;21(8) August 2012.

63. Denny JC., Miller RA, Spickard III A, Schildcrout J, Darbar D, Rosenbloom ST, Peterson JF. Identifying UMLS concepts from ECG Impressions using KnowledgeMap. InAMIA Annual Symposium Proceedings, vol. 2005, p. 196. American Medical Informatics Association, 2005.

64. Dickman S. Tough mining: the challenges of searching the scientific literature. PLoS Biol., 1, E48. 2003.

65. Duncan M, Griffith M, Rutter H, Goldacre JM. Certification of Obesity as a Cause of Death in England 1979-2006. The European Journal of Public Health Advance Access. 2010.

66. Eckel RH, Krauss RM. American Heart Association call to action: obesity as a major risk factor for coronary heart disease. Circulation. 1998;97:2099-2100.

67. Ekman A, Litton JE. New times, new needs; e-epidemiology. Eur J Epidemiol (2007) 22:285-292.

68. Erhardt AR, Schneider R, Blaschke C. Status of text-mining techniques applied to biomedical text. Drug Discovery. Today, 11:315–325. 2006.

69. Exforcys Inc. E/R model, [http://www.exforsys.com/tutorials/datamodeling/entity-relationship-model.html], 2013.

70. Faith SM, Matz PE, Jorge MA. Obesity-depression associations in the population. Journal of Psychosomatic Research 53 (2002) 935-942.

71. Faith MS, Butryn M, Wadden TA, Fabricatore A, Nguyen AM, Heymsfield SB. Evidence for perspective associations among depression and obesity in population-based studies. Obesity reviews (2011) 12, e438-e453.

72. Farhat T, Iannotti JR, Simons-Morton GB. Overweight, obesity, youth, and health-risk behaviors. American journal of preventive medicine, 38(3):258–267, March 2010.

73. Ferreira JD, Pesquita C, Couto FM, Silva MJ. Bringing epidemiology into the Semantic Web. In International Conference on Biomedical Ontologies (ICBO). 2012.

74. Filannino M, Brown G, Nenadic G. ManTIME: Temporal expression identification and normalization in TempEval-3 challenge . Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013).

75. Fiszman M, Chapman W, Aronsky D, Evans SR, Haug JP. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 7 (2000), pp. 593–604.

76. Fiszman M, Rosemblat G, Ahlers CB, Rindflesch TC. Identifying Risk Factors for Metabolic Syndrome in Biomedical Text . AMIA Annu Symp Proc. 2007; 2007: 249–253.

77. Flynn JT, Falkner BE. Obesity hypertension in adolescents: epidemiology, evaluation, and management. The Journal of Clinical Hypertension Vol 13, No 5, May 2011.

78. Ford ES, Mokdad AH. 2008 Epidemiology of obesity in the Western hemisphere. J Clin Endocrinol Metab, November 2008, 93(11):S1-S8

79. Fox C. A stop list for general text. In ACM SIGIR Forum, vol. 24, no. 1-2, pp. 19-21. ACM, 1989.

80. Franks, Hanson, Knowler, Sievers, Bennett, Looker. Childhood obesity, other cardiovascular risk factors, and premature death. Journal of New England of Medicine. 362:485-93, 2010.

81. Frantzi K, Ananiadou S. Automatic Term Recognition using Contextual Cues. Proceedings of 3rd DELOS Workshop, Zurich, Switzerland 1997.

82. Friedman C. (2009) Discovering novel adverse drug events using natural language processing and mining of the electronic health record. Artificial Intelligence in Medicine. 2009:1–5.

83. Friedlin J, McDonald CJA. Natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc (2006), pp. 269–273.

84. Fukuda K, Tamura A, Tsunoda T, Takagi T. Toward information extraction: identifying protein names from biological papers. In: Proceedings of Pacific Symposium on Biocomputations. 1998. pp. 707–18.

85. Gerner M, Nenadic G, Bergman CM. LINNAEUS: a species name identification system for biomedical literature. BMC Bioinformatics 11:85. 2010.

86. Gonzalez LH, Palencia PA, Umana AL, Galindo L, Villafrade AL. Mediates learning experience and concepts maps: a pedagogical tool for achieving meaningful learning in medical physiology students. Advanced Physiology Education . Dec;32(4):312-6. 2008.

