HYPOTHESIS FORMULATION IN MEDICAL RECORDS SPACE

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering

2016

THAMER O BA-DHFARI

SCHOOL OF COMPUTER SCIENCE

Contents

Word count: 39,089

Lists of tables

Lists of figures

Abstract

Patient medical records are a valuable resource that can be used for many purposes including managing and planning for future health needs as well as clinical research. Health databases such as the clinical practice research datalink (CPRD) and many other similar initiatives can provide researchers with a useful data source on which they can test their medical hypotheses. However, this can only be the case when researchers have a good set of hypotheses to test on the data. Conversely, the data may have other equally important areas that remain unexplored. There is a chance that some important signals in the data could be missed. Therefore, further analysis is required to make such hidden areas become more obvious and attainable for future exploration and investigation.

Data mining techniques can be effective tools in discovering patterns and signals in large-scale patient data sets. These techniques have been widely applied to different areas in medical domain. Therefore, analysing patient data using such techniques has the potential to explore the data and to provide a better understanding of the information in patient records. However, the heterogeneity and complexity of medical data can be an obstacle in applying data mining techniques. Much of the potential value of this data therefore goes untapped.

This thesis describes a novel methodology that reduces the dimensionality of primary care data, to make it more amenable to visualisation, mining and clustering. The methodology involves employing a combination of ontology-based semantic similarity and principal component analysis (PCA) to map the data into an appropriate and informative low dimensional space. The aim of this thesis is to develop a novel methodology that provides a visualisation of patient records. This visualisation provides a systematic method that allows the formulation of new and testable hypotheses which can be fed to researchers to carry out the subsequent phases of research. In a small-scale study based on Salford Integrated Record (SIR) data, I have demonstrated that this mapping provides informative views of patient phenotypes across a population and allows the construction of clusters of patients sharing common diagnosis and treatments.

The next phase of the research was to develop this methodology and explore its application using larger patient cohorts. This data contains more precise relationships between features than small-scale data. It also leads to the understanding of distinct population patterns and extracting common features. For such reasons, I applied the mapping methodology to patient records from the CPRD database. The study data set consisted of anonymised patient records for a population of 2.7 million patients. The work done in this analysis shows that methodology scales as $O(n)$ in ways that did not require large computing resources. The low dimensional visualisation of high dimensional patient data allowed the identification of different subpopulations of patients across the study data set, where each subpopulation consisted of patients sharing similar characteristics such as age, gender and certain types of diseases.

A key finding of this research is the wealth of data that can be produced. In the first use case of looking at the stratification of patients with falls, the methodology gave important hypotheses; however, this work has barely scratched the surface of how this mapping could be used. It opens up the possibility of applying a wide range of data mining strategies that have not yet been explored. What the thesis has shown is one strategy that works, but there could be many more. Furthermore, there is no aspect of the implementation of this methodology that restricts it to medical data. The same methodology could equally be applied to the analysis and visualisation of many other sources of data that are described using terms from taxonomies or ontologies.

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 take this opportunity to express my deepest appreciation to all those provided me the possibility to finish this thesis. First of all, special gratitude I give to my supervisor, Andy Brass for constructive supervision throughout my PhD study. Without his guidance, encouragement and constant feedback, this thesis would not have been achievable. I learned a great deal under his direction not only academically but also philosophically in terms of dealing with challenges in both work and life. It has been a true privilege and a pleasure to work with him. I would also like to thank my colleagues in the Bio-Health Informatics Group for their help and support: Muhannad, Oscar, José, Zulkifli, Ahmad and Sahel.

This PhD study would not have been possible without the corporation and support extended by the AstraZeneca team especially James Weatherall and Mark Davies. I also greatly appreciate the support received through the collaborative work undertaken with Health eResearch Centre (HeRC) at the University of Manchester.

A very special thank you to my family for their support they have shown throughout my PhD study. Words cannot express how grateful I am to my mother, father and wife for all of the sacrifices that you have made on my behalf. Thanks are also due to all my friends.

Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
AVG	Average aggregation approach
AZ	AstraZeneca
BC	Before Christ
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CI	Confidence Interval
CLARNS	Clustering large Applications based on Randomized Search
CPRD	Clinical Practice Research Datalink
DAG	Directed Acyclic Graphs
DBSCAN	Density-based Spatial Clustering of Applications with Noise
EHR	Electronic Health Records
EM	Expectation Maximisation
GO	Gene Ontology
GP	General Practitioner
GPRD	General Practice Research Database
HIT	Health Information Technology
HRGs	Healthcare Resource Groups
IC	Information Content
ICA	Independent Component Analysis
ICD	International Classification of Disease
ISO	International Organization for Standardization
LCS	Least Common Subsume

MAX	Maximum aggregation approach
MICA	Most Informative Common Ancestor
MIN	Minimum aggregation approach
MMR	Measles, Mumps and Rubella
NHS	National Health Service
NHS-CCC	British National Health Service Centre for Coding and Classification
NLP	Natural Language Processing
OPCS	Office of Population Censuses and Surveys
OXMIS	Oxford Medical Information System
PCA	Principal Component Analysis
POMR	Problem-Oriented Medical Record
RR	Relative Risk
SD	Standard Deviation
SIR	Salford Integrated Record
SNOMED-CT	Systematized Nomenclature of Medicine Clinical Terms
SOAP	Subjective, Objective, Assessment, Plan
SQL	Structured Query Language
UK	United Kingdom
VAMP	Value Added Medical Products
WHO	World Health Organisation

Chapter 1: Introduction

1.1. Introduction

There is a long history of people recording medical data. Typically, these early records took the form of brief, written case history reports maintained for didactic purposes and to share knowledge between people involved in caring for people with illnesses. Some of the first examples of such documents were discovered in Egypt. Ancient Egyptians did have a tradition of collecting and recording their medical knowledge and practice of medicine as seen preserved in a set of medical papyri [1], [2]. It would appear that one reason for the creation of these records was their religious beliefs in living their eternal life in a healthy body [3], [4]. Egyptian scholars at that time documented their practices on scrolls for their disciples for educational purposes. A few of these scrolls have been discovered. One of the earliest documents has come to be known as Ebers Papyrus [2], [5]. The Ebers Papyrus was purchased by Georg Moritz Ebers, a German Egyptologist, in the 1870s. It is believed that this papyrus text was written circa 1550 BC. The papyrus shows the medical knowledge that the Egyptian scholars had of the human body and its structure. It also has descriptions of treatments and diagnoses of diseases such as stomach diseases, skin diseases, and dental conditions. Another example of medical papyrus was acquired by Edwin Smith, an American Egyptologist, in 1862 [6]–[9]. The document was written on papyrus text and consisted of 48 case reports of injuries, fractures, wounds, dislocations, and tumours that date back to 1600 BC. This shows that from the beginning of recording history, the recoding of knowledge and practice of medicine has been at the centre of human concerns. The recording and capturing of health information about patients' life histories, previous diseases, and previous treatments has always been a fundamental part of health care: the capture of medical data for later re-use has at least a 3,500-year history.

Today, the advent of information technology in the health care industry has transformed the way health care is carried out and documented [10]–[17]. Health Information Technology (HIT) has augmented the traditional techniques used to deal with and manage health care data [18]. The process of documenting patients medical history has evolved rapidly since the introduction of computerised systems to record such information [17]. Electronic patient records play an

important role in modern health care systems, as they keep track of and electronically store a patient's complete medical and prescription histories. According to the International Organization for Standardization (ISO) [19], electronic patient records are defined as repositories of patient data in digital form, generated by one or more encounters in any health care setting, stored and exchanged securely and accessible by multiple authorised users including health care professionals as well as patients. The data within the records contains longitudinal, concurrent and prospective information about patients. This information includes patient demographics, progress notes, medications, vital signs, past medical history and laboratory results.

Besides being repositories of patient health histories, electronic patient records are also used as a means of assessing activity and receiving payments. In the UK, General Practitioners (GPs) are required to provide evidence for the quality of care provided for their patients. GP practices score points according to their level of achievement. This is based on certain indicators including the organisation of the practice, patients experience, and how chronic conditions are managed [20]. The Read coding system, which is used mainly for recording patient medical data, is used to present such evidence. The data captured in patient records is linked with the financial reward that is offered to the GPs. There is a strong evidence base that the use of financial motives has been a driver for improving the quality of care [21]–[27]

Much of the information exchanged in the health care systems is either shared or derived from electronic patient records. According to Berg [28], who considered the electronic record from a sociological perspective, a patient record is not simply a repository of information about a patient, but it also helps to foster communication between doctor and patient and is thus directly relevant to the way that patients' stories unfold. Also, as stated by Rector [29] 'a medical record consists of what clinicians have said about what they have heard, seen, thought, and done'. However, the information recorded in patient histories is not necessarily the complete truth [30], [31]; it can be incomplete, ambiguous, subjective or in some ways unreliable [28], [32], [33]. The aim for gathering such information in patient records is not to establish the truth but to provide a basis for identifying problems and taking actions.

Electronic patient records have helped to provide a better quality of care to patients while also reducing the costs of care [15], [34]–[36]. Furthermore, electronic patient records help health care providers to make more informed clinical decisions as well as improving patient safety [14], [15], [37]. The rapid adoption of electronic patient record systems allows the generation of large amounts of data about patient cohorts. The use of such data for research and discovery is a growing area of investigation in medical research [33], [38]–[44]. There is a great potential for leveraging electronic patient records to solve complex problems in medicine [45]. The greatest potential of electronic patient records, however, lies in the effective use of such data. As the value of information lies in its use [46], the value of the electronic patient records will remain hidden unless a set of approaches are developed that shed light on the data in order to gain new medical insights.

Although the analysis and integration of electronic patient records for research has great potential, there are certain challenges delaying this development. One of these challenges is related to the fact that these records include sensitive and personal information about patients. The use of such data for research is, therefore, subject to both legal and ethical consideration [47]. Often, patient consent is required for accessing personal health data for research [41], [48]. This has an impact on increasing the time and cost needed for carrying out a research [49]. One way to allow sharing this data for research is to provide anonymised patient data – data that has been altered to make it hard to be traced back to individual patients [50], [51]. Many future plans have been proposed to make patient records available for research. In a public address in 2011, the British Prime Minister, David Cameron, announced plans to make the UK patient records available for research – once suitably anonymised [52]. Other initiatives include the EHR4CR European EHR research framework initiative[1], and the eMERGE Network[2] to integrate electronic patient records as a tool for genomic research.

1 http://www.ehr4cr.eu

2 https://emerge.mc.vanderbilt.edu/

1.2. The history of recording patient information

The tradition of recording births, marriages and deaths is very ancient. In response to the Great Plague of London in the 16th century, authorities introduced the custom of issuing a bill of mortality weekly. In 1665, the Great Plague caused the death of 25% of the population of London. Bills of mortality were systematic recordings of deaths, which was vital in keeping records about a population [53]. This was the beginning of the major statistical area of epidemiology, and it led John Graunt (1620–1674) to publish a book based on mortality records, entitled “Natural and Political Observations Made upon the Bills of Mortality”. In this book, Graunt analysed the factors of death from the bills of mortality according to gender, residence, season, and age [54].

After Graunt’s death, little work was done in this area for two centuries, until William Farr (1807–1883) extended Graunt’s ideas to provide a better description of epidemiological issues [54]. Farr devised a new system of diseases classification, or nosology that recorded a person’s cause of death. He also developed an approach for medical data collection and analysis [55]. The detailed records allowed for a more detailed analysis of death risk factors within a general population. The methods Farr used are still in use today [56]. In addition, Farr’s work influenced many public health scholars in the 19th century, such as John Snow, Edwin Chadwick and John Simon [56]. Farr was the greatest epidemiologist and public health statistician of the Victorian era [55]. He can also be considered as the founder of public health [57].

Presenting bills of mortality as a form of tables makes it difficult to interpret statistical data. Subsequently, Florence Nightingale (1820–1910) focused on graphical methods of representing this data rather than using statistical ways. Nightingale is credited with developing polar area diagrams, or the ‘Coxcomb chart’, as she preferred to call it [58]. She used this type of charts to provide a graphical representation of mortality data in 1858 [54]. She also put a great deal of effort into improving data quality and health care standards [59].

What could be regarded as a modern medical recording system was first used in Boston in 1910 [60]. This schema was initiated to improve the quality of hospital records effectiveness in surgical patient treatment. Patients in this schema were followed to evaluate their treatments

and to identify any causes of possible treatment failures; however, the poor quality of recording was an obstacle to implementing outcome management.

In the late 1960s, the British Health and Social Security Department began to investigate the more widespread use of computers in primary care and established two research centres in Oxford and Exeter. The aim was to generate unique records for patients that could be stored centrally by the health care department and remotely accessed by all health care practitioners. It was recognised that the use of coded data would be essential in establishing an effective computerised medical recording system. The Oxford Medical Information System (OXMIS) codes were developed based on the International Classification of Diseases Eighth Revision (ICD-8). OXMIS codes were the most widely used coding system in general practice for representing medical conditions and drugs [61]–[63]. In the early 1980s, James Read, who was working with Abies Informatics, developed the Read codes system to capture patient encounters and record them in a computerised system [61], [64]. The OXMIS differs from the Read codes in that it is designed to aggregate data, whereas Read codes are designed to record data [65]. Many schemas and government initiatives have been undertaken to increase the use of computerised systems to record patient information. There is strong evidence that this involvement has increased the use of electronic patient records in GPs settings. It was reported that the number of computers in general practices increased significantly from 10% to 79% between 1987 and 1993, and in 1996, it reached 96% [66], [67]. Similar patterns of GP computer usage have been seen in Europe, North America, and in Australia; the use of computers for patient management within the UK is still amongst one of the highest worldwide [68]. The National Health Service (NHS) in the UK provides a universal GP coverage of its population. Over 98% of the UK population are registered in the GP system and under the NHS visits to the GP are free of charge [69]. In 2014, it is estimated that there are 40,584 GPs and 7,875 general practices in England [70]. Thus, the GP system in the UK can be a good resource for research as it consists of a large number of patients across long observation periods.

In order for the data recorded in patient records to be used for purposes such as clinical, administrative, financial and research, the potential computerised systems are required to represent patient data in a useable and effective way [33], [71]. Clinical terminologies, along

with the use of many of the established medical coding systems, attempt to provide such a usable form of the records; they establish a foundation of information content in electronic patient records, and they provide a common medical language, which is necessary for storing and retrieving patient data [72], [73].

1.3. Clinical terminology

Communication is an integral component in modern health care systems [46]. However, for good communication between people involved in the health care, we need a controlled and agreed upon language. Many of the communication problems in health care occur when health professionals do not share the same background, which can lead to the use of different words to describe patient encounters. It is important for electronic patient record systems to be able to identify patients with certain medical conditions. Consider, for example, looking in records for patients who have been diagnosed with ‘diabetes’. Simply searching the records for patients with the diagnosis of ‘diabetes’ would probably succeed in most cases. However, other patients with the same medical condition might be recorded into the system using different terms, so searching for the word ‘diabetes’ in the records of patients who have been recorded as being ‘diabetic’, ‘NIDDM’ (noninsulin-dependent diabetes mellitus), or 'adult-onset diabetes' would fail to detect patients who were diagnosed with the same condition.

Controlled clinical terminologies were developed in an attempt to overcome these problems. The aim was to provide a standardised set of terminologies to establish a common language as the basis for a better communication of information among health care professionals [74], [75]. Clinical terminologies consist of a list of terms, aggregated in a systematic way to represent conceptual information that forms a given knowledge domain such as clinical cardiology or paediatric orthopaedics [76]–[79]. A list of primary tasks for clinical terminologies as was suggested by Rector [80]:

1. Capture patient data: a fast and intuitive way to record medical conditions during patient consultations
2. Present medical information related to individual patients
3. Provide the ability to query and retrieve of information at a population level

4. Share and integrate information from multiple health care systems

The use of these tasks in capturing patient data in a standardised manner makes them a key component in the integration of electronic records, decision support systems, and information retrieval systems [81]. Controlled clinical terminologies also play an important role in implementing a structured electronic patient records as they help support the reuse of patient data for various purposes [82],[83]. Clinical terms in any set of terminologies needs to support the capture, storage, and retrieval of the information they represent in ways that preserve and communicate the original information [84]. A standardisation of clinical terminologies was, therefore, needed for safe communication, especially when the data is communicated to and interpreted by machines [85]. It should also be noted that clinical terminologies need to be designed with consideration of the cognitive structures and processes of their users. Without such considerations, the designed terminologies will not be appropriate for people because they are hard to use, although they may or may not be appropriate for machine processing [86]. Coding the clinical terms is required as there are so many ways that a clinical concept can be represented [87]. Provide a standardised coding system is key in reducing the medical errors caused by the misrepresentation and misinterpretation of data [88].

1.4. Coding and classification systems

The medical concepts and terms used in the clinical domain are continuously expanding [17]. This provides a challenge in finding and retrieving specific terms or concepts from the terminology. It is, therefore, essential that the terminology to be organised in such a way that permits concept driven exploration. Medical terms need to be placed into categories that provide a structured grouping of terms and concepts organised on the basis of some common attribute, quality, or property [46].

Generating a classification of medical terms will facilitate the communication between health professionals across health care systems. There is, however, a chance that a term can be cast several different ways in the medical field (one meaning with multiple labels). Medical coding systems were developed to provide a unique set of codes for labelling all terms that convey the same medical concept [74], and through the process of coding, a set of terms describing some

medical concepts is translated into agreed-upon codes. The medical coding process involves converting the natural languages used for describing the concepts related to medical conditions into a set of unique codes (a combination of digits and/or letters). Medical coding systems provide a standard and a common language for a variety of statements with the same concept among health systems [89]. This coding should facilitate the identification of key medical events in a computerised medical record and the aggregation of information across groups of records.

The coding systems can be categorised into two groups based on their purpose: abstracting coding systems and systems that preserve the clinical details in a standardised way [46], [90]. The latter is a comprehensive coding approach for medical recording systems. Although there is no general agreement on what they should be compromised of, they are commonly supposed to help in structuring patient records [91]. In 1968, Larry Weed developed the problem-oriented medical record (POMR) as a strategy for improving the structure of patient records [92], [93]. The POMR was used for handling the complexity of problems in medicine through the supportive organization of data in patient records [68] [69]. This approach gives a readily understandable structure for recording a patient record [94]. The development of the POMR aimed to provide a logical thought process for the documentation and communication of information in patient records. The POMR can be based on the information model for the problem oriented process of care [91]. The combination of coding problems with the POMR meets the demand for both data entry and data retrieval from patient records [95].

A number of systems were built based on one of these methods to code medical terms. Systems such as ICD-10 (the tenth revision of the International Statistical Classification of Diseases and Related Health Problems), OPCS (Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures), and HRGs (Healthcare Resource Groups) aim to abstract the patient records, whereas systems such as Read codes and SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) intend to encode the full details of medical records [90].

ICD-10 and the OPCS are used by clinical coders to encode patient records in hospitals. In contrast, the Read codes system is designed to allow direct coding by the clinician themselves in the primary care setting [90]. There are, therefore, a number of differences in the type of data

that is recorded by the general practitioners (GPs) compared to that in hospitals. The data collected by GPs is accumulated over a number of patients encounters, whilst in hospitals they are collected all at once [96]. According to Coiera [74] the coded data in hospitals can be distorted by differences in the understanding of patient notes between the clinicians who wrote the notes, and medical coders. Also, finding a suitable set of codes that precisely matches what was written initially in patient notes would be problematic for the medical coders [74]. It was therefore suggested that the process of both documenting the medical event and recording it in the system be completed at the same time by one person. This has the potential to improve the quality of the coding process [74].

1.5. Read codes

Read Codes capture patient records in an agreed upon and standardised structured format that is machine readable [64]. The Read codes were originally designed by James Read in 1986 for the purpose of recording the full details of patient data in a structured manner during consultation in computerised patient records [64], [97]–[99]. In the UK, almost all primary care data is captured in electronic patient records. GPs record a significant amount of this data in the form of Read codes. In 1990, the Read code system was purchased by the British National Health Service Centre for Coding and Classification (NHS-CCC). The joint computing group of the Royal College of General Practitioners and the General Medical Services Committee has recognised the use of the Read codes system as the best way to take the most advantage of computers in primary care settings [100]. These records provide a rich dataset as a record of the health of the nation, both now and historically. For example, doctors in secondary care use this information to provide a continuity of care to their patients [101], at a wider level it is an invaluable resource for public health research [102] and for planning for future service provision in the National Health Service (NHS) [103].

Read codes are intended to be a system for GPs to use to record clinical information on computers. In a computerised patient record system, the user is given a list of possible terms to choose. These terms are linked to the codes that are searchable and retrievable by computers [104]. Read codes cover a wide range of details about patient health including data on demographics, lifestyle, symptoms, signs, history of diseases, family-side history of diseases,

social and personal history, diagnoses, therapies, medications, and administrative procedures. As a coded thesaurus, the Read codes enabled the recording of full aspects of patient data in electronic records of GPs [105].

It should be noted that Read codes are used for a wide variety of activities. They can be used to encode different data across the whole set of patient records with primary care ranging from the input, process, results, assessment, therapeutic, and administrative data. The use of the Read code system as a standard language of the NHS for recording patient encounters with their GPs enables electronic patient records functionalities [106].

Read Codes are comprehensive and arranged in taxonomies that reflect a number of levels of detail. The hierarchical structure of Read Codes is based on the 'is-a' relation type between child and parent concepts in different levels of tree. Any term can only have a single parent [107]. For example, a five-byte Read Code provides five levels of detail in such a way that the code has more detail the further it moves away from the root. Table 1.1 demonstrates an example of 5-byte Read code structure encoded with five alphanumeric characters to represent a specific type of asthma. Also, Figure 1.1 shows the hierarchy of Read codes for respiratory system diseases. The H chapter in the Read code structure represents respiratory system diseases, and what appears beyond letter H reveals more detail about the diseases in the same chapter. The different chapter headings, along with the first characters of their Read code are presented in Table 1.2. Chapters beginning with numbers 0 to 9 describe concepts related to medical history, examination, investigations, procedures and administration, while chapters beginning with upper-case letters (A to Z) encode concepts related to patient diagnoses. Furthermore, the chapters starting with lower case letters (a to y) are used to represent medications. The letters 'O', 'o', 'I', and 'i' were not used in Read codes version 2.

Table 1.1. A hierarchy in Read Codes shows the different levels of detail in the 5-byte Read code. These codes are taken from the Read code system provided by the UK Terminology Centre in the Health & Social Care Information Centre (HSCIC).

Hierarchy Level	Read Code	Rubric
1	H....	Respiratory system disease
2	H3...	Chronic obstructive pulmonary disease
3	H33..	Asthma
4	H331.	Intrinsic asthma
5	H3311	Intrinsic asthma with status asthmaticus

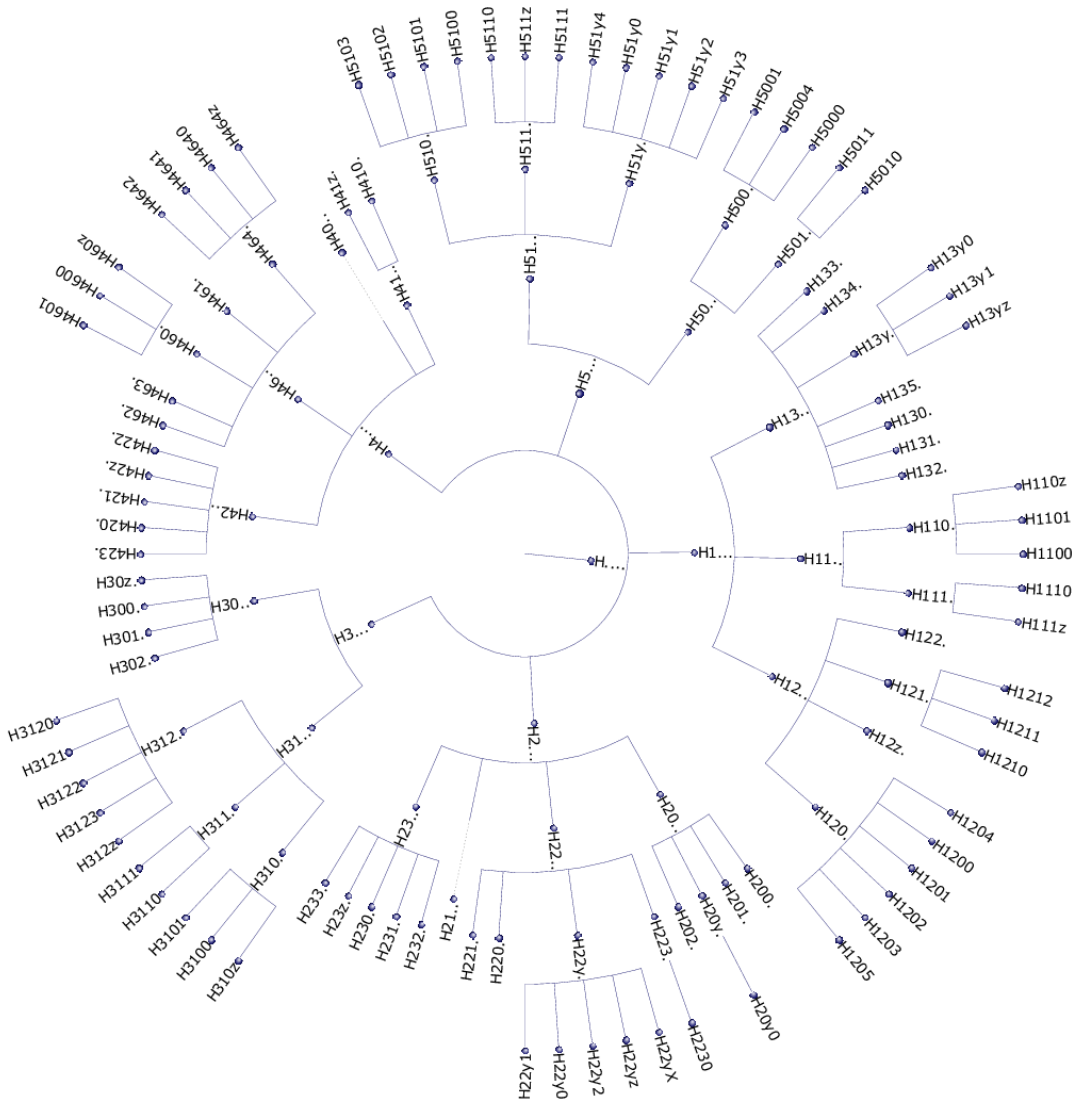


Figure 1.1. A circular tree diagram illustrating the structure of 5-byte Read code hierarchy for the respiratory system diseases. These codes are taken from the Read code system provided by the UK Terminology Centre in the Health & Social Care Information Centre (HSCIC).

Table 1.2. The Read code chapters for the following: (a) procedures of care, (b) diagnosis, and (c) medications. These codes are taken from the Read code system provided by the UK Terminology Centre in the Health & Social Care Information Centre (HSCIC).

A. Read code chapters related to processes of care	
Chapter	**Contents**
0	Occupations
1	History and symptoms
2	Examinations and signs
3	Diagnostic procedures
4	Laboratory procedures
5	Radiology
6	Preventative procedures
7	Operative procedures
8	Other therapeutic procedures
9	Administration
B. Read code chapters related to diagnoses	
Chapter	**Contents**
A	Infectious and parasitic diseases
B	Neoplasms
C	Endocrine, nutritional, metabolic or immunity disorder
D	Diseases of blood and blood-forming organs
E	Mental disorders
F	Nervous system and sense organ diseases
G	Cardiovascular system diseases
H	Respiratory system diseases
J	Digestive system diseases
K	Genitourinary system diseases
L	Pregnancy, childbirth and puerperal disorders
M	Skin and subcutaneous tissue diseases
N	Musculoskeletal and connective tissue diseases
P	Congenital anomalies
Q	Perinatal conditions
R	Symptoms, signs and ill-defined conditions
S	Injury and poisoning
T	Causes of Injury and poisoning
U	External causes of morbidity and mortality
Z	Unspecified conditions
C. Read code chapters related to medications	
Chapter	**Contents**
a	Gastro-intestinal system drugs
b	Cardiovascular system drugs
c	Respiratory system drugs
d	Central nervous system drugs
e	Drugs for infectious diseases
f	Endocrine drugs
g	Obstetric / gynaecological / urinary drugs
h	Malignant & immunosuppressant drugs
i	Nutrition and blood drugs
j	Musculoskeletal & joint drugs
k	Eye drugs
l	Ear, nose & oropharynx drugs
m	Skin drugs
n	Immunology drugs & vaccine
o	Anaesthetics
p	Appliances & reagents
q	Incontinence appliances
s	Stoma appliances
u	Contrast media
y	Drug release administration

The structure of Read codes is inspired by the concept of the POMER that was developed previously [46]. Most of the GP systems in the UK were based on this concept. In these systems, patient records consist of two sections: database and progress notes. The database section includes the identification of patients along with their past medical history, family, and social history, while the progress notes section includes four subsections known as SOAP (subjective, objective, assessment, plan) [46], [94].

1.5.1. Read codes versions

The Read Code system has evolved through three versions. The first version was developed in the early 1980s. This version used alphanumeric codes with four characters and included about 57,128 terms and 40,927 concepts [107]. In 1990, a second version was introduced with the same technical properties as the first version except that the code structure was extended to five bytes – this version was known as the five-byte Read Code. This allowed the system to capture a greater number of concepts and cover more healthcare areas such as secondary care. Furthermore, in its second version, the Read Code system added case sensitivity to its code characters. This led to an expansion in the number of codes stored to reach a total of 125,914 terms and 88,995 concepts. The third version attempted to address some of the technical issues in the earlier versions such as hierarchical relationships between codes. In spite of these improvements, however, GPs still use the second version of Read Codes [87].

1.5.2. Read codes in practice

A patient record created in general practice can be considered as a bag or multi-set of Read codes over a given period of time, in which each code refers to a concept on patient encounter with GPs. Within this bag of Read codes, we can get an overview of the health history of patients. The Read codes were developed in order to represent a wide variety of aspects of patients, such as the procedures undertaken, diagnoses established, medication described, and many other subjective and objective data that can be related to patient care. For example, a single record for a patient can be represented as $p = \{c_1, c_2, \dots, c_n\}$, where p refers to a given patient and c refers to a Read code. An example of this can be as follows (A description of this record is explained in Table 1.3) $p = \{\mathrm{C10F.}, 1372., \mathrm{bd3\,j.}, \mathrm{G20..}, 2469., 246\mathrm{A.}\}$.

Table 1.3. A bag of Read Codes presents a patient encounter with GP. Patient may have more than one type of Codes. Also, some codes might appear in patient encounters more than once based on their condition.

Read Code	Rubric
C10F.	Type II Diabetes Mellitus,
1372.	Trivial smoker < 1 cig/day
bd3j.	Prescription of “Atenolol 25mg tablets
G20..	Essential hypertension
2469.	Measurement of Diastolic Blood Pressure
246A.	Assessment of Diastolic Blood Pressure

The data in patient records were described as a bag rather than a set because each distinct Read code may have been recorded more than once in the patient record over a given period of time. It can also be observed that Read codes describe a number of very different activities, from patient diagnosis (C10F.) to a record of the medication prescribed (bd3j.) to procedures carried out by the GP (2469.). Figure 1.2 illustrates how each record belonging to a given patient in the real world is modelled by a GP based on the Read codes.

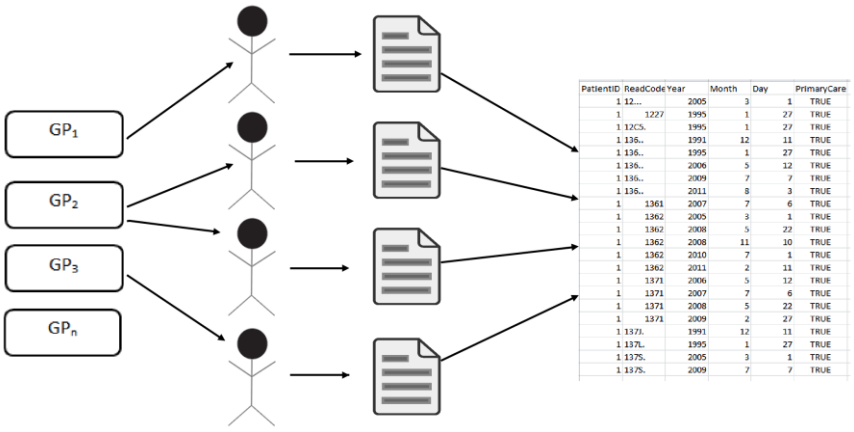


Figure 1.2. A Representation of patient records from the Salford primary care where GPs encode patient details. Each patient has been associated with a bag of Read Codes which are stored as electronic records. These records contain different details about patients including demographics, past history of diseases and treatments.

1.6. UK primary care data: Clinical Practice Research Datalink (CPRD)

In the UK, almost all health care delivery is centred on GPs. GPs functions as gatekeepers for accessing care in the National Health Service (NHS). According to the Office of Health

Economics, the number of interactions between patients and their GPs was 304 million consultations in 2009, with the very young (between birth and 15 years) and the very old (75 years and over) patients recording the largest number of interactions per year [108].

Patient records store such information in a structured format. The format of the data greatly affects its accessibility and quality. Being able to manage this information effectively can have a real impact on patients' health care. Better management of information supports more informed decision making, which will, in turn, allow the better delivery of health care to patients [109]. In the UK, the NHS invested a large amount of effort and resources into collecting patient records [110], [111]. The National Programme for Information Technology aimed to provide NHS health professionals with the information necessary to promote health and enable informed decision making across the health system.

Existing data in electronic patient records should add considerable value to the health care system [112], supporting health managers, health policy-makers, health professionals and most importantly, patients [113]. Methods are required to harness the research to achieve the potential of patient records. The value of data in electronic patient records will remain untapped unless it is explored and translated into practice [114]. Data recorded in patient records can be used for various purposes, including clinical, administrative, financial, and research [33].

Pharmaceutical companies recognised the potential of GPs' data in clinical and market research and noticed that the computerisation of GPs' systems could generate data of real value to the industry. In the late 1980s, the Value Added Medical Products (VAMP) supplied free computerised systems to GPs in return for providing them with anonymised data about morbidity, drug prescriptions and side effects. VAMP then established a database containing the patient records and aimed to make profit by selling the data to pharmaceutical companies [115], [116]. However, in 1993, the database ownership passed to Reuters Health Care, who donated the database to the UK Department of Health. The name of the database was changed to become the General Practice Research Database (GPRD) [38], [117]–[119]. Since 2012, the database has been known as the Clinical Practice Research Datalink (CPRD).

Presently, the CPRD is one of the largest databases of longitudinal medical records from primary care in the world [119]. The database collates anonymised records from 674 primary care practices on a monthly basis. In 2013, the total number of patients included in this database reached over 11 million patients, of which approximately 4.4 million are currently active [119]. Examples of the patient data included in this database are patient demographics, symptoms, tests, diagnoses, therapies, and health-related behaviours. The large number of patients and the accuracy of the data in the CPRD, make the CPRD as a rich source of health data for research [120]–[123].

1.7. Using primary care data for medical research

Data derived from electronic patient records can be of enormous use for researchers in generating new medical knowledge [33], [41], [124]. Health databases can be rich sources for researchers due to the large number of patient cohorts and long term follow-up. This offers a major advantage for epidemiological researchers who have difficulty collecting data with such features. It has been argued that studies carried out using electronic patient records, rather than recruiting individual patients, can be done more quickly and with a larger sample size [124], [125].

A total of 1,450 peer-reviewed journal papers that used the CPRD database have been published since 1988 (assessed in November, 2015). A full list of studies is available online and is updated routinely by the CPRD [126]. These studies cover a wide range of health related research topics. Based on a survey of CPRD papers published between 1995 and 2009, the majority are in pharmacology and pharmacy; this is followed by medicine (general and internal), and finally, public, environmental and occupational health [120]. Several studies have been carried out using the CPRD on epidemiology to assess the associations between diseases and the benefits and risks of certain drugs. Examples of some publications to date include studies showing the absence of an association between the measles, mumps and rubella (MMR) vaccine and autism [127]; cardiovascular risk after acute infection [128]; the lower risk of dementia associated with statin use [129]; the risk of myocardial infarction in patients with psoriasis [130]; the use of oral corticosteroids and risk of fracture [131]; and the association between body mass index and cancer [132].

1.8. Hypothesis, aim and objectives

Traditionally, epidemiological research begins with the hypothesis formulation phase. In this phase, researchers are required to have in-depth knowledge of a subject or disease area, note features of interest, and then formulate a hypothesis. This is followed by conducting a study in order to test the generated hypothesis. The study findings will determine whether or not to support or refute the hypothesis [133].

Such traditional methods have served the community quite well over time. However, with the rapid proliferation of data generated from patient records, the process of formulating new hypotheses from such large-scale data sets has become an overwhelming task for the individual researcher to work through efficiently using traditional methods. Continuing to use the traditional methods could mean that some important signals in the data could be missed. On the other hand, developing new strategies to exploit the new rich sources of data has the potential to uncover some of the hidden signals in the data and make them more obvious, thus, allowing for new discoveries from such data sources.

Health databases have been developed in order to improve the accessibility and usage of patient records for research. One example of such databases is the CPRD. The availability of large patient cohorts over longer observation periods has made the CPRD a valuable data source for epidemiological research [120]–[123]. The CPRD database has been used to produce approximately 1,450 research studies [119], [126]. These studies cover a wide range of research topics such as pharmacology, medicine and public health [120]. Most of these studies were based on hypothesis-driven research. In this type of study, researchers used a particular medical hypothesis to test on the data. For example, one study asked whether having polymyalgia rheumatica made patients more likely to be diagnosed with cancer [134]. This study used the CPRD database to investigate the incidence of new cancer diagnoses in patients having polymyalgia rheumatica. Another study used the CPRD data to assess the association between dementia and obesity [135]. These two studies and other similar ones used specific approaches, including statistical methods, to test their research hypotheses. There is a variety of developed approaches to test a hypothesis; however, little has been done to develop similar approaches for generating hypotheses. The ability to develop such approaches that allow

generating new hypotheses in an automated and unbiased manner would support the data exploration from patient medical records.

Data mining techniques can be effective tools in discovering patterns and signals in large-scale patient data sets. These techniques have been widely applied to different areas in the medical domain [39], [41], [136]–[138]. Therefore, analysing patient data using these techniques has potential in terms of providing a better understanding of the information in these records. However, the data stored in patient records is often complex and high dimensional, as it covers a variety of aspects of patients' histories, such as diagnoses, medications, and laboratory test results [138], [139]; it can also be noisy and incomplete [140], [141]. Most importantly, the data is not numerical; instead it consists of bags of terms chosen from a medical coding system, e.g. Read codes, ICD-9 or ICD-10 terms. Thus, the nature of the data provides a real challenge in effectively interpreting and visualising this data by using established techniques [139], [142]. Therefore, there was a need to develop tools that map electronic patient records into an appropriate and informative low dimensional space. This mapping will support data exploration and hypotheses formulation from such resources.

1.8.1. Aim

The aim of this thesis is to develop a novel computational methodology that provides an effective representation of patient records in a vector space. The ability to present data in this fashion would seem to make it amenable to analysis through more traditional data mining techniques, thus, allowing a more intuitive and straightforward environment for agnostic hypotheses formulation in medical records space. In order to achieve this aim, the following objectives were defined.

1.8.2. Objectives

1. Investigate whether techniques based on the notion of semantic similarity, which has been used successfully in other areas of medical research [143]–[145], could provide a method for mapping patient records into a low-dimension vector space. The objects mapped were the patients, the aim being to provide data representation in which patients who have similar diagnoses were mapped close to each other in a low dimensional space. This objective was

addressed in Chapter 2 and 3 by developing a mapping methodology. The first step was to find an efficient methodology to map patient records into a semantic similarity space; the second step was to develop a strategy for mapping from semantic similarity space to a vector space using principal component analysis (PCA). This was explored in a small data set from Salford.

2. To explore whether the methodology could scale up to much larger population data sets, and to explore the patient stratification generated. This objective was addressed in Chapters 4 and 5 by testing the scalability of the methodology to large-scale patient data sets. The optimisation of the methodology involved a number of steps, each of which provide important challenges. The results can provide a foundation to carry out further analysis of patient records.

3. Present an application of the methodology to identify and stratify patients with a specific disease at the subgroups level. This work allows exploration of the data around a disease and determines patterns in the data. This objective was addressed in Chapter 6 by deploying the methodology to analyse which diseases may be associated with the risk of falls in patients.

1.9. Thesis overview

The structure of this thesis is organised as follows.

- Chapter 2 reviews background knowledge about the strategies used to analyse patient records. This chapter also introduces a novel methodology that maps patient records into a low dimensional vector space.
- Chapter 3 investigates the use of the methodology on a small scale study based on GP data from Salford.
- Chapter 4 describes the challenges of applying the methodology to large scale datasets such as the CPRD and shows how each of the challenges was addressed.
- Chapter 5 explore the mapping of the CPRD patient data to see whether the large scale analysis works and provides any interesting insights on the patient data.
- Chapter 6 presents the application of the methodology to characterise and stratify patients with falls.
- Chapter 7 summarises the achievements and the contributions of this research project.

Chapter 2: A novel methodology to map patient data into low-dimensional space

2.1. Mining electronic patient records: current trends and applications

Much of the work done with medical records has been in understanding disease prevalence, initially for example in determining causes of mortality in a population [53], [54], [146]. This information has been used to inform investment in public health. As records have moved from paper to electronic formats it has been possible to ask more detailed questions. A lot of this work has focussed using epidemiological approaches to ask questions around the prevalence and the incidence of disease and how that changes as a function of region or population subgroups. Another area of intense study has been around the analysis of disease comorbidities, the process of assessing and measuring the associations among frequently co-occurring diseases. For example, Cao et al. studied the relationship between some diseases such as cushingoid facies [147]. Another study by Holmes et al. [148] explored the association for rare diseases such as Kawasaki disease. Similarly, Shin et al. studied the diseases associated with hypertension and diabetes mellitus [149]. These studies have typically [150]–[153] used rule-based approaches. Such approaches are easy to interpret and fast to implement. They do, however, require labour intensive supervision from experienced medical professionals who are able to add and review the rules [154], [155].

As the data sets have become larger and better defined it has been possible to explore the use of data mining and machine learning strategies on this data to explore more detailed questions. The use of mining techniques has the potential to make improvements in clinical research as the data contain more detailed information about large patient populations. Patient records have been successfully applied in many applications in medicine such as patient stratification, missed opportunities detection and risk modelling and understanding of diseases. Various machine learning and data mining techniques have been widely used for applications such as clustering, dimensionality reduction, classification and rule-based approaches. This section provides a

review of some of the current work on the application of data mining and machine learning strategies to the analysis of the large healthcare datasets.

For patient stratification and phenotyping, the main goal is to identify and stratify the patient cohort into different subgroups, so that within each subgroup, patients have similar attributes such as diagnoses, medications, treatment and laboratory test results. Patient stratification has been widely used in several clinical studies and biomedical applications. This process often opens up the possibility of planning for future health needs as well as carrying out future research in areas such as predicting the next complications, adverse event detection and pharmacovigilance. Some studies employed supervised learning techniques in order to develop automated approaches to stratify patients with asthma [156], rheumatoid arthritis [157] and cancer [158]. Wang et al. [159] worked on dividing patients using prior information from clinical experts. They used a cross-sectional design which aimed to investigate the associations between the risk factors and the outcome of interest. Then they used patient subgroups to develop specific risk prediction models for each subgroup. However, such techniques require expert clinical knowledge to define the gold standard and prior information about the patients [154].

Another direct approach for patient stratification is to use unsupervised learning and clustering techniques based on the associated clinical features and temporal patterns. Unsupervised techniques have been widely applied to identifying clusters consisting of patients with similar characteristics. For example, Gotz et al. [160] used a combination of sequential pattern mining and clustering techniques for temporal phenotype identification. One major drawback of this method is the large number phenotypes with an inappropriate support threshold [161]. In addition, a recent topology-based network approach identified three subgroups of patients with diabetes mellitus type 2. They calculated the similarity between patients using cosine similarity, and developed a novel topology data analysis- based approach to perform clustering of patients with selected clinical features [162]. However, this study is disease-centric as the methodology needs to be modified since the clinical features used in clustering are related to a specific disease (diabetes mellitus type 2).