87. Graphic.org. Concept map, [http://www.graphic.org/concept.html], 2013.

88. Gaines RB, Shaw LGM. Concept maps as hypermedia components. Int. J. Hum.-Comput. Stud., 43(3):323–361, September. 1995.

89. Grundy SM. Obesity, metabolic syndrome, and cardiovascular disease. The Journal of Clinical Endocrinology & Metabolism 89(6):2595-2600. 2004.

90. Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of emerging Technologies in Web Intelligence, Vol. 1. No.1 August. 2009.

91. Gupta N, Goel K, Shah P, Misra A. Childhood obesity in developing countries: Epidemiology, determinants, and prevention. Endocrine Reviews, February 2012, 33(1):0000-0000.

92. Gurulingappa H, Klinger R, Hofmann-Apitius M, Fluck J. An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. 2010.

93. Gurulingappa H, Hofmann-Apitius M, Fluck J. Concept identification and assertion classification in patient health records. Workshop Challenges in Natural Language Processing for Clinical Data <4, 2010, Washington/DC>. 2011.

94. Halgrim S, Xia F, Solit I, Cadag E. Extracting information from discharge summaries. Proceedings of the NAACL HLT 2010 second Louhi Workshop on Text and Data Mining of Health Documents, pages 61-67.

95. Hall RH, Dansereau DF, Skaggs LP. Knowledge maps and the presentation of related information domains. The Journal of Experimental Education 61, no. 1 (1992): 5-18.

96. Hammarlund S, Catharina MH, Nilsson MH, Hagell P. Measuring outcomes in Parkinson's disease: a multi-perspective concept mapping study. Quality of Life Research 21, no. 3 (2012): 453-463.

97. Hamon T, Grabar N. Linguistic approach for identification of medication names and related information in clinical narratives. J Am Med Inform Assoc , 17 (5), 549–554. 2010.

98. Hansen JM, Rasmussen ON, Chung G. Extracting Number of Trial Participants from Abstracts of RCTs. Proceedings of TromsØ Telemedicine and e-Health Conference: 9-11 June 2008; TromsØ, Norway 2008.

99. Hara K, Matsumoto Y. Extracting Clinical Trial Design Information from MEDLINE Abstracts. New Generation Computing, 25(2007)263-275.

100. Harrigton B. A semantic network approach to measuring relatedness. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 356-364. 2010.

101. Hartley TR, Barnden AJ. Semantic networks: visualizations of knowledge. Trends in Cognitive Sciences, 1(5):169–175, August 1997.

102. Hay SI, George DB, Moyes CL, Brownstein JS. Big Data Opportunities for Global Infectious Disease surveillance. PLoS medicine 10, no. 4 (2013): e1001413.

103. Hearst M. What is text mining?, [http://people.ischool.berkeley.edu/~hearst/text-mining.html], 2003.

104. Henry SL, Barzel B, Wood-Bradley RJW, Burke SL, Head GA, Armitage JA. Developmental origins of obesity-related hypertension. Clinical and Experimental Pharmacology and Physiology (2012) 39, 799-806.

105. Heyng-Seon O, John-Beom K, Sung-Hyon M. Extracting targets and attributes of medical findings from radiology reports for cross corpus search. ACM-BCM'11 August 1-3, Chicago, IL, USA. 2011.

106. Hong LL, Aurangzeb K, Baharum B, Khairullah K. A review of machine learning algorithms for text documents classification. Journal of advances in Information Technology. Vol. 1, No.1 February 2010.

107. Hossain P, Kawar B, El Nahas M. Obesity and Diabetes in the developing World – A Growing Challenge. Journal of New England of Medicine. Jan 18;356(3):213-5, 2007.

108. Hossein BM. Patent literature trends in Medline throughout 1965-2005. ACIMED, Ciudad de La Habana, v. 20, n. 2, August. 2009.

109. Hotho A, Nurnberger A, Paaß G. A Brief Survey of Text Mining. LDV Forum – GLDV Journal for Computational Linguistics and Language Technology, 20(Suppl 1):19-62. 2005.

110. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP et al. Big data: The future of biocuration. Nature455, no. 7209 (2008): 47-50.

111. Hsiao MY, Chen CC, Chen JH. Using UMLS to construct a generalized hierarchical concept-based dictionary of brain functions for information extraction from the fMRI literature. J Biomed Inform. 2009 Oct;42(5):912-22. Epub 2009 Apr 22.

112. Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, Willett WC. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. Journal of New England of Medicine 345:790–797, 2001.

113. Huybrechts KF, Mikklsen EM, Christensen T, Riis AH, Hatch EE, Wise LA, Sorensen HT, Rothman KJ. A successful implementation of e-epidemiology: the Danish pregnancy planning study “Snart-Gravid”. Eur J Epidemiol (2010) 25:297-304.

114. Kinchin IM. Using concept maps to reveal understanding: A two-tier analysis. School Science Review 81, no. 296 (2000): 41-46.

115. Kinchin IM, Streatfield D, Hay DB. Using concept mapping to enhance the research interview. International Journal of Qualitative Methods 9, no. 1 (2010): 52-68.

116. Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature, Vol 444, 14 December 2006.

117. Jarrar M, Meersman R. Formal Ontology Engineering in the DOGMA Approach, in Liu Ling & Aberer K. (eds.), Proc. of the Internat. Conf. on Ontologies, Databases and Applications of Semantics (ODBase 02), LNCS 2519, Springer Verlag 2002.

118. Jia H, Lubetkin EI. Trends in quality-adjusted life-years lost contributed by smoking and obesity. American Journal of Preventive Medicine, 38/2, p.138-144, 2010.

119. Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. (2008) Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics. 2008;9(S3).

120. Jurafsky D, Martin JH, Kehler A, Linden KV, Ward N. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Vol. 2. Upper Saddle River: Prentice Hall, 2000.

121. Kano Y, Dobson P, Nakanishi M, Tsuji J, Ananiadou S. Text mining meets workflow: Linking U-Compare with Taverna. First published online. August 12, 2010.

122. Karystianis G, Buchan I, Nenadic G. Mining Characteristics of Epidemiological Studies from Medline: A Case Study in Obesity. J BioMed Sem (in press). 2013.

123. Khoury MJ, KL Tram, Ioannidis JPA , Hartge P, Spitz MR, Buring JE, Chanock SJ et al. Transforming epidemiology for 21st century medicine and public health. Cancer Epidemiology Biomarkers & Prevention 22, no. 4 (2013): 508-516.

124. Kim JD, Tsujii J. Corpora and their annotations. 2006. In Sophia Ananiadou and John McNaught, (Eds), Text Mining for Biology and Biomedicine. 46 Gilingham Street, London SW1V 1AH UK, Artech House, 2006. ISBN 1-5053-984-X.

125. Kiritchenko S, De Bruijn B, Carini S, Martin J, Sim I. ExaCT: Automatic Extraction of Clinical Trial Characteristics from Journal Publications. BMC Medical Informatics and Decision Making 2010, 10:56.

126. Krallinger M, Erhardt RAA, Valencia A. Text mining approaches in molecular biology and biomedicine. Drug Discovery. Today, 10, 439–445. 2005.

127. Krishnan S, Rosenberg L, Djousse L, Cupples A, Palmer JR. Overall and central obesity and risk of type 2 diabetes in U.S. black women. Obesity Vol. 15 No. 7. July 2007.

128. Korhonen A, Séaghdha DO, Silins I, Sun L, Högberg J, Stenius U. Text mining for literature review and knowledge discovery in cancer risk assessment and research. PLoS One 7, no. 4 (2012): e33427.

129. Laight DW. Attitudes to concept maps as a teaching/learning activity in undergraduate health professional education: influence of preferred learning style. Medical Teacher 26: 229–233. 2004.

130. Lanzing J. Concept mapping, [http://users.edte.utwente.nl/lanzing/cm_home.htm], 1997.

131. Larsson SC, Wolk A. Obesity and colon and rectal cancer risk: a meta-analysis of prospective studies. Am J Clin Nutri 2007;86:556-65.

132. Last MJ. A Dictionary of Epidemiology. New York: Oxford University Press 2001:180.

133. Lavie LJ, Milani, RV, Venture HO. Obesity and cardiovascular disease. Journal of the American College of Cardiology, Vol.53, No. 21, 2008.