Time plays a significant role in answering medical questions. Moving from static analysis to dynamic trajectories of diseases, conducted in clinical trials within a specific time period are normally used to detect the movement patterns and certain characteristics of diseases such as the sequence of events and the timing between the events. Incorporating a temporal dimension might provide useful insights into missed opportunities for detection, risk modelling and understanding of a disease. Consequently, the patient records are analysed by backtracking the patients' histories before a disease occurs and their records after the disease takes place. This process helps to detect patterns for identifying disease-related comorbidities. Various studies used temporal modelling to provide novel discovery in support of hypothesis generation. Hanauer and Ramakrishnan [163] performed a pairwise association ($X^2\ test$) among all codes included in the ninth revision of ICD-9, and then performed a temporal analysis using a binomial test. Nevertheless, this process is time consuming due to the large number of association calculations among all diseases ($number\ of\ calculations = all\ codes^2$). Another study conducted by Pivovarov et al. [164] analysed patient records for the study of utilization patterns that can quantify the potential overuse of haemoglobin A1c (HbA1c) test for diabetes. They applied a method that compares ordering distribution across time for assessing the inappropriate use over time. The method is specialised in a specific laboratory test (HbA1c) which means that the there is a need for clinical judgement is required to develop the laboratory measurements for the new test.

As mentioned earlier, some previous studies have focused on medical records based on phenotyping, which relies on rule-based approaches. These approaches require significant time and clinical judgement to develop; thus, there is a need for an automated approach for phenotype generation. Furthermore, recent studies have developed strategies to automatically analyse large clinical data sets to identify comorbidities. These studies are often disease-centric and specialise in disease associations.

The work described above typically works with medical records as terminologies, looking at finding associations between terms and sets of terms. There are a whole series of additional machine learning tools and methodologies that cannot be used for these problems, they are based around data that can be represented in a low dimensional vector space. There are many

methods that assume that data can be represented in the form of a vector of numbers. Many powerful and well-understood machine learning and data mining tools cannot therefore be applied to the data in electronic patient records. The goal of our study was therefore to explore a novel strategy that could map electronic patient data from a series of tokens from an ontology into a vector space. The hypothesis that this thesis is investigating is that such a mapping would allow us to apply such methods in patient records, and that this mapping would allow us to use a different set data mining strategies that would allow us to identify and generate multiple new insights (hypotheses) into the data in an automated and unbiased manner.

The key challenge in this task is to find a way of representing the differences between different sets of codes used to describe patients in a numerical form. To do this we will be using the ideas of semantic similarity – a way of capturing numerically the similarity of two terms in a taxonomy or ontology.

2.2. Semantic similarity

Comparing two or more objects is an essential process for information retrieval systems [165]. A suitable metric measure is required to be applied to compare objects for either similarity or distance [166], [167]. The mathematical notions of distance and similarity are used to provide an estimation of the difference and similarity, respectively, between two objects [167]. The notion of similarity indicates how close objects are to each other two. The higher the similarity value, the closer the objects are to each other, and vice versa. The length between two objects is called distance; a larger distance between objects means that they are more different.

There is a broad range of scientific and business studies that considered it convenient to study distance and similarity to identify the relatedness between entities using distance and/or similarity measures [168]. The ability to identify the relatedness between entities is important in various areas of research. For example, in medical research, this can be helpful for identifying patients with similar conditions and ailments, which allows one to apply existing data mining techniques to extract useful information [144].

In many scientific areas, concepts are organised in a hierarchical way using a directed acyclic graph (DAG) in the form of taxonomies and ontologies [169]–[171]. DAG produced a

standardised structure for representing concepts in the domain. The ability to build structures for concepts helps in quantifying the relationship between concepts. Semantic similarity is a method to calculate the topological similarity between taxonomical or ontological concepts [145]. This type of measure includes “is – a” relations to identify concepts with common characteristics [144].

The main goal of semantic similarity is to provide a precise estimation of the similarity between concepts in a way that is similar to human judgement [144]. The concept of semantic similarity has been a part of information retrieval and natural language processing (NLP) [165]. It has been applied in a variety of ways, such as to identify the synonymous characteristics in words within language lexicons – for example, WordNet, an English lexical database. WordNet contains different word forms in the English language (nouns, verbs, adjectives and adverbs) [172], which are compiled into sets of synonyms [173]. Semantic similarity can support information extraction by detecting concepts that are similar to ones that have already been obtained. Ontology learning also relies on semantic similarity. One example is gene ontology (GO), which contains descriptions of genes across all species. The genes can be compared using the similarity of their functions [174]. Word-sense disambiguation [175], [176], automatic hyperlinking [177], spelling error detection [178] and many other areas [179]–[181] can also benefit from accurate similarity estimations. Visualisation and clustering techniques also depend on semantic similarity when grouping objects with similar textual features [182].

Recently, a huge amount of medical data, including patient records, have been made electronically available and have become valuable resources for medical research [45]. However, most of this information is shown in heterogeneous textual formats. Semantic similarity can play a significant role in the integration and classification of such data and can improve the performance and accuracy of information retrieval [144]. For example, using electronic patient records when searching for patients with a specific symptom requires one to exploit a number of different medical concepts, including diagnoses, treatments, medications and similar symptoms. Using automated semantic similarity measures to group related clinical concepts might significantly enhance the query process in patient records [145].

Due to the fact that concepts in the clinical domain use taxonomies (e.g. Read codes, ICD and SNOMED CT) as knowledge bases, the hierarchal relation between objects has most often

been used to determine the score of similarity between clinical terms. As a result, the semantic similarity shows the taxonomical proximity and shared information between terms. For example, flu and chronic bronchitis, with Read codes 'H27..' and 'H31..', respectively, share some characteristics of their meaning: both are diseases related to the respiratory system. In respect to Read codes, they also share a common ancestor, respiratory system disease (Read code: 'H….').

A number of semantic similarity measures have been developed in the last few years in order to calculate the similarities between two concepts or two sets of concepts. These measures can be divided into two groups: node-based (or information content IC) and edge-based (or thesaurus-based), both of which are described in the following section.

2.2.1. Semantic Similarity measures among concepts

There are two types of measures for semantic similarity between two concepts: node-based, which relies on the information content and the properties of the compared concepts; and edge-based, which uses the distance between the concepts [165].

The node-based measures depend on comparing the properties of the concepts and even the properties of their ancestors and descendants. One of the most widely used approaches of node-based to find the conceptual similarity is Information Content (IC), where the calculation of semantic similarity relies on how much information the compared concepts share in common. The Information Content (IC) of a specific concept can be obtained by estimating the probability of frequency of a concept in a given dataset, which is illustrated in the following equation:

$$IC(c) = log^{-1}\, p(c)$$

Where p(c) is the probability of frequency of concept c. P(c) can be calculated as follows:

$$p(c) = \frac{\sum_{w \in w(c)} count(w)}{N}$$

W(c) is the set of concepts in the dataset annotated to c or c's descendent concepts. A higher value of IC for a concept means that the concept is very specific; a lower value shows that the concept is more general. In order to measure the semantic similarity of concepts, IC can be applied to the common ancestor of the concepts which the highest IC score, which is called the

most informative common ancestor (MICA). The general notion of node-based techniques is shown in Figure 2.1.

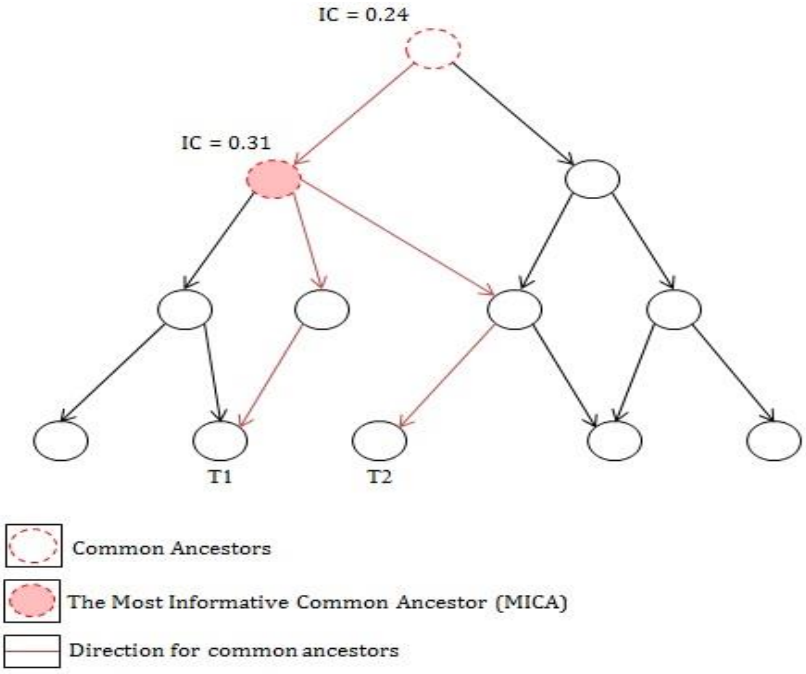


Figure 2.1. The general notion of node-based similarity measures using information content (IC) and the Most Informative Common Ancestor for two concepts (T1 and T2).

Many node-based approaches have been proposed and the most common ones are Resnik, Lin, Jiang and Conrath and Schlicker et al., which have been originally developed for specific applications such as WordNet and GO. Some approaches are briefly described below.

▪ The Resnik measure

Resnik's measure calculates the semantic similarity scores based on the IC of the most informative common ancestor (MICA) of the compared concepts [183], which can be defined as follows:

$$sim_{Res}\,(c1, c2) = IC(c_{MICA})$$

Resnik's measure is efficient at finding the shared information between two concepts, however, it does not show the distance between the concepts. Another problem of this measure is that if pairs of concepts share the same MICA, then they will have the same similarity score.

▪ The Lin measure

To address the problems of Resnik's measure, Lin has developed a similarity measure based on the information content of both the MICA of the concepts and each concept alone [181]. The similarity score in this measure is going to be between 0 and 1. The similarity score can be evaluated by the following formula:

$$sim_{Lin}(c_1, c_2) = \frac{2 \times IC(c_{MICA})}{IC_{(c1)} + IC_{(c2)}}$$

▪ The Jiang and Conrath measure

Jiang and Conrath have developed an approach which is similar to the principal of Lin's measure, however, it starts the process with a calculation of the distance of concepts [165], which is illustrated in the next equation:

$$dis_{J\&C}(c_1, c_2) = \big(IC(c_1) + IC(c_2)\big) - 2 \times IC(c_{MICA})$$

In this case, the distance is the opposite of concept's similarity. Obviously, the higher value of distance between concepts, the less similar they are. The similarity between concepts can be quantified using Jiang and Conrath's measure as follows:

$$sim_{J\&C}(c_1, c_2) = 1 - [IC(c_1) + IC(c_2) - 2 \times IC(c_{MICA})]$$

While the edge-based measures are more direct than node-based measures in calculating semantic similarity in a hierarchy, edge-based approaches rely mainly on the distance or path, which can be done by counting the edges between the concepts (See Figure 2.2). Once the distance is being evaluated, it can be easily translated to a similarity score. The shorter path between the concepts being compared, the more similarity score they will have. A number of measures based on the depth of the concepts have been developed in previous years; some of them are explained below.

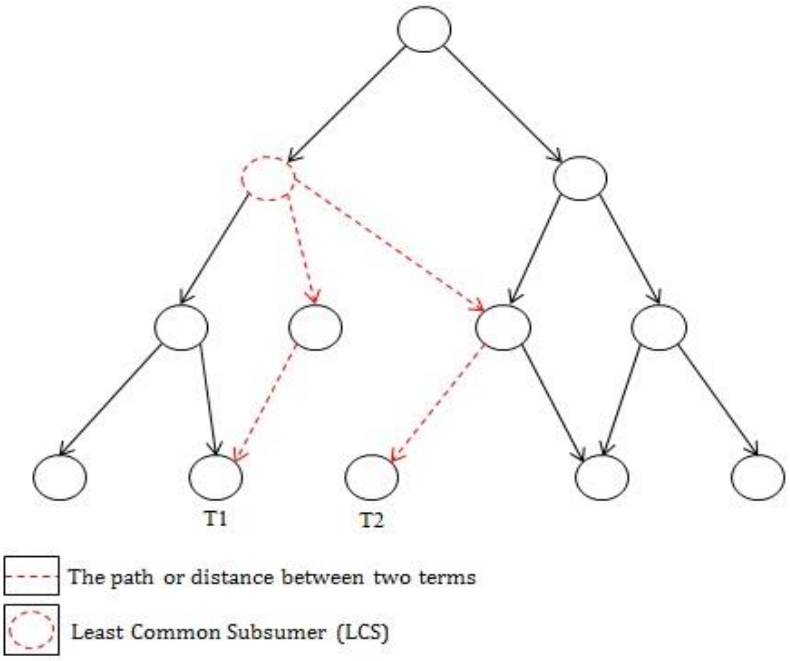


Figure 2.2. The concept of Edge-based similarity measures between two concepts (T1 and T2).

▪ The Rada measure

Rada has proposed a measure to calculate the shortest path possible connecting the two concepts [184], as given in the next expression:

$$dis_{Ra}(c_1, c_2) = |\min\_path(c_1, c_2)|$$

However, this approach is very simple and not really accurate. As a result, several improvements have been proposed.

▪ The Pekar and Staab measure

In Pekar and Staab's measure, we calculate the shortest path between the concepts as well as their Least Common Subsumer (LCS) [185], which can be illustrated as follows:

$$sim_{P\&S}\,(c_1, c_2) = \frac{\delta(root, LCS(c_1, c_2))}{\delta\big(c_1, LCS(c_1, c_2)\big) + \delta\big(c_2, LCS(c_1, c_2)\big) + \delta\big(root, LCS(c_1, c_2)\big)}$$

Where $\delta(a, b)$ stands for the number of edges in the shortest path between the concepts *a* and *b*. The similarity score in this measure is going to be between 0 and 1, where 0 is the minimum similarity score and 1 is the maximum.

▪ The Wu and Palmer measure

Wu and Palmer also considered the path between the two concepts in the hierarchy, but using the depth of LCS [186]. Using LCS in this measure relies on the supposition that the lower concepts in a hierarchy are less differentiated than the higher ones. The similarity score in this measure is going to be between 0 and 1. The similarity between concepts can be quantified using Wu and Palmer measure as follows:

$$sim_{W\&P}\ (c_1, c_2) = \frac{2\ \times\ depth\big(LCS(c_1, c_2)\big)}{|\min\_path(c_1, c_2)| + 2\ \times\ depth(LCS(c_1, c_2))}$$

▪ The Leacock and Chodorow measure

This measure calculates the shortest path between the two compared concepts by counting also the two concepts, which is the minimum path plus one [187]. Also the maximum depth of the hierarchy is being calculated. The expression of this measure can be defined as follows:

$$sim_{L\&C}\ (c_1, c_2) = -log\ \frac{|\min\_path(c_1, c_2)| + 1}{2\ \times\ \max\_depth}$$

Simplicity is the main feature of edge-based measures. They only rely on structure of the hierarchy by counting the nodes between the concepts being compared, which only requires low computational costs when we deal with large datasets. However, many limitations have been raised regarding accuracy of the similarity score because the measures just take the shortest path without any consideration to other paths in the hierarchy [144].

2.2.2. Semantic Similarity measures among sets of concepts

In general, entities in biomedicine are annotated with a number of terms and concepts. As a result, there is a need to rely on sets of concepts instead of single concepts [144]. There are two primary types of measures that can be used for semantic similarity between sets of concepts: pair-wise and group-wise. In pair-wise measures, the lists of concepts are compared individually by calculating semantic similarity between the concepts in the first set and the concepts in the second, using one of the measures for single concepts explained in the previous section. Then, the similarity scores between the concepts are combined in order to

obtain the final semantic similarity score between these two sets. The most widely used pair-wise combination approaches are: average, maximum and minimum, which are explained in detail below [188]. Group-wise measures do not depend on the individual concepts' calculations of the sets being compared, but directly calculate the score using one of the three existing approaches: set, graph and vector.

Owing to the fact that group-wise approaches are not common and many studies have proven that the pair-wise approaches performed much better than the group-wise approaches [144], we will only explain a number of approaches for pair-wise.

▪ Average approach

This approach can be obtained by calculating the average similarity between each concept in the two compared sets (*X*, *Y*), which is shown in the following formula:

$$sim_{AVG(X,Y)} = AVG_{c1 \in X, c2 \in Y} \left(sim(c_1, c_2)\right)$$

Where $sim(c_1, c_2)$ is the semantic similarity score between the two concepts using one of semantic similarity measures among concepts illustrated earlier.

▪ Maximum approach

In this approach, we calculate the maximum similarity between each concept in both sets (*X*, *Y*), as defined in the next equation:

$$sim_{MAX(X,Y)} = MAX_{c1 \in X, c2 \in Y} \left(sim(c_1, c_2)\right)$$

▪ Minimum approach

Similar to average and maximum approaches, this approach calculates the minimum similarity score between each concept in the two compared sets, which can be illustrated as follows:

$$sim_{MIN(X,Y)} = MIN_{c1 \in X, c2 \in Y} \left(sim(c_1, c_2)\right)$$

In short, many studies that intend to apply semantic similarity in their work have difficulty choosing the appropriate measure because each area has its own requirements.

2.3. Interpretation of semantic similarity in medicine

In recent years, semantic similarity has become an increasingly studied topic in the biomedical field [145]. Semantic similarity has been applied to validate results from several studies

investigating topics such as gene clustering [189], gene expression data analysis [190], prediction and validation of molecular interactions, and disease gene prioritization [145]. These studies have tested their biomedical data sets on a wide variety of measures. According to Guo et al. and other studies [189], [191]–[195], after testing a great number of semantic similarity measures among concepts and sets of concepts on medical datasets, their results show that node-based and pair-wise are more suitable and reliable for biomedical datasets due to the fact that edge-based measures suppose that relations of concepts in a data set either have equal distances or a distance as function of the depth. This is not true for the existing datasets in biomedicine. In addition, the most common semantic similarity measures that have been successfully applied to some biomedical domains are the Resnik's measure with pair-wise approaches which have the best performance and provide similarity scores that are very close to physicians judgment [191].

Semantic similarity measures are used to provide an estimation of the similarity between entities. This helps group entities that are similar and close to each other, which then can reduce the number of processes for visualising the data. However, if the dataset being analysed is very large, then visualising and clustering data is still difficult. Consequently, a dimension-reduction process must be performed in order to provide more precise visualisations.

2.4. Principal component analysis (PCA)

There are many datasets that can be considered large matrices in high dimensional spaces. Dealing with such data is a time consuming and complex process. Therefore, one of the critical problems in machine learning is determining how to develop reasonable representations for such complex data in order to ease the analysis and visualisation of the data [196]. As a result, there is a need for an intermediate process in data analysis when the number of variables in the data is overly large for useful analysis. An effective way to deal with this data is to use a dimension-reduction process to map the distances of points in data in high dimensional space into low dimensional spaces.

PCA is one of the most popular dimension reduction techniques and is used in many scientific disciplines [197]. PCA can be defined as a non-parametric technique used to extract important information from complex data in order to present this information as a set of new orthogonal

variables called principal components [197], [198]. The main purpose of PCA is to identify the most meaningful features of the dataset by finding the directions (components) that maximise the variance of the data in order to simplify the data by reducing the dimensions with minimal loss of information [199].

A study by Wood [200] illustrated the process of transforming a high dimensional matrix (*X*) into a low dimensional matrix (*Y*) by calculating the principal components of a dataset. Take the vector *X*, shown as $X = \{x_1, x_2, x_3, \ldots, x_n\}$ with *d* dimensions. The first step is to centralise the data by computing the mean of X and then subtracting the mean from all *X* values, which can be summarised by the following equation:

$$\text{For all } X \text{ values: } \quad x - \bar{x}$$

$$where\ \bar{x} = \ {}^{1}/_{N} \sum_{i=1}^{N} x_i$$

This step produces a dataset with a value of zero for the mean. The subtracting step makes calculating the variance and covariance easier without affecting their values. The covariance matrix (*C*) is calculated as follows:

$$C_{i,j} = {}^{1}/_{N-1} \sum_{q=1}^{N} X_{q,i} \,.\, X_{q,j}$$

Once the covariance matrix is obtained, the next step is to compute eigenvectors and their corresponding eigenvalues, as explained below:

$$Av = \lambda v$$

where A is a square matrix, v is the eigenvector (non-zero vector) and λ is the eigenvalue.

The first principal component of the data set is the eigenvector with the highest eigenvalue. The next step is to place the components in order of significance by sorting the eigenvectors by eigenvalues, highest to lowest. A matrix (*Y*) will be generated with *m* dimensions. The number of dimensions in *X* is the same as *Y*, which means that there is no loss of information. Lastly, one can decide to ignore less significant components, resulting in loss of some information that is not very important.

Although PCA presents datasets in simple and reduced forms, it still has limitations. For instance, it is able to reduce the dimensions only if the original variables are correlated. However, this problem can be addressed by using extended PCA algorithms such as kernel PCA or independent component analysis (ICA).

2.5. Mapping methodology

Visualisation is considered to be an essential step of any exploratory data analysis strategy [201], and effective visualization of high dimensional data requires a dimension reduction strategy [202]. In the first phase of our study we aim to map a complex and high dimensional data such as patient records into a low dimensional space. Multidimensional scaling strategy provides statistical methods such as PCA that map data presented in high dimensional space into a lower dimension space. PCA can be applied to the study dataset for information extraction to reduce the dimensionality. However, the patient records we have in dataset are encoded by Read Codes and this makes it difficult to perform PCA directly. Thus, another transformation step needs to be applied to the data. This step involves calculation the semantic similarity of patient records. Semantic similarity can be calculated either between individual patient records or between pairs of patient records. Both ways of calculations were used in our analysis. Having the semantic similarity scores for patient will allow us to build a vector of similarity scores for each patient in the dataset. These vectors will help to transform the data into a vector space and the PCA can be readily performed.

The strategy we followed to achieve this mapping can be divided into the following steps (Figure 2.3).

- Map patient records into similarity space:
    - Define the semantic similarity between a) individual Read Codes *within* each patient and b) bags of Read Codes *between* pairs of patients.
    - Build a matrix which describes semantic similarity between all pairs of patients.
- Map patient records into vector space:
    - Perform a principal component analysis (PCA) on similarity matrix to map patient records into low dimensional vector space.
- Project patient records onto the low dimensional space:

- Cluster and visualise patient records to gain medical insights.

Each of these steps is described in more details below.

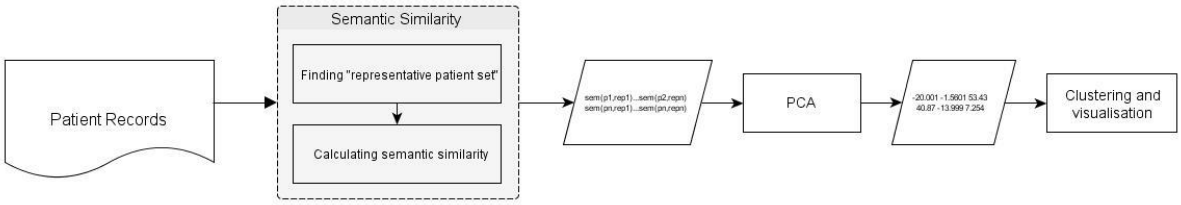


Figure 2.3. An overview of the methodology pipeline, beginning with sets of patient records described as bag of medical codes and ending with PCA coordinates represent patient records in a low dimensional vector space. The methodology consists of three stages: a) map patient records into a similarity space by finding semantic similarity between patients and representative patients. The results of this step are stored in a similarity matrix. Stage (b): map patient records into a vector space by performing principal component analysis (PCA) on the similarity matrix. The PCA transforms the data into low dimensional vector space. Stage (c) involved applying clustering algorithms and visualising the patient records.

2.6. Map patient records into similarity space

The first step of the methodology is to calculate semantic similarity between patient records in order to transform them into a similarity space. In this step we start with patient records in the form of Read codes and ending with patient vectors that hold the correspondent semantic similarity scores.

2.6.1. Semantic similarity calculation

Calculation of semantic similarity for patient records involves two steps. First, we find semantic similarity between any two Read codes within an individual patient where patient p in the dataset has a bag of Read codes such that $p = \{c_1, c_2, \dots, c_n\}$. This is done using the node-based semantic similarity measures mentioned earlier. We chose node-based over edge-based measures in order to leverage the information content of the data set, rather than only ontology structure alone. Second, we find semantic similarity between two patients given that patient p_1 has a bag of codes, and patient p_2 has another bag of codes. For each pair of patients, Minimum, Maximum and Average are calculated.

2.6.2. Generating the semantic similarity matrix

The semantic similarity scores obtained from previous step are then used to construct a matrix, as shown in Table 2.1. To illustrate this, we define $P = \{p_1, p_2, \ldots, p_n\}$ to be a set of n patients, and $\text{sim}(p_i, p_j)$ to be the semantic similarity score between patients i and j. Each patient corresponds to a single row in the similarity matrix, consisting of its pair-wise semantic similarity with all other patients. So, for example, this row for patient p_1 could be written as:

$$sim(p_1, P) = [\text{sim}(p_1, p_1), \text{sim}(p_1, p_2), \ldots, \text{sim}(p_1, p_n)]$$

By representing each patient in the data set as a vector consists of semantic similarity scores, this allowed us to perform the subsequent step which is the dimensionality reduction via PCA.

Table 2.1. An illustration of the similarity matrix of patients. It shows the similarity scores obtained from calculating the semantic similarity of the whole set of patients in the study data set. Each row in the matrix corresponds to a single patient, and is comprised of the similarity scores between that patient and all other patients in the data set.

		Patients (P)			
		p_1	p_2	…	p_n
Patients (P)	p_1	$sim(p_1, p_1)$	$sim(p_1, p_2)$	…	$sim(p_1, p_n)$
	p_2	$sim(p_2, p_1)$	$sim(p_2, p_2)$	…	$sim(p_2, p_n)$
	…	…	…	…	…
	p_n	$sim(p_n, p_1)$	$sim(p_n, p_2)$	…	$sim(p_n, p_n)$

2.7. Map patient records into vector space

Semantic similarity calculation and finding representative set of patients could reduce the amount of process on data. However, in order to be able to visualise and cluster the data, we need to find a low dimensional space that conveys the maximum variation in the data while retaining its structure. It is discussed earlier that one of the methods for reducing the space dimension is PCA. This is a problem well suited to PCA. For that reason, another transformation needs to be performed on data.

2.7.1. Applying principal component analysis

We perform PCA on the resultant similarity matrix. Finding the most important principal components from PCA techniques may result to obtain significant information from the data. An examination of the results through scree plots allows us to see the variance in the data represented by each principal component. In scree plots, the principal components with higher eigenvalues than the others should be considered. Using this approach, we may find the most important dimensions of the patient records for the purpose of analysis and visualisation. Once we find the important principal components, then we can use different clustering algorithms. It is also possible to understand the structure of the data to apply further data mining techniques.

2.8. Project patient records onto low dimensional vector space

After applying the transformations steps abovementioned, each patient is now represented in low dimensional vector space that facilitates clustering and visualisation. Clustering of patient records can be performed either manually or by using automatic clustering algorithms. One of the automatic clustering algorithms used is the k-means algorithm, which aims to partition the data into k clusters, where k is specified by the user input [203], [204]. The other clustering algorithm is expectation maximization (EM). This algorithm performs maximum likelihood estimation for samples in the mixture model by calculating the cluster probabilities in terms of mean and standard deviation for the numeric attributes and value counts [205], [206]. These two clustering algorithms were selected in order to demonstrate the methodology. Both algorithms have a straightforward implementation and can be easily run on the data. The results we obtained from either the manual clustering or when using any of automatic clustering algorithms will be assessed and evaluated. Further clustering algorithms are explored and discussed in chapter 4.

2.9. Implementation tools

To apply the work strategy, we have two calculation steps to apply to the patient records. In order for us to calculate semantic similarity we have used software developed by Statham [207]. A number of features have been added to the existing system to work on patient records described by Read Codes. The software is also capable to perform principal component

analysis. For clustering and visualisation, we have used MATLAB and Weka for clustering and to plot the figures needed for the study.

2.10. Methodology optimisation

The processing time needed for our methodology to be applied to patient records is computationally expensive owing to the calculation of semantic similarity and principal component analysis. Another challenge we faced during the study concerns the computational resources needed for processing this task. These challenges become an issue when performing the methodology to larger data sets. Therefore, a number of solutions have been proposed for optimisation.

2.10.1. Selecting the representative set of patients

Having defined the semantic similarity between individual patients, we could represent each patient as a vector of the semantic similarities to all other patients in the data set. However, this pair-wise comparison of patients is overwhelming and it becomes more complicated to deal with, in particular, when the given data set is large. Therefore, we have introduced an alternative strategy to reduce the number of comparisons between patients. This is achieved using the idea of a "covering set" of patients, which we define as a representative set of patients. For the covering set we need to select a subset of patients that can adequately represent the whole set of patients in the data. We can associate all patients with greater than some fixed similarity to a patient P as the set belonging to P. Note that it is possible for a given patient to belong to multiple sets. If we have n patients in the records, then we also now have n sets (one for each patient). The covering set would then be the subset of these sets such that every patient in the group belongs to at least one set in the covering group. Note that this is a much stronger statement than simply randomly picking a subset of patients. We are guaranteeing that every patient in the sample is covered by at least one representative patient set. Building a covering set is a known NP-complete problem – so there is no perfect way of achieving a covering set [208], [209]. The actually methodology we have used is described below.

The set of representative patients will act as a covering set for the whole space of patients in the data. This means that every patient in the data is within a certain similarity to one member of

the covering set. It is also worth mentioning that there will be cases an individual patient is being covered by one or more representative patients. Nonetheless, this behaviour is understandable and expected in our case as there are common patterns of diseases across the data. For example, a group of patients who have been diagnosed with both 'type II diabetes mellitus' and 'chest infection', those patients are covered by patients who have similar diagnoses. The choice of representative patients in our study is based on how patients are similar to each other in terms of their diseases.

In this process, what we are trying to achieve is to select representative patients that allow us to cover the whole space of patients, and not trying to partition that space into unique areas. However, if there are two representative patients that are quite similar to each other (based on the patients covered by each of them) all this means is that we just added a dimension to the data that does not add any new information. The use of PCA, in the subsequent step, is to help us to reduce this redundancy by taking away the dimensions with similar variability.

The representative patients are used when generating similarity scores for each patient in the data. This is important for these reasons: a) by comparing each patient to the representative patients rather than every patient in the data, thus the number of patient similarity scores for each patient is much more manageable, b) the number of comparisons needed are reduced allowing larger data sets to be used. The characteristics of a good set of representative patients would be points spread throughout the data such that every patient is within the threshold of at least one representative patient.

Representative patients are found by calculating the semantic similarity between each patient in the data set with each patient in the initially empty representative patients set. If a patient does not have a similarity score higher than a cut-off value k with any of the representative patients, then this patient is added as a representative patient. Otherwise, this patient has already been covered by one of the existing representative patients. The similarity cut-off value k is calculated by finding the average similarity between a random sample of patients in the data. Figure 2.4 illustrates how representative patients are chosen from a data set.

The number of selected representative patients may be vital in determining how patient records mapped into low dimensional space. An evaluation will be carried out to study the behaviour of

the method when using different number of representative patients to map patient records. Once the representative patients set is obtained, we generate a similarity matrix. However, this time the similarity matrix will be consisted of similarity scores calculated between patients P and the representative patients $Reps$. Table 2.2 below shows the similarity matrix after the optimisation step. To illustrate this, let $P = \{p_1, p_2, \dots, p_n\}$ be a set of n patient records and $Reps = \{rep_1, rep_2, \dots, rep_m\}$ be a representative set of k patient records, where $|Reps| \leq |P|$. All the similarity scores between P and Rep are stored in this similarity matrix, so that each patient in the dataset can be represented as a row vector of its pair-wise semantic similarity $sem(p_i, rep_j)$ to the patient records from the representative set $Reps$ where:

$$sim(p_i, Reps) = [\mathrm{sim}(p_1, \mathrm{rep}_1), \mathrm{sim}(p_1, rep_2), \dots, \mathrm{sim}(p_1, rep_m)]$$

Table 2.2. An illustration of similarity matrix between patients and representative patients. This matrix shows the similarity scores obtained from calculating the semantic similarity of patient records set P and representative set of patients Rep where each row in the matrix represents one patient record. Each patient record is represented as a vector of similarity scores.

		Representative patient set ($Reps$)			
		rep_1	rep_2	…	rep_m
Patients (P)	p_1	$sim(p_1, rep_1)$	$sim(p_1, rep_2)$	…	$sim(p_1, rep_m)$
	p_2	$sim(p_2, rep_1)$	$sim(p_2, rep_2)$	…	$sim(p_2, rep_m)$
	…	…	…	…	…
	p_n	$sim(p_n, rep_1)$	$sim(p_n, rep_2)$	…	$sim(p_n, rep_m)$

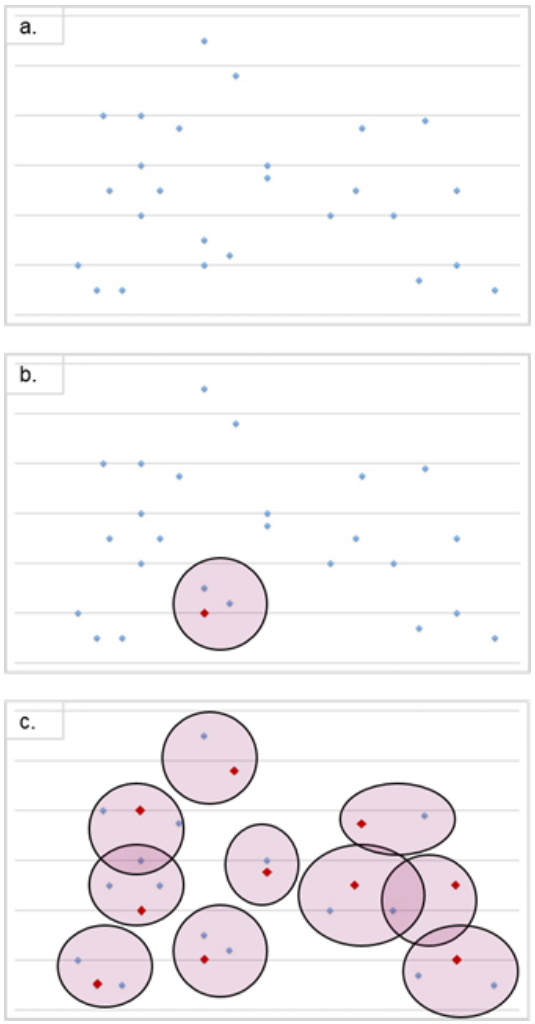


Figure 2.4. An illustration of the process of selecting representative patients from a data set. The process consists of the following steps: (a) each patient in the data set can be represented in a graph with the Euclidian distance between two points inversely proportional to the similarity of two patients. An initial empty representative patients set is created. (b) each point is then selected in turn. If the selected point is not within a threshold distance to a representative point it is added to the representative patient set. As the representative patient set is initially empty the first point chosen is always added to the set. (c) the process is repeated until there is no point on the graph which is not within the threshold distance to a representative point.

2.10.2. Comparing representative patients against themselves

Although finding a set of representative patients reduces the time of processing, the size of the obtained similarity matrix, in some cases, is still large and therefore performing PCA on such matrix is not always feasible. However, given that those representative patients are initially selected to cover the whole space of patients in the data, this subset can be used as an equivalent subset to the main set. An illustration of the similarity matrix of calculating the

semantic similarity scores between the representative patients against themselves is shown in Table 2.3. Visualising this matrix through PCA will give us an idea of how the representative patients cover the patient space and to show the common patterns of diseases found in the data.

Table 2.3. An illustration of similarity matrix of the representative patients. This matrix shows the similarity scores obtained from calculating the semantic similarity of the representative patients against themselves $Rep_n vs.\ Rep_n$ where each row in the matrix represents one representative patient. Each representative patient record is represented as a vector of similarity scores.

		Representative patient set $(\boldsymbol{Rep})$			
Representative patients $(\boldsymbol{Rep})$		rep_1	rep_2	...	rep_n
	rep_1	$sim(rep_1, rep_1)$	$sim(\mathrm{rep}_1, rep_2)$	...	$sim(rep_1, rep_n)$
	rep_2	$sim(\mathrm{rep}_2, rep_1)$	$sim(rep_2, rep_2)$	...	$sim(rep_2, rep_n)$
	...	...	...	...	...
	rep_n	$sim(\mathrm{rep}_n, rep_1)$	$sim(\mathrm{rep}_n, rep_2)$	...	$sim(rep_n, rep_n)$

2.11. Methodology: Development and application

The thesis divides the work done on the methodology across a number of chapters. This chapter has introduced the basic ideas behind the strategy. The first test of the methodology is shown in Chapter 3. This is a study done on a relatively small data set (patient records obtained from the Salford Integrated Record database). The results of this study indicate that the methodology does provide a mapping of records into a low dimensional vector space in which the position in the space to which a patient is mapped does have a useful medical interpretation. This indicated that the methodology warranted further analysis. However, we still have a number of questions that need to be answered about the methodology. For example: are we using the most appropriate similarity measure, is our method for determining representative patients appropriate, does the methodology scale to larger data sets and is it reproducible. We have also not looked at what might be appropriate methods for clustering this data and whether such clustering does provide useful new hypotheses.

These issues are addressed in Chapter 4. At this stage of the research, we had access to patient records from the CPRD database. This data consisted of 2,754,367 patients with 7,408,369 medical records. With data of this size, we were able to test the scalability behaviour of the methodology as well as looking to the other challenges. The first step of the methodology was to calculate the semantic similarity between patients. For this step, we evaluated a number of similarity measures with the aim of finding an appropriate measure for this kind of data. We also looked into the process of finding representative patients in more detail and investigated whether varying the number of representative patients could affect the mapping of patients. Another step of the methodology was to cluster patients based on their similarities. To cluster the patients, we applied the DBSCAN algorithm on the results obtained from the PCA. Using this algorithm for clustering this data seemed to produce clusters that make medical sense.

Once the methodology had been more rigorously assessed we then explored its application in chapters 5 and 6. Chapter 5 looks specifically around the stratification we seen across a large population cohort. Chapter 6 applies the methodology to hypothesis generation in the area of falls in the elderly.

Chapter 3: Taming EHR data: using semantic similarity to reduce dimensionality

Leila Kalankesh[a], James Weatherall[b], Thamer Ba-Dhfari[a], Iain Buchan[c], Andy Brass[a]

a School of Computer Science, University of Manchester, Manchester M13 9PL, UK

b AstraZeneca R&D, Alderley Park, Cheshire SK10 4TG, UK

c Institute of Population Health, University of Manchester, Manchester M13 9PL, UK

Address for correspondence:

Dr J.Weatherall

AstraZeneca R&D, Parklands

Alderley Park, Macclesfield, Cheshire SK10 4TG, UK

james.weatherall@astrazeneca.com

Author contributions

Leila Kalankesh	Edited manuscript and designed methods
James Weatherall	Edited manuscript and designed methods
Thamer Ba-Dhfari	Wrote manuscript, designed methods and analysed data
Iain Buchan	Edited manuscript and provided data access
Andy Brass	Edited manuscript and designed methods

Abstract

Medical care data is a valuable resource that can be used for many purposes including managing and planning for future health needs as well as clinical research. However, the heterogeneity and complexity of medical data can be an obstacle in applying data mining techniques. Much of the potential value of this data therefore goes untapped. In this paper we have developed a methodology that reduces the dimensionality of primary care data, in order to make it more amenable to visualisation, mining and clustering. The methodology involves employing a combination of ontology-based semantic similarity and principal component analysis (PCA) to map the data into an appropriate and informative low dimensional space. Throughout the study, we had access to anonymised patient data from primary care in Salford, UK. The results of our application of this methodology show that diagnosis codes in primary care data can be used to map patients into an informative low dimensional space, which in turn provides the opportunity to support further data exploration and medical hypothesis formulation.

Keywords:

Electronic Health Records, Semantics, Principal Component Analysis, Data Mining, Primary Health Care.

3.1. Introduction

Electronic Health Records (EHRs) are a valuable resource for health service providers, as well as biomedical researchers [102]. Owing to the complex structure and high dimensionality of medical records, it is a challenge to interpret and visualise these data and computerized techniques are often required [139]. Visualisation is an essential step of any exploratory data analysis strategy, and effective visualization of high dimensional data requires a dimension reduction strategy [202]. In this study, we investigate whether techniques based on the notion of semantic similarity, which has been used successfully in other areas of life science research [143]–[145], could provide a method for achieving this visualisation. The objects that will be mapped are the patients, the aim being to provide a data representation in which patients that have similar descriptions are mapped close to each other in a low dimensional vector space.

3.1.1. Read codes

In the UK, almost all primary care data is captured in electronic patient records. General practitioners (GPs) record a significant amount of this data in the form of Read Codes. Read Codes capture patient records in an agreed and standard, structured format which is machine readable [64]. These records provide a rich dataset as a record of the health of the nation. For example, doctors in secondary care use this information to provide a continuity of care to their patients [101], and at a wider level it is an invaluable resource for public health research [102] as well as for planning for future service provision in the National Health Service (NHS) [103]. Read Codes are comprehensive and are arranged in taxonomies that reflect a number of levels of increasing detail.

GPs effectively record patient encounters as a bag or multi-set of Read Codes $p = \{c_1, c_2, \ldots, c_n\}$ where p refers to a given patient record and c refers to the Read Code. For example, consider the following patient record (Table 3.1): $p = \{\text{C10F.}, \text{1372.}, \text{bd3 j.}, \text{G20..}, \text{2469.}, \text{246A.}\}$

Table 3.1. An example of a typical GP-patient encounter described by a bag of Read Codes.

Read Code	Rubric
C10F.	Type II Diabetes Mellitus,
1372.	Trivial smoker < 1 cig/day
bd3j.	Prescription of “Atenolol 25mg tablets”
G20..	Essential hypertension
2469.	Measurement of Diastolic Blood Pressure
246A.	Assessment of Diastolic Blood Pressure

It can also be seen that Read Codes describe a number of very different activities, from patient diagnosis (C10F.) to a record of the medication prescribed (bd3j.) to procedures carried out by the GP (2469.).

Although the data in the EHRs has been encoded in a standard way, its complexity and high dimensionality makes it difficult to analyse and interpret using many established data mining techniques [139]. Therefore, in this paper, we have explored whether it is possible to take high dimensional patient data and map it into a low dimensional vector space in a way that facilitates meaningful interpretation of the data. We believe that the ability to present data in this fashion will make it amenable to analysis through more traditional data mining strategies, as well as allowing a much more intuitive and straightforward environment for exploring hypotheses.

3.2. Materials and methods

3.2.1. Data set

The study dataset contains primary care information from Salford, a city in the North West of England. It consists of anonymised records from a population of 23,900 patients, with 442,589 individual Read Code entries. Ethical permission for the study was granted by the North West e-Health Board.

3.2.2. Semantic similarity

Semantic similarity is used to calculate how likely it is that concepts are similar to each other in terms of their meaning or semantic content. It is commonly used for ontology learning and information retrieval [210], [211]. The idea of using semantic similarity on patient records emerged from the fact that they are used as a representative model of primary care. We can calculate semantic similarity if there is any form of shared information between two objects, and so in this case we can essentially relate patient records to each other in terms of the content they share.

This representation has the capability to cope with the very heterogeneous data in patient records as it offers a featureless model, which does not ignore features of objects as these have already been captured in the Read Codes. [212] argues that the notion of proximity (whether through similarity or dissimilarity) is more fundamental than notion of feature. In addition, the hierarchical structure of Read Codes makes it possible to apply similarity methods based on ontological structure. It is also helpful that we have such a large corpus of data as that allows us to accurately measure the information content of the Read Codes in the Salford data.

The notion of semantic similarity is helpful in studying patterns in data as it can be used to express the shared and common properties of data [213] and therefore support identification of objects or entities that are conceptually close. A number of measures have been developed to find semantic similarity either between two concepts or between two sets (or bags) of concepts.

The concept of semantic similarity has been successfully applied in various applications. One of these is gene ontology (GO), where it has been used to compare genes and proteins based on the similarity of their functions [143], [188]. Another is WordNet, which is a lexical database for the English language [172]. Recent studies have applied semantic similarity to clinical records and found that the results generated by semantic similarity calculation matched the similarity scores obtained by physicians [144].

3.2.2.1. Measures to calculate similarity between two concepts

Measures developed to identify semantic similarity among concepts fall loosely into two categories: edge-based (link-based) and node-based (information-content based).

In edge-based measures, links and types of concepts are considered to be the data sources. This approach often depends on the depth of concepts in an ontological hierarchy and will select the shortest path from all the possible paths between the concepts. Two such examples are those developed by Wu and Palmer [186], and Leacock and Chodorow [187].