134. Leake A, leake D. Jump-starting concept map construction with knowledge extracted from documents. In In Proceedings of the Second International Conference on Concept Mapping (CMC. 2006).

135. Leaman R and Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 2008:652-63.

136. Leaman R, Miller C, Gonzalez G. Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark. In 2009 Symposium on Languages in Biology and Medicine, Seowipo-si, Jeju island, South Korea, November 12-13, 2009.

137. Lee CH, Lee GG, Leu Y. Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning. Expert Systems with Applications 36, no. 2 (2009): 1675-1684.Lin X. Automatically-generated Concept Maps as a Learning Tool. WASET 2010 66 (2010).

138. Li D, Kipper-Schuler K, Savova G. Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 94-95.

139. Li Z, Liu F, Antieau L, Cao Y, Yu H. Lancet: a high precision medication event extraction system for clinical text. Journal of the American Medical Informatics Association.2010;17:563–567. doi: 10.1136/jamia.2010.004077.

140. Lin X, Morton L. Visual mapping for medical concepts. AMIA 2003 Symposium Proceedings, P912.Lambiotte JG, Dansereau DF. Effects of knowledge maps and prior knowledge on recall of science lecture content. The Journal of experimental education 60, no. 3 (1992): 189-201.

141. Lin X. Automatically-generated Concept Maps as a Learning Tool. WASET 2010 66 (2010).

142. Lorenzo C, Serrano-Rios M, Martinez-Larrad MT, Gonzalez-Villapando C, Williams K, Gabriel R, Stern MP, Haffner SM. Which obesity index best explains prevalence differences in type 2 diabetes mellitus? Obesity Vol. 15 No. 5 May 2007.

143. Low S, Chin MC, Deurenberg-Yap M. Review on epidemic of obesity. Ann Acad Med Singapore. 2009 Jan;38(1):57-9.

144. Luo Z, Johnson SB, Lai AM, Weng C. Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. In AMIA Annual Symposium Proceedings, vol. 2011, p. 843. American Medical Informatics Association, 2011.

145. Luo Z, Miotto R, Weng C. A Human-Computer Collaborative Approach to Identifying Common Data Elements in Clinical Trial Eligibility Criteria. Journal of Biomedical Informatics (2012).

146. Luppino FS, de Wit LM, Bouvy PF, Stijnen T, Cuijpers Pim, Penninx BWJH, Zitman FG. Overweight, obesity and depression. Arch Gen Psychiatry. 2010;67(3):220-229.

147. Malik VS, Popkin BM, Bray GA, Despres JP, Hu FB. Sugar-sweetened beverages, obesity, type 2 diabetes mellitus, and cardiovascular disease risk. Circulation. 2010;121:1356-1364.

148. Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc 2003:420–424.

149. Mansouri A, Affendey Suriani L, Mamat A. Named Entity Recognition Approaches. IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2, February. 2008.

150. Marrero M, Sánchez-Cuadrado S, Lara JM, Andreadakis G. Evaluation of Named Entity Extraction Systems. Research in Computing Science, 41:47–58. 2009.

151. Marinou K, Tousoulis D, Antonopoulos AS, Stefanidi E, Stefanidis C. Obesity and cardiovascular disease: from pathophysiology to risk stratification. International Journal of Cardiology 138 (2010) 308.

152. Markham, KM, Mintzes JJ, Jones GM. The concept map as a research and evaluation tool: Further evidence of validity. Journal of Research in Science Teaching 31, no. 1 (1994): 91-101.

153. Markow PG, Lonning RA. Usefulness of concept maps in college chemistry laboratories: Students' perceptions and effects on achievement. Journal of Research in Science Teaching 35, no. 9 (1998): 1015-1029.

154. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization. 1998.

155. McClure JR, Sonak B, Suen HK. Concept map assessment of classroom learning: Reliability, validity, and logistical practicability. Journal of Research in Science Teaching, 36:475-492,1999.

156. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Studies in health technology and informatics 1 (2001): 216-220.

157. Mento AJ, Martinelli P, Jones RM. Mind Mapping in Executive Education: Applications and Outcomes. Journal of Management Development , 18(4), 390– 407. 1999.

158. Mertens IL, Van Gaal LF. Overweight, obesity, and blood pressure: the effects of modest weight reduction. Obesity Research Vol. 8 No. 3 May 2000.

159. Meystre MS, Savova KG, Kipper-Schuler CK, Hurdle FJ. Extracting Information from Textual Documents in the Electronic Health Record: a Review of Recent Research. Methods Inf Med 2008, 47(Suppl 1):128-144.

160. Miller LC, Rosas SR, Hall K. Using concept mapping to describe sources of information for public health and school nursing practice. Journal of Research in Nursing 17, no. 5 (2012): 466-481.

161. Mind-Mapping. Mind-maps, [http://www.mind-mapping.co.uk/], 2013.

162. Molaison FE, Taylor AK, Erickson D, Connell LC. The use and perception of concept mapping as a learning tool by dietetic internship students and preceptors. Journal of Allied Health. Vol 23, Number 3. November 2008.

163. Monasta L, Batty GD, Cattaneo A, Lutje V, Ronfani L, Van Lenthe FJ, Brug J. Early-life determinants of overweight and obesity: a review of systematic reviews. Obes Rev. 2010 Oct;11(10):695-708. doi: 10.1111/j.1467-789X.2010.00735.x.

164. Moro A, Navigli R. WiSeNet: building a wikipedia-based semantic network with ontologized relations. In Proceedings of the 21st ACM international conference on Information and Knowledge management. pp. 1672-1676. ACM. 2012.

165. Muslea I. Extraction Patterns for Information Extraction Tasks: A Survey. In Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction (Orlando, FL, 1999), pp. 1-6.

166. Mykowiecka A, Marciniak M, Kup´s´c A. Rule-based information extraction from patients' clinical data. J Biomed Inform Oct 2009;42(5):923–36.

167. Navigli R, Ponzetto SP. BabelNet: Building a very large multilingual semantic network. In Proceedings of the 48th annual meeting of the association for computational linguistics, pp. 216-225. Association for Computational Linguistics, 2010.

168. Krauthammer M, Nenadic G. Term identification in the biomedical literature. Journal of Biomedical Informatics, 37, 512-526. 2004.

169. Nenadić G, Ananiadou S, McNaught J. (2004). Enhancing automatic term recognition through recognition of variation. In Proceedings of COLING 2004. Geneva. 604--610.

170. Nesbit JC, Adesope OO. Learning with concept and knowledge maps: A meta-analysis. Review of Educational Research, 76, 413-448. 2006.

171. Nguyen DM, El-Serag HB. The epidemiology of obesity. Gastroenterol Clin North Am. 2010 Mar;39(1):1-7.

172. Noy FN, McGuinness LD. Ontology development 101: A guide to creating your first ontology. Technical Report SMI-2001-0880, Stanford Medical Informatics. 2001.

173. Novak JD, Canas JA. The theory underlying concept maps and how to construct them. [http://cmap.coginst.uwf.edu/info/], 2008.

174. Ogden LC, Yanovski ZS, Carroll DM, Flegal MK. The epidemiology of obesity Gastroenterology 132:2087-2102, 2007.

175. Ogden CL, Lamb MM, Carroll MD, Flegal KM. Obesity and socioeconomic status in adults: United states 1988-1994 and 2005-2008. 2010 NCHS data brief no 50. Hyattsville, MD: National Center for health statistics. 2010.00735.

176. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of obesity in the united states, 2009-2010.NCHS data brief no.82 January 2012.

177. Oliveira A, Pereira FC, Cardoso A. Automatic reading and learning from text. In Proceedings of the International Symposium on Artificial Intelligence (ISAI). 2001.

178. Onyike CU, Crum RM, Lee HB, Lyketsos CG, Eaton WW. Is obesity associated with major depression? Results from the third national health and nutrition examination survey. Am J Epidemiol 2003;158:1139-1147.

179. Petten SB, Williams JVA, Lavorato DH, Brown L, McLaren L, Eliasziw M. Major depression, antidepressant medication and the risk of obesity. Psychother Psychosom 2009;78-182-186.