In node-based measures, the position of the node is ignored and the content and properties of the node are considered as the data source. Node-based measures are information content (IC) based and the calculation of semantic similarity depends on the frequencies of occurrence of the two concepts as well as that of their most informative common ancestor (MICA). Examples of node-based measures are those developed by Resnik [214], Jiang and Conrath [215], and Lin [181].

3.2.2.2. Measures to calculate similarity between sets of concepts

There are two aggregation approaches that can be used to calculate semantic similarity scores between two bags or sets of concepts. The first approach is pair-wise, where the concepts in one set are paired with the concepts in the other set and semantic similarity is calculated within each pair. The second approach is group-wise, where semantic similarity scores are calculated between the two bags of terms directly using set, graph or vector measures. In this study, we focus solely on pair-wise similarity, based on the findings in [145]. Examples of pair-wise similarity are the Maximum, Minimum and Average approaches, as explained in [143], [212]. As the names suggest, these aggregate functions use the maximum, minimum or average of all the pair-wise similarity values between two sets, as a proxy for similarity between the sets themselves.

3.2.3. Analysis Strategy

The strategy we used can be divided into the following steps:

1. Map patient records into a similarity space
    - Calculate the semantic similarity score for every possible pair of patients (including cases where the pair is the same patient twice)
    - Construct a matrix of the scores, where both the rows and the columns of the matrix are indexed by patient (see Table 3.2). We call this the similarity matrix
2. Map patient records into a vector space
    - Perform a principal component analysis (PCA) on the similarity matrix in order to map patients into a low dimensional vector space
3. Project patient records onto the low dimensional vector space
    - Cluster and visualise patient records to gain medical insights
4. Optimisation of the process

Each of these stages is described in more detail below.

3.2.4. Map Patient Records into Similarity Space

3.2.4.1. Semantic similarity calculation

Calculation of semantic similarity for patient records involves two steps. First, we find the semantic similarity between all pairs of Read Codes within an individual patient, where patient p in the dataset has a series of Read Codes c such that $p=\{c_1,c_2,\ldots,c_n\}$. This is done using the node-based semantic similarity measures mentioned earlier. We chose node-based over edge-based measures in order to leverage the information content of the data set, rather than ontology structure alone. Second, we find semantic similarity between pairs of patients given that patient $p_1=\{c_1,c_2,\ldots,c_n\}$, and patient $p_2=\{c_1,c_2,\ldots,c_m\}$. For each pair of patients, Maximum, Minimum and Average are calculated.

3.2.4.2. Generating the semantic similarity matrix

The semantic similarity scores obtained from previous steps were then used to construct a matrix, as shown in Table 3.2. To illustrate this, we define $P=\{p_1,p_2,\ldots,p_n\}$ to be a set of n patients, and $\text{sim}(p_i,p_j)$ to be the semantic similarity score between patients i and j. Each patient

corresponds to a single row in the similarity matrix, consisting of its pair-wise semantic similarity with all other patients. So, for example, this row for patient p_1 could be written as:

$$sim(p_1, P) = [\mathrm{sim}(p_1, p_1), \mathrm{sim}(p_1, p_2), \ldots, \mathrm{sim}(p_1, p_n)]$$

Table 3.2. The similarity matrix. Each row corresponds to a single patient, and is comprised of the similarity scores between that patient and all other patients in the data set.

	p_1	p_2	…	p_n
p_1	$sim(p_1,p_1)$	$sim(p_1,p_2)$	…	$sim(p_1,p_n)$
p_2	$sim(p_2,p_1)$	$sim(p_2,p_2)$	…	$sim(p_2,p_n)$
…	…	…	…	…
p_n	$sim(p_n,p_1)$	$sim(p_n,p_2)$	…	$sim(p_n,p_n)$

3.2.5. Map Patient Records into a Vector Space

3.2.5.1. Principal component analysis (PCA)

Patient records represented in similarity space are still represented in a very high number of dimensions. In order to visualise and get insights into such records we need to find a low dimensional space that conveys maximum variability in the data whilst retaining the key elements of its structure. This is a problem well suited to PCA. Thus, we perform PCA on the similarity matrix shown in Table 3.2. An examination of the PCA results through scree plots allows us to determine the most important principal components and thus the effective dimensionality of the data.

3.2.6. Projecting patient records onto a low dimensional vector space

After applying the transformation steps above, each patient can now be thought of as being represented in a low dimensional vector space that facilitates clustering and visualisation.

Clustering of patient records can be performed either manually or by using an automated clustering algorithm. We employed two automated clustering algorithms: (i) simple k-means, where the data is partitioned into *k* clusters, where *k* is specified by the user [204]; (ii) expectation maximization (EM), which performs maximum likelihood estimation by calculating the cluster probabilities in terms of the mean and standard deviation of their attributes [206].

3.2.7. Optimising the semantic similarity calculation process

Given the size of the data set in this analysis, the pair-wise comparison of patient records quickly becomes computationally overbearing. Therefore, we introduced an alternative strategy to reduce the number of comparisons between patient records. The step is to find a subset of patients that can adequately represent the whole space of patient records. We define the so-called representative set of patient records (a covering set of patient records) where $Rep = \{rep_1, rep_2, \ldots, rep_m\}$. This strategy has been successfully applied to handle similar situations [216], [217].

3.2.7.1. Selecting the representative set of patients

The strategy for selecting representative patients is as follows. We first define a similarity cut-off *x* such that if $sim(p_1,p_2) > x$ then we can consider patient p_2 to be similar to patient p_1 and will be represented by p_1 instead. The process continues until all patients in the dataset have a representative patient. The representative set of patient records were then used to calculate a reduced similarity matrix.

Selecting a set of representative patients helps to overcome the computational hurdle of the processing time needed to calculate semantic similarity within and between all patient records. As a result, finding an adequate number of representative patients is important. At the same time, we also need to maintain a balance such that the number of representative patients still provides an adequate description of the full set of patient records. Determining the optimal number of representative patients involves testing a range of values for the similarity cut-off *x*. Having found that optimal setting we can then perform the subsequent stages of the analysis.

3.2.8. Test data - projecting patients records into diagnosis space

To demonstrate the method, we have taken the records of all patients with Read Codes corresponding to diagnosis (codes starting with the capital letters A-Z). This subset of the data contains 22,931 Read Codes, corresponding to 1,737 distinct diagnoses, for a total of 5,327 patients.

3.3. Results

3.3.1. Optimising the semantic similarity calculation

We introduced two modifications to our methodology for optimisation. Firstly, we tested a range of values to find the optimal number of representative patients. Secondly, we computed the semantic similarity matrix by comparing the set of representative patients against the same set, rather than the whole set of patients.

To find representative patients we selected a range of values (based on similarity cut-off *x*) and applied two different semantic similarity measures: Resnik and Lin. The Jiang and Conrath measure was ultimately not used as it did not yield clear clusters via PCA.

Figure 3.1 shows the representation of diagnosis codes when using the Lin measure with Average. We tested a range of possible numbers of representative patients. Here we set the number of representative patients to 960 and 2,034 in blue/diamonds and green/crosses respectively. It should be noted that these two plots are similar in representing the diagnosis codes for patient records. We obtained similar results when performing the semantic similarity calculation using the Resnik measure with Average.

The second optimisation step for the methodology is done by obtaining similarity scores by comparing the representative patients set against the same set. In this step, we applied two semantic similarity measures: Resnik and Lin measures with Maximum and Average approaches, respectively (as these combinations yielded the clearest clusters via PCA). We compare the results we obtained from using this optimisation step with the results from using the whole set of patient records. The results from Lin and Resnik measures were similar.

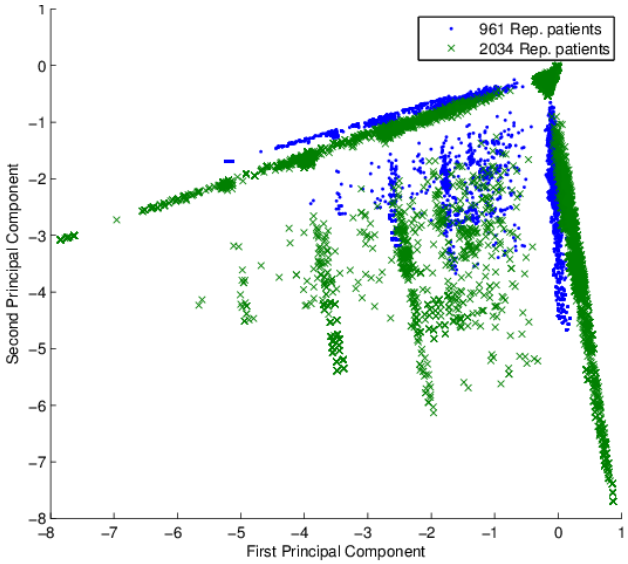


Figure 3.1. A PCA representation of diagnosis codes for patient records obtained using the Lin semantic similarity measure with the Average approach. We set the number of representative patients to 960 and 2,034 patients in blue/diamonds and green/crosses respectively.

3.3.2. Representation of patient records

On completion of the methodology optimisation step, we conducted further tests in order to evaluate the semantic similarity measures that were used. This was done by performing similarity calculations using different similarity measures. Each similarity measure has been applied along with the three aggregation approaches: Average, Maximum and Minimum. Figure 3.2 shows the results produced from using the Resnik measure with the Maximum approach. It can be seen here that the first two principal components capture nearly all the variation in diagnosis data. Visualisation of patient records based on the first two principal components was therefore deemed sufficient to demonstrate the methodology.

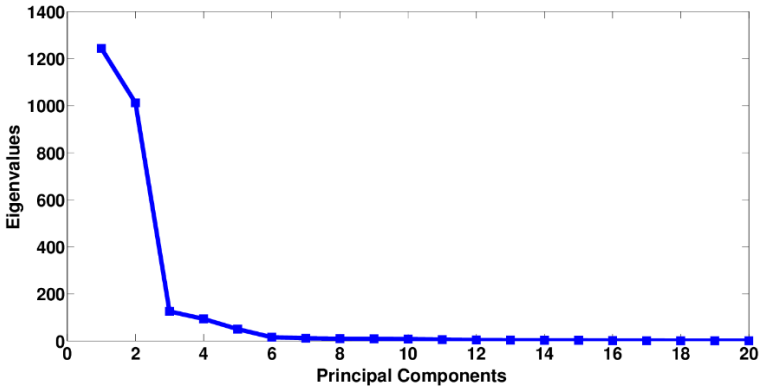


Figure 3.2. A scree plot showing the degree of variation in diagnosis codes that is described by the first 20 principal components.

3.3.3. Clustering analysis

In manual clustering, we identified 12 clusters. A cluster analysis was performed to identify the dominant diagnosis codes within each cluster (Figure 3.3). The analysis shows significant disease patterns where e.g. patients with 'chest infection', 'upper respiratory infection', 'asthma' and a collection of other diseases were grouped together towards the bottom of the plot. To the left of the figure patients who have circulatory system diseases such as 'angina pectoris', 'hypertensive disease' and 'stroke' are placed close to each other. The right side of the figure has patients who have been diagnosed with diabetes and related diseases. An interesting finding in this figure is that the patients clustered towards the top-middle of the plot were diagnosed as having both circulatory system diseases and diabetes. For instance, there is a cluster of patients who were diagnosed with both diabetes mellitus and angina pectoris. The association between these two diseases is well known, with diabetes being regarded as a risk factor for angina due to its accelerating effect on atherosclerosis [218].

We also applied two automated clustering algorithms on the representation of patient records found in Figure 3.3. These clustering algorithms are the simple k-means algorithm and Expectation Maximization (EM) (see Appendix A for the cluster analysis of k-means and EM results). Cluster analysis produced similar results to the manual clustering. However, the difference here concerns the number of clusters. In manual clustering, 12 clusters were identified, while the EM algorithm generated 24 clusters. Therefore, it is possible that the EM

algorithm might reveal more details regarding the classification and patterns across patient records by providing a more granular clustering of the data.

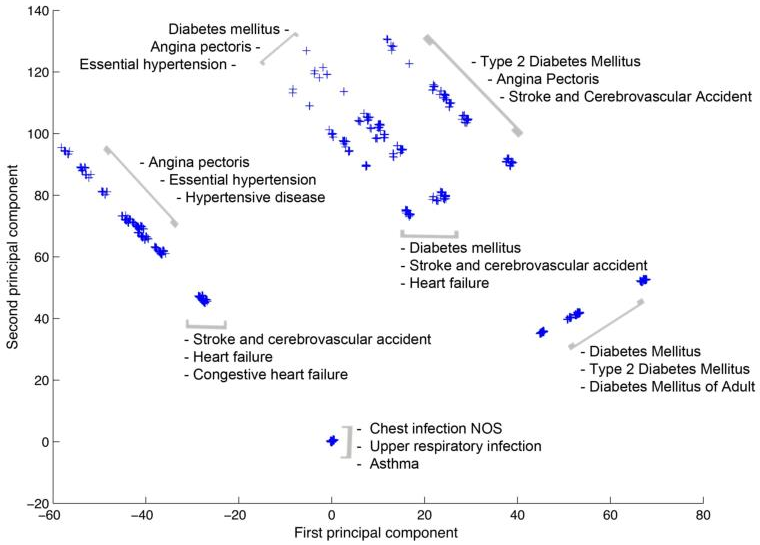


Figure 3.3. A cluster analysis of the PCA results (Resnik + Maximum approach), where 12 clusters have been manually identified. The analysis shows that patients with similar diagnosis codes were grouped together in nearby clusters.

3.4. Discussion

The results obtained in this study show that patients with similar sets of diagnoses are grouped together. This was one of the key objectives of this work, to develop a mapping from a similarity space to a low dimensional vector space that placed patients with similar diagnoses close together.

This finding may provide a foundation for further analysis of electronic healthcare record collections. In particular dimensionality reduction via semantic similarity could provide an environment for the generation of new biomedical hypotheses. In this work we identified associations such as that between diabetes and angina, which is relatively well known. However, it should also be possible to identify new disease associations, which in turn yield hypotheses to be proven or refuted via laboratory experiments, or even in-silico network biology analytics [42].

During the implementation of our methodology, we encountered a number of computational challenges. As a result, we introduced two modifications to our methodology for optimisation purposes. One of these modifications was to find representative patients. The second

modification was to obtain the semantic similarity scores by comparing the whole set of representative patients against the same set. Introducing these modifications allowed us to streamline the methodology and reduce the overall time needed to obtain the study results.

The first stage of our methodology was to map patient records to similarity space and to do so we applied a number of semantic similarity measures. The variety of the measures raised a key question as to how each measure would capture the essential semantic similarities within the EHR collection. To address this question, we applied five semantic similarity measures with three similarity aggregation approaches (Average, Maximum and Minimum) to the dataset. The results show that there are different representations of the records for each measure. They also show that these measures capture patient records in different ways. As a result, we chose to continue with measures such as Resnik and Lin that showed clear and significant patterns of diagnosis codes. The other measures were excluded for one of two reasons. First, the scree plot of the PCA analysis did not provide significant variance in the data to enable representation in a low dimensional space. Second, the figures generated for the PCA analysis did not cluster patient records as clearly as the chosen measures.

3.5. Conclusion

This study was undertaken to explore and evaluate certain techniques such as semantic similarity and principal component analysis, in order to investigate whether they could be beneficial in the interpretation of large and complex patient record collections. We have proposed an approach that uses semantic similarity to map patient records into similarity space and then applies PCA in order to further reduce the dimensionality of the data. This work allowed us to characterize data at the population level. We have shown that it is possible to take data in the form of Read Codes and map it into a low dimensional space in ways such that distance relates to similarity in patient records. It is clear that mapping the patient data into a vector space opens up the possibility of applying a wide range of data mining strategies which have not yet been explored. It is also worth noting that nothing in the implementation restricts the methodology to Read Code data. The same methodology could equally be applied to the analysis and visualisation of many other sources of data which are described using terms from taxonomies or ontologies.

Acknowledgements

The authors would like to thank Parivash Ashrafi for continuing this work academically and Sajan Khosla at AZ for providing helpful input.

Chapter 4: Testing the scalability behaviour of the methodology at large-scale data sets

4.1. Introduction

The data sets used in health informatics, whether in the form of structured or unstructured data, are growing rapidly. In part this is because the technology used to capture them is developing rapidly, whether through ubiquitous remote sensing, increasing use of electronic patient records and genome sequencing [219]–[221]. Big data is a term describing data that is typically high volume, volatile and with complex structures and high dimensionality [45], [222], [223]. Businesses have raised questions as to how best to manage and leverage these very large data sets and capitalise on their advantages. Analyse of big data may lead to scientific discoveries and economic benefits [224]–[229]. In medical research, analysing big data sets of patient records can provide accurate predictions of future observations, as large-scale data contains more precise relationships between features than small scale data [230]. It could also provide a better understanding of distinct population patterns and allows for the extraction of common features. Furthermore, big data may be able to explore hidden structures, which do not exist or cannot be classified as outliers in small data, across different subpopulations of patients [224], [231].

While big data provides new opportunities, it also raises challenges for data scientists regarding data analysis approaches [226], [232]. Such challenges are derived from the four characteristics of big data: volume (size of data), variety (heterogeneity), velocity (processing speed) and value (accuracy) [233]. Big data cannot be scaled using traditional data analysis methods and systems due to these characteristics [234]. Big data also raises computational issues, such as scalability and storage performance bottlenecks [228], [235]. In addition, the high dimensionality of big data can introduce noise accumulation measurement errors and incidental homogeneity [224]. The combination of patient records and genome data is a significant challenge, due to the huge volume, variety, data quality problems and velocity issue regarding the use of real time data [236].

As presented in Chapter 2 and 3, we developed a novel methodology that builds upon the idea of semantic similarity, taking patient data in the form of codes and mapping it into a low dimensional vector space in which distance relates to the similarity of patient phenotypes. In Chapter 3, we demonstrated the use of this methodology on small scale patient data set. The next phase of the research was to develop this methodology further and explore its application in significantly larger data sets such as the Clinical Practice Research Datalink (CPRD)CPRD database. The CPRD database is one of the largest databases that holds complete electronic records for over 11 million patients from primary care practitioners across the UK [38], [118], [119]. The CPRD provides anonymised patients records for health research purposes. Working on data with such size gave us the opportunity to test the scalability of the methodology as well as looking to the other challenges. These challenges include choosing an appropriate similarity measure, finding an effective set of representative patients for data at this scale, and finding a clustering algorithm that provides sensible clustering of patient data. The work in this chapter presents a discussion of how these challenges were addressed.

4.2. Materials and methods

4.2.1. CPRD patient data set

The CPRD data set used in this study consisted of anonymised patient records from 2011. In this year, there were 72,928,339 medical records for a total of 4,491,207 patients. The medical records were described using the Read codes system and include codes related to prescribed drugs, laboratory results and diagnoses. For the purpose of this study, only records with Read codes corresponding to diagnosis codes (codes starting with capital letters A-Z) were included in the study. This subset contained 2,754,367 patients with over 7,408,369 diagnosis entries. Other information about patients, such as age and gender, was also provided. The ethical permission for the study was given by the Independent Scientific Advisory Committee of the Medicines and Healthcare Products Regulatory Agency (ISAC-MHRA) reference number: 15_249.

The size of the data poses a significant challenge to directly applying the methodology. With such large-scale data, pair-wise comparisons of patients quickly become computationally

overbearing. For example, if we want to generate a similarity matrix for this data set, the size of the matrix will be extremely large and require a large amount of computing power. For a data set with 2,754,367 patients, the size of the similarity matrix will be $7.59e10^{12}$. Even if we found representative patients, the size of the similarity matrix would still be large, making it difficult to handle. To determine whether an alternative method would work, we trialled the process with patient subsets. We split the data set into groups based on patient age and gender. There are 16 age groups for male and female patients. This made a total of 32 patient groups. The age groups are based on five-year age intervals, except for the first group, which contained patients between 0 and 17 years, and the last group, which contained patients aged 90 or above. Table 4.1 presents the distribution of patients across the 32 groups as well as the number of records in each group. Splitting the data set into different age/gender groups helped us to compare the results between different groups (e.g. comparing common disease patterns between male and female children).

There are a number of parameters we need to optimise; the semantic similarity measure, the algorithm for choosing the representative patients and the clustering strategy. The assumption is made that these are relatively independent so that the choice of similarity measure will not be a function of the method used for choosing the representative patients.

4.2.2. Choice of semantic similarity measures

The first step of the methodology involves computing semantic similarity between patient records and to do so a number of semantic similarity measures were implemented. To address this question, we conducted an evaluation using three semantic similarity measures the Resnik, the Lin and the Jiang and Conrath. This analysis also examined three different methods for calculating the similarity between bags of patient codes. As before, we are looking for a similarity measure that will provide an informative clustering of the data in a low dimensional space. This was done using the same method for choosing representative patients as described in chapter 2. The test data used for this evaluation was the Salford GP data that has been used before in Chapter 3.

Table 4.1. A summary of the study data set. The data set consists of anonymised patient records from the CPRD database. The records were registered in 2011. The data consists of 7,408,369 records for a total of 2,754,367 patients. We divided the data into 32 groups based on the age and gender of patients.

Patient gender	Age	Number of patients (N = 2,754,367)	(%)	Number of records (N = 7,408,369)	(%)	Records/patient (Average = 2.87)
Male patients	Birth-17	270,334	(22.31)	616,723	(20.18)	2.28
	18-24	88,927	(7.34)	173,595	(5.68)	1.95
	25-29	62,428	(5.15)	127,102	(4.16)	2.04
	30-34	65,022	(5.37)	137,862	(4.51)	2.12
	35-39	69,824	(5.76)	153,929	(5.04)	2.20
	40-44	79,956	(6.60)	183,634	(6.01)	2.30
	45-49	84,105	(6.94)	199,218	(6.52)	2.37
	50-54	79,656	(6.57)	200,268	(6.55)	2.51
	55-59	74,293	(6.13)	197,084	(6.45)	2.65
	60-64	82,505	(6.81)	231,228	(7.57)	2.80
	65-69	75,733	(6.25)	224,815	(7.36)	2.97
	70-74	59,795	(4.93)	189,241	(6.19)	3.16
	75-79	50,942	(4.20)	174,032	(5.69)	3.42
	80-84	36,730	(3.03)	131,662	(4.31)	3.58
	85-89	21,571	(1.78)	79,444	(2.60)	3.68
	≥ 90	9,932	(0.82)	36,064	(1.18)	3.63
Female patients	Birth-17	270,733	(17.55)	622,191	(14.30)	2.30
	18-24	134,561	(8.72)	333,667	(7.67)	2.48
	25-29	99,765	(6.47)	259,538	(5.96)	2.60
	30-34	100,566	(6.52)	267,074	(6.14)	2.66
	35-39	98,656	(6.40)	266,430	(6.12)	2.70
	40-44	107,625	(6.98)	294,573	(6.77)	2.74
	45-49	110,807	(7.18)	312,626	(7.18)	2.82
	50-54	101,275	(6.57)	292,589	(6.72)	2.89
	55-59	89,854	(5.82)	264,115	(6.07)	2.94
	60-64	94,306	(6.11)	284,283	(6.53)	3.01
	65-69	85,381	(5.53)	270,808	(6.22)	3.17
	70-74	69,938	(4.53)	236,014	(5.42)	3.37
	75-79	62,849	(4.07)	226,957	(5.21)	3.61
	80-84	53,002	(3.44)	193,480	(4.45)	3.65
	85-89	37,647	(2.44)	137,524	(3.16)	3.65
	≥ 90	25,649	(1.66)	90,599	(2.08)	3.53

4.2.3. Selecting representative patients

The second stage of the optimisation was to explore different strategies for selecting representative patients using the similarity measure selected in section 4.2.2. A number of issues were explored. The first step was to explore how large the covering set (the number of representative patients) needed to be. To achieve this, we altered the cut-off values associated with the similarity values that described the set associated with a representative patient. We

then applied the methodology using each of the generated representative patients sets to a test data. This data consisted of records of 82,505 male patients aged between 60 and 64 years (see Table 4.1). The assumption made here was that we had a large enough covering set at the point when adding more representative patients no longer changed the final clustering pattern observed. Further analysis was done on the chosen set of representative patients. This include finding how many patients were covered by each representative patient (what is the distribution of set sizes in the covering set?), and the mapping of the representative patients in low dimensional space.

We also explored the scaling performance by asking whether we would expect to find more representatives in larger data sets. In essence, how many patients would you need to sample before you had sensibility covered the diagnosis space. In order to test the performance of the method, we conducted an evaluation to see its scalability behaviour when applied to large scale data sets. For the purpose of the evaluation, we split the CPRD data in different way. In here, the data set (n=2,754,367 patients) was split into 14 subsets, where 13 subsets consisted of 200,000 patients and one subset had the rest of 154,367 patients. Patient records were assigned to the subsets in a random way. It should be noted that there is no link between these subsets and the 32 patient groups mentioned in Table 4.1. Figure 4.1 gives pseudo code for the algorithm followed to evaluate the process of selecting representative patients.

1: Set i = 1

2: $patient\ subset_{all} = \{\}$

3: $reps_{all} = \{\}$

4: split CPRD data set into 14 subsets

5: Choose a random subset from the 14 subsets

$$patient\ subset_{all} = \{patient\ subset_i\}$$

6: Find representative patients in $patient\ subset_{all}$ and save them as $reps_{all}$

7: Choose another random subset and merge it with the previous subset(s):

$$patient\ subset_{all} = \{patient\ subset_{all} + patient\ subset_{i+1}\}$$

8: Find representative patients in $patient\ subset_{all}$ and save them as $reps_{all}$

9: i++

10: Repeat steps 7 to 9 for all 14 subsets

Figure 4.1. The algorithm used to evaluate the process of selecting representative patients.

From this analysis, we were able to see the sample size at which adding new patients was unlikely to generate more representative patients.

The final issue that was investigated was the reproducibility of the method. As discussed in chapter 2, generating a covering set is an NP-complete problem. It is possible that the results seen are a function of the order in which the representative sets were chosen. Analyses were therefore performed in which the data was shuffled. This will lead to a different starting point for the generation of the representative patients. The final clustered data was examined for analyses that had been run on shuffled data sets to explore how sensitive the final interpretation was to the precise choice of representative patients.

4.2.4. Clustering patient records

Once the data has been mapped into a low dimensional vector space we would like to apply some basic data mining strategies, such as clustering, to the data. In the Salford study (Chapter 3), we applied two clustering algorithms: the expectation maximization (EM) and k-means algorithms. Both algorithms performed well on the Salford data set. However, these algorithms do not work properly on large scale data sets because of complexity and computational costs of the data [237]–[239] [228]. One of the drawbacks of using the k-means is that the number of clusters has to be specified by the user beforehand [238], for this methodology, we needed an algorithm that allows automatic generation of clustering. K-means is also less accurate when dealing with a large amount of outliers [238]. Moreover, the efficiency of this clustering algorithm is low when there are a large number of clusters in the data [239].

Thus, there was a need to test other clustering methods to scale and speed up the clustering process and at the same time maintain good clustering quality. Consequently, researchers have always targeted the scalability and the speed of clustering algorithms, which has resulted in the devolvement of a number of big data clustering techniques, such as sampling-based, hierarchical and density-based algorithms. Big data sampling-based algorithms such as CLARNS (Clustering large Applications based on Randomized Search) [240] and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [241] perform clustering on data

sample and then use iterative control strategies to optimise an objective function. Such algorithms make clustering more effective and efficient than traditional techniques. However, the run time is unreasonable for large databases with a huge number of objects [242]. Hierarchal clustering techniques build a hierarchical decomposition of the objects, represented by a dendrogram. In the hierarchy, each node represents a cluster of the data sets. Hierarchical clustering does not require number of clusters k as an input, but a termination condition should be defined to identify when the merge process should be finished. The main challenge with this type of clustering is the difficulty of obtaining suitable parameters for a termination condition [242].

Another algorithm that may work effectively on large data sets is DBSCAN (Density-based Spatial Clustering of Applications with Noise) [242]. DBSCAN is a typical density-based clustering technique that has been adopted in a number of applications in different areas [243]. In DBSCAN, the density associated with a point is obtained by counting the number of points in a region with a specified radius. Points with a density above the specified threshold are constructed into clusters, while points with a low density are marked as outliers. DBSCAN requires two parameters epsilon (*eps*) and the minimum number of points (*minPts*) required to form a dense region [242].

As the size of patient records increases, using DBSCAN to cluster such large data sets is suitable [243]. One feature of DBSCAN that made it beneficial to our analysis is the ability to perform the clustering without specifying the number of clusters as it depends on the nature of the data [243]. Also, as this algorithm is based on the idea of clustering points based on their density, this feature will help to discover clusters of patients who have high frequent diseases in the data. For these reasons, we think that perform this algorithm on the study data could provide us with interesting insights about patients.

One way to test and evaluate the clustering produced by the DBSCAN is to use the Silhouette coefficient measure [244]. Silhouette measure is used to assess the parameters used to perform the clustering. It measures the distance between the resulting clusters, where a higher Silhouette score relates to a model with better defined clusters, and lower scores indicate that the model has too many or too few clusters. Silhouette scores range between -1 and 1.

4.2.5. Analytical tools and high performance computing infrastructure

Throughout the study, we have been given access to some of the high performance computing infrastructure at AstraZeneca to perform the analysis [245]. The CPRD data set was stored in IBM Netezza data warehouse, a powerful parallelised data warehousing system. The medical data was extracted using the Aginity Workbench[1] software which provides basic SQL (Structured Query Language) queries along with more advanced features for querying large volumes of data.

In the Salford study (Chapter 3), most of the computational, statistical calculations and visualisation of data was carried out by using the programming language Java and the Weka software. Although these tools perform well on Salford study, the computational costs of analysing CPRD data were significantly greater. Python can be an effective alternative to using both Java and Weka for such analysis. Python is an effective programming language for analysing big scientific data and has been used in a wide range of applications [246]–[255]. For this reason, we have used python for the statistical calculations, clustering and visualisation of patient records in order to generate clusters of patient and to produce graphs needed to generate hypotheses. This also provided useful test of the correctness of the coding as the results from the python and Java/Weka codes could be compared to check that they gave equivalent results.

4.3. Results

4.3.1. Semantic similarity measures evaluation

This evaluation was done by performing similarity calculation using three measures: Resnik, Lin and Jiang and Conrath. The measures included in this evaluation are all node-based measures. Each similarity measure has been applied along with three aggregation approaches: Average, Maximum and Minimum. The results of this evaluation showed that each of the three similarity measures captured the data in different way. There were different representations of the records for each measure. Even though, applying one similarity measure with different

[1] http://www.aginity.com/workbench/netezza/

aggregation approach would produce different results. The results of this calculation are presented in Figure 4.3, Figure 4.4 and Figure 4.4 for the following measures Resnik, Lin and Jiang and Conrath, respectively. It can be seen from (Figure 4.3e and f, Figure 4.4e and f, and Figure 4.4e and f) that using the Minimum approach with any of the three measures did not provide a clear representation of the records compared to the other two approaches: Average and Maximum, as well as their scree plots did not show significant variance in the data. Likewise, applying the Jiang and Conrath measure with any of the three aggregation approaches (Figure 4.4) did not yield clear representation of the data. Noticeably, the combination of the Resnik measure with the Maximum approach (Figure 4.3) showed clear and significant patterns of the data. As a result, we chose to continue with these measures. The other measures were excluded for one of two reasons. First, the scree plots generated by the PCA analysis did not provide significant variance in the data to enable representation in a low dimensional space. Second, the PCA plots did not cluster patient records as clearly as the chosen measure.

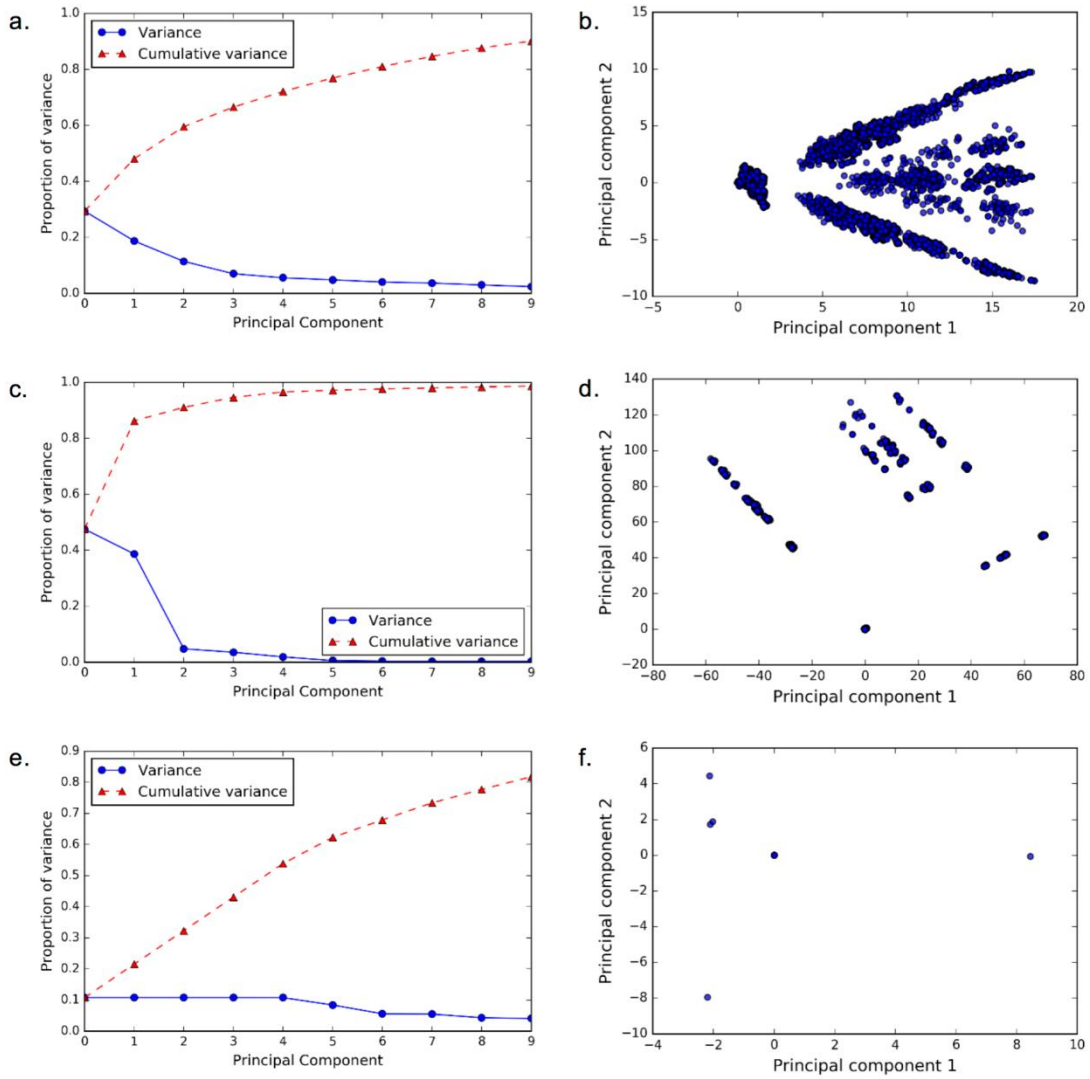


Figure 4.2. Mapping patient records into low dimensional vector space using the Resnik measure along with three aggregation approaches: Average (a and b), Maximum (c and d) and Minimum (e and f). (a), (c) and (e) present the scree plots showing the variation in the data captured using the first 10 principal components. Scatterplots (b), (d) and (f) present the principal component analysis (PCA) representation of patient records in low dimensional space; x-axis: 1st principal component; y-axis: 2nd principal component. The scree plots in (a), (c) and (e) show that the first three principal components capture 59.5%, 91.0% and 32.3% of the variation in the data using the Average, Maximum and Minimum, respectively. The combination of the Resnik and the Maximum approach (c and d) seems to show significant patterns of the patient records.

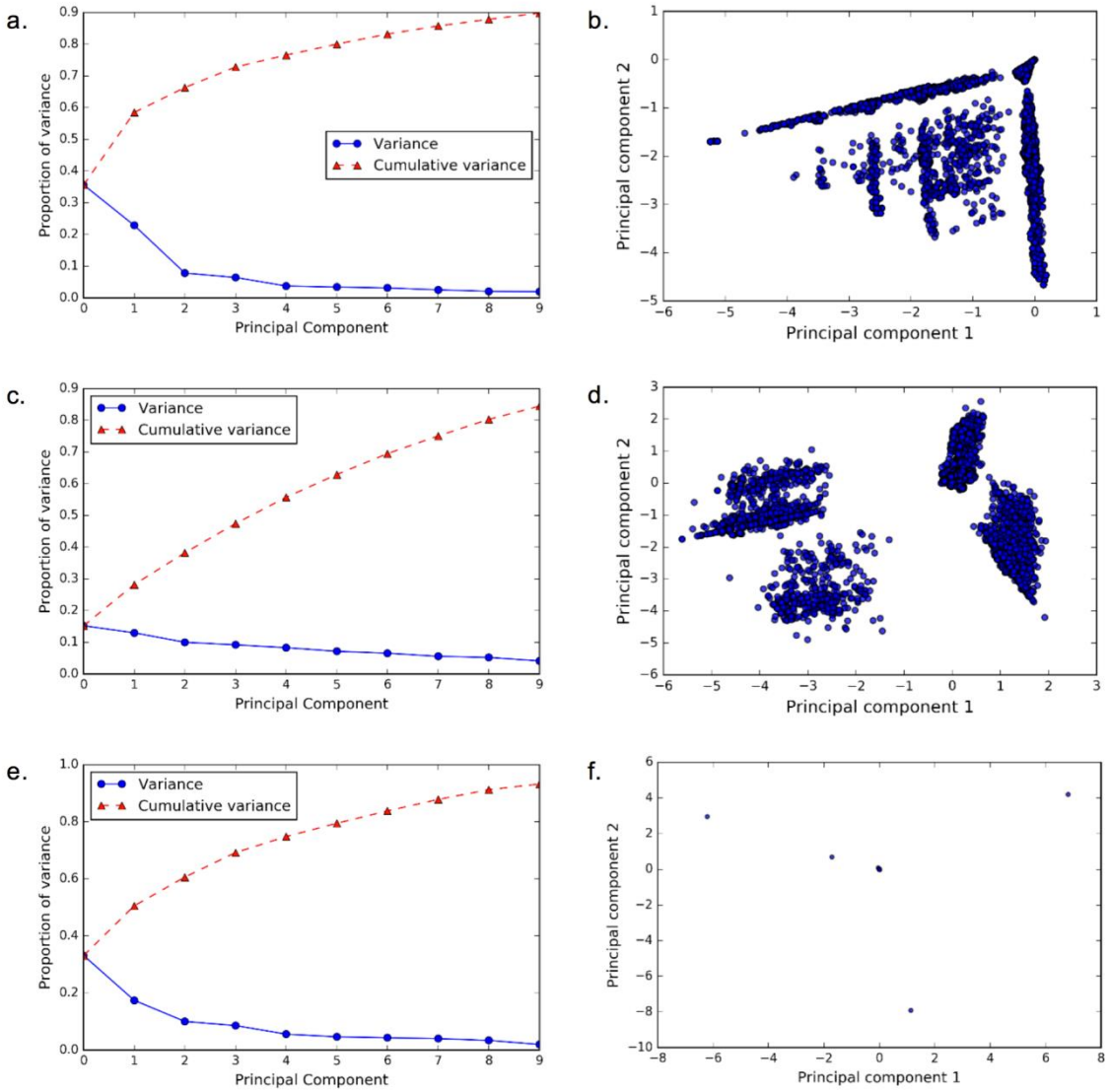


Figure 4.3. Mapping patient records into low dimensional vector space using the Lin measure along with three aggregation approaches: Average (a and b), Maximum (c and d) and Minimum (e and f). (a), (c) and (e) present the scree plots showing the variation in the data captured using the first 10 principal components. Scatterplots (b), (d) and (f) present the principal component analysis (PCA) representation of patient records in low dimensional space; x-axis: 1st principal component; y-axis: 2nd principal component. The scree plots in (a), (c) and (e) show that the first three principal components capture 66.4%, 38.2% and 60.6% of the variation in the data using the Average, Maximum and Minimum, respectively.

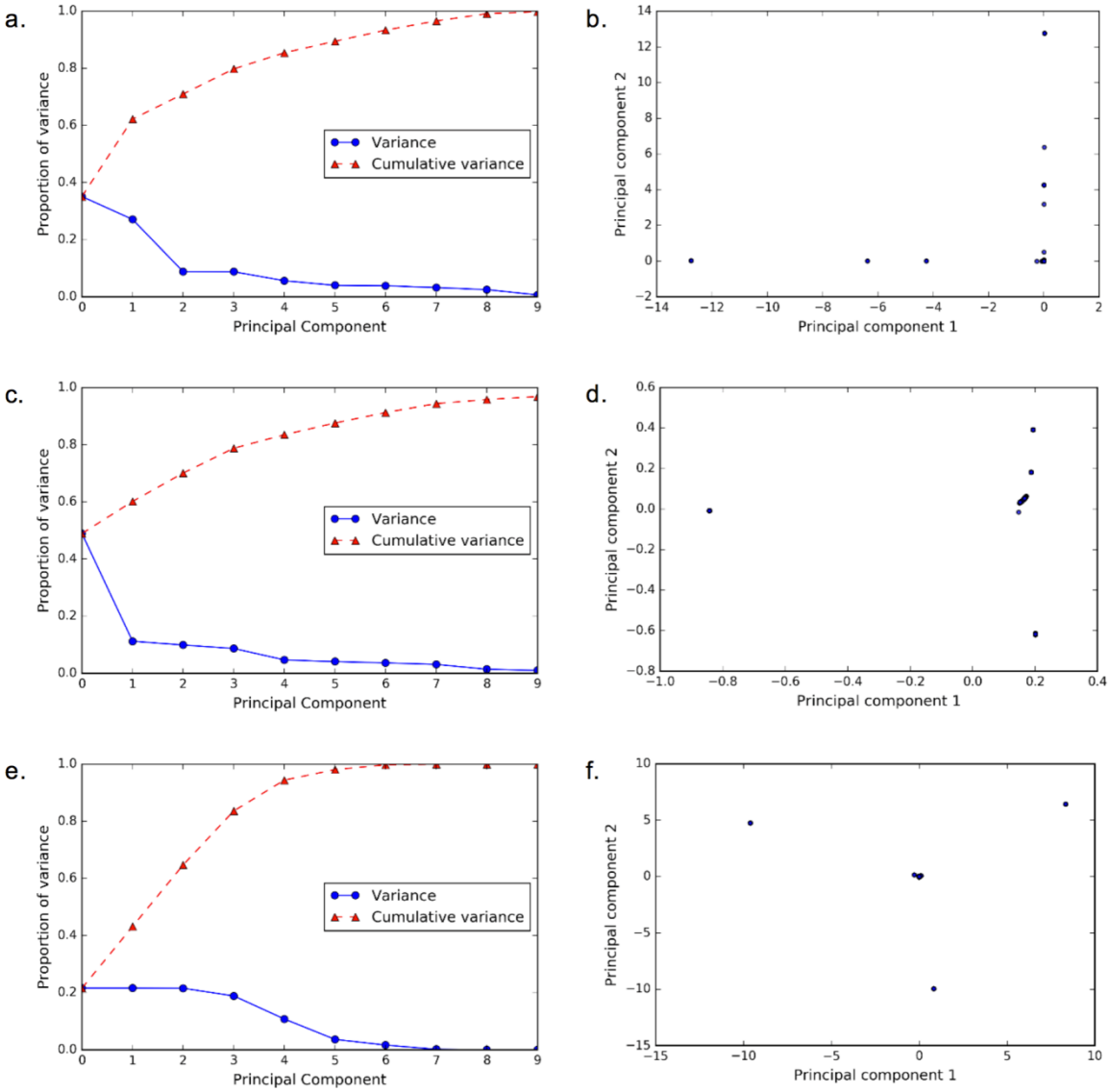


Figure 4.4. Mapping patient records into low dimensional vector space using the Jiang and Conarth measure along with three aggregation approaches: Average (a and b), Maximum (c and d) and Minimum (e and f). (a), (c) and (e) present the scree plots showing the variation in the data captured using the first 10 principal components. Scatterplots (b), (d) and (f) present the principal component analysis (PCA) representation of patient records in low dimensional space; x-axis: 1st principal component; y-axis: 2nd principal component. The scree plots in (a), (c) and (e) show that the first three principal components capture 71.0%, 70.1% and 64.7% of the variation in the data using the Average, Maximum and Minimum, respectively.

4.3.2. Selecting representative patients

In order to the find representative patients, firstly we need to define a similarity cut-off value k. This value is defined by finding the average similarity for a random sample of patients. As the study data set is large, we selected a random sample of 10% of patients from the study data set (n = 275,436 patients). The average similarity score for this sample was 3.13. We used this value as a cut-off to select representative patients. By using this cut-off, there were a total of 3,500 patients identified as representative patients.