180. Pi-Sunyer FX. The obesity epidemic: Pathophysiology and consequences of obesity. Obesity, 10(S12):97S–104S, December 2002.

181. Plotnick E. Concept Mapping: a Graphical System for Understanding the Relationship Between Concepts,
[http://ericir.syr.edu/ithome/digests/mapping.html], 1997.

182. PubMed. PubMed definition, [http://www.ncbi.nlm.nih.gov/pubmed], 2013.

183. Puhl MR, Heuer AC. The stigma of obesity: A review and update. Obesity(2009) 17, 941-964. doi:10.1038/oby.636.

184. Radovanovic M, Ivanovic M. Text mining: approaches and applications. Vol. 38, No. 3, 2008, 227-234 Novi Sad J. Math. Vol. 38, No. 3, 2008, 227-234.

185. Rajman M, Besancon NR, Besancon R. Text mining: Natural language techniques and text mining applications. In Proceedings of the 7th IFIP Working Conference on Database. 1997.

186. Rani P, Reddy R, Mathur D, Bandyopadhyay S, Laha A. Compositional information extraction methodology from medical reports. DASFAA'11 Proceedings of the 16th international conference on Database systems for advanced applications: Part II Springer-Verlag Berlin, Heidelberg, 2011.

187. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nature Reviews Genetics (2012).

188. Reichherzer T, Hayes P, Eskridge CT, Saavedra R, Mehrotra M, Bobrovnikoff D. Collaborative Knowledge Capture in Ontologies. InK-CAP 05. Banff, Canada. 2004.

189. Reichherzer T, Leake D. Towards Automatic Support for Augmenting Concept Maps with Documents. In Concept Maps: Theory, Methodology, Technology. Proceedings of the Second International Conference on Concept Mapping, vol. 1. 2006.

190. Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. Lancet 2008;371:569-78.

191. Rodriguez-Esteban R. Biomedical text mining and its applications. PLoS Computational Biology 5(12). 2009.

192. Roberts L. Using concept maps to measure statistical understanding. International Journal of Mathematical Education in Science and Technology 30, no. 5 (1999): 707-717.

193. Ryan MA, Duong M, Healy L, Ryan AS, Parekh N., Reynolds VJ, Power GD. Obesity, metabolic syndrome and esophageal adenocarcinoma: Epidemiology, etiology and new targets. Cancer Epidemiology 25 (2011) 309-319.

194. Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campell EM, Cattuto C, Khandelwal S, Mabry PL, Vespignani A. Digital Epidemiology. PloS Comput Biol *(7): e10002616. Doi:10/1371/journal.pcbi.1002616. 2012.

195. Safayeni F, Derbentseva N, Cañas AJ. Concept maps: A theoretical note on concepts and the need for cyclic concept maps. Manuscript submitted for publication.(pdf). Retrieved August 4 (2003): 2010.

196. Sauter LV. E/R model, [http://www.umsl.edu/~sauterv/analysis/er/er_intro.html], 2000.

197. Schadow G, McDonald CJ. Extracting structured information from free text pathology reports. AMIA Annu Symp Proc. 2003:584-8.

198. Schmidt JH. Alternative approaches to concept mapping and implications for medical education: Commentary on reliability, validity and future research directions. Advances in Health Sciences Education 9: 251–256. 2004.

199. SearchSQLServer E/R model, [http://searchsqlserver.techtarget.com/sDefinition/0,,sid87_gci949503,00.html], 2013.

200. Seasr. Text Mining overview, [http://seasr.org/documentation/text-mining-overview/], 2010.

201. Settles B. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. Bioinformatics, 21(14):3191-3192, 2005.

202. Shafiei M, Milios E. Model-based overlapping co-clustering. In: Proceedings of the Fourth Workshop on Text Mining, Sixth SIAM International Conference on Data Mining, Bethesda, Maryland, April 22. 2006.

203. Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. J Comput Biol, 10(6):821–855. 2003.

204. Sirohi E, Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. Pac Symp Biocomput; 2005. pp. 308–18.

205. SmallStock. Concept Maps and Mindmaps, [http://www.smallstock.info/info/mindmap.htm], 2013.

206. Solt I, Gerner M, Thomas P, Nenadic G, Bergman CM, Leser U, Hakenberg J. Gene mention normalization in full texts using GNAT and LINNAEUS. Proceedings of the BioCreative III Workshop. 2010.

207. Soyibo K. Using concept maps to analyze textbook presentations of respiration. The American Biology Teacher (1995): 344-351.

208. Sowa FJ. (1987) Semantic Networks. Wiley, Second edition (1992).

209. Spasić I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text.. In: Briefings in Bioinformatics, 2005, 6(3), 239–251.

210. Spasić I, Sarafraz F, Keane AJ, Nenadić G. Medication information extraction with linguistic pattern matching and semantic rules. Journal of the American Medical Informatics Association, Vol. 17, No. 5, pp. 532-535. 2010.

211. Sperrin M, Marshall AD, Higgins V, Buchan I, Renehan AG. Slwoing down of adult body mass index trend increases in England: a latent class analysis of cross-sectional surveys (1992-2010). International Journal of Obesity (2013).

212. Stanfill HM, Williams M, Fenton HS, Jenders AR, Hersh RW. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Asoc 2010;17:646-651. doi:10.1136/jamia.2009.0010024.

213. Stavrianou A, Andritsos P, Nicoloyannis N. Overview and semantic issues of Text Mining, SIGMOD Record, 36(3), p. 23-34. 2007.

214. Stone AA, Broderick EJ. Obesity and pain associated in the United States. Obesity(2012) doi:10.1038/oby.2011.397.

215. Stunkard AJ, Faith MS, Allison KC. Depression and obesity. Society of Biological Psychiatry 2003.

216. Suakkaphong N, Zhang Z, Chen H. Disease named entity recognition using semi-supervised learning and conditional random fields. Journal of the American society for Information Science and Technology, 62(4):727-737, 2011.

217. Sure Y, Staab S, Angele J, Wenke D, Maedche A. OntoEdit: Guiding ontology development by methodology and inferencing. Submitted to: Prestigious Applications of Intelligent Systems (PAIS), in conjunction with ECAI 2002, July 21–26 2002, Lyon, France.

218. Takenobu T, Hironori O, Hozumi T. Effectiveness of complex index term in information retrieval. The 6th RIAO Conference. Pp.1322-1331. 2000.

219. Tan A. Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, pages 65–70.

220. Thew S, Sutcliffe A, Procter R, de Bruijn O, McNaught J, Venters CC, Buchan I. Requirements engineering for e-Science: experiences in epidemiology. Software, IEEE 26, no. 1 (2009): 80-87.

221. Tiago Tresoldi, [http://code.google.com/p/nltk/source/browse/trunk/nltk_contrib/nltk_contrib/stringcomp.py], 2009.

222. Torre DM, Daley B, Stark-Schweitzer T, Siddartha S, Petkova J, Ziebert M. A qualitative evaluation of medical student learning with concept maps. Medical Teacher; 29:949–55 2007.

223. Toutanova K, Manning CD. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of_speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70. 2000.

224. Trasande L, Cronk C, Durkin M, Weiss M, Schoeller DA, Gall EA. Environment and obesity in the National Children's Study. Environmental Health Perspectives, 117(2), 159–166, 2009.

225. Trochim, William MK. An introduction to concept mapping for planning and evaluation. Evaluation and program planning 12, no. 1 (1989): 1-16.

226. Tu SW, Carini S, Rector A, Maccallum P, Toujilov I, Harris S. OCRe: an ontology of clinical research**.** 2009.

227. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. J Am Med Inform Assoc. 2006; 13: 691–695.

228. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge summaries. J Am Med Inform Assoc, 15 (1) (2007), pp. 14–24.

229. Van Gaal LF, Mertens IL, de Block CE. Mechanisms linking obesity with cardiovascular disease. Nature, Vol. 444, 13 December 2006.

230. Villalon JJ, Rafael AC. Concept Map Mining: A definition and a framework for its evaluation. In Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on, vol. 3, pp. 357-360. IEEE, 2008.

231. Vucenik I, Stains JP. Obesity and cancer risk: evidence, mechanisms, and recommendations. Ann. N. Y. Acad. Sci. 1271 (2012) 37-43.