To ensure that the selection of those representative patients was not because of any bias or internal structure in the data (e.g. patient records ordered in certain criteria), we shuffled patient records to remove such structure. We repeated the shuffling twice on the data and in each time we calculated the representative patients. We used the same cut-off value to find representative patients in each run. As a result, two different sets of representative were obtained from these two runs. One set was with 2,622 representative patients and the second with 3,589 representative patients.

We also tested the method using another cut-off, this time we set it to 1.22, and found that 601 patients were identified as representative patients. To compare this set with the previously obtained sets, we checked for an overlapping between the representative patients across the four sets (Figure 4. 5).

The overlapping between the four representative patients sets shows that the method we used is quite consistent when picking up patients who can cover the whole space of patients. By looking for an intersection between the sets, we found that all four sets shared 572 patients in common. While the intersection between the 2,622 set, the 3,500 set and the 3,589 set was 2,524 representative patients. That is 96.3% of the patients in the 2,622 representative patients set. Furthermore, the intersection between the set of 3,500 representative patients the set of 2,622 representative patients was 2,526 patients; whereas with the third set it was 3,366 patients.

Notably, the number of the overlaps between the four sets of representative patients increased in the bigger sets. When we reached to the 3,589 set this number became stable and less

sensitive to change. This suggests that at this point we have enough coverage of the data and that the method had selected most of the patients who can be used as representatives. Furthermore, the overlapping between the sets indicate that the patient space is a mixture of high dense areas where patients had common diseases patterns; and other sparse areas where patients with rare patterns. The representative patients in the smaller sets would not be enough to cover all patients as the method was only picking up the common patients (i.e. patients who have high similarity scores with other patients). However, the bigger sets seemed to include those representative patients in the smaller sets as well as representatives for patients with less density (outliers). By this way we ensured that we provided a reasonable covering set for that space.

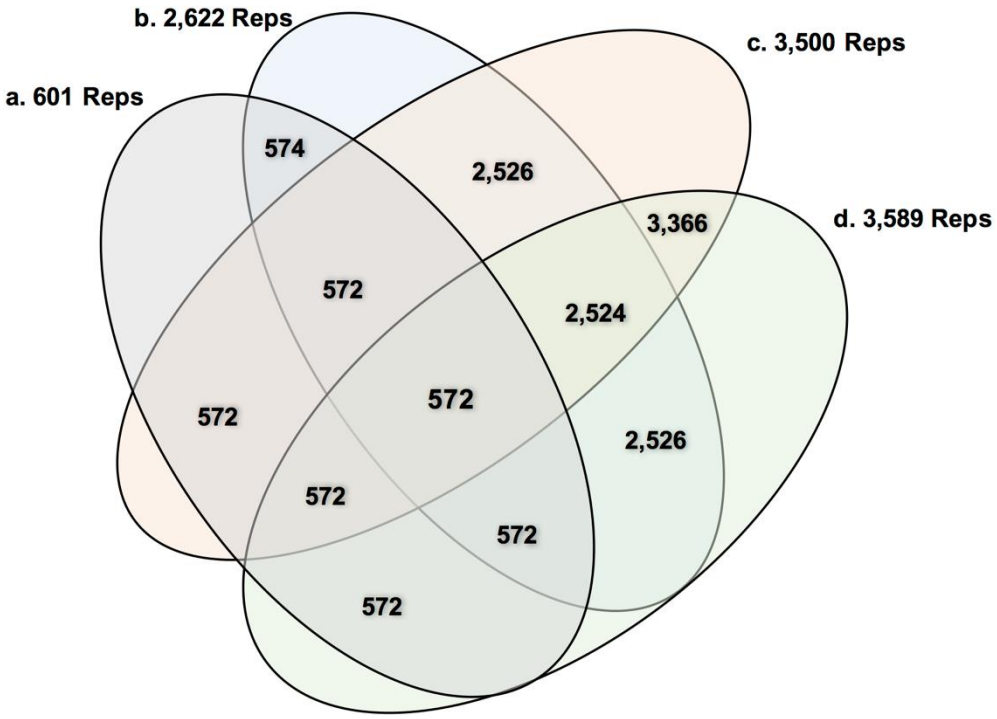


Figure 4. 5. The overlapping between four representative patients sets: (a) 601 set overlaps, (b) 2,622 set, (c) 3,500 set and (d) 3, 589 set. All four sets have 572 representative patients in common.

4.3.2.1. Testing four sets of representative patients on a test data

Using any of the four representative patients sets obtained previously to map patient records would help to minimise the number of the pair-wise comparisons needed to construct the similarity matrix. However, the low dimensional representation of patient records is expected to behave differently when using different number of representative patients. For this reason, we applied the methodology to a test data using each of the four representative patients sets. The Resnik measure and the Maximum approach were used to compute the semantic similarity.

The results of mapping the patient records using the four representative patients set is shown in Figure 4.6. It can be seen from this figure that the PCA representation of the data shows four clear groupings of patients in each of the four plots. The annotations in each plot in this figure show the labels of each group. The representation of patients in Figure 4.6c, Figure 4.6d would appear to have more similar structure than the other two. In order to investigate these results in more detail, we performed two-step analysis. We started by tracking the movement of patients between groups when using different representative patients set. This was followed by identifying the common disease patterns in each group.

In the first step of this analysis, we looked at each group of patient in Figure 4.6 in order to identify the patients. Table 4. 2 shows the size of each group in the four plots. By tracking patient movements, we found that a number of patients had moved between groups when applying each of the four representative patients set. However, when using the 3,500 set (Figure 4.6c) and the 3,589 set (Figure 4.6d) there was less movement between the groups. For example, when we traced patients in group 1 in Figure 4.6c, we found that they split in Figure 4.6d by 94.1% in group 1, and 5.9% in group 3. The same was observed with patients in group 2 where they split in Figure 4.6d by 92.5% in group 2 and 7.5% in group 4. While patients in group 3 split by 90.2% in group 1 and 9.8% in group 3. Patients in group 4 split in Figure 4.6d by 93% in group 2 and 7% in group 4.

The second step of the analysis involved analysing the medical codes defined for patients in each of the four groups (Figure 4.6a - Figure 4.6d). The overall analysis shows that all four mappings had similar patterns of diseases. This can be as a result of the nature of the data,

where the common disease patterns were related to both musculoskeletal and skin diseases. However, there were some variability between the groups in each of the four mappings.

The patient groups in Figure 4.6c and Figure 4.6d appeared to have almost similar disease patterns to each other. For example, patients in group 1 in both mappings had high frequency of diseases related to skin such as cellulitis and actinic keratosis. While the common diseases in patients in group 4 in both mappings were related to pain, these include: pain in limb, pain in cervical spine and acute back pain. Furthermore, patients in group 3 in both mappings had a combination of diseases that were common in patients of both group 1 and 4. The common types of diseases found in patients in group 2 were chest infections, essential hypertension and impotence.

Similarly, when using the 2,622 representative patients set (Figure 4.6b), we found similar patterns. For instance, patients in group 1, 2 and 3 in Figure 4.6b had disease patterns related to skin and heart problems. The same patterns were found in groups 1, 2 and 3 in Figure 4.6c. While patients in group 4 had diseases such as infective otitis externa, pain in limb and wax in ear. These patterns were not common in the other sets.

The patterns found when using the 601 representative patients set (Figure 4.6a) were slightly different to what we got from the other sets. Most of the patterns were related to the two common disease types in the data, these include diseases related to both musculoskeletal and skin. However, there were a number of patients with high frequency of respiratory related diseases such as chest infection, upper respiratory infection and sinusitis. These patterns were common in patients in groups 1-3. Such diseases were not observed with this frequency in (Figure 4.6b - Figure 4.6d). However, we found that patients in group 2 has similar disease patterns to the patients in group 2 in (Figure 4.6b - Figure 4.6d), this group of patients had high number of cases with essential hypertension, impotence and diabetes.

Overall, these results indicate that varying the number of representative patients have an effect in mapping patient records into low dimensional space. These mappings were based on how the representative patients cover the patient space. As discussed earlier, the smaller representative patients sets covered patients who are common in the data. This was reflected

on the patterns found when using the 601 representative patients set. However, when the number of representative patients increased, we found much more detail about the structure of the data. The reason for such behaviour is that with higher number of representative patients we had a wide coverage of all patients, including patients with common diseases and patients with less frequent diseases.

The mappings obtained from using the two sets of 3,500 and 3,589 representative patients seemed to have very similar structure. This shows that with this number of representative patients, the mapping of patients became stable and reproducible. Therefore, we chose to continue our analysis with using the set of 3,500 representative patients as this set demonstrated to provide a sensible covering for all patients.

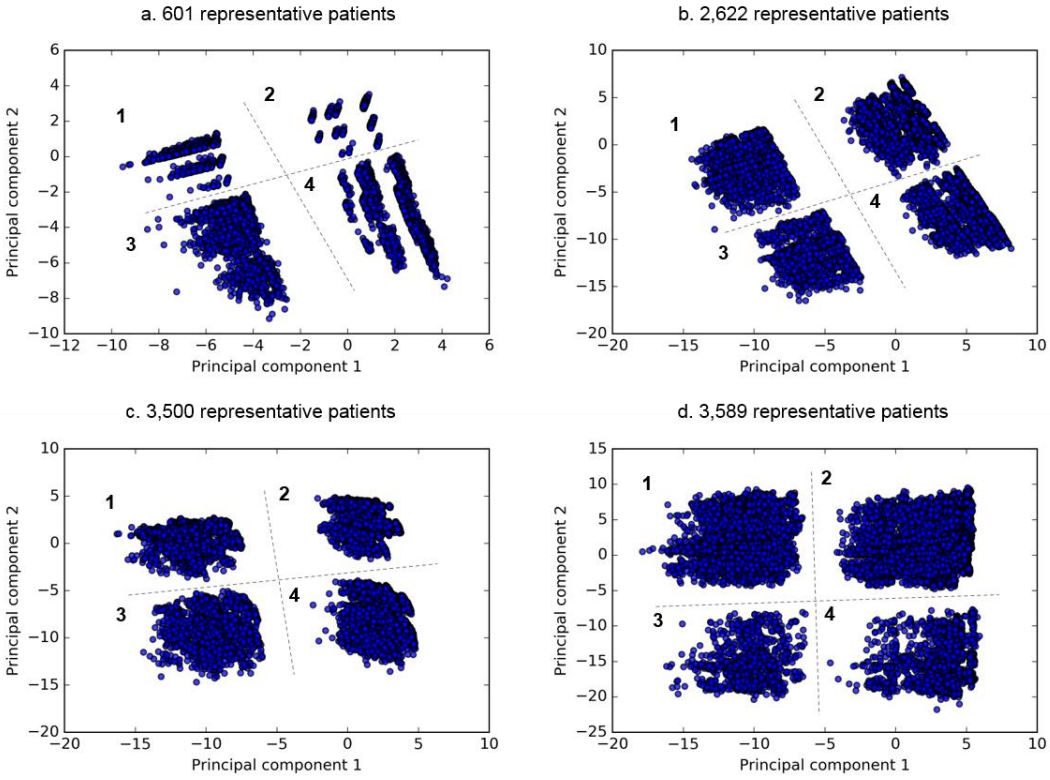


Figure 4.6. Mapping patients records using four sets of representative patients. (a) mapping patients using the 601 representative patients. (b) mapping patients using the 2,622 representative patients. (c) mapping patients using the 3,500 representative patients. (d) mapping patients using the 3,589 representative patients. The labels inside each plot show the group number for that particular plot. The scatter plots show the PCA representation of the records; x-axis: 1st principal component; y-axis: 2nd principal component.

Table 4. 2. The number of patients in each group (see Figure 4.6).

	Group 1	Group 2	Group 3	Group 4
601 representative patients (Figure 4.6a)	10,349	30,959	6,720	34,477
2,622 representative patients (Figure 4.6b)	12,375	51,161	3,125	15,844
3,500 representative patients (Figure 4.6c)	13,415	41,312	5,554	22,224
3,589 representative patients (Figure 4.6d)	17,631	58,893	1,338	4,643

4.3.2.2. Representative patient coverage

We analysed the set of the 3,500 representative patients to know how many patients were being covered by each representative set. This distribution is presented in Figure 4.7. The heavily tailed distribution in this figure suggests that there are few representative patients who cover large number of patients in the data, and the rest of representative patients cover small number of patients. For example, one of the common patterns in the data was related to respiratory system diseases where 26.5% of the study population have been diagnosed with diseases such as upper respiratory infection, chest infection and asthma. We found that those patients were being covered by 6% of the representative patients. Furthermore, this distribution also tells us that, besides the very common patients, there are other patients with much less frequent disease patterns. Those patients were also being covered by the representative patients. By covering the common patients in the data as well as the ones with less frequent patterns, we ensured that this set has sampled the whole space of patients. This shows that this set of representative patients can be thought of as a good covering set and representative to all patients in the data either the ones with very common patterns or the ones with less common patterns.

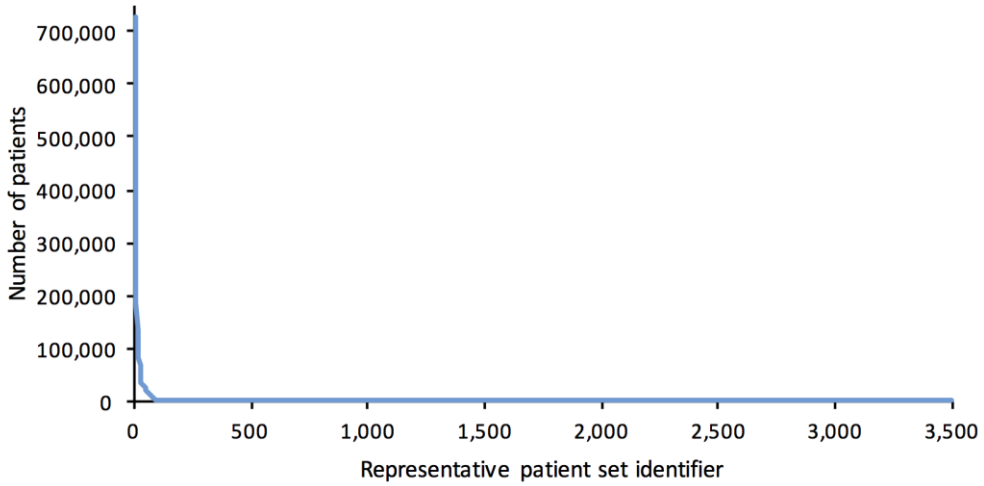


Figure 4.7. The distribution of patients covered by each representative patient in the 3,500 set of representative patients.

4.3.2.3. Comparing representative patients against themselves

We applied the methodology to the 3,500 representative patients set in order to map them into a low dimensional space. In this analysis, we applied the Resnik measure with the Maximum approach. Figure 4.8 shows the results of performing the PCA analysis on the representative patients. The analysis showed that most of the diagnosis codes associated with the representatives were generic Read diagnosis codes, as most of them describes high-level Read codes such as skin and subcutaneous tissue diseases (Read code: M....), respiratory system diseases (Read code: H....) and musculoskeletal and connective tissue diseases (Read code: N....). Interestingly, we saw the essentially the same group structure in all the analyses – this particularly patient group shows four main groups. The difference we see when we look at all the data, rather than just the representative patients, is that you have a feeling for the more detailed size and richness in structure of these groups.

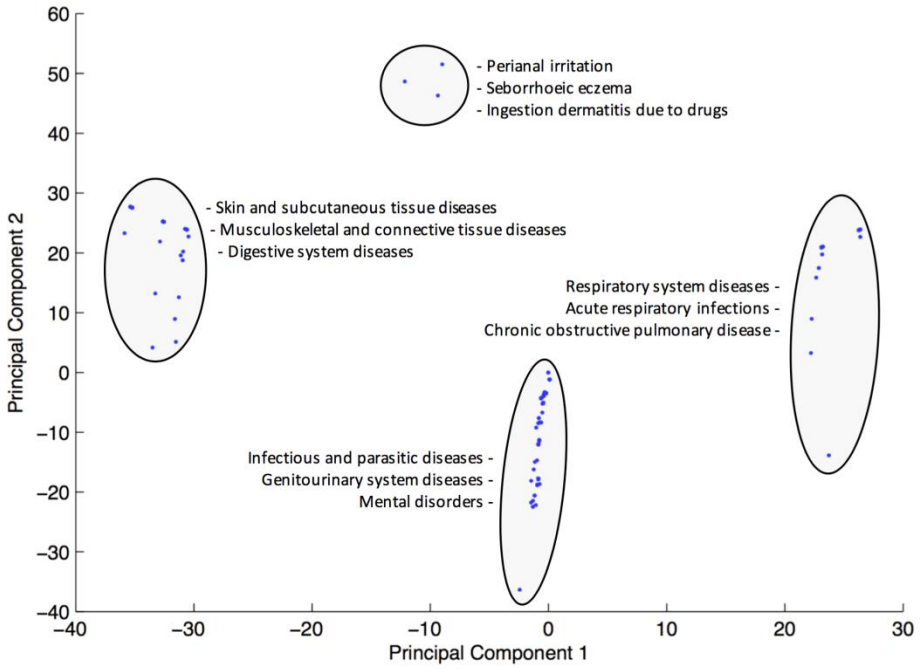


Figure 4.8. The PCA representation of the set of 3,500 representative patients mapped into low dimensional vector space. The scatterplot shows four clusters of the representatives along with the top three frequent Read diagnosis codes associated with them.

4.3.2.4. Evaluating the process of selecting representative patients

The semantic similarity calculation for this evaluation was done using the Resnik measure with the Maximum approach. The cut-off used for this analysis was set to 3.13. This process is presented in Figure 4.9, we showed the process of finding representative patients for each of the 14 subsets. In this figure, we can see the number of new representative patients found in each subset; and the cumulative sum of new representative patients. This process starts by selecting representative patients from one subset. In the first subset, there were 732 patients found to be as representative patients. Then, we combined the first subset with another random subset and looked for representative patients. This time there were 354 new representative patients found.

Figure 4.9 presents the number of new representative patients and the cumulative sums of the new representative patients found in each of the 14 subsets. In this analysis, we started by finding representative patients in one random subset. In this subset, there were 732 patients found to be as representative patients. We combined the first subset with another random

subset and looked for representative patients. This time we found 354 new representative patients. These were then added to the set of representative patients identified earlier. This process was repeated for the rest of the subsets. The last three subsets contained 81, 49 and 10 new representative patients and these were added to the final list of the representative patients. The total number of representative patients found in this data was 3,500 patients.

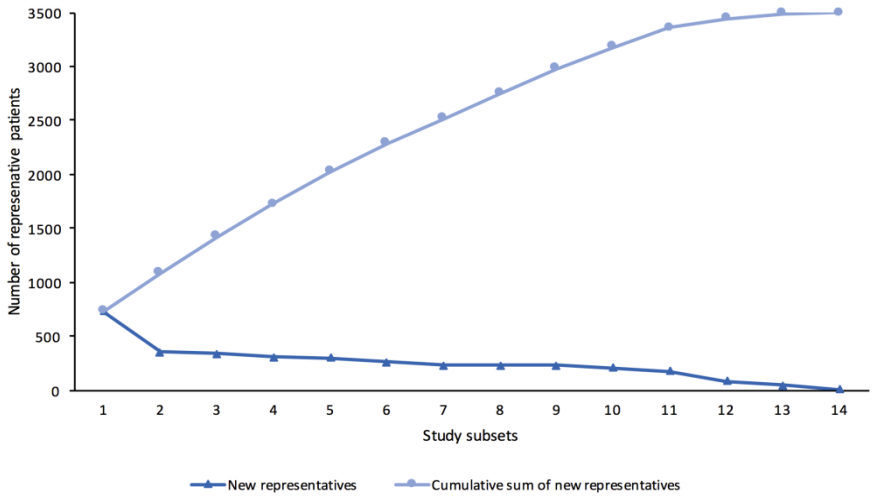


Figure 4.9. Evaluating the process of selecting representative patients from the study data set. A total of 3,500 patients were identified as representative patients among all patients in data (n= 2,754,367 patients). For this process, the data set was divided into 14 subsets, each of which consisted of 200,000 patients – apart from one subset that had 154,367 patients. This plot shows the number of new representatives found in each subset in the data, and their cumulative sum.

4.3.3. Mapping CPRD patient records

To map patients in the 32 patients groups we used the set of 3,500 representative patients. The distribution of these representative patients across the 32 patient groups is given in Figure 4.10, where 54.2% of the representative patients were from the male groups and 45.8% from the female groups. The Resnik measure and the Maximum approach were used in the calculation of the semantic similarity. Once we built the similarity matrices for the 32 patient groups, we performed the PCA to further reduce the data dimensionality. The PCA scree plots for most of the patient groups indicate that over 45% of the variation in the data was being captured by the first two principal components. Projecting patient records based on the first two principal components was therefore deemed sufficient to demonstrate the methodology.

Figure 4.11 shows the representation for both male and female patient groups. The results in this figure show that patient records maintain a clear structure when being mapped into a low dimensional space using such methods. It can also be seen that the PCA representation of patient records was almost identical for a number of patient groups. For instance, the groups of male patients aged between 18 and 44 years share similar PCA representations (Figure 4.11a). Other similar patterns were also found in other patient groups. This could indicate that patients in certain groups might have similar disease patterns. To investigate this more, we performed cluster analysis in order to understand underlying structure of the data.

4.3.4. Clustering analysis using DBSCAN

The DBSCAN clustering algorithm was applied to the patient representation found in Figure 4.11. For the DBSCAN parameters, we set *eps=8* and *minPts=10*. These values were tested and evaluated by using the Silhouette coefficient measure. By using these values with the DBSCAN parameters, we were able to obtain high scores of Silhouette measure. The Silhouette scores ranged between 0.5 and 0.8 for most of patient groups. This suggests that we obtained a good number of clusters for each patient group.

The number of clusters generated by the DBSCAN for each patient group is presented in Table 4.3. A preliminary analysis of DBSCAN clusters showed that patients with semantically similar diagnosis codes were grouped together in nearby clusters. It also revealed significant patterns of diseases across the patient groups (more analysis of this clustering in Chapter 5).

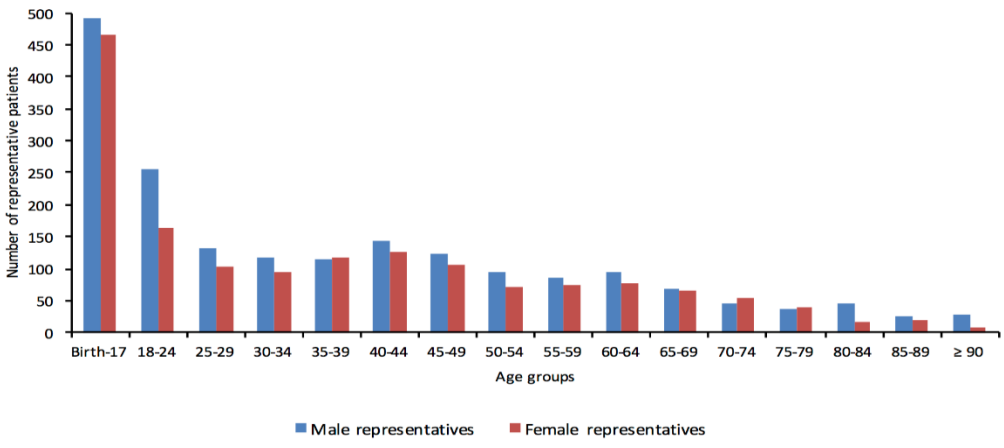


Figure 4.10. The distribution of the 3,500 representative patients across the 32 patient groups.

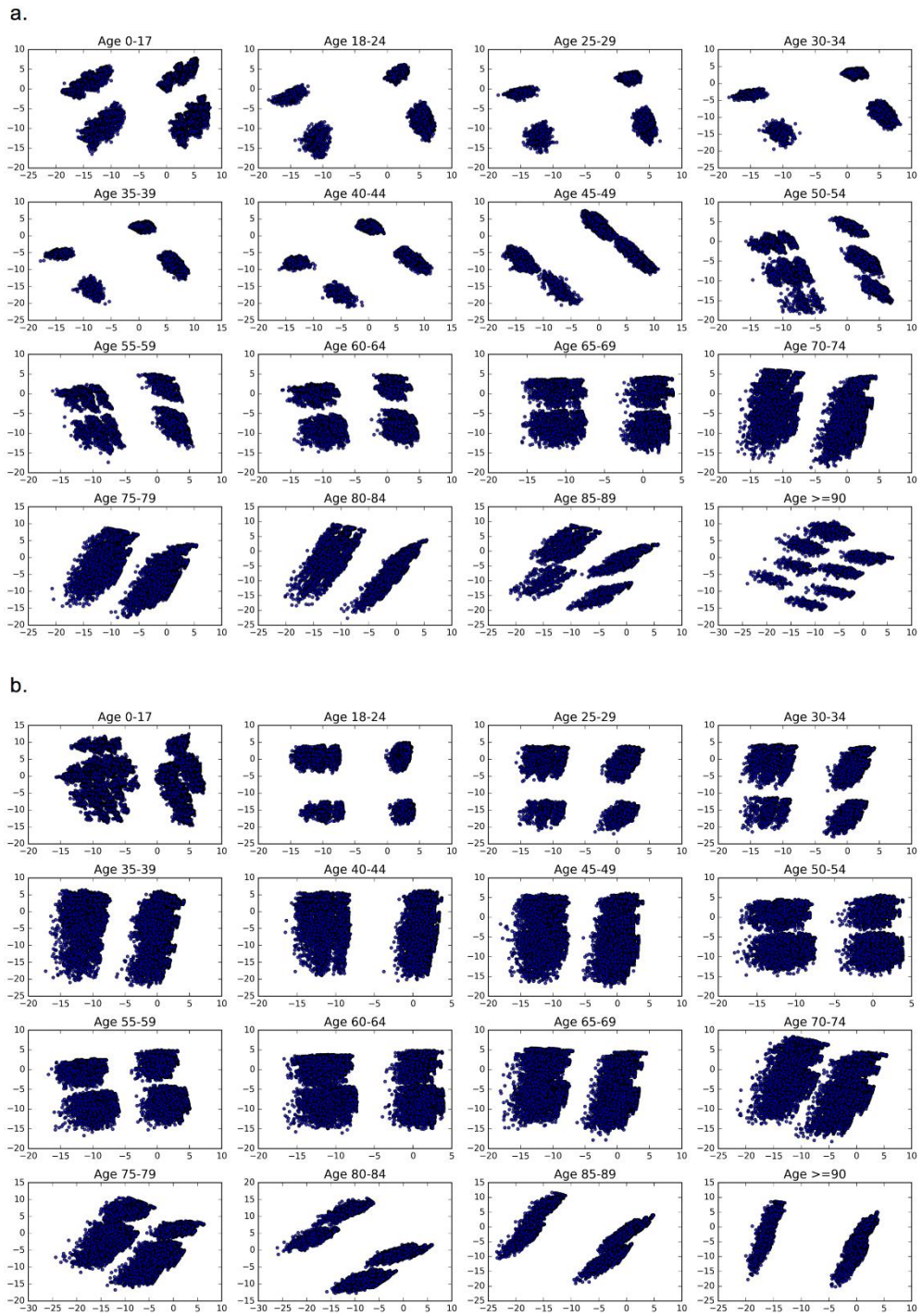


Figure 4.11. Mapping patient records into low dimensional vector space using the Resnik measure with the Maximum approach. (a) shows the representation of male patient groups and (b) shows the representation of female patient groups The scatter plots show the PCA representation of the records; x-axis: 1st principal component; y-axis: 2nd principal component.

Table 4.3. The number of clusters generated by the DBSCAN clustering algorithm (*eps*=8, *minPts*=10). A total of 1,412 clusters were produced for the study data set.

Age group	Number of clusters in male groups	Number of clusters in female groups
0-17	62	41
18-24	52	46
25-29	50	51
30-34	50	33
35-39	46	27
40-44	49	35
45-49	59	36
50-54	39	67
55-59	46	53
60-64	76	60
65-69	62	37
70-74	58	41
75-79	42	43
80-84	27	26
85-89	23	33
> 90	17	25

4.4. Discussion and conclusion

The results obtained in this chapter show that the methodology can take large scale patient cohorts and map them into low dimensional space in a way that facilitate meaningful interpretation of the data. This was one of the key objectives of this work in order to develop the methodology and test its scalability in larger patient data sets.

The data set we had at this stage of the research was obtained from the CPRD database. This data was split into 32 groups based on age and gender of patients, where each group was treated as a separate data set. Although the mapping analysis was done independently for each patient group, the PCA produced almost similar representation for a number of patient groups. Further analysis on these results showed that patients in certain age groups had similar disease patterns. This can be a way of validating the methodology as the results obtained were reproducible for a number of patients group.

The development of the methodology in this chapter involved a number of steps, each of which provides important challenges. One of the key challenges in this work was choosing an

appropriate semantic similarity measure which allowed a clear representation of patient records. In this chapter, we performed an evaluation between three node-based measures to test how each measure would capture the records. Based on their PCA results, we chose the Resnik measure with the Maximum approach. This combination of measures provides sufficient clustering of patient records in a low dimensional space.

Another challenge we faced in applying the methodology to the CPRD data was finding a set of representative patients. With data of this scale, it becomes difficult to find a set of patients that can effectively cover the whole space of patients in the data. In this chapter, we tested whether the number of representative patients could affect the mapping of patients. There were four sets of representative patients identified in the study data set. However, we chose to continue our analysis with the set that consists of 3,500 representative patients. The reason for choosing this set over the others was based on the results obtained from using this set to map patient records. We noticed that the PCA representation of patients had become stable at the point of using this set and did not change after using bigger sets. This set of representative patients has also shown to be a sufficient covering set for all patients in the data. Using this set to map patients from the 32 groups seemed to produce informative representation of the data.

The distribution of patients covered by each representative patient in the 3,500 set suggest that we have two types of patients in the data: a) patients who have diseases that were very common in the data, and b) patients with less common diseases. Those two types of patients were covered by the representative patients. In this way, we guaranteed that this method of sampling patients would result in finding patients who can sufficiently cover all the patients in the space. Therefore, this suggests that using this method is considered better than sampling patients in random way. Random sampling of patients would not provide a good representative set of patients as it does not cover all types of patients in the data. As the distribution was skewed towards patients with common disease patterns, a random sampling of patients would tend to bias the choice to these patients and we might not properly sample some of the less densely populated areas of disease space.

One of the optimisation steps introduced in Chapter 2 was to build the similarity matrix by computing the similarity of the representative patients against themselves. This representation

of the data gave us an idea of types of disease that are common in the representative patients. However, using this way to represent patients did not seem to give much detail about the data as a whole. This was due to the high coverage of some representative patients. Consequently, we did not consider this way to construct the similarity matrix.

It was shown in this chapter that the number of representative patients grows slowly when the data size increases. In other words, this means that whenever a data set gets new patients, these patients might be similar to the existing ones and might already being covered by representative patients. So, the new patients simply mean more patients in the data but not necessarily more different patients. The method used to determine the representative patients scales well as the data volume increases. Although this process is scalable, it requires a long processing time to generate the set of representative patients. Fortunately, however, representative patients only need to be found once; the set can then be reused for any runs of the data.

A feature of DBSCAN that made it beneficial to our analysis was the ability to discover clusters based on the density of the points to cluster. Thus, the high coverage of some representative patients did not influence the clustering process. In fact, DBSCAN returned clusters of patients based on the density, where high dense clusters were corresponding to the common diseases in the data and vice versa. The initial analysis of the clustering showed that applying such clustering with the semantic similarity to this type of data produced clusters that make medical sense. Therefore, it is possible that the DBSCAN algorithm might reveal more details regarding the classification and patterns across patient records by providing a more granular clustering of the data.

Chapter 5: Mapping patient cohort data from clinical coding space into distance space: novel tools for hypothesis generation, stratification and cohort identification

5.1. Introduction

Electronic patient records play an important role in modern health care systems. Their use in collecting and storing patients' complete medical histories has helped in providing better quality of care for patients [15], [34]–[36], [41], [256], [257]. Electronic patient records are also a valuable resource for medical research. Data derived from such records can be of enormous benefit to researchers in gaining new medical knowledge [33], [39], [40], [44], [45], [102], [258]. Such data can lead to various medical discoveries that help in understanding the diagnosis and treatment of specific diseases [127]–[129], [259]–[261].

Health databases have been developed in order to improve the accessibility and usage of patient records for research. Such databases capture a variety of information about patients, including diagnoses, medications, and laboratory results. An example of such databases is the UK Clinical Practice Research Datalink (CPRD). The CPRD database is one of the largest health databases in the world as it consists of anonymised records of a population of more than 11 million patients [119].

The increasing availability of large patient cohorts across longer observation periods has made the CPRD a valuable source of data for epidemiological research [120]–[123]. The CPRD database has been used to produce about 1,450 research studies [119], [126]. These studies cover a wide range of research topics such as pharmacology, medicine, and public health [120]. Most of these studies were based on hypothesis-driven research. In this type of study, researchers used a prior hypothesis to test on the data. For example, one study asked whether having polymyalgia rheumatica makes patients more likely to be diagnosed with cancer [134]. This study used the CPRD database to investigate the incidence of new cancer diagnoses in patients having polymyalgia rheumatica. Another study used the CPRD data to assess the

association between dementia and obesity [135]. It is clear that using such sources of patient data can lead to new discoveries about diseases, medications and clinical practice.

The CPRD and many other similar initiatives can provide researchers with a useful data source on which to test their hypotheses. However, the most common way for researchers to use CPRD has been to query the data to answer very specific questions. For example, does treatment with a specific drug cause some patients to experience a particular rare side effect? If researchers have good questions, then CPRD can provide good answers. But are we missing some opportunities? Are there many other equally important questions that could be asked of the data that people have not yet thought to ask? Therefore, further analysis is required in order to make such hidden areas become more obvious and attainable for future exploration and investigation.

Data mining techniques can be effective tools in discovering patterns and signals in patient medical records. These techniques have been widely used in different applications in the medical domain [39], [41], [136]–[138]. Using such techniques to analyse patient records has the potential in terms of providing a better understanding of the information in these records. However, the data stored in patient records is often complex and high dimensional, as it covers a variety of aspects of patients' information, such as diagnoses, medications, and treatments [138]–[141]. Most importantly, the data is not numerical; instead it consists of bags of terms chosen from a medical coding system, e.g. Read codes, ICD-9 or ICD-10 terms. The nature of the data thus provides a real challenge in effectively interpreting and visualising this data using many established techniques [139], [142]. Therefore, in this chapter, we have investigated whether it is possible to take high dimensional patient data such as the CPRD data and map it into a low dimensional vector space in a way that facilitates meaningful interpretation of the data. Having the data represented in such a space would make it amenable to analysis through more traditional data mining techniques. Here, we applied the mapping methodology along with the modifications introduced in Chapter 4 to the CPRD data with two objectives in mind. The first was to ask if such a mapping is possible in such a large data set – does the methodology scale well with large data sets. The second was to ask whether such mapping provides any useful insights into health data.

5.2. Materials and methods

5.2.1. Study population

The study dataset contains primary care information from the CPRD database. It consists of anonymised records from the year of 2011. The data is described using the Read codes system. It consists of a population of 2,754,367 patients with 7,408,369 Read diagnoses entries (see appendix for the distribution of Read code chapters in the data). We split the study data into 32 patient groups based on age and gender, where we have 16 age groups for each male and female patients (see Table 4.1). Besides patient identifiers and Read codes, we have also been given access to the age and gender of patients.

5.2.2. Map patient records into a low dimensional vector space

A two-step computationally methodology has been applied to study data set. The methodology has been described in more detail in chapter 4. This methodology is based on the notion of similarity representation and dimensionality reduction techniques. In the first stage of the methodology, we perform a calculation of semantic similarity between patients records in order to map the data to a similarity space. For this calculation, we apply the Resnik node-based measure with the Maximum approach. Following on from this, we find a representative set of patients who will be used to represent all patients in the study data set. As the data set was split into 32 patient groups, a separate similarity matrix will be generated for each patient group to store the semantic similarity scores.

The patient records that have been transformed to the similarity space are still represented in high dimensional space. To visualise the data and get detailed insight, we need to find a low dimensional space that conveys the maximum variability in the data whilst retaining the key elements of its structure. This is a problem well-suited to PCA. Thus, in the second stage of the methodology, we perform the PCA analysis on each similarity matrix to further reduce the data dimensionality. The PCA representation of patient records will then be used for clustering through the DBSCAN algorithm.

To look for an enrichment in the resultant clusters, we analyse each cluster separately and look at the sets of patients defined in these clusters. The analysis is not done from the processed data that leads to the clustering. Instead, we return to the actual patient records to put an interpretation on the stratification by looking for enriched Read codes. One way in which the codes in these clusters could be examined is to find their occurrences in the data. This gives us an idea of the significant codes in these groups of patients.

5.2.3. Construction of patient archetypes

The process of mapping patient records into low dimensional space along with the clustering step allows the stratification of patients based on the medical records. Using the semantic similarity calculation and the DBSCAN clustering algorithm provide us with the ability to identify different subpopulation of patients in the data. So that within each subpopulation, patients have similar characteristics such as sage, gender and certain types of diseases. Such classification of patients can be referred to as *patient archetypes*. The concept of patient archetype was used in this thesis to refer to a group of patients who have the same age and gender, and diagnosed with similar types of diseases. Such grouping can be found in the clusters obtained from the DBSCAN algorithm. Since we cluster patients based on the similarity of diseases, and that each cluster consisted of patients with same age group and gender, then these clusters were considered to be patient archetypes. So, a patient archetype represents an average patient from that cluster.

In this analysis, each patient archetype will contain all medical codes from every patient defined in a corresponding cluster. Figure 5.1 presents the process followed to construct the archetypes from patient clusters. The number of patient archetypes for each patient group depends on the number of clusters produced for this particular group. For example, if the DBSCAN algorithm produced n clusters for $patient\ group_{\mathrm{i}}$, then there will be corresponding n patient archetypes. By obtaining a set of patient archetypes for any patient group, this group can be redefined by its patient archetypes such that:

$$patient\ group_{\mathrm{i}} = \{patient\ archetype_1, patient\ archetype_2, \dots, patient\ archetype_{\mathrm{n}}\}.$$

This was applied to all 32 patient groups in our study. Representing patient groups in such way allowed us to reapply the methodology on this new form of the data. The aim for this was to study the different subpopulations within each patient group and to assess their similarities with each other.

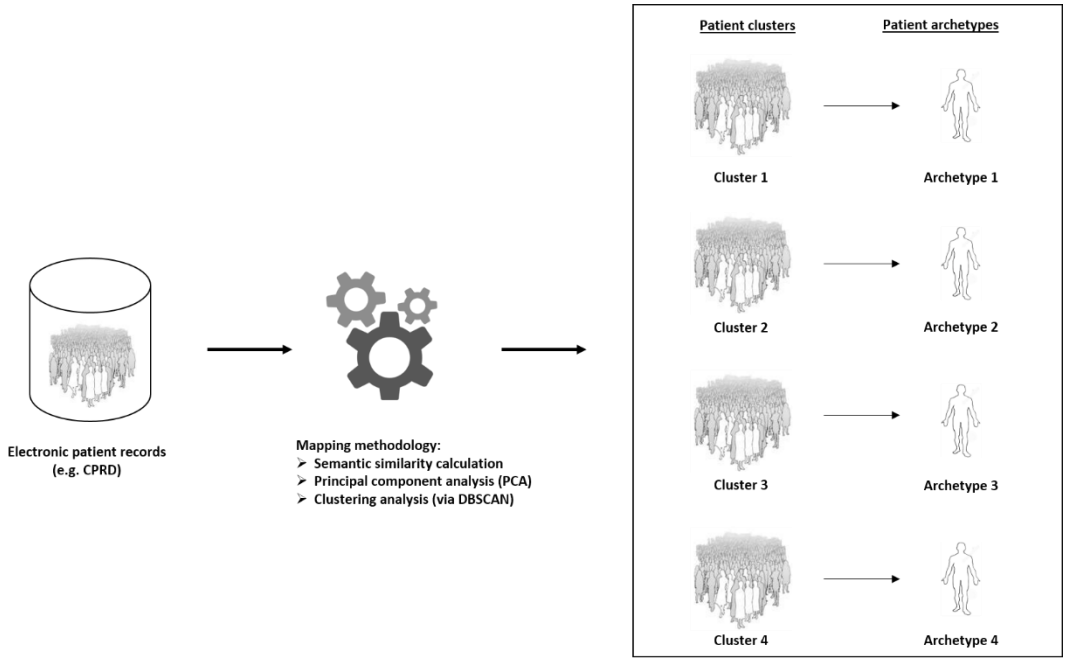


Figure 5.1. The process of constructing patient archetypes. The process used to create patient archetypes is as follows: 1) the methodology begins with finding the semantic similarity between patients and applying the PCA on the similarity matrices; 2) then perform clustering analysis using the DBSCAN algorithm; 3) archetypes will be created using all medical codes for every patient in the clusters.

5.3. Results

5.3.1. Representation of patient records into diagnosis space

Patient records were mapped into a low dimensional space using a set of 3,500 representative patients. These representatives were used in the calculation of the semantic similarity with other patients. For this calculation, we used both the Resnik and the Maximum approach. Since the study data set was split into 32 patient groups, a separate similarity matrix was built for each group. Following on from this, we performed PCA and clustering analysis through DBSCAN algorithm (this analysis has been discussed in more detail in chapter 4). The clustering for male and female patient groups is presented in Figure 5.2. A preliminary analysis on the types of diseases in these clusters showed that the clusters consists of patients with semantically similar diagnosis codes, and those patients were grouped together in nearby clusters. In order to

investigate this more, we performed further analysis on the clusters to identify the classification and patterns across patient groups.

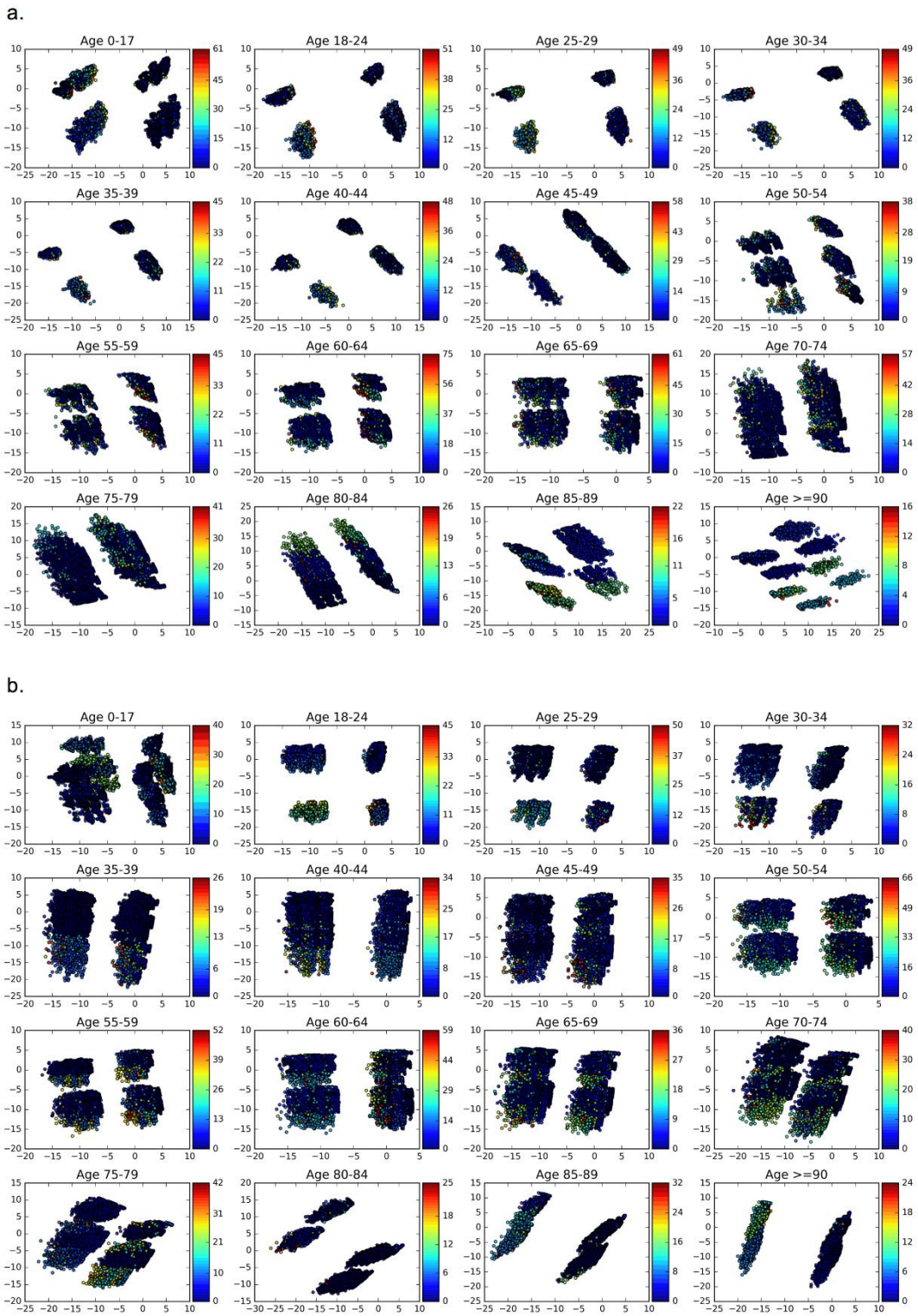


Figure 5.2. The DBSCAN cluster analysis of the 32 patient groups. (a) shows the clustering of male patient groups and (b) shows the clustering of female patient groups. The figure shows the PCA representation of patient records in a low dimensional space; x-axis: 1st principal component; and y-axis:

2nd principal component. The colour bar shows the number of clusters obtained in each patient group (N.B. cluster numbers start from zero in DBSCAN).

5.3.2. Patient stratification

The overall analysis of the clusters revealed significant disease patterns across the study population. These patterns were specific to the age and the gender of patients. We identified distinct patterns in patients at the following age groups: children and adolescents (birth to 17 years), adults (18 to 39 years), middle aged (males: 40 to 64 years, females: 40 to 54 years) and older patients (males: $\geq$ 65 years, females: $\geq$ 55 years). This grouping can also be seen in the representation of patient records shown in Figure 5.2. Patients in these four groups had almost identical PCA representation to each other. For example, in Figure 5.2a we can see that patient in the four male groups aged between 18 and 39 shared a very similar representation.

Based on this grouping, we looked into the cluster of patients in order to identify the common disease patterns. Figure 5.3 presents the top ten common diseases in each of the four ages of males and females. In each age group, we found that patients have been diagnosed with similar types of diseases. However, patients at the transition points between the age groups (males at age: 18, 39 or 64 years; females at age: 17, 39 or 54 years) become more susceptible to have new types of diseases that appear more often in patients at the subsequent age group. For example, a clear signal of the kind of change that happens in the transitions came through in male patients. The common pattern of diseases in males aged between 18 and 39 years is around pains and injuries related to sport activities. The pattern for males changed in the subsequent age group as more patients have been diagnosed with cardiovascular diseases such as essential hypertension. The following section describes the patterns identified in each of the four ages of male and female patients.

In children and adolescents, the common disease categories were related to respiratory system infections (49.40% of patients) and skin/subcutaneous tissue diseases (32.75% of patients) including atopic dermatitis/eczema and acne vulgaris. Both male and female patients in this

group were highly related with diseases including upper respiratory infection, throat infection, acute conjunctivitis, chest infection, ear pain and infection ear.

For adults, the disease categories found were similar to the ones in children and adolescents. However, there was a significant increase in the number of male patients with musculoskeletal/connective tissue diseases (birth-17 years: 7.43%, 17-39 years: 22.88%). Similarly, female patients had an increase in the number of patients with genitourinary diseases (birth-17 years: 8.64%, 17-39 years: 26.80%). Adult females at this age were also found to have diseases such as pain in limb, benign neoplasm of skin, hay fever, thrush, sinusitis and cystitis.

In middle aged patients, 40.52% of males and 42.57% of females have been diagnosed with musculoskeletal/connective tissue diseases. Female patients at this age were also reported with more cases related to menopause such as heavy periods, thrush and hot flushes-menopausal. While for males at the same age, the number of patients with cardiovascular diseases had increased by 6.76% compared to male patients at earlier age groups (birth-17 years: 0.46%, 17-39 years: 3.81%, 40-64 years: 11.03%). Diseases such as Acute back pain, Pain in cervical spine and Chest infection appeared highly in both sexes in middle aged patients.

For older patient groups, almost half of the population at this age (53.87% of females and 43.67% of males) had musculoskeletal/connective tissue diseases and injuries. They also had essential hypertension and accidental fall. Male older patients were highly reported with type II diabetes mellitus, actinic keratosis and gout, whereas female older patients were associated with urinary tract infection and Arthralgia of hip.

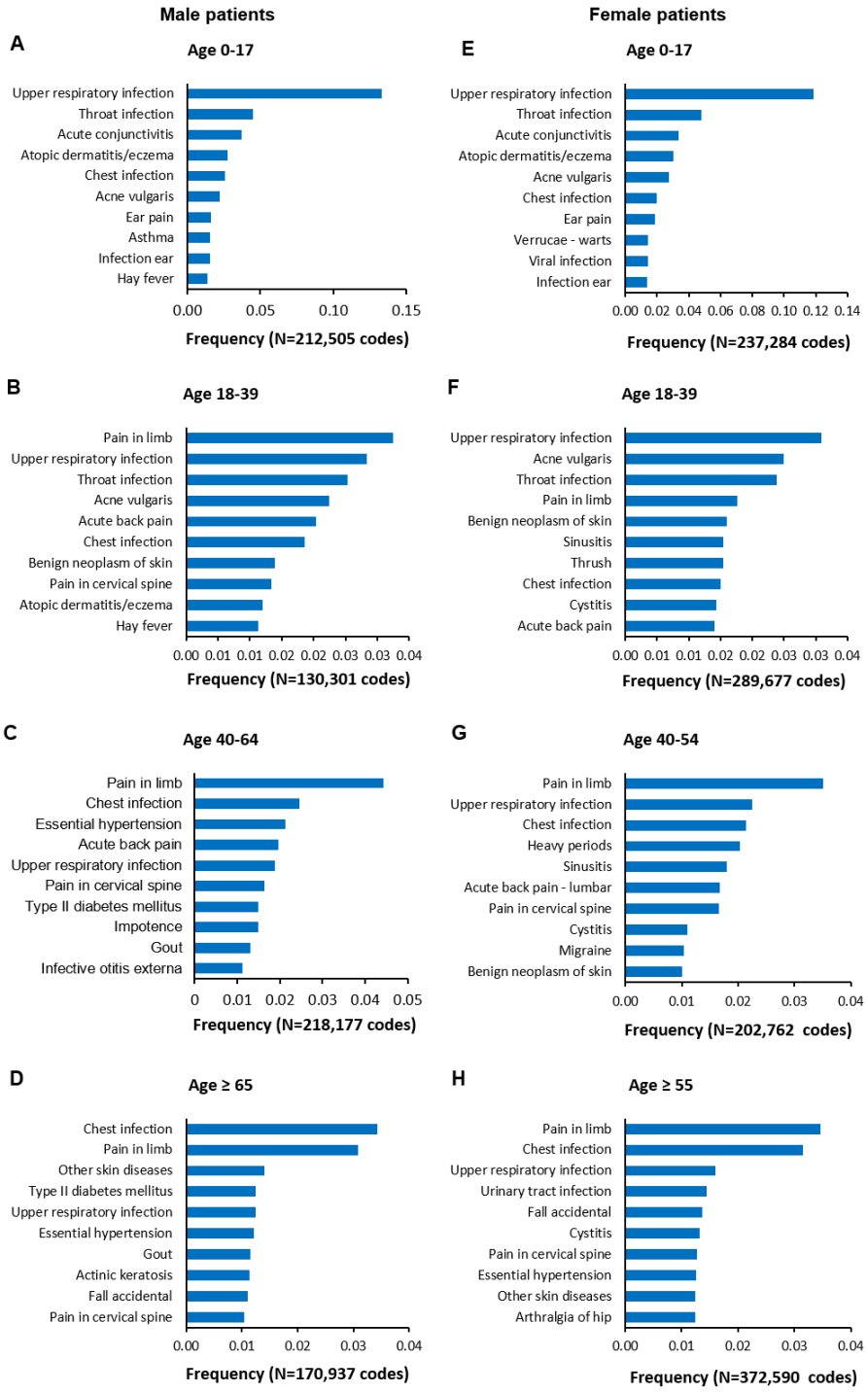


Figure 5.3. A list of the top ten common diseases for the 4 ages of male and female patients**.** Four ages of male patients were at: birth to 17 years, 18 to 39 years, 40 to 64 years and ≥ 65 years; whereas females: birth to 17 years, 18 to 39 years, 40 to 54 years and ≥ 55 years. (A)-(D) present common diseases in male patients; and (E)-(H) present common diseases in female patients.

One of the interesting patterns found in this analysis is related to mental health disorders. This pattern was found in 240,982 patients (8.75% of the study population) at different age groups. The distribution of patients with mental disorders based on age and gender is shown in Figure 5.4. The prevalence of mental disorders (mean (SD)) was slightly higher for males (8.30% (4.89%)) than for females (7.86% (5.26%)). Male children and adolescent patients had higher incidence rates than females for behaviour disorders including attention deficit hyperactivity disorder (ADHD), childhood autism and behaviour disorder. While their female counterparts had more cases of anxiety states and panic disorders (Figure 5.4c). The pattern continued in female patients aged between 18-39 years with 12.37% of cases were reported with anxiety states. In male patients at this age, there were higher incidences with impotence and alcohol problem drinking (Figure 5.4d). For patients aged 40 and above, 27.40% of cases in male patients have been reported with impotence, whereas 15.62% of cases in female patients were anxiety states (Figure 5.4e).

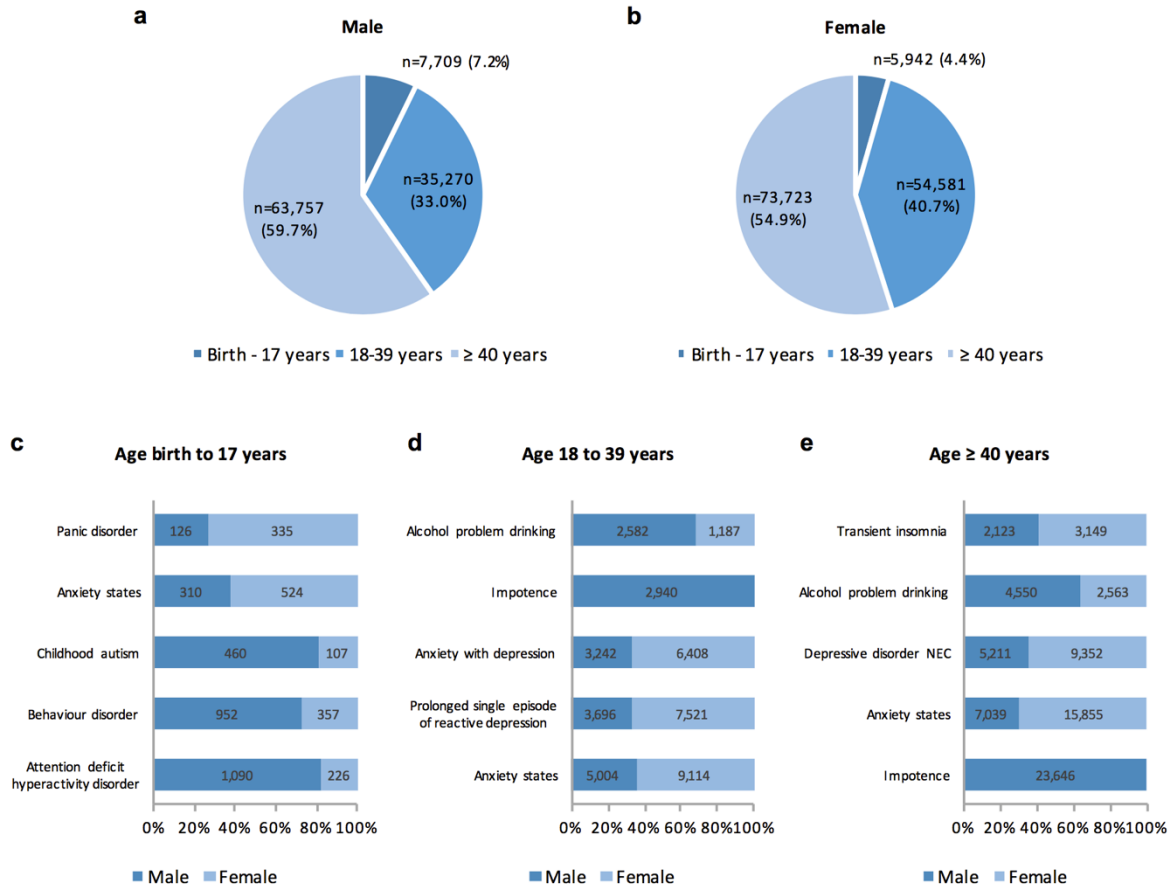


Figure 5.4. The distribution of mental health disorders in patients data during 2011. Plots (a) and (b) show the number of male and female patients, respectiveley, who have been diagnosied with mental health disorders based on three age groups birth to 17 years, 18 to 39 years and 40 years or above. Plots (c), (d) and (e) show the top five common disorders in male and female patients. Males under 18 having more behavioural disorders such as attention deficit hyperactivity disorder (ADHD) and autism more than females at the same age. Another common disorder in males above 18 is impotence. The number of cases with impotence increased significantly at age of 40 and above. On the other hand, females have been associated with higher incidence of anxiety throughout the three age groups.

5.3.3. Patient archetypes

Figure 5.5 presents the results of applying the methodology to the sets of patient archetypes in each patient group. This figure shows how patient archetypes were placed in the space depending on their semantic similarity scores. Patient archetypes which exhibited very high scores of similarity were grouped close to each other, whereas archetypes with low scores were placed far from the rest. Further analysis on the latter showed that the types of diseases defined in these archetypes were not common among the rest. One interesting finding in this analysis was around mental health in male and female patients aged between birth to 17 years. In males, we found archetypes that strongly enriched with mental and behavioural disorders such

as attention deficit hyperactivity disorder (ADHD), autistic psychopathy and Kanner's syndrome. These archetypes of patients can be seen in the first plot in (Figure 5.5a), they are located towards the left side of the plot. This strong signal of mental health in males disappears in females as we found these conditions appear alongside other types of diseases, and not as strong as in males.

A cluster analysis was also performed on the patient archetypes. To demonstrate this clustering, we analysed eight patient groups (four male groups and four female groups) at the following age groups: a) birth to 17 years, b) 25 to 29 years, c) 50 to 54 years, and d) 85 to 89 years. As a result, a total of 36 clusters were emerged from these groups (Figure 5.6 and Figure 5.7). A common type of clusters across three male groups was associated strongly with mental health disorders (Figure 5.6a: cluster 1, Figure 5.6b: cluster 1, and Figure 5.6c: cluster 1). A number of overlapped diseases were also found in certain patient groups, e.g. diseases such as upper respiratory infection, acne vulgaris and atopic dermatitis/eczema in male and female patients aged between birth and 17 years (Figure 5.6a and Figure 5.7a).

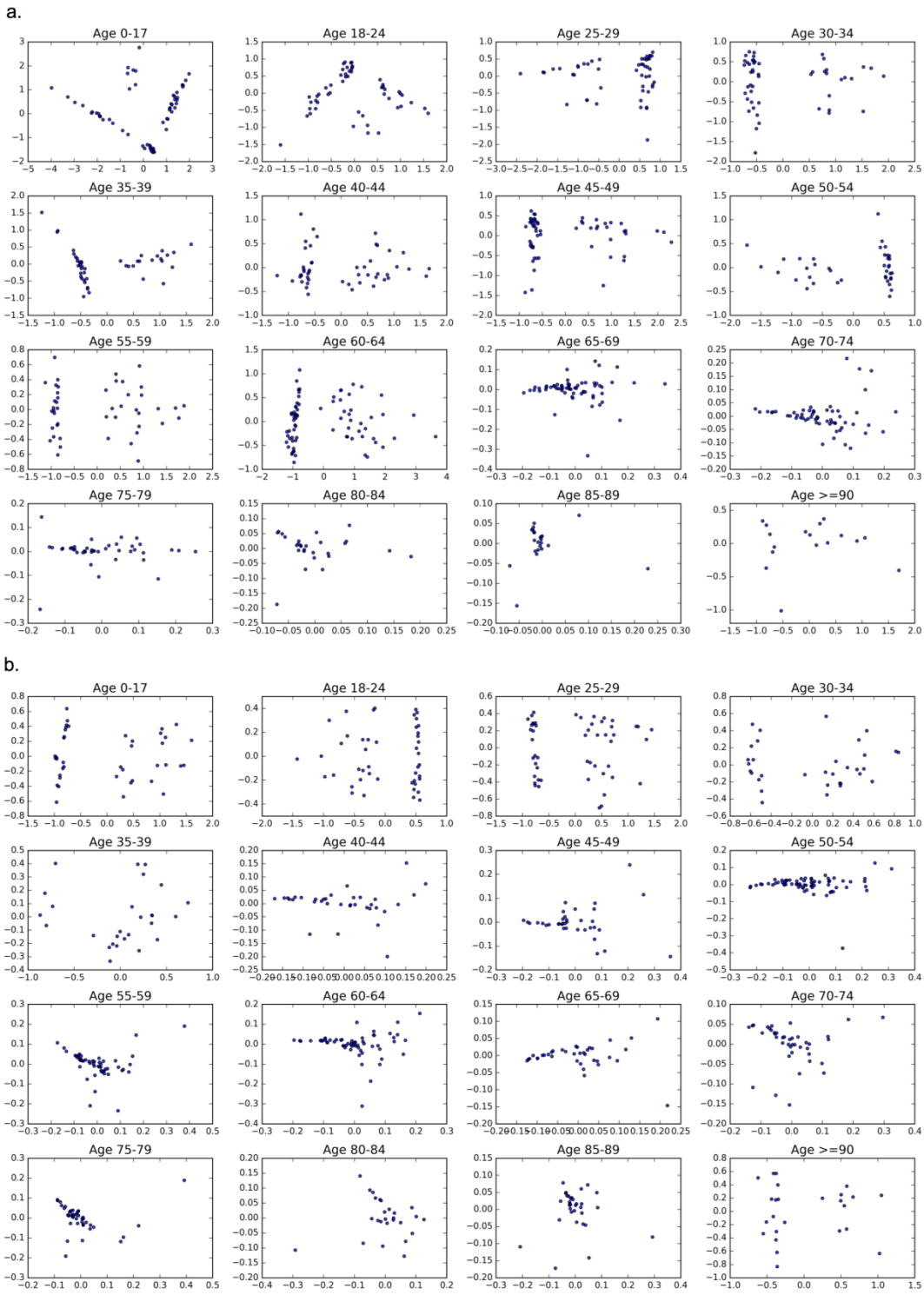


Figure 5.5. The PCA represenation of patient archetypes in a low dimensional space (x-axis: 1st principal component, y-axis: 2nd principal component). (a) shows the male patients archetypes and (b) shows the female patients archetypes. The label above each plot presents the age group of patients.

In children and adolescents (Figure 5.6a and Figure 5.7a), nine clusters were identified (four for males and five for females). The largest two clusters (male cluster 4 and female cluster 5) were more likely to associate with respiratory system diseases and skin/subcutaneous tissue diseases. While female cluster 1 was the only cluster in this group enriched with digestive system diseases.

For adults aged between 25 and 29 years (Figure 5.6b and Figure 5.7b), two clusters (male cluster 2 and female cluster 2) are enriched with injuries and pains such as ankle sprains, pain in limb and whiplash injuries. There was a distinct cluster (female cluster 4), where patients were often diagnosed with genitourinary system diseases such as urinary tract infection, cystitis and amenorrhoea. This cluster comes also with thrush and upper respiratory infections.

In middle aged patients between 50 and 54 years (Figure 5.6c and Figure 5.7c), most of female clusters were associated with menopause including hot flushes – menopausal and heavy periods; within these clusters, we also found some diseases that related with women at menopause age, these include respiratory infections (female cluster 1-4) [262], essential hypertension (female cluster 4) [263], and cystitis (female cluster 2) [264]. In males, cardiovascular diseases significantly increased in this age group compared with the other two groups. Diseases such as essential hypertension and raised blood pressure were common in males (male cluster 4 and 5). Furthermore, there were some similarities in clusters obtained from patient archetypes in (Figure 5.6b: clusters 2-4) and in (Figure 5.6c: clusters 2-4). These clusters consisted of similar types of diseases.

In patients aged between 85 and 89 (Figure 5.6d and Figure 5.7d), most clusters for both sexes were highly associated with chest infections, pain in limb, accidental falls and urinary tract infections. Though, certain diseases appeared more frequently in some clusters more than the others. For example, actinic keratosis (male cluster 1); atrial fibrillation (male cluster 2); gout (male cluster 3); trophic leg ulcer and malignant neoplasm of sweat gland (male cluster 4). For female clusters, acute conjunctivitis (female cluster 1); polymyalgia rheumatica (female cluster 2); trophic leg ulcer and essential hypertension (female cluster 3); cellulitis (female cluster 4).

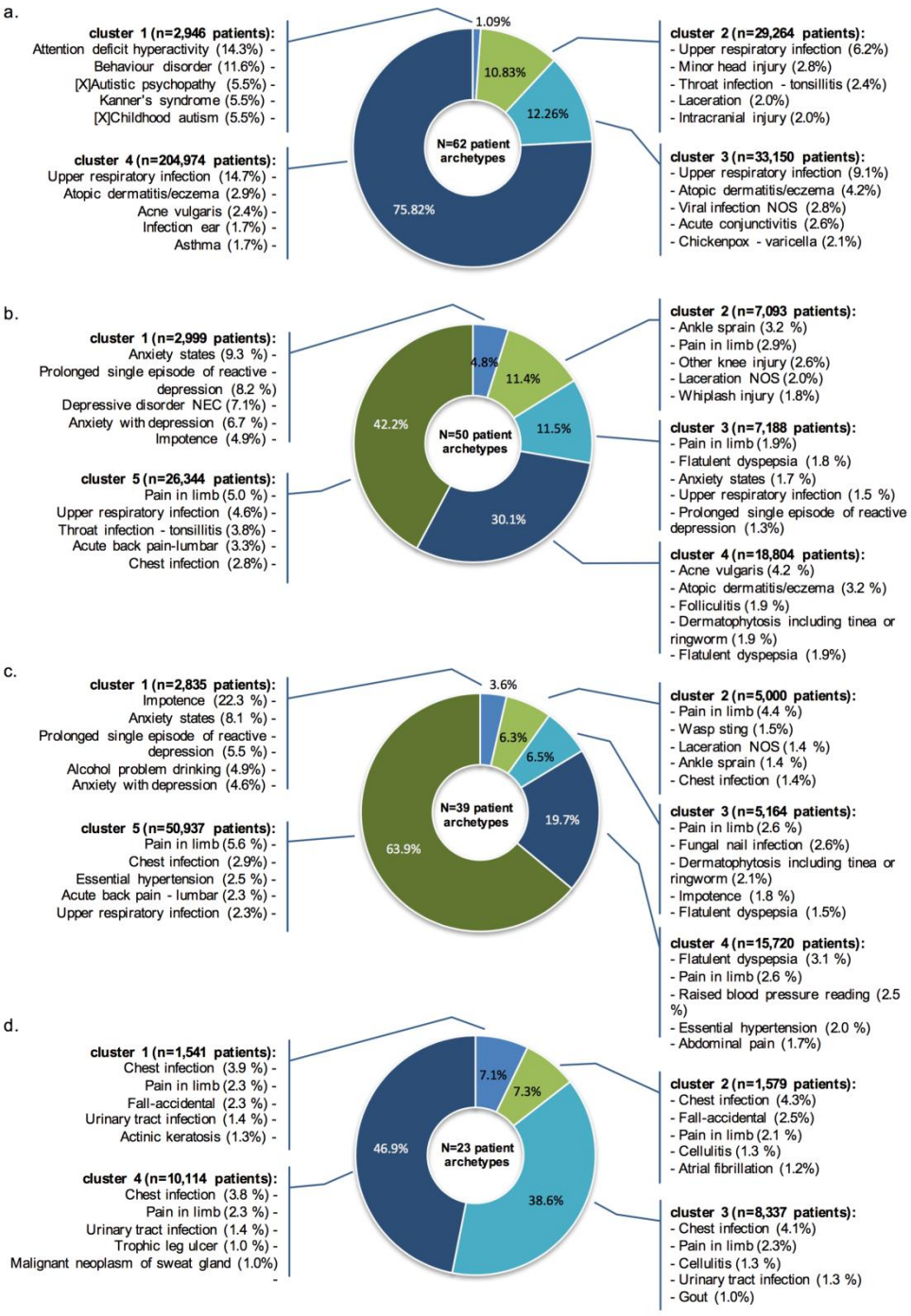


Figure 5.6. Cluster analysis of male patient archetypes along with a list of the top five diseases in each cluster. Each plot shows the clusters found in patient at four age groups: (a) birth to 17 years, (b) 25 to 29 years, (c) 50 to 54 years, and (d) 85 to 89 years. A common disease pattern across (a), (b) and (c) was mental health disorders. In (a: cluster 1) 14.3% of male patients in this cluster had attention deficit hyperactivity disorder (ADHD). While in (b: cluster 1) there were 9.3% of patients in this cluster diagnosed with anxiety states. In the older group (c: cluster 1), there were high number of patients with impotence (22.3%).

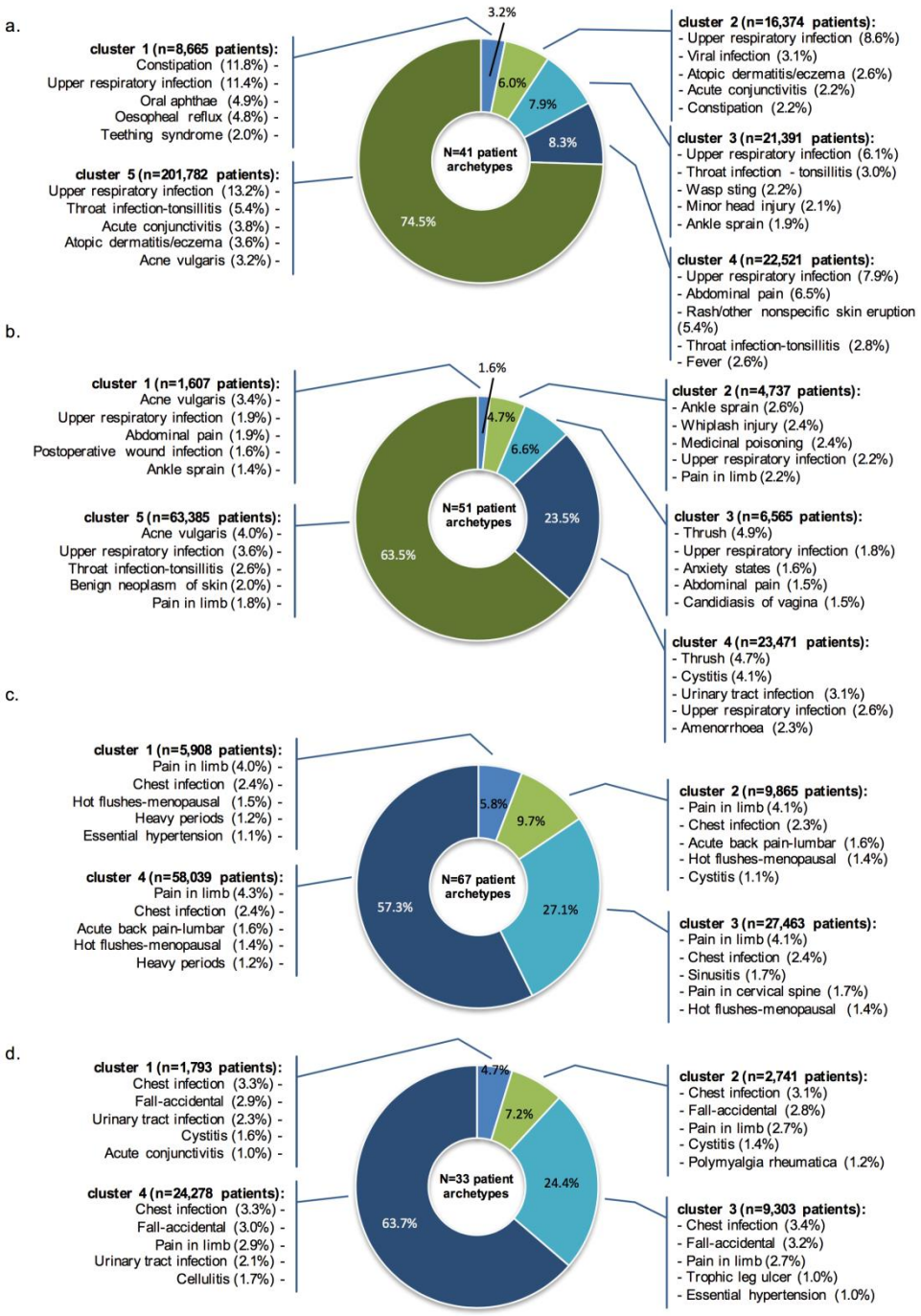


Figure 5.7. Cluster analysis of female patient archetypes along with a list of the top five diseases in each cluster. Each plot shows the clusters found in patient at four age groups: (a) birth to 17 years, (b) 25 to 29 years, (c) 50 to 54 years, and (d) 85 to 89 years. One of the common disease patterns in female group (c) was related to menopause, in all four clusters we found high number of patients with heavy periods and/or hot flushes-menopausal.

5.4. Discussion and conclusion

In this chapter, it was shown that it is possible to take high dimensional patient data, described as bags of medical codes, and map it into a low dimensional space in ways that distance relates to similarity in patient records. Using semantic similarity along with cluster analysis through DBSCAN algorithm provides a way to characterize and stratify patients. This analysis allowed the identification of different subpopulations of patients across the study data set, where each subpopulation consisted of patients sharing similar characteristics such as age, gender and certain types of diseases.

In the first stage of this work, we started by looking at the representation of patient records in a low dimensional space. The challenge at this stage was whether any particular patterns or classifications could be inferred from the similarity calculation and cluster analysis. Using the initial data from the clustering allowed us to identify that both men and women can be classified into four age groups based on their medical records. The types of disease found in these four groups were related to specific age and gender.

The second stage of the work addressed the challenges presented by making sense of the clusters produced by the DBSCAN algorithm. Clustering patient records through DBSCAN provided clusters of patients with similar types of diseases. This shows that using such a clustering method on patient data produced clusters that make medical sense. In this study, we also used the concept of patient archetypes. These represent groups of patients with certain medical conditions, and with similar age and gender. Using this idea of patient archetypes can help to provide a better understanding of patient characteristics.

In conclusion, the work described in this chapter shows that using the mapping methodology on large scale patient data sets does provide interesting insights about the data. The work in the next chapter investigates in more detail one particular part of the results obtained in this chapter to investigate it in more detail.

Chapter 6: Identification of disease subgroups through semantic similarity analysis: falls in the very elderly as case study

6.1. Introduction

Falls are a serious health issue worldwide and can lead to a number of major health problems including injury, disability, and mortality [265], [266]. Falls are defined as a sudden, uncontrolled change in position that denies an individual the opportunity to restore balance, which results in the individual moving from a standing or sitting position to a lower level [267], [268]. Physicians diagnose falls on the basis intrinsic or external factors. External factors include dealing with the environmental orientation including foreign objects on the walking surface, a design flaw in the walking surface, slippery surfaces and an individual's impaired physical condition, while intrinsic factors include things pertaining to ailments [269]. Dykes et al. [270] stated that biological changes that come with age are the major cause of falls and injuries related to falls. An individual's diagnosis is also associated with falls in people with some diseases such as anaemia, neoplasms, congestive heart failure, stroke and cerebrovascular accident being highly associated with fall risks [271]. There has been a sharp rise in the medicalisation of older adults, not only in diagnosis but also in prescribing and hospitalisation [272].

According to a World Health Organization (WHO) prediction, 28%–35% of people aged above 65 fall every year, while those aged above 70 have an increased rate of falls ranging from 32%–42% [273]. Evidence of increasing costs affecting national budgets and deaths caused by falls have been at the forefront for establishing mechanisms to detect and prevent falls. It is predicted that if no measures are put in place, injuries caused by falls are likely to increase at a very high rate. A recent study indicated that the estimated annual falls cost in the UK NHS is over £2.3 billion [274]. It is therefore essential to explore a more complete set of comorbidities that might be associated with falls. This opens up a better opportunity to identify patients at risk of falls, helping guide policy so as to reduce falls.

Population-based studies on falls or other diseases usually use standard statistics methods for testing specific *a priori* defined hypotheses, such as [275], which used Clinical Practice Research Datalink (CPRD) datasets to study the association between falls and mortality. However, it is difficult to use these standard statistical techniques to search for unknown new hypothesis.

This research is a population-based cohort study using the CPRD datasets. Normally traditional strategies have used CPRD data to ask specific questions in order to test hypotheses like does having diabetes make a patient more likely to fall. However, we want to see from studying the data are we asking the right question of the data? Also can we find better questions? The purpose of this chapter is to explore a novel analysis strategy, introduced in chapter 4 that will attempt to find some good questions and hypotheses about an important medical issue, falls in the elderly. The associations found can be then tested using more traditional comorbidity measures. Therefore, the outcome of the study will give us a better insight into diseases associated with falls.

6.2. Methods

6.2.1. Data source

Medical information was gained from CPRD. This database is the world's largest longitudinal and anonymised clinical research database comprising electronic medical records from primary care in the UK [276]. In 2013, over 11 million patients from about 700 primary care practices in the UK are included in this database [108]. The data includes demographic information, clinical information, medical history (including diagnosis, treatment and medications), referrals, laboratory results and hospital admissions. The medical history information in CPRD is recorded using Read coding systems [277].

6.2.2. Study population

We analysed all CPRD patient records in 2011. To facilitate the statistical analysis, we have divided the patients into 32 patient groups based on age and gender, where we have 16 age

groups for each male and female patients (see Table 4.1). The exposures of interest were diagnosis codes taken from the falls cases.

6.2.3. Falls coding identification

The term ‘falls’ is used in this study to identify events recorded by general practitioners (GPs) using Read codes. While there are different codes for falls within the Read codes, GPs usually encode a fall using the general codes for falls. Therefore, we have taken the most frequent codes recorded in CPRD for falls, which are: Accidental falls (TC…), Other falls (TCy..) and Accidental falls NOS (TCz..). See Appendix C for all falls codes in Read code system.

6.2.4. Mapping falls-patient records

Two-step computational methodology has been applied to study the data set. The methodology has been described in more detail in chapter 4. This methodology is based on the notion of similarity representation and dimensionality reduction techniques. In the first stage of the methodology, we perform a calculation of semantic similarity between patients records in order to map the data to similarity space. For this calculation, we apply the Resnik node-based measure with MAX. Following on from this, we find a representative set of patients who will be used to represent all patients in the study data set. In the second stage, we perform the PCA analysis on each of the similarity matrices to further reduce the data dimensionality. The PCA representation of patient records will be then used for clustering and visualisation through the DBSCAN algorithm.

6.2.5. Statistical analysis

The distribution of patients based on age and sex is calculated to identify the changes of falls in different age groups and whether there is any sex difference. Age and sex are considered to be the main modifiers in disease and medication risk and treatment [278].
In addition, after mapping elderly patients’ records into low dimensional space, we analyse each cluster separately and look at the sets of patients defined in these clusters. This helps in identifying the conditions that have been associated with falls. Then, a p-value of less than 0.05

is to be considered statistically significant regarding which clusters have more falls codes than would be expected by chance.

After visualising clusters of faller and non-faller patients, the common conditions that substantially appear with falls were studied. There is typically considered to be a relationship between two conditions whenever they affect the same patient significantly more than chance alone. To assess the possibility of associations between diseases, Relative Risk (*RR*) and *Φ*-correlation were used. These measures are widely used in clinical literature [279], [280].

Table 6.1. 2*2 contingency table used to calculate RR and Φ-correlation.

	with Disease	without Disease	total
Falls	a	b	e
without falls	c	d	f
total	g	h	N

The Relative Risk means the ratio of observed co-occurrence in an exposed group (fallers) to that of a non-exposed group (non-fallers). The possibility of associations between diseases can be assessed by calculating the RR with a pair of diseases using contingency table (Table 6.1), according to Altman [281]:

$$RR = \frac{a\ /\ e}{c\ /f}$$

When two diseases have an RR value greater than 1, it means that these two diseases have an agreement, while if the RR value is less than 1, it means they have a disagreement. A relative risk value of 1 means there is no association between these two diseases. Then, the standard error of the natural logarithm of the risk relative is calculated as follows:

$$\mathrm{SE}\{\ln(RR)\} = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}}$$

The standard error is used then to compute the 95% confidence interval which measures the significance of relative risk. The 95% confidence interval can be calculated as follows:

$$95\%\, CI = \exp\ (\ln(RR)\ \pm\ 1.96\ \times\ \mathrm{SE}\{\ln(RR)\}\ )$$

Another way to quantify the association strength between two diseases is phi-correlation (*Φ*-correlation). This measure can be calculated according to this formula, using contingency table:

$$\Phi = \frac{ad - bc}{\sqrt{e * f * g * h}}$$

The values of *Φ*-correlation range from -1 to +1, where -1 indicates perfect negative relationship and +1 indicates perfect positive relationship (See Appendix C for more information about the relationship between the RR and *Φ*-correlation measures).

At first, falls-associations in the population as a whole and very elderly population were searched, and then falls-associations in each of the subgroups that have been identified in the clustering in order to identify different risks in different subgroups of patients. An association with p value less than 0.05 was considered as statistically significant.

6.3. Results

6.3.1. Sample demographics

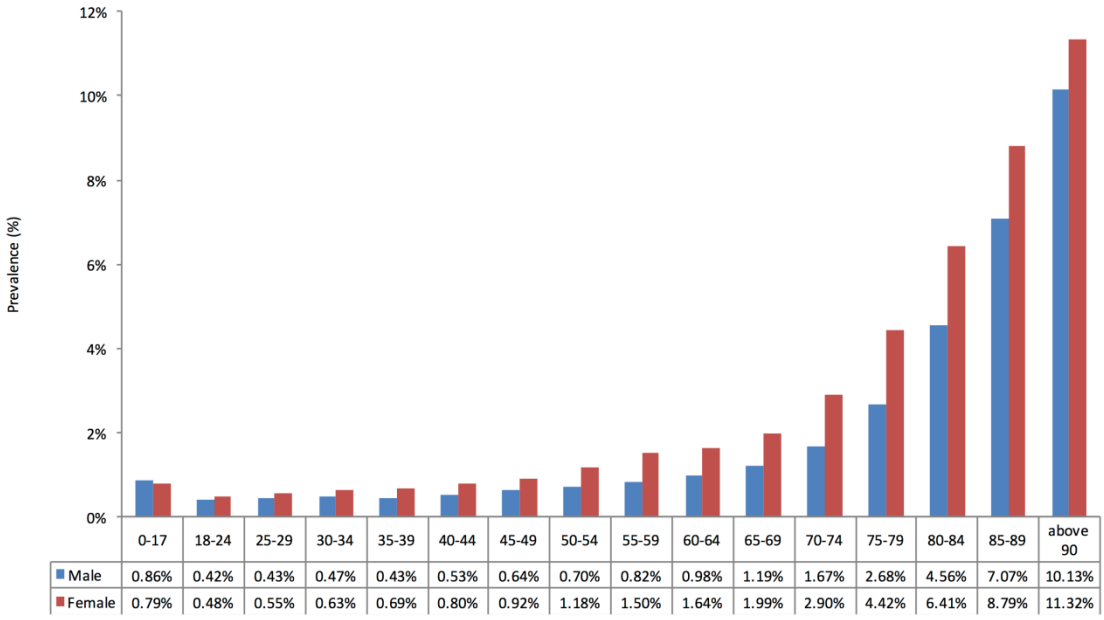
A total of 7,408,369 patient records from 2,754,367 patients were analysed. Among these, 46,055 patients reported one or more falls.

Figure 6.1 contains the distribution of falls in all patients based on age and sex. From the population distribution falls results have been divided into three main stages: childhood (aged from birth to 17 years), adulthood (aged 18-64 years), and geriatric (above 64 years). In childhood, 0.86% of men and 0.79% of women fell over. In early adulthood (18-24 years), the proportions of patients with falls dropped to 0.42% for men and 0.48% for women. For adults between 18 to 49 years, a similar increase was observed for both sexes between age groups. Nevertheless, there was no important age difference found for men and women (maximum difference is 0.12%). For adults aged between 50 to 64 years, the proportion of falls in women

age groups starts increasing rapidly. The proportion of falls was slightly smaller in men than women over adulthood groups.

Geriatric falls were remarkably reported more than falls in children and adults. More than 57% of patients with falls are aged above 64 years. The rate of falls increased with age particularly for patients aged 80 years or above. In men aged 85-89 years, 7.07% had experienced falls, and for men 90 years or above this rose to 10.13%. The results of geriatric falls show more differences than adults regarding sex. The prevalence of geriatric falls was higher for women (mean: 5.97%; SD: 3.59) than for men (mean: 4.55%; SD: 3.48).

As reported, falls are not serious condition for children and adults; it is a serious health issue for elderly patients. These results are supported in the clinical literature [265], [266]. As a results, the mapping methodology will be applied on all CPRD elderly patients (aged 64 years or above) in 2011.



	0-17	18-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	above 90
Male	0.86%	0.42%	0.43%	0.47%	0.43%	0.53%	0.64%	0.70%	0.82%	0.98%	1.19%	1.67%	2.68%	4.56%	7.07%	10.13%
Female	0.79%	0.48%	0.55%	0.63%	0.69%	0.80%	0.92%	1.18%	1.50%	1.64%	1.99%	2.90%	4.42%	6.41%	8.79%	11.32%

Figure 6.1. The distribution of falls in the study data set based on the age and sex of patients.

6.3.2. Mapping patients records into diagnosis space

We developed a semantic similarity approach that uses patient records to facilitate the visualisation of patients. Thus patients with similar diagnoses are placed together. After

mapping all elderly patient records into a low dimensional space, 434 clusters were identified. By filtering out using p value less than 0.05 for all falls codes, 38 clusters were selected: 22 clusters for men and 16 for women (see Appendix C for the analysis of these clusters). 15 clusters out of 38 clusters were for the very elderly patients, aged 90 years or above. In the full elderly cohort, strong relationships were observed between falls and five disease categories that have been highly identified with falls: infectious, cardiovascular, injury, musculoskeletal and digestive system diseases (Figure 6.2).

6.3.3. Falls-associated diseases in the elderly population

In the general population, 160 comorbidities have appeared significantly with falls in such clusters (using P-value less than 0.05 with all diseases codes). Using RR and Φ-correlation we have tested the association between these diseases and falls (see Appendix C for more information).

Table 6.2 shows the association between falls and 48 diseases, which have shown high scores in both measures. These diseases were grouped into six main disease categories: injuries, infections, cardiovascular, digestive, musculoskeletal and other diseases. In general, total falls-associated diseases rates were higher in men than women. However, women were at higher risk of open wound of leg and retention of urine than men (RR: 2.21 and 2.89 for men; RR: 1.13 and 2.24 for women respectively).

Injuries showed the highest rates for both men and women (RR: 3.87 and 3.02 respectively). In injuries, head injury (RR: 6.71 for men, RR: 4.83 for women) and intracranial injury (RR: 6.05 for men, RR: 4.75 for women) presented the highest rates among all diseases. Age was a significant covariate in the relationship between falls and injuries, where most cases were within the very elderly patients.

There were also high scores for cardiovascular and infectious diseases (RR: 2.20 and 1.62 respectively). Among these diseases, postural hypotension and urinary tract infection (UTI) showed higher scores than other diseases of these disease categories in men (RR: 3.59 and 2.75 respectively) and women (RR: 3.35 and 1.74 respectively). On observation of all other diseases, senile confusion and pressure sore have been significantly reported for men (RR:

4.17, 3.38 respectively) and women (RR: 3.05, 2.45 respectively). Moreover, hypothyroidism shows a positive score with falls in men (RR: 1.42), while there was no association in women (RR: 0.96).