232. W3C. OWL Web ontologies, [http://www.w3.org/TR/owl-guide/], 2013.

233. Wafula B. Automatic Construction of Concept Maps. (2006).

234. Wang Y, Beydoun MA, Liang L, Caballero B, Kumanyika SK. Will All Americans Become Overweight or Obese? Estimating the Progression and Cost of the US Obesity Epidemic Obesity (2008a) 16 10, 2323–2330. doi:10.1038/oby.2008.351.

235. Wang X, Chused AEN, Friedman C, Markatou M. Automated Knowledge Acquisition from Clinical Narrative Reports. Proc AMIA Symp 2008b, 6:783-787.

236. Wang Y. Annotating and Recognising Named Entities in Clinical Notes. Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, August, Suntec, Singapore. 2009.

237. Watson M, Smith A, Watter S. Leximancer concept mapping of patient case studies. In Knowledge-Based Intelligent Information and Engineering Systems, pp. 179-179. Springer Berlin/Heidelberg, 2005.

238. Whitlock G, Lewington S, Sherliker P, Clarke R, Emberson J, Halsey J, Qizilbash N, Collins R, Peto R. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. Prospective Studies Collaboration, Lancet. 2009 Mar 28;373(9669):1083-96. Epub 2009 Mar 18.

239. Willemsen AM, Jansen GA, Komen JC, Van Hooff S, Waterham HR, Brites PTM, Wanders RJA, van Kampen AHC. Organization and integration of biomedical knowledge with concept maps for key peroxisomal pathways. Bioinformatics 24, no. 16 (2008): i21-i27.

240. Williams CG. Using concept maps to assess conceptual knowledge of function. Journal for Research in Mathematics Education (1998): 414-421.

241. Wolin KY, Colditz GA. Can weight loss prevent cancer&quest. British journal of cancer 99, no. 7 (2008): 995-999.

242. World Cancer Research Fund (WCRF). Obesity, [http://www.wcrf.org/index.php], 2013.

243. World Health Organization (WHO). Definition of obesity, risk factors, complications, epidemiology, [http://www.who.int/en/], 2013.

244. World Health Organization (WHO). [http://www.who.int/mediacentre/factsheets/fs311/en/index.html], 2013.

245. Xiong W, Song M, deVersterre LW. A comparative study of an unsupervised word sense disambiguation approach. In Bioinformatics: Concepts, Methodologies, Tools, and Applications. Chapter 66, pp 1306. 2013.

246. Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Medinfo, 2004 pp. 565–572.

247. Xu R, Gatern Y, Superkar SK, Das KA, Altman BR, Garber MA. Extracting Subject Demographic Information from Abstracts of Randomized Clinical Trial Reports. In Proc. 12th World Congress on Health (Medical) Informatics 2007; 550-554.

248. Yang Y, Liu X. A Re-examination of Text Categorization Methods. In Proceedings of ACM SIGIR '99 conference, pp 42-49.

249. Yang XQ, Yuan SS, Chun L, Zhao L, Peng S. Faster Algorithm of String Comparison. (2001).

250. Yang H, Spasic I, Keane AJ, Nenadic G. A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries. Journal of the American Medical Informatics Association. Vol. 16, issue 4, July-August, 596-600. 2009.

251. Yiheng C, Bing Q, Ting L, Yuanchao L, Sheng L. The comparison of SOM and K-Means for text clustering. Computer and information Science, Volume 3, No 2. 2010.

252. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006:30.

253. Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medical knowledge discovery: A survey. Journal of Biomedical Informatics, Volume 43, Issue 4, August 2010, Pages 650-660.

254. Zhu, Fei,Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, Vongsangnak W, Shen B. Biomedical text mining and its applications in cancer research. Journal of Biomedical Informatics 2012.

255. Žubrinic K. Automatic creation of a concept map. University of Dubrovnik. 2011. Available in [http://www.ztel.fer.unizg.hr/_download/repository/KZubrinic-KvalifikacijskiRad.pdf].

256. Zubrinic K, Kalpic D, Milicevic M. The automatic creation of concept maps from documents written using morphologically rich languages. Expert Systems with Applications (2012).

257. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Briefings in Bioinformatics, 8(5):358-375, 2007.