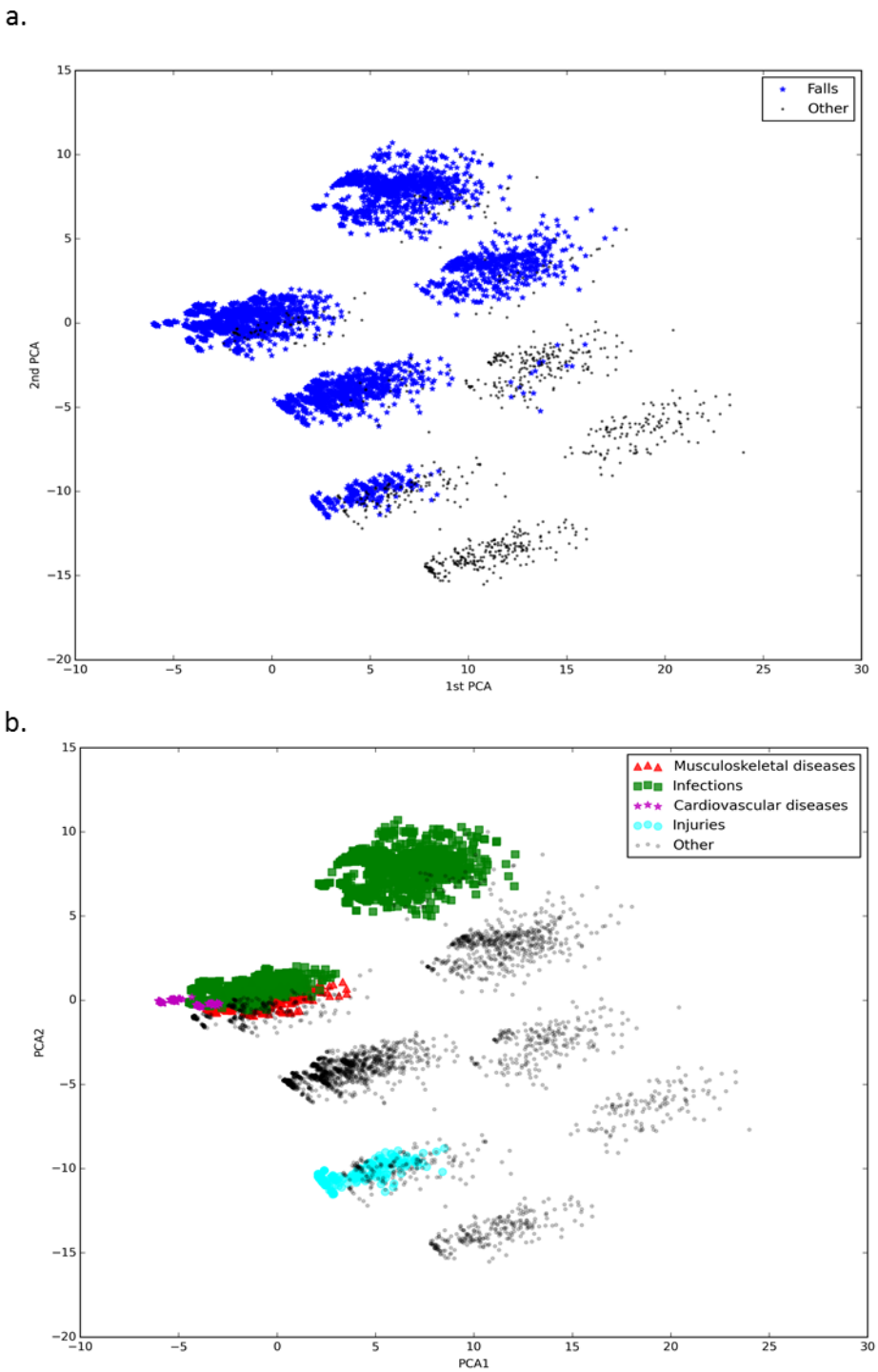


Figure 6.2. The cluster analysis of men patients aged 90 years or above (n=9,932). Scatter plots (a) and (b) show the PCA representation of patient records in a low dimensional space; x-axis: 1st principal component; y-axis: 2nd principal component. Nodes are patients; node colour identifies diseases or diseases categories. (A) shows the clusters enriched with falls in the data. (B) shows the other types of diseases that appear significantly with falls. These include disease categories such as infectious, cardiovascular, injury, musculoskeletal and digestive system diseases.

6.3.4. Falls-associated diseases in very elderly patients

Due to the fact that most of the clusters enriched with falls (15 clusters out of 38 clusters) were for the very elderly patients, we decided to take into account only very elderly patients. These patients were divided into two groups based on sex: men patients (*n*= 9932) and women patients (*n*= 25649). We re-applied the comorbidity measures to these two groups. For men patients, 13 diseases showed positive associations with falls. Most of the diseases in this group showed lower scores than the ones in the whole elderly population. We also found that faller patients in this group were more likely to associate with rectal bleeding (RR: 2.30) and atrial fibrillation and flutter (RR: 1.64) (Table 6.3A). While the list of falls-association comorbidities in very elderly women patients consisted of 24 diseases. Injury diseases showed the highest rates for this group as in the whole elderly population. We found that faller patients in this group were more likely to associate with other injuries and musculoskeletal diseases that did not appear significantly in the general population level. The increase for pain in limb (RR: 1.33) and acute lower respiratory tract infection (RR: 1.64) was observed in this group of patients compared with the general population (RR: 1.23, 1.25 respectively), whereas aching leg syndrome (RR: 1.42) and osteoporosis (RR: 1.42) were generally constant (Table 6.3B).

Table 6.2. Significant associated diseases with falls in elderly population level (≥ 65 years). Tables (A) and (B) present the associations in men and women, respectively. The associated diseases have six disease categories: injuries, infections, cardiovascular, digestive, musculoskeletal and other diseases. Injuries showed the highest rates for both men and women. Hypothyroidism shows a positive score with falls in men (RR: 1.42), while there was no association in women (RR: 0.96). LCI, lower confidence interval; UCI, upper confidence interval.

A. Significant associated diseases with falls in men fallers population

Disease	*Φ*-correlation	RR	95% LCI	95% UCI	P value
Injury diseases					
Head injury	0.033	6.713	5.219	8.635	<0.0001
Intracranial injury	0.029	6.050	4.642	7.885	<0.0001
Fracture of hip, trauma and humerus	0.035	4.826	4.007	5.812	<0.0001
Laceration NOS	0.026	3.860	3.129	4.761	<0.0001
Leg bruise	0.011	3.842	2.361	6.252	<0.0001
Haematoma with intact skin	0.019	3.703	2.806	4.887	<0.0001
Post-traumatic wound infection	0.013	3.151	2.230	4.451	<0.0001
Implant complications	0.012	2.924	2.051	4.170	<0.0001
Wasp sting	0.014	2.472	1.912	3.196	<0.0001
Open wound of leg	0.001	1.131	0.502	2.551	0.8**
Heart diseases					
Stroke and cerebrovascular accident	0.013	1.993	1.622	2.447	<0.0001
Postural hypotension	0.032	3.589	3.057	4.213	<0.0001
Cardiac failure	0.010	1.960	1.516	2.535	<0.0001
DVT - Deep vein thrombosis	0.011	2.068	1.605	2.664	<0.0001
Infectious diseases					
Skin and subcutaneous tissue infections	0.013	1.677	1.441	1.952	<0.0001
Respiratory tract infection	0.015	1.719	1.509	1.960	<0.0001
Chest infection	0.011	1.276	1.192	1.365	<0.0001
Urinary tract infection	0.038	2.750	2.485	3.044	<0.0001
Cystitis	0.012	1.797	1.484	2.176	<0.0001
Cellulitis	0.021	1.931	1.723	2.164	<0.0001
Digestive system diseases					
Bowel obstruction	0.005	1.885	1.158	3.071	0.007
Constipation NOS	0.017	1.900	1.642	2.200	<0.0001
Gastrointestinal haemorrhage	0.004	1.740	1.018	2.976	0.03
Musculoskeletal Diseases					
Swelling of calf	0.019	2.571	2.121	3.117	<0.0001
Osteoporosis	0.016	2.772	2.142	3.588	<0.0001
Aching leg syndrome	0.008	1.397	1.200	1.628	<0.0001
Acute back pain - lumbar	0.009	1.339	1.183	1.515	<0.0001
Pain in limb	0.010	1.243	1.157	1.336	<0.0001
Others					
Pressure sore	0.020	3.375	2.656	4.288	<0.0001
Restlessness and agitation	0.009	3.311	1.983	5.529	<0.0001
Hyponatraemia	0.017	2.923	2.288	3.734	<0.0001
Difficulty in swallowing	0.012	2.677	1.942	3.691	<0.0001
Dependent oedema	0.007	2.269	1.475	3.492	<0.0001
Retention of urine	0.017	2.241	1.874	2.680	<0.0001
Chronic renal failure	0.007	1.939	1.326	2.836	<0.0001
Insomnia	0.008	1.840	1.393	2.430	<0.0001
Dizziness	0.008	1.642	1.311	2.055	<0.0001
Shortness of breath	0.009	1.606	1.325	1.947	<0.0001
Hypothyroidism	0.005	1.415	1.099	1.822	0.005
Vitamin B12 deficiency	0.004	1.406	1.045	1.893	0.02
Haematuria	0.006	1.300	1.095	1.544	0.002
Anaemia	0.023	2.342	2.027	2.706	<0.0001
Trophic leg ulcer	0.019	2.614	2.150	3.178	<0.0001

Intertrigo	0.004	1.377	1.039	1.826	0.02
Senile confusion	0.027	4.168	3.361	5.170	<0.0001
Collapse	0.020	3.137	2.487	3.957	<0.0001
Abnormal loss of weight	0.015	2.487	1.946	3.178	<0.0001

B. Significant associated diseases with falls in women fallers population

Disease	*Φ*-correlation	RR	95% LCI	95% UCI	P value
Injury diseases					
Head injury	0.031	4.833	4.021	5.810	<0.0001
Intracranial injury	0.029	4.752	3.902	5.787	<0.0001
Fracture of hip, trauma and humerus	0.040	3.225	2.919	3.564	<0.0001
Laceration NOS	0.028	3.347	2.885	3.883	<0.0001
Leg bruise	0.013	2.527	1.989	3.210	<0.0001
Haematoma with intact skin	0.022	3.080	2.587	3.667	<0.0001
Post-traumatic wound infection	0.006	1.718	1.277	2.313	0.0001
Implant complications	0.004	2.148	1.120	4.119	0.01**
Wasp sting	0.007	1.558	1.273	1.909	<0.0001
Open wound of leg	0.011	2.211	1.732	2.823	<0.0001
Heart diseases					
Stroke and cerebrovascular accident	0.011	1.698	1.452	1.986	<0.0001
Postural hypotension	0.032	3.352	2.938	3.825	<0.0001
Cardiac failure	0.005	1.430	1.127	1.815	0.002
DVT - Deep vein thrombosis	0.006	1.480	1.198	1.827	0.0001
Infectious diseases					
Skin and subcutaneous tissue infections	0.009	1.343	1.203	1.500	<0.0001
Respiratory tract infection	0.007	1.246	1.122	1.382	<0.0001
Chest infection	0.002	1.055	1.003	1.110	0.02**
Urinary tract infection	0.029	1.743	1.646	1.846	<0.0001
Cystitis	0.006	1.152	1.068	1.242	0.0001
Cellulitis	0.026	1.785	1.662	1.917	<0.0001
Digestive system diseases					
Bowel obstruction	0.001	1.213	0.789	1.864	0.3**
Constipation NOS	0.016	1.733	1.554	1.933	<0.0001
Gastrointestinal haemorrhage	0.005	1.844	1.192	2.851	0.003
Musculoskeletal Diseases					
Swelling of calf	0.018	1.984	1.742	2.259	<0.0001
Osteoporosis	0.010	1.407	1.264	1.565	<0.0001
Aching leg syndrome	0.012	1.403	1.281	1.535	<0.0001
Acute back pain - lumbar	0.006	1.185	1.090	1.289	<0.0001
Pain in limb	0.012	1.226	1.170	1.285	<0.0001
Others					
Pressure sore	0.017	2.449	2.052	2.923	<0.0001
Restlessness and agitation	0.006	2.010	1.391	2.905	<0.0001
Hyponatraemia	0.011	1.782	1.506	2.109	<0.0001
Difficulty in swallowing	0.008	1.825	1.430	2.329	<0.0001
Dependent oedema	0.010	2.050	1.601	2.624	<0.0001
Retention of urine	0.011	2.893	2.070	4.043	<0.0001
Chronic renal failure	0.005	1.786	1.252	2.548	0.0006
Insomnia	0.007	1.498	1.243	1.806	<0.0001
Dizziness	0.005	1.242	1.062	1.452	0.004
Shortness of breath	0.003	1.142	0.971	1.342	0.09**
Hypothyroidism	-0.001	0.962	0.834	1.111	0.7**
Vitamin B12 deficiency	0.007	1.555	1.268	1.906	<0.0001
Haematuria	0.002	1.180	0.942	1.477	0.1**
Anaemia	0.014	1.673	1.476	1.896	<0.0001
Trophic leg ulcer	0.014	1.850	1.605	2.134	<0.0001
Intertrigo	0.010	1.488	1.302	1.701	<0.0001

Senile confusion	0.024	3.047	2.597	3.575	<0.0001
Collapse	0.019	2.767	2.308	3.319	<0.0001
Abnormal loss of weight	0.013	2.080	1.732	2.499	<0.0001

Table 6.3. Significant associated diseases with falls in very elderly patients (≥ 90 years). Tables (A) and (B) present the associations in men and women, respectively. In (A) 13 diseases showed positive associations with falls – with highest associations in rectal bleeding (RR: 2.30) and atrial fibrillation and flutter (RR: 1.64). In (B) 24 diseases showed positive associations with falls – with highest associations in injuries. LCI, lower confidence interval; UCI, upper confidence interval.

A. Significant associated diseases with falls in men very elderly patients

Disease	***Φ*-correlation**	**RR**	**95% LCI**	**95% UCI**	**P value**
[D] Senile confusion	0.04	2.15	1.45	3.20	<0.0001
Urinary tract infection	0.03	1.57	1.24	1.98	<0.0001
Fracture of unspecified bones	0.03	3.36	1.58	7.16	<0.0001
Pain in limb	0.03	1.46	1.18	1.81	0.0002
Laceration NOS	0.03	2.30	1.38	3.83	0.0002
Postural hypotension	0.03	1.94	1.25	3.03	0.0009
Bleeding PR	0.03	2.30	1.26	4.19	0.002
Fracture of humerus	0.03	2.96	1.27	6.88	0.002
Hip fracture	0.02	1.64	1.10	2.45	0.007
[D]Retention of urine	0.02	1.51	1.00	2.29	0.03
Cellulitis NOS	0.02	1.31	1.00	1.72	0.03
Atrial fibrillation and flutter	0.02	1.64	0.97	2.77	0.04
Respiratory tract infection	0.02	1.34	0.98	1.84	0.04

B. Significant associated diseases with falls in women very elderly patients

Disease	***Φ*-correlation**	**RR**	**95% LCI**	**95% UCI**	**P value**
Hip fracture	0.034	1.722	1.440	2.059	<0.0001
Urinary tract infection	0.031	1.416	1.266	1.584	<0.0001
Head injury	0.031	2.596	1.802	3.740	<0.0001
Intracranial injury NOS	0.026	2.704	1.696	4.310	<0.0001
Laceration NOS	0.026	1.964	1.444	2.672	<0.0001
Arthralgia of hip	0.025	1.553	1.278	1.887	<0.0001
Pain in limb	0.023	1.332	1.171	1.515	<0.0001
Haematoma with intact skin	0.022	1.951	1.368	2.780	<0.0001
Postural hypotension	0.022	1.801	1.319	2.460	<0.0001
[D]Groin pain	0.020	2.320	1.419	3.793	<0.0001
Fracture of humerus	0.020	1.996	1.326	3.005	0.0001
Cellulitis NOS	0.018	1.282	1.112	1.479	0.0002
Fracture of unspecified bones	0.019	2.001	1.309	3.060	0.0002
Minimal trauma fracture	0.018	1.520	1.161	1.990	0.0006
Acute lower respiratory tract infection	0.017	1.644	1.180	2.290	0.0009
Aching leg syndrome	0.017	1.423	1.125	1.799	0.001
Swelling of calf	0.016	1.439	1.110	1.865	0.002
Leg bruise	0.016	1.912	1.180	3.097	0.003
[D] Senile confusion	0.015	1.442	1.097	1.894	0.003
Osteoporosis	0.014	1.419	1.067	1.887	0.007
Closed fracture pelvis, single pubic ramus	0.014	1.947	1.114	3.404	0.007
[D]Collapse	0.014	1.538	1.076	2.198	0.008
Closed fracture of radius (alone)	0.012	1.898	1.018	3.540	0.02
Congestive heart failure	0.011	1.340	1.001	1.795	0.03

6.3.5. Falls-associated diseases in distinct subgroups of very elderly patients

We looked into the whole elderly and very elderly population level; then, we re-applied the comorbidity measures to falls clusters to see how associations might change in subgroups. From clustering, we identified three distinct falls subgroups of very elderly patients for each sex that enrich with falls different associated comorbidities. These subgroups consisted of around 70% of the overall very elderly patients.

We found that patients in men-subgroup 1 (n= 4,435) were more associated with diabetes (RR: 10.70), depression (RR: 14.26), musculoskeletal diseases (RR: 3.71), cardiovascular diseases (RR: 3.62) and urinary tract infection (RR: 3.89) (Table 6.4A). Patients in men-subgroup 2 (n= 2,296) were more diagnosed with type 2 diabetes mellitus (RR: 11.25), infectious diseases (RR: 5.20), mortality (RR: 7.50), rhabdomyolysis (RR: 11.25) and urinary tract infection (RR: 6.43) (Table 6.4B). Patients in men-subgroup 3 (n= 1,196) showed notable associations between falls and Vitamin B12 deficiency anaemia (4.82), anaemia (1.81), malignant neoplasms (2.41) and cardiovascular diseases (2.07) (Table 6.4C).

Most falls-associated diseases in women-subgroup 1 (n=12,065) are around cardiovascular diseases (RR: 5.68), musculoskeletal diseases (RR: 3.80) and infectious diseases (RR: 3.09). They also enrich with chronic confusional state (RR: 8.38), macular cyst or hole (RR: 8.38) and detrusor instability (RR: 6.28) (Table 6.5A). Patients in women-subgroup 2 (n=3,116) were often diagnosed with adverse reaction to beta-blockers (RR: 20.72), statin causing adverse effect in therapeutic use (RR: 20.72), anaemia (RR: 20.72), infectious diseases (RR: 8.60) and cardiovascular diseases (RR: 11.51) (Table 6.5B). Patients in women-subgroup 3 (1,560) had falls significantly associated with nervous system diseases (RR: 3.73), cardiovascular diseases (RR: 4.44), urinary tract infection (RR: 1.41), mixed venous and arterial leg ulcer (RR: 8.69) and anaemia (RR: 1.90) (Table 6.5C).

Table 6.4. Falls-associated diseases in distinct subgroups of very elderly men patients. Tables (A), (B) and(C) present the associations in men-subgroup 1, men-subgroup 2 and men-subgroup 3, respectively. In (A) patients are enriched with diabetes, depression and macular scars. In (B) patients are enriched with type 2 diabetes, Nutritional deficiencies, Keratosis, Rhabdomyolysis and mortality. In (C) patients are enriched with anaemia.

A. Significant associated diseases with falls in men-subgroup 1

Disease	*Φ*-correlation	RR	95% LCI	95% UCI	P value
Type 2 diabetes mellitus	0.03	7.13	0.65	78.42	0.02
Type 1 diabetes mellitus	0.04	14.26	0.89	227.46	0.0004
Depression	0.04	14.26	0.89	227.46	0.0004
Other macular scars	0.04	14.26	0.89	227.46	0.0004
Atrial fibrillation and flutter	0.05	3.47	1.69	7.12	<0.0001
Stroke due to cerebral arterial occlusion	0.03	4.28	1.18	15.47	0.006
DVT - Deep vein thrombosis	0.03	3.00	1.03	8.77	0.02
Hypotension	0.03	3.89	1.09	13.87	0.01
Postural hypotension	0.04	3.44	1.52	7.80	0.0005
Cystitis	0.03	1.94	0.98	3.87	0.04
Urinary tract infection	0.03	3.89	1.09	13.87	0.01
Osteoarthritis NOS	0.04	3.96	1.48	10.60	0.0009
Difficulty in walking	0.03	4.75	0.96	23.46	0.01
Backache, unspecified	0.04	4.19	1.56	11.29	0.0005
Pain in limb	0.05	1.92	1.44	2.57	<0.0001

B. Significant associated diseases with falls in men-subgroup 2

Disease	*Φ*-correlation	RR	95% LCI	95% UCI	P value
Malignant neoplasm of prostate	0.05	2.65	1.24	5.65	0.004
Toxic goitre	0.05	5.63	1.42	22.34	0.0007
Type 2 diabetes mellitus - poor control	0.04	11.25	0.71	179.29	0.002
Nutritional deficiencies	0.06	11.25	1.59	79.49	<0.0001
Cardiac failure	0.04	2.46	1.10	5.51	0.01
Stroke	0.05	7.50	1.26	44.65	0.0008
Postural hypotension	0.05	2.30	1.14	4.65	0.01
Inguinal hernia	0.04	2.39	1.07	5.33	0.02
Renal failure unspecified	0.05	3.46	1.14	10.53	0.008
Urinary tract infection	0.07	6.43	1.90	21.79	<0.0001
Skin and subcutaneous tissue infections	0.05	2.53	1.25	5.15	0.004
Cellulitis NOS	0.05	2.01	1.18	3.42	0.006
Dermatitis NOS	0.05	3.75	1.22	11.53	0.004
Keratosis	0.06	11.25	1.59	79.49	<0.0001
Rhabdomyolysis	0.06	11.25	1.59	79.49	<0.0001
Pain in limb	0.04	1.62	1.03	2.54	0.03
Senile confusion	0.07	2.29	1.46	3.59	<0.0001
Slurred speech	0.07	5.12	1.79	14.58	<0.0001
Death, cause unknown	0.03	3.75	0.76	18.47	0.04
Found dead	0.04	11.25	0.71	179.29	0.002

C. Significant associated diseases with falls in men-subgroup 3

Disease	*Φ*-correlation	RR	95% LCI	95% UCI	P value
Vitamin B12 deficiency anaemia	0.03	4.82	0.30	76.76	0.05
Refractory anaemia	0.02	2.41	0.22	26.46	0.05
Malignant neoplasm of pancreas	0.02	2.41	0.22	26.46	0.05
Dilatation - cardiac	0.02	2.41	0.22	26.46	0.05
Malignant neoplasm of oesophagus	0.02	2.41	0.22	26.46	0.05
Rickets	0.02	2.41	0.22	26.46	0.05
Malignant neoplasm of trachea, bronchus and lung	0.02	2.41	0.22	26.46	0.05
Acute heart failure	0.02	2.41	0.22	26.46	0.05
Prolonged P-R interval	0.02	2.41	0.22	26.46	0.05
DVT - Deep vein thrombosis	0.03	2.07	0.54	7.93	0.05
Stroke due to intracerebral haemorrhage	0.02	1.93	0.38	9.87	0.05

Adverse reaction to ramipril	0.01	1.38	0.29	6.58	0.05
Sideropenic anaemia	0.00	1.20	0.14	10.73	0.05
External haemorrhoids, simple	0.00	1.20	0.14	10.73	0.05
Chronic renal failure	0.00	1.07	0.23	4.92	0.05

Table 6.5. Falls-associated diseases in distinct subgroups of very elderly women patients. Tables (A), (B) and(C) present the associations in women-subgroup 1, women-subgroup 2 and women-subgroup 3, respectively. In (A) patients are enriched with cardiovascular, musculoskeletal and infectious diseases as well as confusional chronic state, macular cyst or hole and detrusor instability. In (B) patients are enriched with adverse reaction to beta-blockers, statin causing adverse effect in therapeutic use and anaemia. In (C) patients are enriched with nervous system diseases and mixed venous and arterial leg ulcer.

A. Significant associated diseases with falls in women-subgroup 1

Disease	***Φ*-correlation**	**RR**	**95% LCI**	**95% UCI**	**P value**
Viral gastroenteritis	0.02	3.31	1.24	8.84	0.005
Norwegian scabies	0.02	2.26	1.01	5.03	0.03
Hyponatraemia	0.02	1.84	1.01	3.36	0.03
Hypokalaemia	0.02	2.87	1.34	6.18	0.002
Other specified disease of blood	0.02	3.35	1.11	10.08	0.01
Chronic confusional state	0.02	8.38	1.40	50.08	0.0003
Macular cyst or hole	0.02	8.38	1.40	50.08	0.0003
Keratitis	0.02	5.39	1.40	20.80	0.001
Aortic stenosis, non-rheumatic	0.02	3.77	1.04	13.68	0.01
Small vessel cerebrovascular disease	0.02	4.19	1.35	12.96	0.002
Aortic aneurysm NOS	0.03	12.57	1.77	89.12	<0.0001
Raynaud's syndrome	0.02	6.28	1.15	34.26	0.003
Temporal arteritis	0.02	6.28	1.15	34.26	0.003
Haemorrhoids NOS	0.03	3.93	1.44	10.70	0.001
Hypotension	0.02	2.73	1.04	7.17	0.02
Chest infection	0.02	2.69	1.12	6.49	0.01
Diverticulosis of the colon	0.02	4.19	1.14	15.45	0.007
Detrusor instability	0.02	6.28	1.15	34.26	0.003
Urinary tract infection	0.04	4.94	2.47	9.89	<0.0001
Skin and subcutaneous tissue infections	0.02	1.42	1.02	1.98	0.03
Cellulitis NOS	0.04	1.63	1.35	1.96	<0.0001
Intertrigo	0.02	1.54	1.02	2.32	0.03
Ingrowing great toe nail	0.02	3.28	1.34	8.03	0.002
Urticaria	0.02	2.99	1.13	7.92	0.01
Osteoporosis	0.03	2.83	1.38	5.81	0.001
Arthralgia of hip	0.02	1.49	1.10	2.02	0.007
Other joint symptoms	0.02	6.28	1.15	34.26	0.003
Spinal stenosis NOS	0.02	8.38	1.40	50.08	0.0003
Adhesive capsulitis of the shoulder	0.03	2.67	1.36	5.28	0.001
Acquired trigger thumb	0.02	6.28	1.15	34.26	0.003

B. Significant associated diseases with falls in women-subgroup 2

Disease	***Φ*-correlation**	**RR**	**95% LCI**	**95% UCI**	**P value**
Nits - head lice	0.18	13.81	2.44	78.06	<0.0001
Microcytic hypochromic anaemia	0.15	20.72	1.35	318.77	<0.0001
Adverse reaction to betablockers	0.15	20.72	1.35	318.77	<0.0001
Statin causing adverse effect in therapeutic use	0.15	20.72	1.35	318.77	<0.0001
Cardio-respiratory arrest	0.15	20.72	1.35	318.77	<0.0001
E.coli infection	0.12	10.36	0.98	109.28	0.004
Urinary tract infection, site not specified	0.11	3.34	1.43	7.79	0.004
Atrial fibrillation	0.10	6.91	0.75	63.33	0.02
Stroke due to cerebral arterial occlusion	0.10	6.91	0.75	63.33	0.02
Newcastle conjunctivitis	0.10	6.91	0.75	63.33	0.02

Cystitis	0.09	4.14	0.97	17.67	0.03

C. Significant associated diseases with falls in women-subgroup 3

Disease	**Φ-correlation**	**RR**	**95% LCI**	**95% UCI**	**P value**
Thrush of mouth and oesophagus	0.04	2.54	1.10	5.84	0.01
anaemia	0.04	1.90	1.08	3.35	0.01
Blurred vision NOS	0.05	4.35	1.32	14.35	0.001
Cellulitis	0.06	2.51	1.43	4.40	0.0002
Otitis externa NOS	0.05	3.73	1.44	9.62	0.0005
Benign paroxysmal positional vertigo	0.06	4.35	1.64	11.50	<0.0001
Primary pulmonary hypertension	0.05	8.69	1.23	61.49	0.0002
Congestive heart failure	0.06	2.56	1.46	4.49	0.0001
External haemorrhoids, simple	0.04	3.48	1.10	11.02	0.008
Postural hypotension	0.07	3.03	1.70	5.39	<0.0001
Functional gastrointestinal tract disorders	0.05	5.22	1.25	21.72	0.001
Urinary tract infection, site not specified	0.03	1.41	1.04	1.89	0.02
Mixed venous and arterial leg ulcer	0.05	8.69	1.23	61.49	0.0002
Arthralgia of hip	0.08	2.70	1.72	4.25	<0.0001
Synovial cyst of popliteal space	0.05	5.22	1.25	21.72	0.001
Pain in limb	0.04	1.57	1.14	2.15	0.004
Tingling of skin	0.05	8.69	1.23	61.49	0.0002
Face ache	0.04	3.86	1.20	12.47	0.004
Slurred speech	0.07	4.68	1.88	11.64	<0.0001

6.4. Discussion

Some previous studies have focused on medical records based on phenotyping, which rely on rule-based approaches. These approaches require significant time and clinical judgement to develop [282], [283]. Thus, there is a need for an automated approach for phenotype generation. Furthermore, some recent studies developed a variety of machining learning and clustering techniques to automatically analyse large clinical data sets to identify comorbidities [154], [279]. These studies are often disease-centric and specialise in disease associations. The goal of our study, in contrast to previous studies, is to map patient records into a low dimensional space to characterize and stratify patients with a specific disease at the subgroups level.

In this large population-based study of the national UK general practice, the distribution of the dataset suggests that while falls are not serious condition for children and adults, it is a serious health issue for elderly patients. At the age of menopause, falls starts increasing rapidly for women, whereas in men this increase started at 65 years. The results of geriatric falls showed more important sex and age differences than in adults, where falls rates are higher in women patients and in older age groups. This result is similar to some of the results reported by [284].

After clustering analysis, we found 48 diseases notably associated with falls. The association around injuries [285]–[296], infections [297]–[300], cardiovascular [301]–[307] and musculoskeletal diseases [265], [308], [309] are supported in the clinical literature. Digestive system diseases have an indirect relationship with falls as reported in [310]. Studies of the connections between falls and anaemia [311]–[313] and senile confusion [314], [315] provide similar results to our study. In addition, the strengths of most of the associations were higher in men than women. This could be related to the fact that the number of women is greater than men (334,466 and 254,703 respectively).

The results showed that the relative attribution of falls incidences increased with age, particularly very elderly patients, as demonstrated in [287], [316], [317]. As a result, we focused our analysis on very elderly falls-patients, who are at higher risk of falls than other patients in the elderly population. We found 13 diseases for men and 24 for women that are notably associated with falls; most of these diseases are a subset of the 48 diseases at the elderly population level. However, the associations between falls and atrial fibrillation and flutter were observed only in very elderly men patients. A previous study proved the association between the two diseases at this age, however, no sex differences were found [318].

Observing the distinct differences in disease associations between different falls-subgroups might uncover useful characterisations of falls-patients. We identified three completely independent subgroups for each sex enriched of falls codes. Men subgroups consisted of ~80% of the overall very elderly men patients, while women subgroups comprised 65% of the overall very elderly women patients.

Patients in men-subgroup 1 are mostly associated with both types of diabetes [319]–[323], depression [324]–[326], musculoskeletal diseases, cardiovascular diseases and urinary tract infection [327]. In this subgroup, patients were strongly associated with both depression and hypotension. However, no direct associations have been reported in the clinical literature review. Patients in men-subgroup 2 were more likely to associate with type 2 diabetes mellitus, infectious diseases and malignant neoplasm of prostate [328] as well as musculoskeletal diseases, cardiovascular diseases and urinary tract infection. Moreover, a study reports that age and the male sex are independent risk factors for falls-related deaths [329], as found in this

subgroup. Although the association between renal failure and senile confusion together with falls was observed in this subgroup, there are no studies directly linking these diseases to falls. However, no studies have reported the link between both diseases and falls. Patients in men-subgroup 3 were enriched for vitamin B12 deficiency anaemia [311], [330], [331] and myeloid leukaemia [324]–[326].

In women-subgroup 1, falls had strong connections with hyponatremia, hypokalaemia, nervous, cardiovascular and muscular diseases. A previous study has demonstrated an association between falls and all these diseases [332]. Women-subgroup 2 patients were more likely to be diagnosed with infections, anaemia and cardiovascular diseases. In addition, beta-blockers and statin are two types of medications for cardiovascular diseases [333], [334]. Results of previous studies suggest adverse reaction to these medications have strong connections with falls in the elderly [335], [336]. These results are very similar to those in this subgroup. Patients in women-subgroup 3 have been strongly associated with cardiovascular diseases, infections, anaemia and musculoskeletal diseases. Patients were also associated with functional gastrointestinal tract disorder through benign paroxysmal positional vertigo (BPPV), as demonstrating in [310].

The only falls-associated disease category all subgroups had in common was cardiovascular diseases. In addition, urinary tract infection appeared in all subgroups except in men-subgroup 1. Epidemiological studies have proved that both diseases [297]–[300] and [327] are independently associated with a risk increase for falls. In addition, depressive symptoms and musculoskeletal conditions are significantly associated with falls in men-subgroup 1 and women-subgroup 1. Previous study proved the association between falls and depression, mediated by musculoskeletal diseases [337].

The outcome of this study is to help generating hypotheses. These outputs will be examined by experts in the domain to see the extent to which they are hypotheses that have been validated by previous work. These will provide a test for the methods. We also found other falls-associated diseases for example, cystitis; diverticulosis of the colon and gastrointestinal haemorrhage. However, to the best of our knowledge, these associations were not previously reported in the literature. Any un-validated hypotheses that were generated, and which were

believed to be of interest to the domain experts, might become the basis of a further study in order to test them.

Our study has some limitations. Although CPRD is the largest data set in the UK and include patients in residential and nursing care homes, some of the chronic diagnoses were underestimated such as musculoskeletal diseases [338]. Another limitation of this study is that the exposures of interest were diagnosis codes taken from the cases. One possible extension to this work is to include other types of medical codes such as medications, treatments and measures as well as linking to genotypes and biomarkers [42], [162]. This will facilitate the translation of data and lead to a more precise stratification of patients with a specific disease. In this study, disease trajectory was not considered. It is not clear when the first diagnosis of a fall or other diseases took place, particularly, which diseases were recorded beforehand and in which order.

In conclusion, this chapter used the mapping methodology to characterize and stratify patients with falls at subgroups level. This work allows exploration of the data around falls and determines interesting patterns or correlations in the data. At first, falls-associations in the population as a whole and in the very elderly population were studied. Strong relationships were observed between falls and five disease categories: infectious, cardiovascular, injury, musculoskeletal and digestive system diseases, which are well reported in the literature. Then falls-associations were looked within each of the subgroups identified in the clustering. Useful and distinct characterisations of falls-patients were successfully found, some of which are well defined in the literature. However, interesting novel hypotheses will be passed to epidemiologists for further exploration.

Acknowledgments

I would like to thank Muhannad Almohaimeed for his contribution on providing the statistical analysis of this study and Tjeerd Van Staa for providing helpful background information on falls in the elderly.

Chapter 7: Conclusion and future directions

7.1. Overall discussion

This thesis aimed to develop a novel methodology that can provide a way to map patient records into a low dimensional vector space. The initial hypothesis explored in this thesis was to investigate that such a mapping would allow us to apply a wide range of data mining strategies that would make the data easier to understand and hopefully supports data exploration and hypothesis generation. To achieve this, a novel methodology was developed to makes it possible to represent patient data into low dimensional vector space. This methodology builds upon the idea of semantic similarity to take patient data from a diagnosis space and map it to a low dimensional vector space in a way that helps to make useful medical interpretation of the data.

To the best of our knowledge, there is no similar work in existing literature that uses the notion of similarity representation of patients for mapping patients into low dimensional space vector space at population level. Most of the current research on electronic patient records is dealing with medical terms by trying to find associations between certain terms with other sets of terms. These studies use a number of machine learning and data mining techniques [160]–[164]. However, there are other techniques that cannot be directly applied to patient data with its original format. The work done in developing the mapping methodology has provided a way to transform patient data into a space that can be suitable for any machine learning or data mining strategies to be applied to.

The research phases and achievements of this thesis were presented in a logical sequence. The first phase of the research was to find a novel methodology to transform electronic patient records from medical records space into a low dimensional vector space. This was done in two transformation steps. The first step was to find a strategy to map patients records into a similarity space. For this, we applied the semantic similarity in order to find the similarities between sets of patients. The second step was to develop a strategy to take the patients from the similarity space and transform them into a lower dimensional space. For this step, we used

the PCA to reduce the data dimensionality. Having patient records represented in a low dimensional vector space allows many traditional data mining techniques to be applied to the data. This was tested on a small scale study based on patient records from the NHS Salford integrated care record. The results of this study show that the methodology does appear to provide this mapping effectively and in way that gives useful medical interpretations.

The second phase was to explore the use of the methodology at much larger patient data sets. In this phase, we looked at issues related to the scalability behaviour of the methods. At this phase of the research, we had access to a data set of 2.7 million patients derived from the CPRD database. Applying the methodology to data of this scale posed a number of challenges. One of these challenges was related to choosing a suitable similarity measure. It was shown that each measure has its own way to represent patient records. In an evaluation of a number of similarity measures, we shown in Chapter 4 that using the combination of the Resnik measure with the Maximum approach could reveal more details about the classification in patient records. It is, therefore, recommend to use these measures when calculating similarity between patients.

Another important challenge faced was whether the method developed to select representative patients scales with the data size increases. Such a process becomes difficult when performed in large scale data. In this analysis, it was demonstrated that this method scaled well as the data volumes increase. The analysis described in Chapter 4 showed that the methodology was efficient and scalable. In terms of memory, the method scales well in large patient data sets. This work validated the mathematical strategies in the methodology, showing that the analysis is straightforward and useable on a large scale. It should be noted that this validation was not only for the semantic similarity analysis; it was able to use semantic similarity at scale and then to employ mapping from similarity space to distance space.

The third phase focused on one particular disease as a use case. In this analysis, the methodology was applied to characterize and stratify patients with falls at a population level. This work allowed the exploration of the data around falls and to determine interesting patterns or correlations. It also identified distinct subgroups of patients with falls, where each subgroup consisted of a number of diseases that have been significantly associated with falls. Several of

these associations was well-documented in the medical literature. This was a validation of the methodology as it retrieves some of the known associations. However, there were other associations that were not reported in the literature. An example of this is the association between falls and cystitis. Such hypotheses need further exploration using classical epidemiology.

The findings of this work provide a snapshot of a disease processes over a large number of patients. This does not generate full hypotheses as the current work only focuses on one side of the story (focuses only on what happened in one year) and disease trajectory was not considered. So, it is not clear when the first diagnosis of a fall or any other diseases took place, particularly, which diseases were recorded beforehand and in which order. This does not allow us to analyse the temporal nature of disease, which is important for modelling more detailed predictions.

The methodology developed in this thesis can be considered as an enabling technology of big health data to characterise and understand patients interactions with the healthcare system at a population level. Also, this methodology can be used to reveal some hidden signals in the data that were not known and make them more obvious for further exploration and testing. A key finding of this research is the wealth of data that can be produced. In the first use case it gave important hypotheses.

In summary, the overall outcome from the thesis is the development of novel methodology that can be used for data exploration and hypothesis generation from patient records. The results have proven that the methodology is computationally tractable. What was previously thought of as a complex and large problem is very manageable through this analysis.

7.2. Limitations

Although this thesis successfully provided novel strategies that support data exploration from patient medical records, there are a number of limitations regarding the data and the methods used in this work. Firstly, we are working directly in this research with data as it exists in the Salford database and the CPRD database. It is known that there are significant differences between different GP practices in how electronic patient records are being recorded. We are

therefore working with data that is inherently noisy. Consequently, some real signals can be missed, and spurious signals generated. Specifically, in the case of some of the chronic diagnoses such as musculoskeletal diseases and falls we know that they are under reported [338]. This is because that the level of recording in patient records may vary between practices (some recording more details than others). This can be also because patients and GPs define conditions differently. For example, patients and GPs fail to perceive falls as a condition, or they consider it an unavoidable part of the ageing process. As a result, we know that there will be some patients who are wrongly included in the control (no fall recorded) group. This reduces the power of our analysis meaning we might miss some of the less strong associations. Also, the clustering approach aims to detect patients with Read codes that are frequently correlated. If a practice would record very few Read codes, this practice would be included into a cluster with low correlations. The planned description of the characteristics of each cluster would identify such practices but of course the interpretation of what is happening within such practices with low quality recording will be limited.

Secondly, there are other limitations regarding the methods used in the methodology. In this research, the idea of finding a set of representative patients was introduced to address the problem of the pairwise comparison between patients in the data. Identifying representative patients from a data sets has helped to reduce the overall processing time. However, as representative patients act as 'covering set' for all patients in the data, there are well-known problems associated with building a covering set [208], [209]. One of these problems is related to the overrepresentation of some of the representative patients. However, as we perform the PCA on the similarity matrix, the redundancy of the representative patients should be taken out. Nonetheless, an improvement should be introduced to this method to address this problem. The techniques used for the dimensionality reduction and the clustering such as PCA and DBSCAN were performed on the data for the purpose of demonstrating the methodology. These techniques have shown to produce results that are informative and reproducible. This was one way to validate these results. However, there might be better strategies to do either the dimensionality reduction or the clustering. What was presented in this thesis is just the start of the process. The data in its new format become more amenable to much wider range of approaches.

There is also a possibility of the method itself introducing noise into the signal. The process of dimensionally reduction will mean that some information is lost. We also know that the precise way in which we perform the analysis – either in dimensional reduction or clustering phases - can introduce some uncertainty. There could well be hypotheses in the data that we will falsely reject.

7.3. Future work

The methodology developed in this thesis could be extended and applied in different ways. First, it could be extended to explore patient trajectories over a period of time. The work presented in Chapter 6 was part of a larger study using the methodology developed in this thesis to investigate the trajectories of patients with falls. The work done so far was to stratify patient with falls. One area of future work could be to focus on incorporating temporal dimensions which might provide useful insights into missed opportunities detection, risk modelling and understanding of a disease. One way of achieving this is to use the mapping methodology to identify a set of patients who had falls in one year. These patients will be matched to a control set of patients. The controls will have no records of falls. The history of all patients will be tracked back for missed opportunity detection, and following years in order to find the consequences of falls. The timeline will be divided into several time dimensional vectors depending on the number of years are being analysed. In each time dimensional vector (e.g. year), we will capture all codes that have been recorded. Then the mapping methodology will be applied on each vector separately. Essentially, the same approach can be applied towards the study the trajectory of other diseases.

One aspect of the methodology that was not covered in this study, but would be very useful to investigate, would be to analyse secondary care data. For many diseases, the diagnostic pathway is examined in both primary and secondary care. Primary care data contains an overview of a patient's medical journey, while secondary care data provides detailed events about specific conditions. Analysing linked primary and secondary care patient records could, therefore, provide a better understanding of different health conditions, trends in disease occurrence and management, effectiveness of prevention and treatment, and side effects of treatment [339].

In the current work, the exposures of interest were diagnosis codes taken from patient records. One such direction would be to include other types of medical codes such as medications, treatments and measures as well as linking to genotypes and biomarkers. This would facilitate the translation of data and lead to a more precise understanding of patients with specific diseases.

In addition, in terms of applications of the mapping methods, the methodology has recently been deployed on the CPRD data sets in different projects to stratify and identify patients with different diseases and medications. Firstly, the methodology was applied on antibiotics users as an initial process of a project in collaboration with Farr Institute for Health Informatics Research at the University of Manchester. This project aims to achieve this by improving data analytics by developing the capabilities to identify interventions and target population with the potentially highest impact and by implementing at least two simple interventions to reduce antibiotic prescribing in Greater Manchester. The methodology was deployed on the overall cohort of incidental antibiotics users in the CPRD data sets during year 2000 to 2016 in order to provide descriptive analyses of characteristics in patients with incidental use of antibiotics. Secondly, the methodology was also applied to the CPRD data set to identify different inflammatory bowel disease (IBD) patients with different disease associations. The analysis was conducted on six different diseases of IBD, which are abdominal pain, rectal bleeding, coeliac disease, constipation, irritable bowel syndrome and other non-infective inflammatory gastroenteritis and colitis. This project is collaboration with the Gastroenterology and Nutrition group at the University of Manchester.

Finally, there are a number of possible areas outside healthcare. These include applying the methodology to data that comes from media. The methodology can be employed to analyse large amounts of data described by tokens from taxonomies or ontologies. In a recent collaboration between the University of Manchester and the BBC, we started applying the methodology on data from BBC news articles. A preliminary analysis of this data showed promising results on clustering of news articles based on their similarities.

7.4. Conclusion

The use of patient records in medical research has shown great potential to discover new medical insights. However, the data is complex and high dimensional and this makes many traditional data mining strategies difficult to implement. Much of the potential value of this data, therefore, goes untapped. This thesis focused on developing a novel methodology that allows the mapping of patient records into a low dimensional vector space. We believe that the ability to represent the data in such way would seem to make it amenable to analysis through more traditional data mining techniques, thus, allowing a more intuitive and straightforward environment for hypotheses formulation. We have shown throughout the thesis that the methodology does appear to provide this mapping effectively and in a way that appears to be novel. The work presented in this thesis has barely scratched the surface of what this mapping could be used for. It opens up the possibility of applying a wide range of data mining strategies which have not yet been explored. What the thesis has shown is one strategy that works, but there could be many more. It is also worth noting that no aspect of the implementation of this methodology that restricts it to medical codes described by Read codes; it could equally be applied to the analysis and visualisation of many other data sources, which are described using terms from taxonomies or ontologies.

References

[1] G. S. el-Assal, “Ancient Egyptian medicine.,” *Lancet*, vol. 2, pp. 272–274, 1972.

[2] J. F. Nunn, *Ancient egyptian medicine*. University of Oklahoma Press, 2002.

[3] M. E. Salem and G. Eknoyan, “The kidney in ancient Egyptian medicine: where does it stand?,” *Am. J. Nephrol.*, vol. 19, no. 2, pp. 140–7, 1999.

[4] A. Saber, “Ancient Egyptian surgical heritage.,” *J. Invest. Surg.*, vol. 23, no. 6, pp. 327–34, Dec. 2010.

[5] C. Von KLEIN, “The medical features of the papyrus Ebers,” *J. Am. Med. Assoc.*, vol. XLV, no. 26, pp. 1928–1935, Dec. 1905.

[6] R. F. Gillum, “From Papyrus to the Electronic Tablet: A Brief History of the Clinical Medical Record with Lessons for the Digital Age,” *Am. J. Med.*, vol. 126, no. 10, pp. 853–857, 2013.

[7] M. Stiefel, A. Shaner, and S. D. Schaefer, “The Edwin Smith Papyrus: The Birth of Analytical Thinking in Medicine and Otolaryngology,” *Laryngoscope*, vol. 116, no. 2, pp. 182–188, Feb. 2006.

[8] Q. Al-Awqati, “How to write a case report: lessons from 1600 B.C.,” *Kidney Int*, vol. 69, no. 12, pp. 2113–2114, 2006.

[9] J. H. Breasted, *The Edwin Smith Surgical Papyrus Published in Facsimile and Hieroglyphic Transliteration*. Chicago: University of Chicago Press, 1930.

[10] L. T. Kohn, J. M. Corrigan, and M. S. Donaldson, *To err is human:: building a Safer Health System*, vol. 6. National Academies Press, 2000.

[11] a. Wolfe, “Institute of Medicine Report: Crossing the Quality Chasm: A New Health Care System for the 21st Century,” *Policy, Polit. Nurs. Pract.*, vol. 2, no. 3, pp. 233–235, 2001.

[12] T. G. Thompson, D. J. Brailer, and D. J. Braile, “The Decade of Health Information Technology: Delivering Consumer-centric and Information-rich Health Care,” 2004.

[13] A. M. Epstein, T. H. Lee, and M. B. Hamel, “Paying physicians for high-quality care.,” *N. Engl. J. Med.*, vol. 350, no. 18, pp. 1910-1912-1912, 2004.

[14] W. F. Stewart, N. R. Shah, M. J. Selna, R. a Paulus, and J. M. Walker, “Bridging the inferential gap: the electronic health record and clinical evidence.,” *Health Aff. (Millwood).*, vol. 26, no. 2, pp. w181-91, 2007.

[15] R. Hillestad, J. Bigelow, A. Bower, F. Girosi, R. Meili, R. Scoville, and R. Taylor, “Can electronic medical record systems transform health care? Potential health benefits,

savings, and costs.," *Health Aff. (Millwood).*, vol. 24, no. 5, pp. 1103–17, 2005.

[16] J. M. Walker, "Electronic medical records and health care transformation.," *Health Aff. (Millwood).*, vol. 24, no. 5, pp. 1118–20, 2005.

[17] E. H. Shortliffe, "The evolution of electronic medical records.," *Acad. Med.*, vol. 74, pp. 414–419, 1999.

[18] M. Conrick, *Health informatics: Transforming healthcare with technology*. Thomson Learning Australia.

[19] ISO/TC 215 Health informatics, *Health informatics -- Electronic health record -- Definition, scope and context*. 2005.

[20] S. Gillam and A. N. Siriwardena, *The Quality and Outcomes Framework: Qof-Transforming General Practice*. Radcliffe Publishing.

[21] M. Roland and S. Campbell, "Successes and failures of pay for performance in the United Kingdom.," *N. Engl. J. Med.*, vol. 370, no. 20, pp. 1944–9, 2014.

[22] S. Campbell, A. Steiner, J. Robison, D. Webb, A. Raven, and M. Roland, "Is the quality of care in general medical practice improving? Results of a longitudinal observational study.," *Br. J. Gen. Pract.*, vol. 53, no. 489, pp. 298–304, Apr. 2003.

[23] S. Campbell, D. Reeves, E. Kontopantelis, E. Middleton, B. Sibbald, and M. Roland, "Quality of primary care in England with the introduction of pay for performance," *N. Engl. J. Med.*, vol. 357, no. 2, pp. 181–190, Jul. 2007.

[24] M. Roland, "Linking Physicians' Pay to the Quality of Care — A Major Experiment in the United Kingdom," *N. Engl. J. Med.*, vol. 351, no. 14, pp. 1448–1454, Sep. 2004.

[25] T. Doran, C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh, and M. Roland, "Pay-for-Performance Programs in Family Practices in the United Kingdom," *N. Engl. J. Med.*, vol. 355, no. 4, pp. 375–384, Jul. 2006.

[26] S. M. Campbell, M. O. Roland, E. Middleton, and D. Reeves, "Improvements in quality of clinical care in English general practice 1998-2003: longitudinal observational study," *BMJ*, vol. 331, no. 7525, p. 1121, Nov. 2005.

[27] S. Campbell, M. Roland, and D. Wilkin, "Improving the quality of care through clinical governance," *BMJ*, vol. 322, no. 7302, pp. 1580–1582, Jun. 2001.

[28] M. Berg, "Medical work and the computer-based patient record: a sociological perspective.," *Methods Inf. Med.*, vol. 37, no. 3, pp. 294–301, 1998.

[29] A. L. Rector, W. A. Nowlan, S. Kay, W. a Nolan, and S. Kay, "Foundations for an Electronic Medical Record," *Methods Inf Med*, vol. 30, no. 3, pp. 179–186, 1991.

[30] M. A. Musen, Y. Shahar, and E. H. Shortliffe, “Biomedial Informatics: computer applications in health care and biomedicine,” Springer Science & Business Media, 2006, pp. 698–736.

[31] J. Merrill, Z. Camacho, L. Laux, R. Lorimor, J. Thornby, and C. Vallbona, “Uncertainties and ambiguities: measuring how medical students cope,” *Med. Educ.*, vol. 28, no. 4, pp. 316–322, 1994.

[32] J. S. Ash, M. Berg, and E. Coiera, “Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors,” *J. Am. Med. Informatics Assoc.*, vol. 11, no. 2, pp. 104–112, Mar. 2004.

[33] P. Taylor, *From patient data to medical knowledge: the principles and practice of health informatics*. Blackwell Publishing Ltd, 2008.

[34] B. Chaundhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton, and P. G. Shekelle, “Systematic Review: Impact of Health Information Technology on,” *Ann. Intern. Med.*, vol. 144, no. 10, pp. 742–752, 2006.

[35] Y. Bar-Dayan, H. Saed, M. Boaz, Y. Misch, T. Shahar, I. Husiascky, and O. Blumenfeld, “Using electronic health records to save money,” *J. Am. Med. Informatics Assoc.*, vol. 20, no. e1, pp. e17–e20, Jun. 2013.

[36] P. G. Shekelle, S. C. S. Morton, and E. E. B. Keeler, “Costs and benefits of health information technology.,” *Evid. Rep. Technol. Assess. (Full. Rep).*, vol. 132, pp. 1–71, 2006.

[37] D. C. Kaelber and D. W. Bates, “Health information exchange and patient safety,” *J. Biomed. Inform.*, vol. 40, no. 6, pp. S40–S45, Dec. 2007.

[38] L. A. García Rodríguez and S. Pérez Gutthann, “Use of the UK General Practice Research Database for pharmacoepidemiology,” *Br. J. Clin. Pharmacol.*, vol. 45, no. 5, pp. 419–425, May 1998.

[39] R. L. Richesson, M. M. Horvath, and S. a Rusincovitch, “Clinical research informatics and electronic health record data.,” *Yearb. Med. Inform.*, vol. 9, pp. 215–23, 2014.

[40] P. J. Embi and P. R. O. Payne, “Clinical Research Informatics: Challenges, Opportunities and Definition for an Emerging Domain,” *J. Am. Med. Informatics Assoc.*, vol. 16, no. 3, pp. 316–327, May 2009.

[41] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care.,” *Nat Rev Genet*, vol. 13, no. 6, pp. 395–405, 2012.

[42] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, and T. Werge, “Using electronic patient records to discover

disease correlations and stratify patient cohorts," *PLoS Comput Biol*, vol. 7, no. 8, p. e1002141, 2011.

[43] S. Hoffman, "Electronic Health Records and Research: Privacy Versus Scientific Priorities," *Am. J. Bioeth.*, vol. 10, no. 9, pp. 19–20, Sep. 2010.

[44] H.-U. Prokosch and T. Ganslandt, "Perspectives for medical informatics," *Methods Inf Med*, vol. 48, no. 1, pp. 38–44, 2009.

[45] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.

[46] T. Benson, *Principles of health interoperability HL7 and SNOMED*, vol. 36. London: Springer Verlag, 2010.

[47] P. Coorevits, M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, G. De Moor, and D. Kalra, "Electronic health records: new opportunities for clinical research," *J. Intern. Med.*, vol. 274, no. 6, pp. 547–560, Dec. 2013.

[48] F. F. Ozair, N. Jamshed, A. Sharma, and P. Aggarwal, "Ethical issues in electronic health records: A general overview," *Perspect. Clin. Res.*, vol. 6, no. 2, pp. 73–76, 2015.

[49] S. Noble, J. Donovan, E. Turner, C. Metcalfe, A. Lane, M. A. Rowlands, D. Neal, F. Hamdy, Y. Ben-Shlomo, and R. Martin, "Feasibility and cost of obtaining informed consent for essential review of medical records in large-scale health services research," *J. Health Serv. Res. Policy*, vol. 14, no. 2, pp. 77–81, 2009.

[50] K. El Emam, S. Rodgers, and B. Malin, "Anonymising and sharing individual patient data," *BMJ*, vol. 350, Mar. 2015.

[51] D. J. Willison, C. Emerson, K. V Szala-Meneok, E. Gibson, L. Schwartz, K. M. Weisbaum, F. Fournier, K. Brazil, and M. D. Coughlin, "Access to medical records for research purposes: varying perceptions across research ethics boards," *J. Med. Ethics*, vol. 34, no. 4, pp. 308–314, Apr. 2008.

[52] Cabinet Office Prime Minister's Office 10 Downing Street and The Rt Hon David Cameron MP, "PM speech on life sciences and opening up the NHS," 2011. .

[53] G. Rahmathulla, H. G. Deen, J. A. Dokken, S. M. Pirris, M. A. Pichelmann, E. W. Nottmeier, R. Reimer, and R. E. Wharen Jr, "Migration to the ICD-10 coding system: A primer for spine surgeons (Part 1)," *Surg. Neurol. Int.*, vol. 5, no. Suppl 3, p. S185, 2014.

[54] R. M. Merrill, *Introduction to epidemiology*. Jones & Bartlett Publishers, 2013.

[55] P. Dunn, "Dr William Farr of Shropshire (1807–1883): obstetric mortality and training," *Arch. Dis. Childhood-Fetal Neonatal Ed.*, vol. 87, no. 1, pp. F67–F69, 2002.

[56] M. Whitehead, “William Farr’s legacy to the study of inequalities in health,” *Bull. Heal. Organ.*, vol. 78, no. 1, pp. 86–87, 2000.

[57] W. Farr, *Vital statistics: a memorial volume of selections from the reports and writings of William Farr*. Royal Sanitary Institute, 1885.

[58] J. M. Eyler, “Health Statistics in Historical Perspective,” *Heal. Stat. Shap. Policy Pract. to Improv. Popul. Heal.*, p. 24, 2005.

[59] F. Nightingale, *Florence Nightingale: measuring hospital care outcomes*. Joint Commission Resources, 1999.

[60] “Standardization of Hospital Records — Medical Legislation: The Registration of Drug-Stores in Massachusetts — An Important Discovery — Medical Notes,” *Bost. Med. Surg. J.*, vol. 170, no. 2, pp. 63–66, Jan. 1914.

[61] G. M. Hayes and D. E. Barnett, *UK health computing: recollections and reflections*. British Computer Society, 2008.

[62] J. L. Painter, “Toward Automating an Inference Model on Unstructured Terminologies: OXMIS Case Study,” Springer, 2010, pp. 645–651.

[63] J. Perry, *Oxmis Problem Codes for Primary Medical Care: A Coding System Devised for Participants in the Oxford Community Health Project*. Oxmis Publications, 1978.

[64] R. Read, “The Read clinical classification,” vol. 300, no. 6742, p. 45.

[65] P. Moreton, *The Very Stuff of General Practice*. Radcliffe, 1999.

[66] Department of Health, *Computerisation in GP practices 1993 survey*. Leeds: NHS Management Executive, 1993.

[67] L. Simpson and P. Robinson, *E-clinical Governance: A Guide for Primary Care*. Radcliffe Medical Press, 2002.

[68] P. Preece, *The Use of Computers in General Practice*, 4th ed. Elsevier Health Sciences, 2000.

[69] Health and Social Care Information Centre, “Attribution Data Set GP-Registered Populations Scaled to ONS Population Estimates – 2011,” 2012. .

[70] Health and Social Care Information Centre, “General and Personal Medical Services, England - 2004-2014, As at 30 September,” 2015. .

[71] J. J. Cimino, “From Data to Knowledge through Concept-oriented Terminologies: Experience with the Medical Entities Dictionary,” *J. Am. Med. Inform. Assoc.*, vol. 7, no. 3, pp. 288–297, 2000.

[72] AHIMA and AMIA Terminology and Classification Policy Task Force Members,

Healthcare Terminologies and Classifications: An Action Agenda for the United States. AHIMA and AMIA, 2006.

[73] K. E. Campbell and K. Giannangelo, “Language barrier: getting past the classifications and terminologies roadblock,” *JOURNAL-AHIMA*, vol. 78, no. 2, p. 44, 2007.

[74] E. Coiera, *Guide to Medical Informatics, the Internet and Telemedicine*. Chapman and Hall Medical, Ltd., 1997.

[75] WHO EXECUTIVE BOARD, “eHealth : standardized terminology,” 2006.

[76] S. T. Rosenbloom, R. A. Miller, K. B. Johnson, P. L. Elkin, and S. H. Brown, “Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems,” *J. Am. Med. Informatics Assoc.*, vol. 13, pp. 277–288, 2006.

[77] D. A. Evans, J. J. Cimino, W. R. Hersh, S. M. Huff, and D. S. Bell, “Toward a medical-concept representation language. The Canon Group.,” *J. Am. Med. Informatics Assoc.*, vol. 1, no. 3, pp. 207–217, 1994.

[78] K. E. Campbell, D. E. Oliver, K. A. Spackman, and E. H. Shortliffe, “Representing Thoughts, Words, and Things in the UMLS,” *J. Am. Med. Informatics Assoc.*, vol. 5, no. 5, pp. 421–431, Sep. 1998.

[79] J. J. Cimino, “The concepts of language and the language of concepts.,” *Methods Inf Med*, vol. 37, no. 4, p. 311, 1998.

[80] A. L. Rector, “Clinical terminology: Why is it so hard?,” *Methods Inf. Med.*, vol. 38, no. 4, pp. 239–252, 1999.

[81] A. L. Rector, “Thesauri and formal classifications: Terminologies for people and machines,” *Methods Inf. Med.*, vol. 37, no. 4–5, pp. 501–509, 1998.

[82] R. Qamar, “Semantic mapping of clinical model data to biomedical terminologies to facilitate interoperability,” 2008.

[83] J. J. Cimino, “Collect once, use many. Enabling the reuse of clinical data through controlled terminologies.,” *J. AHIMA*, vol. 78, no. 2, pp. 24-29-32, 2007.

[84] J. J. Cimino, “In defense of the Desiderata.,” *J. Biomed. Inform.*, vol. 39, no. 3, pp. 299–306, Jun. 2006.

[85] R. Engelbrecht and J. Ingenerf, “Relevance of Terminological Standards and Services in Telemedicine,” in *Information Technology Solutions for Healthcare*, Springer, 2006, pp. 110–134.

[86] J. Zhang, “Representations of health concepts: a cognitive perspective,” *J. Biomed. Inform.*, vol. 35, no. 1, pp. 17–24, Feb. 2002.

[87] S. de Lusignan, “Codes, classifications, terminologies and nomenclatures: definition, development and application in practice,” *Inform. Prim. Care*, vol. 13, no. 1, pp. 65–70, 2005.

[88] R. Qamar, J. Kola, and A. L. Rector, “Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies.,” *AMIA Annu. Symp. Proc.*, vol. 2007, pp. 608–613, 2007.

[89] P. Aalseth, *Medical coding: what it is and how it works*. Jones & Bartlett Publishers, 2005.

[90] J. Cimino, “Review paper: coding systems in health care,” *Methods Inf. Med. der Inf. der Medizin*, vol. 35, no. 4, pp. 273–284.

[91] A. L. Rector, W. A. Nowlan, and S. Kay, “Foundations for an electronic medical record.,” *Methods Inf Med*, vol. 30, no. 3, pp. 179–186, 1991.

[92] L. Jacobs, “Interview with Lawrence Weed, MD— The Father of the Problem-Oriented Medical Record Looks Ahead,” *Perm. J.*, vol. 13, no. 3, pp. 84–89, 2009.

[93] L. L. Weed, “Medical Records That Guide and Teach,” *N. Engl. J. Med.*, vol. 278, no. 12, pp. 652–657, Mar. 1968.

[94] P. Salmon, A. Rappaport, M. Bainbridge, G. Hayes, and J. Williams, “Taking the problem oriented medical record forward.,” *Proc. AMIA Annu. Fall Symp.*, pp. 463–467, 1996.

[95] G. M. Hayes, “Computers in the consultation. The UK experience.,” *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 103–106, 1993.

[96] M. J. Ball and M. F. Collen, *Aspects of the computer-based patient record*. Springer Science & Business Media, 1992.

[97] P. Taylor, “Guide to Medical Informatics, the Internet and Telemedicine; Cybermedicine,” *BMJ*, vol. 316, no. 7125, p. 158, Jan. 1998.

[98] L. Toews, “An evaluation methodology for clinical vocabularies and evaluation of the read codes,” vol. 36, pp. 30–43.

[99] D. A. R. Obinson, D. I. P. C. Omp, E. R. S. Chulz, P. H. B. Rown, C. O. P. Rice, M. P. Hil, D. Robinson, E. Schulz, P. Brown, C. Price, D. A. R. Obinson, D. I. P. C. Omp, E. R. S. Chulz, P. H. B. Rown, C. O. P. Rice, and M. P. Hil, “Updating the Read Codes : User-interactive Maintenance of a Dynamic Clinical Vocabulary,” *J. Am. Med. Inform. Assoc.*, vol. 4, no. 6, pp. 465–472, Nov. 1997.

[100] N. Smith, A. Wilson, and T. Weekes, “Use of Read codes in development of a standard data set,” *Br. Med. J.*, vol. 311, no. 7000, pp. 313–315, 1995.

[101] K. Giannangelo, “Making the connection between standard terminologies, use cases and

mapping," *HIM J.*, vol. 35, no. 3, pp. 8–12, 2006.

[102] R. Lawrenson, T. Williams, and R. Farmer, "Clinical information for research; the use of general practice databases," *J.Public Heal. Med.*, vol. 21, pp. 299–304, 1999.

[103] S. Gnani and A. Majeed, *A user's guide to data collected in primary care in England*. ERPHO., 2006.

[104] N. Booth, "What are the Read Codes?," vol. 11, no. 3, pp. 177–182.

[105] T. Benson, "Why general practitioners use computers and hospital doctors do not," *BMJ*, vol. 325, no. November, pp. 1090–1093, 2002.

[106] House of Commons: Health Committee, *The electronic patient record: Sixth Report of Session 2006-07*, no. September. 2007.

[107] T. E. Bentley, C. Price, and P. J. B. Brown, "Structural and lexical features of successive versions of the Read Codes," in *proceedings of the 1996 Annual Conference of the Primary Health Care Specialist Group of the British Computer Society.*, pp. 91–103.

[108] L. Hawe, E. and Cockcroft, *OHE Guide to UK Health and Health Care Statistics*. Office of Health Economics, 2013.

[109] D. Atkins, K. Fink, and J. Slutsky, "Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality.," *Ann. Intern. Med.*, vol. 142, no. 12 Pt 2, pp. 1035–41, 2005.

[110] S. Brennan, *The NHS IT Project: The Biggest Computer Programme in the World-Ever!* Radcliffe Publishing, 2005.

[111] C. Chantler, T. Clarke, and R. Granger, "Information technology in the English National Health Service.," *JAMA*, vol. 296, no. 18, pp. 2255–8, 2006.

[112] W. R. Hersh, "Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance," *Am J Manag Care*, vol. 81, pp. 126–28, 2007.

[113] S. Svenningsen, "Electronic patient records and medical practice," *Reorganization Roles, Responsib. Risks. Copenhagen Copenhagen Bus. Sch. Thesis*, 2002.

[114] D. Cooksey, *A review of UK health research funding*. The Stationery Office, 2006.

[115] J. Parkinson, S. Davis, and T. van Staa, "The general practice research database: now and the future," *Pharmacovigilance, Second Ed.*, pp. 341–348, 2007.

[116] L. Wood and R. Coulson, "Revitalizing the General Practice Research Database: plans, challenges, and opportunities," *Pharmacoepidemiol. Drug Saf.*, vol. 10, no. 5, pp. 379–383, 2001.

[117] A. A. Kousoulis, I. Rafi, and S. de Lusignan, “The CPRD and the RCGP: building on research success by enhancing benefits for patients and practices,” *Br. J. Gen. Pract.*, vol. 65, no. 631, pp. 54–55, Jan. 2015.

[118] T. Williams, T. van Staa, S. Puri, and S. Eaton, “Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource,” *Ther. Adv. Drug Saf.*, vol. 3, no. 2, pp. 89–99, 2012.

[119] E. Herrett, A. M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. van Staa, and L. Smeeth, “Data Resource Profile: Clinical Practice Research Datalink (CPRD),” *Int. J. Epidemiol.*, vol. 44, no. 3, pp. 827–836, 2015.

[120] Y.-C. Chen, J.-C. Wu, I. Haschler, A. Majeed, T.-J. Chen, and T. Wetter, “Academic Impact of a Public Electronic Health Database: Bibliometric Analysis of Studies Using the General Practice Research Database,” *PLoS One*, vol. 6, no. 6, p. e21404, Jun. 2011.

[121] L. Gould, A. Walker, P. Ryan, S. Schneeweiss, N. Santanello, and S. Sacks, “Using Databases for Both Hypothesis Generating and Hypothesis Confirmation: One Database or Two?,” 2009, vol. 18, pp. S205–S206.

[122] A. Majeed, “Sources, uses, strengths and limitations of data collected in primary care in England.,” *Health Stat. Q.*, no. 21, pp. 5–14, 2004.

[123] E. Herrett, S. L. Thomas, W. M. Schoonen, L. Smeeth, and A. J. Hall, “Validation and validity of diagnoses in the General Practice Research Database: a systematic review,” *Br. J. Clin. Pharmacol.*, vol. 69, no. 1, pp. 4–14, Jan. 2010.

[124] S. Yu, K. P. Liao, S. Y. Shaw, V. S. Gainer, S. E. Churchill, P. Szolovits, S. N. Murphy, I. S. Kohane, and T. Cai, “Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources,” *J. Am. Med. Informatics Assoc.*, vol. 22, no. 617, pp. 1–10, Sep. 2015.

[125] F. Kurreeman, K. Liao, L. Chibnik, B. Hickey, E. Stahl, V. Gainer, G. Li, L. Bry, S. Mahan, K. Ardlie, B. Thomson, P. Szolovits, S. Churchill, S. N. Murphy, T. Cai, S. Raychaudhuri, I. Kohane, E. Karlson, and R. M. Plenge, “Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records.,” *Am. J. Hum. Genet.*, vol. 88, no. 1, pp. 57–69, 2011.

[126] Clinical Practice Research Datalink, “CPRD Bibliography.” .

[127] L. Smeeth, C. Cook, P. E. Fombonne, L. Heavey, P. L. C. Rodrigues, P. P. G. Smith, and P. A. J. Hall, “MMR vaccination and pervasive developmental disorders: A case-control study,” *Lancet*, vol. 364, no. 9438, pp. 963–969, 2004.

[128] L. Smeeth, S. L. Thomas, A. J. Hall, R. Hubbard, P. Farrington, and P. Vallance, “Risk of myocardial infarction and stroke after acute infection or vaccination.,” *N. Engl. J. Med.*,

vol. 351, no. 25, pp. 2611–2618, 2004.

[129] H. Jick, G. L. Zornberg, S. S. Jick, S. Seshadri, and D. a Drachman, “Statins and the risk of dementia.,” *Lancet*, vol. 356, no. 9242, pp. 1627–1631, 2000.

[130] J. M. Gelfand, A. L. Neimann, D. B. Shin, X. Wang, D. J. Margolis, and A. B. Troxel, “Risk of myocardial infarction in patients with psoriasis.,” *JAMA*, vol. 296, no. 14, pp. 1735–1741, 2006.

[131] T. P. Van Staa, H. G. M. Leufkens, L. Abenhaim, B. Zhang, and C. Cooper, “Use of oral corticosteroids and risk of fractures. June, 2000.,” *J. Bone Miner. Res.*, vol. 20, no. 8, p. 1487–1494; discussion 1486, 2005.

[132] K. Bhaskaran, I. Douglas, H. Forbes, I. dos-Santos-Silva, D. a. Leon, and L. Smeeth, “Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults,” *Lancet*, vol. 384, no. 9945, pp. 755–765, 2014.

[133] D. G. Kleinbaum, L. Kupper L., and H. Morgenstern, *Epidemiologic research - principles and quantitative methods*. John Wiley & Sons, 1982.

[134] S. Muller, S. L. Hider, J. Belcher, T. Helliwell, and C. D. Mallen, “Is cancer associated with polymyalgia rheumatica? A cohort study in the General Practice Research Database.,” *Ann. Rheum. Dis.*, pp. 1–5, 2013.

[135] N. Qizilbash, J. Gregson, M. E. Johnson, N. Pearce, I. Douglas, K. Wing, S. J. W. Evans, and S. J. Pocock, “BMI and risk of dementia in two million people over two decades: a retrospective cohort study,” *Lancet Diabetes Endocrinol.*, vol. 8587, no. 15, pp. 1–21, 2015.

[136] R. Bellazzi and B. Zupan, “Predictive data mining in clinical medicine: Current issues and guidelines,” *Int. J. Med. Inform.*, vol. 77, no. 2, pp. 81–97, 2008.

[137] R. Bellazzi, F. Ferrazzi, and L. Sacchi, “Predictive data mining in clinical medicine: a focus on selected methods and applications,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 5, pp. 416–430, Sep. 2011.

[138] N. Lavrač, “Selected techniques for data mining in medicine,” *Artif. Intell. Med.*, vol. 16, no. 1, pp. 3–23, 1999.

[139] K. J. Cios and G. William Moore, “Uniqueness of medical data mining,” *Artif. Intell. Med.*, vol. 26, pp. 1–24, 2002.

[140] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, “Secondary Use of EHR: Data Quality Issues and Informatics Opportunities.,” *AMIA Summits Transl. Sci. Proc.*, vol. 2010, pp. 1–5, 2010.

[141] J. Wu, J. Roy, and W. F. Stewart, “Prediction Modeling Using EHR Data - Challenges,

Strategies, and a Comparison of Machine Learning Approaches," *Med. Care*, vol. 48, no. 6 Suppl 1, pp. S106–S113, 2010.

[142] S. Tsumoto, "Problems with mining medical data," in *Computer Software and Applications Conference, 2000. COMPSAC 2000. The 24th Annual International*, 2000, pp. 467–468.

[143] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Semantic similarity measures as tools for exploring the gene ontology.," *Pac. Symp. Biocomput.*, vol. 8, pp. 601–612, 2003.

[144] D. Sánchez and M. Batet, "Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective," *J. Biomed. Inform.*, vol. 44, pp. 749–759, 2011.

[145] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies.," *PLoS Comput. Biol.*, vol. 5, p. e1000443, 2009.

[146] H. W. Jones, "JOHN GRAUNT AND HIS BILLS OF MORTALITY," *Bull. Med. Libr. Assoc.*, vol. 33, no. 1, pp. 3–4, Jan. 1945.

[147] H. Cao, M. Markatou, G. B. Melton, M. F. Chiang, and G. Hripcsak, "Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics.," 2005.

[148] A. B. Holmes, A. Hawson, F. Liu, C. Friedman, H. Khiabanian, and R. Rabadan, "Discovering disease associations by integrating electronic clinical data and medical literature," *PLoS One*, vol. 6, no. 6, p. e21132, 2011.

[149] A. M. Shin, I. H. Lee, G. H. Lee, H. J. Park, H. S. Park, K. Il Yoon, J. J. Lee, and Y. N. Kim, "Diagnostic analysis of patients with essential hypertension using association rule mining," *Healthc. Inform. Res.*, vol. 16, no. 2, pp. 77–81, 2010.

[150] L. Penberthy, R. Brown, F. Puma, and B. Dahman, "Automated matching software for clinical trials eligibility: measuring efficiency and flexibility," *Contemp. Clin. Trials*, vol. 31, no. 3, pp. 207–217, 2010.

[151] M. Klompas, G. Haney, D. Church, R. Lazarus, X. Hou, and R. Platt, "Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance," *PLoS One*, vol. 3, no. 7, p. e2626, 2008.

[152] M. Schmiedeskamp, S. Harpe, R. Polk, M. Oinonen, and A. Pakyz, "Use of International Classification of Diseases, Ninth Revision Clinical Modification Codes and Medication Use Data to Identify Nosocomial Clostridium difficile Infection," *Infect. Control Hosp. Epidemiol.*, vol. 30, no. 11, pp. 1070–1076, 2009.

[153] A. Wright, J. Pang, J. C. Feblowitz, F. L. Maloney, A. R. Wilcox, H. Z. Ramelson, L. I.

Schneider, and D. W. Bates, “A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record,” *J. Am. Med. Informatics Assoc.*, vol. 18, no. 6, pp. 859–867, 2011.

[154] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun, “Limestone: High-throughput candidate phenotype generation via tensor factorization,” *J. Biomed. Inform.*, vol. 52, pp. 199–211, 2014.

[155] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai, “A review of approaches to identifying patient phenotype cohorts using electronic health records,” *J. Am. Med. Informatics Assoc.*, vol. 21, no. 2, pp. 221–230, 2014.

[156] S. M. Meystre, V. G. Deshmukh, and J. Mitchell, “A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations,” 2009, vol. 2009, p. 442.

[157] K. P. Liao, T. Cai, V. Gainer, S. Goryachev, Q. Zeng- treitler, S. Raychaudhuri, P. Szolovits, S. Churchill, S. Murphy, and I. Kohane, “Electronic medical records for discovery research in rheumatoid arthritis,” *Arthritis Care Res. (Hoboken).*, vol. 62, no. 8, pp. 1120–1127, 2010.

[158] M. A. Al-Haddad, J. A. Waters, J. J. R. Aguilar-Saavedra, J. Kesterson, and M. M. Schmidt, “Comparing methods for identifying pancreatic cancer patients using electronic data sources,” 2010.

[159] X. Wang, F. Wang, J. Wang, B. Qian, and J. Hu, “Exploring patient risk groups with incomplete knowledge,” 2013, pp. 1223–1228.

[160] D. Gotz, F. Wang, and A. Perer, “A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data,” *J. Biomed. Inform.*, vol. 48, pp. 148–159, 2014.

[161] F. Wang, C. Liu, Y. Wang, J. Hu, and G. Yu, “A Graph Based Methodology for Temporal Signature Identification from HER,” 2015, vol. 2015, p. 1269.

[162] L. Li, W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, and J. T. Dudley, “Identification of type 2 diabetes subgroups through topological analysis of patient similarity,” *Sci. Transl. Med.*, vol. 7, no. 311, p. 311ra174-311ra174, 2015.

[163] D. A. Hanauer and N. Ramakrishnan, “Modeling temporal relationships in large scale clinical associations,” *J. Am. Med. Informatics Assoc.*, vol. 20, no. 2, pp. 332–341, 2013.

[164] R. Pivovarov, D. J. Albers, G. Hripcsak, J. L. Sepulveda, and N. Elhadad, “Temporal trends of hemoglobin A1c testing,” *J. Am. Med. Informatics Assoc.*, vol. 21, no. 6, pp. 1038–1044, 2014.

[165] J. J. Jiang and D. W. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy,” *arXiv Prepr. C.*, 1997.

[166] M. R. Anderberg, “Cluster analysis for applications. 1973,” *Acad. New York*.

[167] F. Fratev, O. Polansky, A. Mehlhorn, and V. Monev, “Application of distance and similarity measures. The comparison of molecular electronic structures in arbitrary electronic states,” *J. Mol. Struct.*, vol. 56, pp. 245–253, 1979.

[168] C. Seung-Seok, C. Sung-Hyuk, and C. C. Tappert, “A survey of binary similarity and distance measures.,” *J. Syst. Cybern. Informatics*, vol. 8, no. 1, pp. 43–48, 2010.

[169] J. J. Goeman and U. Mansmann, “Multiple testing on the directed acyclic graph of gene ontology,” *Bioinformatics*, vol. 24, no. 4, pp. 537–544, 2008.

[170] J. Bard, S. Y. Rhee, and M. Ashburner, “An ontology for cell types.,” *Genome Biol.*, vol. 6, no. 2, p. R21, 2005.

[171] B. Zhang, D. Schmoyer, S. Kirov, and J. Snoddy, “GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.,” *BMC Bioinformatics*, vol. 5, no. 1, p. 16, 2004.

[172] C. Fellbaum, *WordNet: An Electronic Lexical Database*, vol. 71. 1998.

[173] Z. Zhou, Y. Wang, and J. Gu, “A new model of information content for semantic similarity in WordNet,” in *Future Generation Communication and Networking Symposia, 2008. FGCNS’08. Second International Conference on*, 2008, vol. 3, pp. 85–89.

[174] G. O. Consortium, “The Gene Ontology (GO) project in 2006.,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D322-6, 2006.

[175] S. Patwardhan, S. Banerjee, and T. Pedersen, “Using Measures of Semantic Relatedness for Word Sense Disambiguation,” in *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 2003, vol. 4, pp. 241–257.

[176] G. Leroy and T. C. Rindflesch, “Effects of information and machine learning algorithms on word sense disambiguation with small datasets,” *Int. J. Med. Inform.*, vol. 74, pp. 573–585, 2005.

[177] S. J. Green, “Building hypertext links by computing semantic similarity,” *IEEE Trans. Knowl. Data Eng.*, vol. 11, pp. 713–730, 1999.

[178] A. Budanitsky and G. Hirst, “Semantic distance in WordNet : An experimental , application-oriented evaluation of five measures,” *Evaluation*, vol. 2, pp. 29–34, 1998.

[179] R. L. Cilibrasi and P. M. B. Vitányi, “The Google similarity distance,” *IEEE Trans. Knowl. Data Eng.*, vol. 19, pp. 370–383, 2007.

[180] M. Stevenson and M. a. Greenwood, “A semantic approach to IE pattern induction,” *Proc. 43rd Annu. Meet. Assoc. Comput. Linguist. - ACL ’05*, pp. 379–386, 2005.

[181] D. Lin, “An Information-Theoretic Definition of Similarity,” in *Proceedings of ICML*, 1998, pp. 296–304.

[182] S. Aseervatham and Y. Bennani, “Semi-structured document categorization with a semantic kernel,” *Pattern Recognit.*, vol. 42, pp. 2067–2076, 2009.

[183] P. Resnik, “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language,” *J. Artif. Intell. Res.*, vol. 11, pp. 95–130, 1999.

[184] R. Rada, H. Mili, E. Bicknell, and M. Blettner, “Development and application of a metric on semantic nets,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19. pp. 17–30, 1989.

[185] V. Pekar and S. Staab, “Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision,” *Proceedings of the 19th international conference on Computational linguistics - Volume 1*. pp. 1–7, 2002.

[186] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994, pp. 133–138.

[187] C. Leacock and M. Chodorow, “Combining Local Context and WordNet Similarity for Word Sense Identification,” in *WordNet: An electronic lexical database.*, 1998, pp. 265–283.

[188] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, “Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation,” *Bioinformatics*, vol. 19, pp. 1275–1283, 2003.

[189] J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. a. Martínez-Cruz, F. J. Corrales, A. Rubio, L. A. Martinez-Cruz, F. J. Corrales, A. Rubio, L. a. Martínez-Cruz, F. J. Corrales, and A. Rubio, “Correlation between gene expression and GO semantic similarity,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 2, no. 4, pp. 330–337, 2005.

[190] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. F. Chen, “A new method to measure the semantic similarity of GO terms.,” *Bioinformatics*, vol. 23, no. 10, pp. 1274–81, 2007.

[191] X. Guo, R. Liu, C. D. Shriver, H. Hu, and M. N. Liebman, “Assessing semantic similarity measures for the characterization of human regulatory pathways.,” *Bioinformatics*, vol. 22, pp. 967–973, 2006.

[192] F. M. Couto, M. J. Silva, and P. M. Coutinho, “Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors,” in *Proceedings of the*

14th ACM, 2005, pp. 343–344.

[193] F. Azuaje, H. Wang, and O. Bodenreider, “Ontology-driven similarity approaches to supporting gene functional assessment,” in *Proceedings of the ISMB’2005 SIG meeting on Bio-ontologies*, 2005, pp. 9–10.

[194] Y. Tao, L. Sam, J. Li, C. Friedman, and Y. A. Lussier, “Information theory applied to the sparse gene ontology annotation network to predict novel gene function,” in *Bioinformatics*, 2007, vol. 23.

[195] R. M. Riensche, B. L. Baddeley, A. P. Sanfilippo, C. Posse, and B. Gopalan, “XOA: Web-enabled cross-ontological analytics,” in *Proceedings - 2007 IEEE Congress on Services, SERVICES 2007*, 2007, pp. 99–105.

[196] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, “Dimension Reduction Techniques,” *A Distrib. Theory Nonparametric Regres.*, pp. 448–458, 2002.

[197] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2. pp. 433–459, 2010.

[198] I. Jolliffe, *Principal component analysis*, vol. 98. 2005.

[199] J. Shlens, “A Tutorial on Principal Component Analysis,” *Measurement*, vol. 51, p. 52, 2005.

[200] F. Wood, “Principal component analysis,” 2009.

[201] R. L. Somorjai, B. Dolenko, and M. Mandelzweig, “Direct classification of high-dimensional data in low-dimensional projected feature spaces—Comparison of several classification methodologies,” *J. Biomed. Inform.*, vol. 40, no. 2, pp. 131–138, Apr. 2007.

[202] K. Ahmad and B. Vrusias, “Learning to visualise high-dimensional data,” in *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, 2004, pp. 507–512.

[203] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” *Appl. Stat.*, pp. 100–108, 1979.

[204] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, *The analysis of a simple k-means clustering algorithm*. .

[205] P. S. Bradley, U. Fayyad, and C. Reina, “Scaling EM (expectation-maximization) clustering to large databases,” 1998.

[206] C. Wu, J. R. Steinbauer, and G. M. Kuo, “EM clustering analysis of diabetes patients basic diagnosis index,” p. 1158.

[207] David Statham, “Read Code Processing A Software Tool for Data Mining General

Practice Datasets," Manchester, UK, 2009.

[208] A. Caprara, P. Toth, and M. Fischetti, "Algorithms for the Set Covering Problem," *Ann. Oper. Res.*, vol. 98, no. 1, pp. 353–371, 2000.

[209] V. Chvatal, "A Greedy Heuristic for the Set-Covering Problem*," *Math. Oper. Res.*, vol. 4, no. 3, pp. 1973–1976, 1979.

[210] P. Cimiano, A. Hotho, and S. Staab, "Learning concept hierarchies from text corpora using formal concept analysis," vol. 24, no. 1, pp. 305–339.

[211] C. Muller, I. Gurevych, and M. Muhlhauser, *Integrating Semantic Knowledge into Text Similarity and Information Retrieval*. {IEEE} Computer Society.

[212] E. Pekalska, "Dissimilarity representations in pattern recognition: concepts, theory, and applications," in *Advanced School for Computing and Imaging, Delft University of Technology*, 2005, p. 344.

[213] I. Ntoutsi, "The Notion of Similarity in Data and Pattern Spaces.," 2004.

[214] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *J. Artif. Intell. Res.*, vol. 11, pp. 95–130, 1999.

[215] J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," p. 9008.

[216] E. Pekalska and R. P. W. Duin, "Automatic pattern recognition by similarity representations," *Electron. Lett.*, vol. 37, no. 3, pp. 159–160, 2001.

[217] R. P. W. Duin, "Relational discriminant analysis and its large sample size problem," in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, vol. 1, pp. 445--449 vol.1.

[218] B. JA, C. MA, and L. P, "Diabetes and atherosclerosis: Epidemiology, pathophysiology, and management," *JAMA*, vol. 287, no. 19, pp. 2570–2581, 2002.

[219] A. Jacobs, "The pathologies of big data," *Commun. ACM*, vol. 52, no. 8, pp. 36–44, 2009.

[220] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering," 2013, pp. 1434–1453.

[221] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science (80-.).*, vol. 332, no. 6025, pp. 60–65, 2011.

[222] S. Sagiroglu and D. Sinanc, "Big data: A review," *Int. Conf. Collab. Technol. Syst.*, pp. 42–47, May 2013.

[223] J. S. Ward and A. Barker, “Undefined by data: a survey of big data definitions,” *arXiv Prepr. arXiv1309.5821*, 2013.

[224] J. Fan, F. Han, and H. Liu, “Challenges of Big Data analysis,” *Natl. Sci. Rev.*, pp. 1–38, 2014.

[225] F. Mazzocchi, “Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science.,” *EMBO Rep.*, vol. 16, no. 10, pp. 1250–1255, 2015.

[226] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” *Nat. Rev. Genet.*, vol. 13, no. 6, pp. 395–405, 2012.

[227] I. N. Sarkar, “Biomedical informatics and translational medicine.,” *J. Transl. Med.*, vol. 8, p. 22, 2010.

[228] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, “Big Data Clustering: A Review,” pp. 707–720, 2014.

[229] D. Howe, “Big data: the future of biocuration,” *Nature*, vol. 455, no. 7209, pp. 47–50, 2008.

[230] J. Fan, F. Han, and H. Liu, “Challenges of Big Data analysis,” *Natl. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, Jun. 2014.

[231] J. Fan and J. Lv, “Sure independence screening for ultrahigh dimensional feature space,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 70, no. 5, pp. 849–911, 2008.

[232] P. Taylor, “Personal genomes: When consent gets in the way.,” *Nature*, vol. 456, no. 7218, pp. 32–33, 2008.

[233] H. Chen, R. H. L. Chiang, and V. C. Storey, “Business Intelligence and Analytics: From Big Data To Big Impact,” *Mis Q.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[234] O. Kwon, N. Lee, and B. Shin, “Data quality management, data usage experience and acquisition intention of big data analytics,” *Int. J. Inf. Manage.*, vol. 34, no. 3, pp. 387–394, 2014.

[235] K. Feldman and N. V Chawla, “Scaling personalized healthcare with big data,” 2014.

[236] L. J. Frey, L. Lenert, and G. Lopez-Campos, “EHR Big Data Deep Phenotyping. Contribution of the IMIA Genomic Medicine Working Group.,” *Yearb. Med. Inform.*, vol. 9, no. 1, pp. 206–211, 2014.

[237] N. Alldrin, A. Smith, and D. Turnbull, “Clustering with EM and K-means,” *Univ. San Diego, California, Tech Rep.*, pp. 261–95, 2003.

[238] M. Khalilian and N. Mustapha, “Data stream clustering: Challenges and issues,” *arXiv Prepr. arXiv1006.5261*, 2010.

[239] A. McCallum, K. Nigam, and L. H. Ungar, “Efficient clustering of high-dimensional data sets with application to reference matching,” 2000, pp. 169–178.

[240] R. T. Ng and J. Han, “Clarans: A method for clustering objects for spatial data mining,” *Knowl. Data Eng. IEEE Trans.*, vol. 14, no. 5, pp. 1003–1016, 2002.

[241] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: an efficient data clustering method for very large databases,” in *ACM SIGMOD Record*, 1996, vol. 25, no. 2, pp. 103–114.

[242] M. Ester, H. H. P. Kriegel, J. J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in *Second International Conference on Knowledge Discovery and Data Mining*, 1996, vol. 2, pp. 226–231.

[243] Y. He, H. Tan, W. Luo, S. Feng, and J. Fan, “MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data,” *Front. Comput. Sci.*, vol. 8, no. 1, pp. 83–99, 2014.

[244] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[245] M. Singh and B. Leonhardi, “Introduction to the ibm netezza warehouse appliance,” 2011, pp. 385–386.

[246] T. E. Oliphant, “Python for scientific computing,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 10–20, 2007.

[247] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[248] R. A. Ince, R. S. Petersen, D. C. Swan, and S. Panzeri, “Python for information theoretic analysis of neural data,” *Front. Neuroinform.*, vol. 3, 2009.

[249] H. P. Langtangen, *Python scripting for computational science*, vol. 3. Springer, 2006.

[250] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, and B. Wilczynski, “Biopython: freely available Python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[251] B. Chapman and J. Chang, “Biopython: Python tools for computational biology,” *ACM SIGBIO Newsl.*, vol. 20, no. 2, pp. 15–19, 2000.

[252] J. M. Bjørndalen, B. Vinter, and O. J. Anshus, “PyCSP-Communicating Sequential Processes for Python.,” 2007, pp. 229–248.

[253] M. J. De Hoon, B. Chapman, and I. Friedberg, “Bioinformatics and computational biology with Biopython,” *Genome Informatics*, vol. 14, pp. 298–299, 2003.

[254] F. S. Bao, X. Liu, and C. Zhang, “PyEEG: an open source python module for EEG/MEG feature extraction,” *Comput. Intell. Neurosci.*, vol. 2011, 2011.

[255] N. Rey-Villamizar, V. Somasundar, M. Megjhani, Y. Xu, Y. Lu, R. Padmanabhan, K. Trett, W. Shain, and B. Roysam, “Large-scale automated image analysis for computational profiling of brain tissue surrounding implanted neuroprosthetic devices using Python,” *Front. Neuroinform.*, vol. 8, 2014.

[256] S. L. West, C. Blake, Zhiwen Liu, J. N. McKoy, M. D. Oertel, and T. S. Carey, “Reflections on the use of electronic health record data for clinical research,” *Health Informatics J.*, vol. 15, no. 2, pp. 108–121, Jun. 2009.

[257] C. Weng, P. Appelbaum, G. Hripcsak, I. Kronish, L. Busacca, K. W. Davidson, and J. T. Bigger, “Using EHRs to integrate research with patient care: promises and challenges,” *J. Am. Med. Informatics Assoc.*, vol. 19, no. 5, pp. 684–687, 2012.

[258] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, L. J. Jensen, and S. Brunak, “Using electronic patient records to discover disease correlations and stratify patient cohorts.,” *PLoS Comput. Biol.*, vol. 7, no. 8, p. e1002141, 2011.

[259] H. Singh, K. Hirani, H. Kadiyala, O. Rudomiotov, T. Davis, M. M. Khan, and T. L. Wahls, “Characteristics and predictors of missed opportunities in lung cancer diagnosis: an electronic health record-based study.,” *J. Clin. Oncol.*, vol. 28, no. 20, pp. 3307–15, Jul. 2010.

[260] A. L. Masica, E. Ewen, Y. A. Daoud, D. Cheng, N. Franceschini, R. E. Kudyakov, J. R. Bowen, E. S. Brouwer, D. Wallace, N. S. Fleming, and S. L. West, “Comparative effectiveness research using electronic health records: impacts of oral antidiabetic drugs on the development of chronic kidney disease,” *Pharmacoepidemiol. Drug Saf.*, vol. 22, no. 4, pp. 413–422, Apr. 2013.

[261] S. B. Stakic and S. Tasic, “Secondary use of EHR data for correlated comorbidity prevalence estimate,” in *32nd Annual International Conference of the IEEE EMBS Buenos Aires, Argentina,* 2010, pp. 3907–3910.

[262] F. G. Real, C. Svanes, E. R. Omenaas, J. M. Antò, E. Plana, D. Jarvis, C. Janson, F. Neukirch, E. Zemp, J. Dratva, M. Wjst, K. Svanes, B. Leynaert, and J. Sunyer, “Lung function, respiratory symptoms, and the menopausal transition,” *J. Allergy Clin. Immunol.*, vol. 121, no. 1, p. 72–80.e3.

[263] R. Lima, M. Wofford, and J. F. Reckelhoff, “Hypertension in Postmenopausal Women,” *Curr. Hypertens. Rep.*, vol. 14, no. 3, pp. 254–260, Jun. 2012.

[264] S. L. Jackson, E. J. Boyko, D. Scholes, L. Abraham, K. Gupta, and S. D. Fihn, “Predictors of urinary tract infection after menopause: A prospective study,” *Am. J. Med.*, vol. 117, no. 12, pp. 903–911, Dec. 2004.

[265] R. W. Sattin, “Falls among older persons: a public health perspective.,” *Annu. Rev. Public Health*, vol. 13, pp. 489–508, 1992.

[266] A. Ungar, M. Rafanelli, I. Iacomelli, M. A. Brunetti, A. Ceccofiglio, F. Tesi, and N. Marchionni, “Fall prevention in the elderly,” *Clinical Cases in Mineral and Bone Metabolism*, vol. 10. pp. 91–95, 2013.

[267] E. K. Stanmore, J. Oldham, D. A. Skelton, T. O’Neill, M. Pilling, A. J. Campbell, and C. Todd, “Fall incidence and outcomes of falls in a prospective study of adults with rheumatoid arthritis,” *Arthritis Care Res.*, vol. 65, pp. 737–744, 2013.

[268] L. Zuyev, A. N. Benoit, F. Y. Chang, and P. C. Dykes, “Tailored prevention of inpatient falls: development and usability testing of the fall TIPS toolkit.,” *Comput. Inform. Nurs.*, vol. 29, p. TC21-C28, 2011.

[269] S. M. Friedman, B. Munoz, S. K. West, G. S. Rubin, and L. P. Fried, “Falls and fear of falling: Which comes first? A longitudinal prediction model suggests strategies for primary and secondary prevention,” *J. Am. Geriatr. Soc.*, vol. 50, pp. 1329–1335, 2002.

[270] P. C. Dykes, D. L. Carroll, A. Hurley, S. Lipsitz, A. Benoit, F. Chang, S. Meltzer, R. Tsurikova, L. Zuyov, and B. Middleton, “Fall prevention in acute care hospitals: a randomized trial.,” *JAMA*, vol. 304, pp. 1912–1918, 2010.

[271] C. Stern and R. Jayasekara, “Interventions to reduce the incidence of falls in older adult patients in acute-care hospitals: a systematic review.,” *Int. J. Evid. Based. Healthc.*, vol. 7, pp. 243–249, 2009.

[272] D. Melzer, B. Tavakoly, R. E. Winder, J. A. H. Masoli, W. E. Henley, A. Ble, and S. H. Richards, “Much more medicine for the oldest old: trends in UK electronic clinical records.,” *Age Ageing*, vol. 44, no. 1, pp. 46–53, 2015.

[273] “WHO | Falls.” [Online]. Available: http://www.who.int/mediacentre/factsheets/fs344/en/. [Accessed: 23-Nov-2015].

[274] J. Craig, A. Murray, S. Mitchell, S. Clark, L. Saunders, and L. Burleigh, “The high cost to health and social care of managing falls in older adults living in the community in Scotland,” *Scott. Med. J.*, vol. 58, pp. 198–203, 2013.

[275] J. Gribbin, R. Hubbard, C. Smith, J. Gladman, and S. Lewis, “Incidence and mortality of falls amongst older people in primary care in the United Kingdom.,” *QJM*, vol. 102, pp. 477–483, 2009.

[276] A. M. Pfeil, P. Imfeld, R. Pettengell, S. S. Jick, T. D. Szucs, C. R. Meier, and M.

Schwenkglenks, “Trends in incidence and medical resource utilisation in patients with chronic lymphocytic leukaemia: insights from the UK Clinical Practice Research Datalink (CPRD),” *Ann. Hematol.*, pp. 1–9, 2014.

[277] C. J. Currie, C. D. Poole, M. Evans, J. R. Peters, and L. M. Christopher, “Mortality and other important diabetes-related outcomes with insulin vs other antihyperglycemic therapies in type 2 diabetes,” *J. Clin. Endocrinol. Metab.*, vol. 98, pp. 668–677, 2013.

[278] A. P. Arnold, “Promoting the understanding of sex differences to enhance equity and excellence in biomedical science.,” *Biol. Sex Differ.*, vol. 1, no. 1, p. 1, 2010.

[279] C. a Hidalgo, N. Blumm, A.-L. Barabási, and N. a Christakis, “A dynamic network approach for the study of human phenotypes.,” *PLoS Comput. Biol.*, vol. 5, no. 4, p. e1000353, Apr. 2009.

[280] J. Zhang and Y. KF, “What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes,” *JAMA*, vol. 280, no. 19, pp. 1690–1691, Nov. 1998.

[281] D. G. Altman, *Practical statistics for medical research*. CRC press, 1990.

[282] G. Hripcsak and D. J. Albers, “Next-generation phenotyping of electronic health records,” *J. Am. Med. Informatics Assoc.*, vol. 20, no. 1, pp. 117–121, 2013.

[283] Y. Chen, R. J. Carroll, E. R. M. Hinz, A. Shah, A. E. Eyler, J. C. Denny, and H. Xu, “Applying active learning to high-throughput phenotyping algorithms for electronic health records data,” *J. Am. Med. Informatics Assoc.*, vol. 20, no. e2, pp. e253–e259, 2013.

[284] C. I. Gryfe, a Amies, and M. J. Ashley, “A longitudinal study of falls in an elderly population: I. Incidence and morbidity.,” *Age Ageing*, vol. 6, no. 4, pp. 201–10, 1977.

[285] C. M. Patino, R. McKean-Cowdin, S. P. Azen, J. C. Allison, F. Choudhury, and R. Varma, “Central and peripheral visual impairment and the risk of falls and falls with injury.,” *Ophthalmology*, vol. 117, no. 2, p. 199–206.e1, 2010.

[286] M. E. Tinetti, D. I. Baker, G. McAvay, E. B. Claus, P. Garrett, M. Gottschalk, M. L. Koch, K. Trainor, and R. I. Horwitz, “A multifactorial intervention to reduce the risk of falling among elderly people living in the community.,” 1994.

[287] J. H. Downton and K. Andrews, “Prevalence, characteristics and factors associated with falls among the elderly living at home,” *Aging Clin. Exp. Res.*, vol. 3, no. 3, pp. 219–228, 1991.

[288] S. Eriksson, S. Strandberg, Y. Gustafson, and L. Lundin-Olsson, “Circumstances surrounding falls in patients with dementia in a psychogeriatric ward,” *Arch. Gerontol. Geriatr.*, vol. 49, pp. 80–87, 2009.

[289] V. M., V. R., S. J.C., B. R., and A. S., “The relationship of falls to injury among hospital in-patients,” *International Journal of Clinical Practice*, vol. 59, no. 1. pp. 17–20, 2005.

[290] L. Clemson, R. G. Cumming, and M. Roland, “Case-control study of hazards in the home and risk of falls and hip fractures,” *Age Ageing*, vol. 25, pp. 97–101, 1996.

[291] J. B. Lauritzen, M. M. Petersen, and B. Lund, “Effect of external hip protectors on hip fractures.,” *Lancet*, vol. 341, no. 8836, pp. 11–13, 1993.

[292] M. C. Kim, H. Cho, S. Sunwoo, S. W. Kim, and H. J. Cho, “Prevalence and associated factors of fall among the elderly in nursing home,” *J. Korean Geriatr. Soc.*, vol. 3, no. 4, pp. 29–38, 1999.

[293] R. W. Bolt and P. G. Watts, “The relationship between aetiology and distribution of facial lacerations,” *Inj. Extra*, vol. 35, no. 1, pp. 6–11, 2004.

[294] M. C. Nevitt, S. R. Cummings, and E. S. Hudes, “Risk factors for injurious falls: a prospective study,” *J. Gerontol.*, vol. 46, no. 5, pp. M164-70, 1991.

[295] N. J. M. D. Hartshorne, R. C. M. D. Harruff, and E. C. J. Alvord, “Fatal Head Injuries in Ground-Level Falls.,” *American Journal of Forensic Medicine & Pathology*, vol. 18, no. 3. pp. 258–264, 1997.

[296] P. Kannus, H. Sievänen, M. Palvanen, T. Järvinen, and J. Parkkari, “Prevention of falls and consequent injuries in elderly people,” *Lancet*, vol. 366, no. 9500, pp. 1885–1893, 2005.

[297] K. Kallin, J. Jensen, L. L. Olsson, L. Nyberg, and Y. Gustafson, “Why the elderly fall in residential care facilities, and suggested remedies.,” *The Journal of family practice*, vol. 53. pp. 41–52, 2004.

[298] V. F. Trewin, C. J. Lawrence, and G. B. Veitch, “An investigation of the association of benzodiazepines and other hypnotics with the incidence of falls in the elderly,” *J Clin Pharm Ther*, vol. 17, no. 2, pp. 129–133, 1992.

[299] M. Pigłowska, J. Kostka, and T. Kostka, “Association between respiratory tract infections and incidence of falls in nursing home residents,” *Pol. Arch. Med. Wewn.*, vol. 123, pp. 371–377, 2013.

[300] R. A. Fox, “Atypical presentation of geriatric infections.,” *Geriatrics*, vol. 43, no. 5, pp. 58–59, 1988.

[301] A. Forster and J. Young, “Incidence and consequence of falls due to stroke: a systematic inquiry,” *Bmj Clin. Res. Ed.*, vol. 311, pp. 83–86, 1995.

[302] D. Hyndman, A. Ashburn, and E. Stack, “Fall events among people with stroke living in the community: Circumstances of falls and characteristics of fallers,” *Arch. Phys. Med.*

Rehabil., vol. 83, pp. 165–170, 2002.

[303] K. L. Hollands, D. Agnihotri, and S. F. Tyson, “Effects of dual task on turning ability in stroke survivors and older adults,” *Gait Posture*, vol. 40, no. 4, pp. 564–569, 2014.

[304] E. L. Inness, A. Mansfield, B. Lakhani, M. Bayley, and W. E. McIlroy, “Impaired reactive stepping among patients ready for discharge from inpatient stroke rehabilitation,” *Phys. Ther.*, vol. 94, no. 12, pp. 1755–1764, 2014.

[305] L. G. Jacobs, H. H. Billett, K. Freeman, C. Dinglas, and L. Jumaquio, “Anticoagulation for stroke prevention in elderly patients with atrial fibrillation, including those with falls and/or early-stage dementia: a single-center, retrospective, observational study,” *Am. J. Geriatr. Pharmacother.*, vol. 7, no. 3, pp. 159–166, 2009.

[306] T. S. H. Jørgensen, A. H. Hansen, M. Sahlberg, G. H. Gislason, C. Torp-Pedersen, C. Andersson, and E. Holm, “Falls and comorbidity: The pathway to fractures.,” *Scand. J. Public Health*, p. 1403494813516831-, 2014.

[307] Y. Gerber, L. J. Melton, S. A. Weston, and V. L. Roger, “Osteoporotic fractures and heart failure in the community,” *Am. J. Med.*, vol. 124, pp. 418–425, 2011.

[308] S. G. Leveille, J. Bean, K. Bandeen-Roche, R. Jones, M. Hochberg, and J. M. Guralnik, “Musculoskeletal Pain and Risk for Falls in Older Disabled Women Living in the Community,” *J. Am. Geriatr. Soc.*, vol. 50, no. 4, pp. 671–678, 2002.

[309] G. Ziere, J. P. Dieleman, a Hofman, H. a P. Pols, T. J. M. van der Cammen, and B. H. C. Stricker, “Polypharmacy and falls in the middle age and elderly population.,” *Br. J. Clin. Pharmacol.*, vol. 61, no. 2, pp. 218–23, 2006.

[310] V. B. Pothula, F. Chew, T. H. J. Lesser, and a. K. Sharma, “Falls and vestibular impairment,” *Clin. Otolaryngol. Allied Sci.*, vol. 29, no. 2, pp. 179–182, 2004.

[311] T. S. Dharmarajan, S. Avula, and E. P. Norkus, “Anemia increases risk for falls in hospitalized older adults: an evaluation of falls in 362 hospitalized, ambulatory, long-term care, and community patients,” *J. Am. Med. Dir. Assoc.*, vol. 8, no. 3, pp. e9–e15, 2007.

[312] N. Pandya, B. Bookhart, S. H. Mody, P. A. Funk Orsini, and G. Reardon, “Study of anemia in long-term care (SALT): prevalence of anemia and its relationship with the risk of falls in nursing home residents,” *Curr. Med. Res. Opin.*, vol. 24, no. 8, pp. 2139–2149, 2008.

[313] B. W. J. H. Penninx, S. M. F. Pluijm, P. Lips, R. Woodman, K. Miedema, J. M. Guralnik, and D. J. H. Deeg, “Late-Life Anemia Is Associated with Increased Risk of Recurrent Falls,” *J. Am. Geriatr. Soc.*, vol. 53, no. 12, pp. 2106–2111, 2005.

[314] E. M. Brody, M. H. Kleban, M. S. Moss, and F. Kleban, “Predictors of falls among institutionalized women with Alzheimer’s disease.,” *J. Am. Geriatr. Soc.*, 1984.

[315] J. C. Morris, E. H. Rubin, E. J. Morris, and S. A. Mandel, “Senile dementia of the Alzheimer’s type: an important risk factor for serious falls,” *J. Gerontol.*, vol. 42, no. 4, pp. 412–417, 1987.

[316] P. A. Stalenhoef, J. P. M. Diederiks, J. A. Knottnerus, A. D. M. Kester, and H. Crebolder, “A risk model for the prediction of recurrent falls in community-dwelling elderly: a prospective cohort study,” *J. Clin. Epidemiol.*, vol. 55, no. 11, pp. 1088–1094, 2002.

[317] M. E. Tinetti, M. Speechley, and S. F. Ginter, “Risk factors for falls among elderly persons living in the community,” *N. Engl. J. Med.*, vol. 319, no. 26, pp. 1701–1707, 1988.

[318] N. A. Sanders, J. A. Ganguly, T. L. Jetter, M. Daccarett, S. L. Wasmund, M. Brignole, and M. H. Hamdan, “Atrial fibrillation: an independent risk factor for nonaccidental falls in older patients.,” *Pacing Clin. Electrophysiol.*, vol. 35, no. 8, pp. 973–9, 2012.

[319] S. Morrison, S. R. Colberg, M. Mariano, H. K. Parson, and A. I. Vinik, “Balance training reduces falls risk in older individuals with type 2 diabetes,” *Diabetes Care*, vol. 33, no. 4, pp. 748–750, 2010.

[320] V. Carnevale, E. Romagnoli, L. D’Erasmo, and E. D’Erasmo, “Bone damage in type 2 diabetes mellitus,” *Nutr. Metab. Cardiovasc. Dis.*, 2014.

[321] M. S. Maurer, J. Burcham, and H. Cheng, “Diabetes mellitus is associated with an increased risk of falls in elderly residents of a long-term care facility,” *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.*, vol. 60, no. 9, pp. 1157–1162, 2005.

[322] A. V Schwartz, T. A. Hillier, D. E. Sellmeyer, H. E. Resnick, E. Gregg, K. E. Ensrud, P. J. Schreiner, K. L. Margolis, J. A. Cauley, and M. C. Nevitt, “Older women with diabetes have a higher risk of falls A prospective study,” *Diabetes Care*, vol. 25, no. 10, pp. 1749–1754, 2002.

[323] A. I. Vinik, E. J. Vinik, S. R. Colberg, and S. Morrison, “Falls Risk in Older Adults with Type 2 Diabetes,” *Clin. Geriatr. Med.*, vol. 31, no. 1, pp. 89–99, 2015.

[324] A. Biderman, J. Cwikel, A. Fried, and D. Galinsky, “Depression and falls among community dwelling elderly people: a search for common risk factors,” *J. Epidemiol. Community Health*, vol. 56, no. 8, pp. 631–636, 2002.

[325] A. Turcu, S. Toubin, F. Mourey, P. D’Athis, P. Manckoundia, and P. Pfitzenmeyer, “Falls and depression in older people,” *Gerontology*, vol. 50, no. 5, pp. 303–308, 2004.

[326] L. Quach, F. M. Yang, S. D. Berry, E. Newton, R. N. Jones, J. A. Burr, and L. A. Lipsitz, “Depression, antidepressants, and falls among community-dwelling elderly people: the MOBILIZE Boston study,” *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.*, vol. 68, no. 12, pp. 1575–1581, 2013.

[327] J. Rhoads, A. Clayman, and S. Nelson, “The relationship of urinary tract infections and falls in a nursing home.,” *Director*, vol. 15, no. 1, pp. 22–26, 2006.

[328] K. Bylow, W. Dale, K. Mustian, W. M. Stadler, M. Rodin, W. Hall, M. Lachs, and S. G. Mohile, “Falls and physical performance deficits in older patients with prostate cancer undergoing androgen deprivation therapy,” *Urology*, vol. 72, no. 2, pp. 422–427, 2008.

[329] W. S. Aronow and C. Ahn, “Association of Postprandial Hypotension With Incidence of Falls, Syncope, Coronary Events, Stroke, and Total Mortality at 29-Month Follow-Up in 499 Older Nursing Home Residents,” *J. Am. Geriatr. Soc.*, vol. 45, no. 9, pp. 1051–1053, 1997.

[330] T. Dharmarajan and E. P. Norkus, “Mild anemia and the risk of falls in older adults from nursing homes and the community,” *J. Am. Med. Dir. Assoc.*, vol. 5, no. 6, pp. 395–400, 2004.

[331] M. S. Duh, S. H. Mody, P. Lefebvre, R. C. Woodman, S. Buteau, and C. T. Viech, “Anaemia and the risk of injurious falls in a community-dwelling elderly population,” *Drugs Aging*, vol. 25, no. 4, pp. 325–334, 2008.

[332] T. Tachi, T. Yokoi, C. Goto, M. Umeda, Y. Noguchi, M. Yasuda, M. Minamitani, T. Mizui, T. Tsuchiya, and H. Teramachi, “Hyponatremia and hypokalemia as risk factors for falls,” *Eur. J. Clin. Nutr.*, 2014.

[333] S. Bangalore, F. H. Messerli, J. B. Kostis, and C. J. Pepine, “Cardiovascular protection using beta-blockers: a critical review of the evidence,” *J. Am. Coll. Cardiol.*, vol. 50, no. 7, pp. 563–572, 2007.

[334] C. P. Cannon, B. A. Steinberg, S. A. Murphy, J. L. Mega, and E. Braunwald, “Meta-analysis of cardiovascular outcomes trials comparing intensive versus moderate statin therapy,” *J. Am. Coll. Cardiol.*, vol. 48, no. 3, pp. 438–445, 2006.

[335] J. Gribbin, R. Hubbard, J. R. Gladman, C. Smith, and S. Lewis, “Risk of falls associated with antihypertensive medication: population-based case–control study,” *Age Ageing*, p. afq092, 2010.

[336] D. Scott, L. Blizzard, J. Fell, and G. Jones, “Statin therapy, muscle function and falls risk in community-dwelling older adults,” *QJM*, p. hcp093, 2009.

[337] L. H. Eggermont, B. W. Penninx, R. N. Jones, and S. G. Leveille, “Depressive Symptoms, Chronic Pain, and Falls in Older Community-Dwelling Adults: The MOBILIZE Boston Study,” *J. Am. Geriatr. Soc.*, vol. 60, no. 2, pp. 230–237, 2012.

[338] N. F. Khan, S. E. Harrison, and P. W. Rose, “Validity of diagnostic coding within the General Practice Research Database: a systematic review.,” *Br. J. Gen. Pract.*, vol. 60, no. 572, pp. e128-36, 2010.

[339] C. Crooks, J. West, and T. Card, “A comparison of the recording of comorbidity in primary and secondary care by using the Charlson Index to predict short-term and long-term survival in a routine linked data cohort,” *BMJ Open*, vol. 5, no. 6, p. e007974, 2015.

Appendix A: Salford patient data: cluster analysis

Two clustering algorithms were applied to the data: simple k-means (Figure A.1) and Expectation Maximisation (EM) (Figure A.3). Both algorithms were performed to the results obtained using the Resnik measure with the Maximum approach. The cluster analysis for the k-means and EM algorithms is presented in Figure A.2 and Figure A.4 respectively.

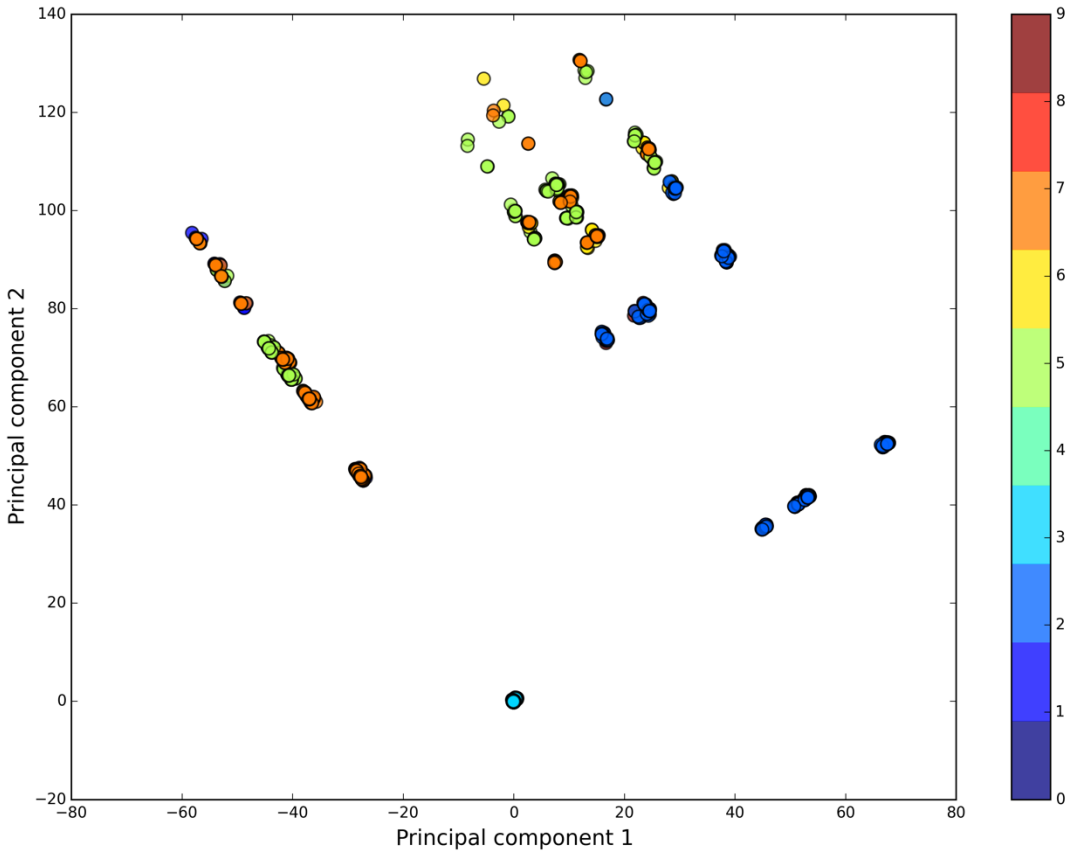
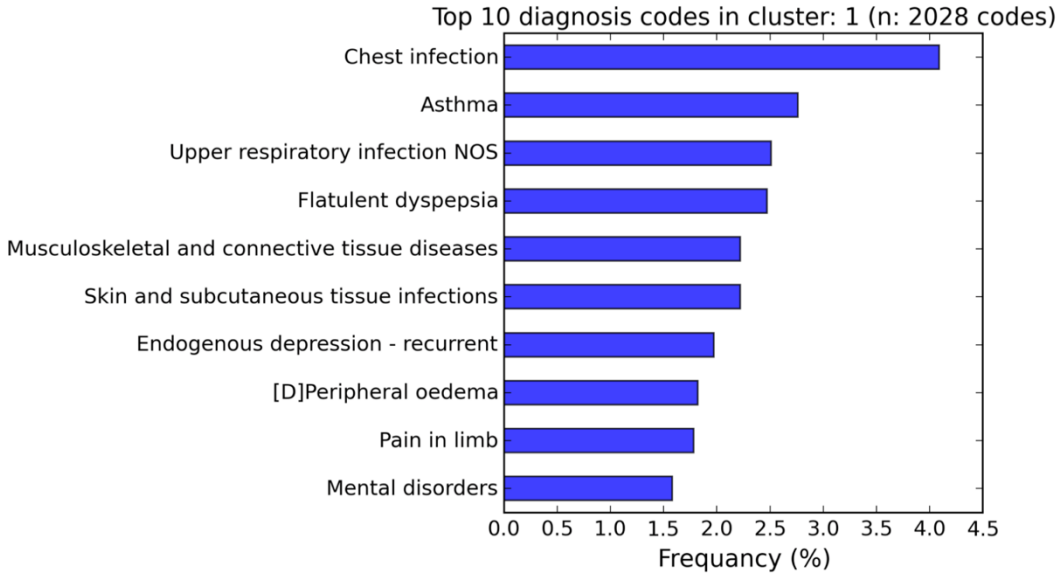
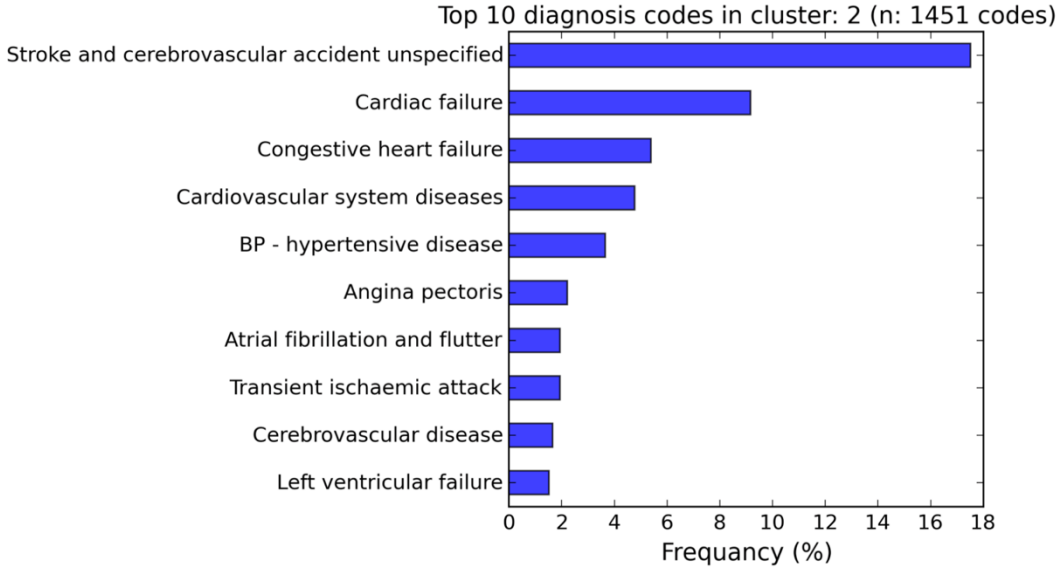
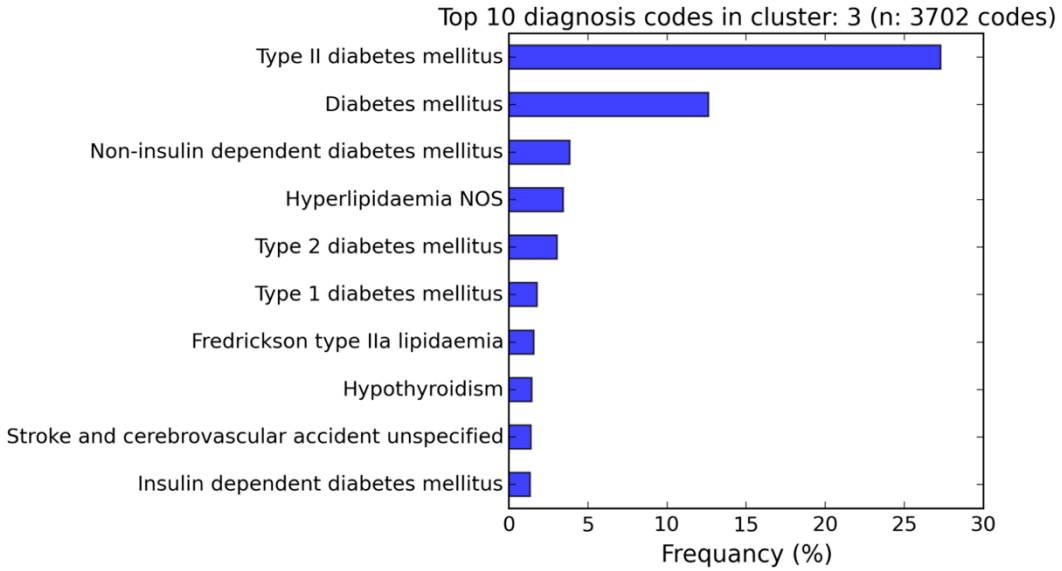
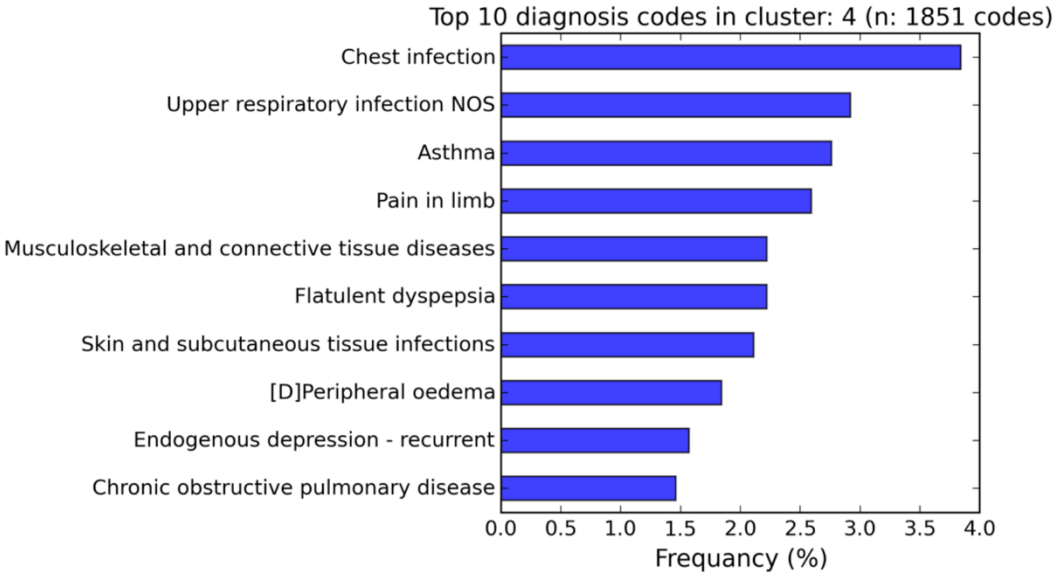


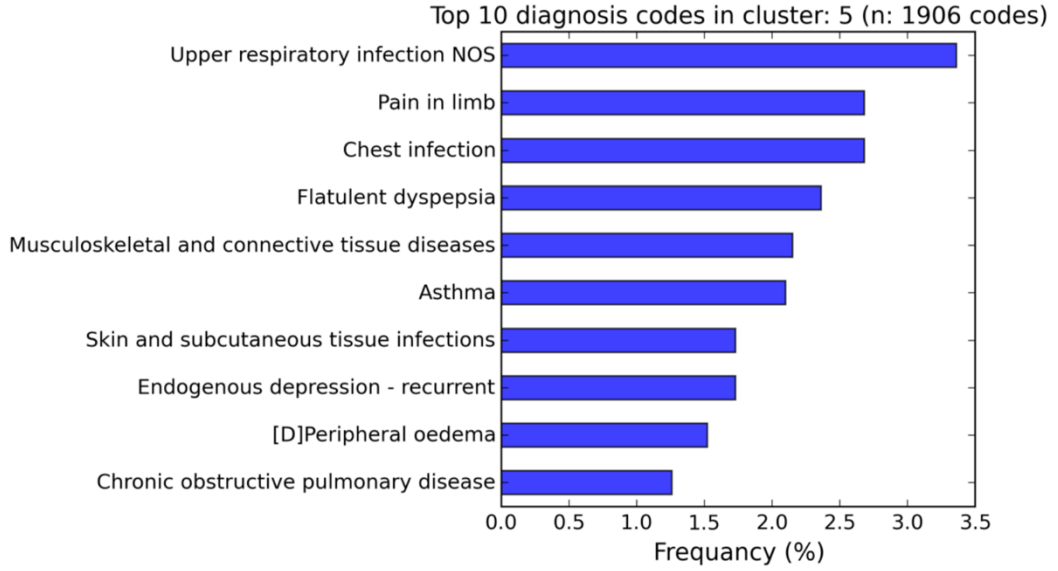
Figure A.1. Clustering analysis of patients records using the k-means clustering algorithm (*k=10*). The PCA representation of patient records was calculated using the Resnik measure with the Maximum approach. 10 clusters were generated by k-means algorithm.

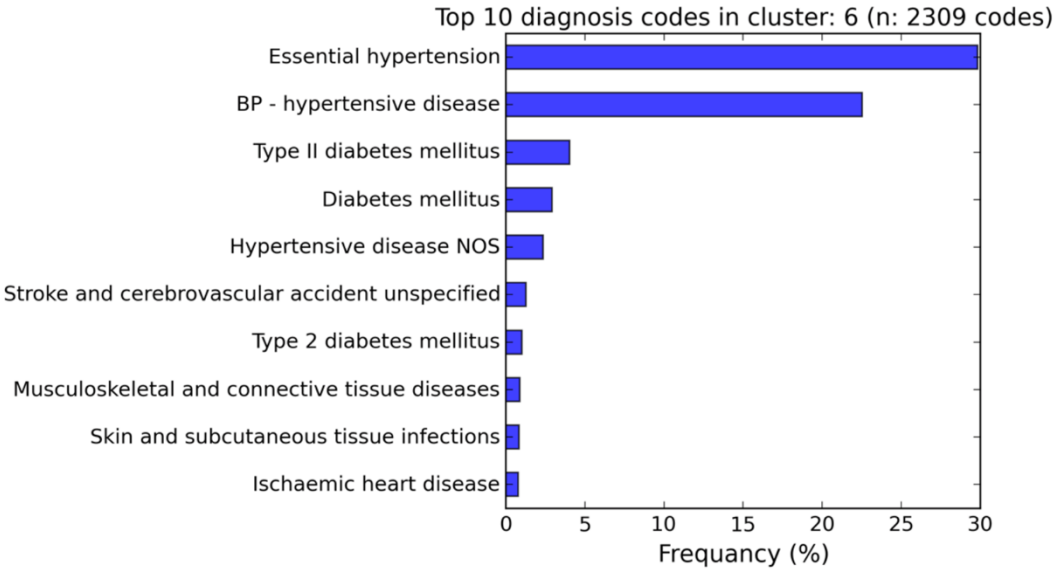

Top 10 diagnosis codes in cluster: 1 (n: 2028 codes)
Chest infection
Asthma
Upper respiratory infection NOS
Flatulent dyspepsia
Musculoskeletal and connective tissue diseases
Skin and subcutaneous tissue infections
Endogenous depression - recurrent
[D]Peripheral oedema
Pain in limb
Mental disorders
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5
Frequancy (%)

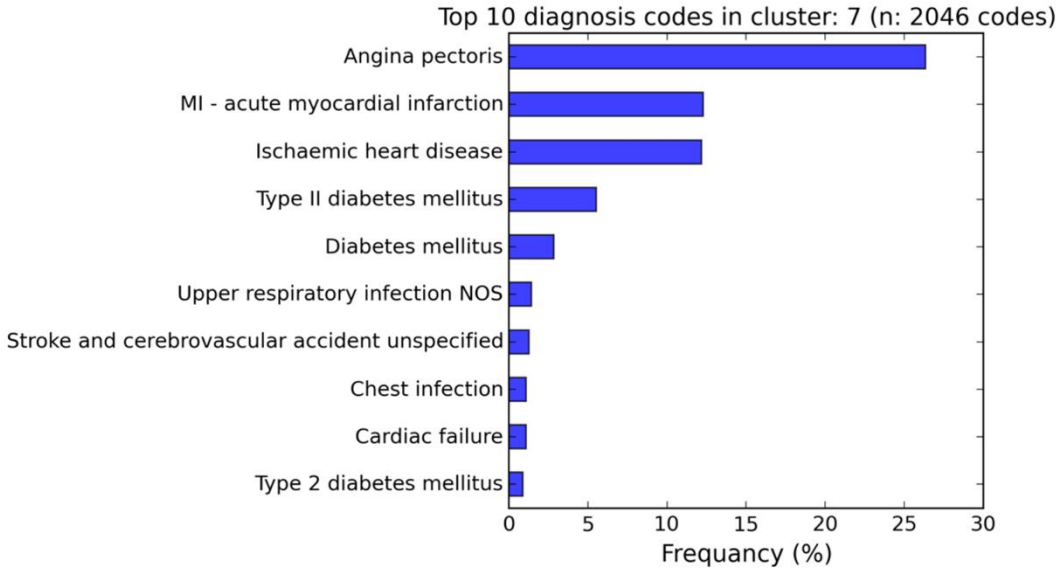

Top 10 diagnosis codes in cluster: 2 (n: 1451 codes)
Stroke and cerebrovascular accident unspecified
Cardiac failure
Congestive heart failure
Cardiovascular system diseases
BP - hypertensive disease
Angina pectoris
Atrial fibrillation and flutter
Transient ischaemic attack
Cerebrovascular disease
Left ventricular failure
0 2 4 6 8 10 12 14 16 18
Frequancy (%)

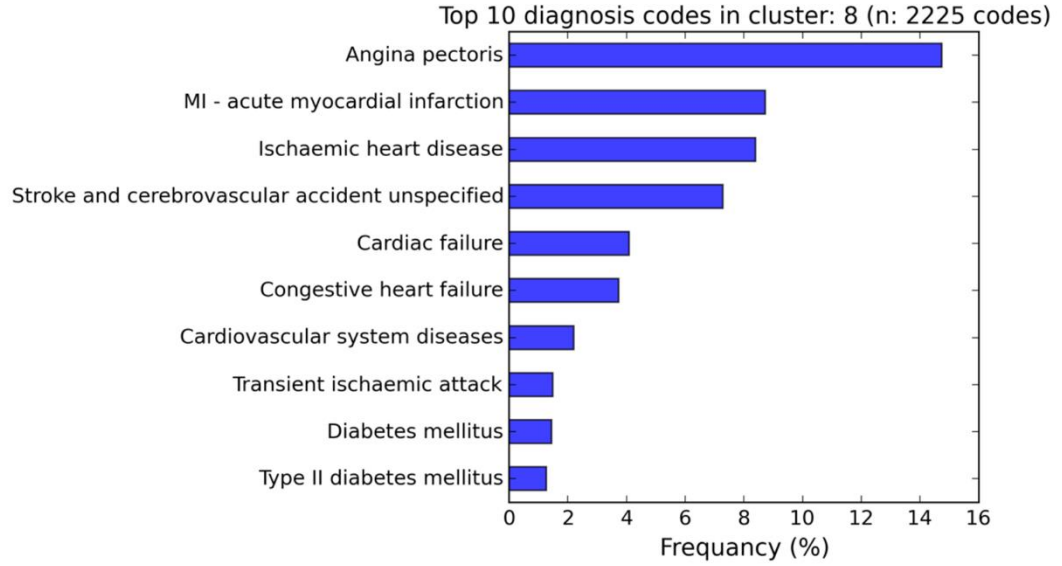

Top 10 diagnosis codes in cluster: 3 (n: 3702 codes)
Type II diabetes mellitus
Diabetes mellitus
Non-insulin dependent diabetes mellitus
Hyperlipidaemia NOS
Type 2 diabetes mellitus
Type 1 diabetes mellitus
Fredrickson type IIa lipidaemia
Hypothyroidism
Stroke and cerebrovascular accident unspecified
Insulin dependent diabetes mellitus
0 5 10 15 20 25 30
Frequancy (%)

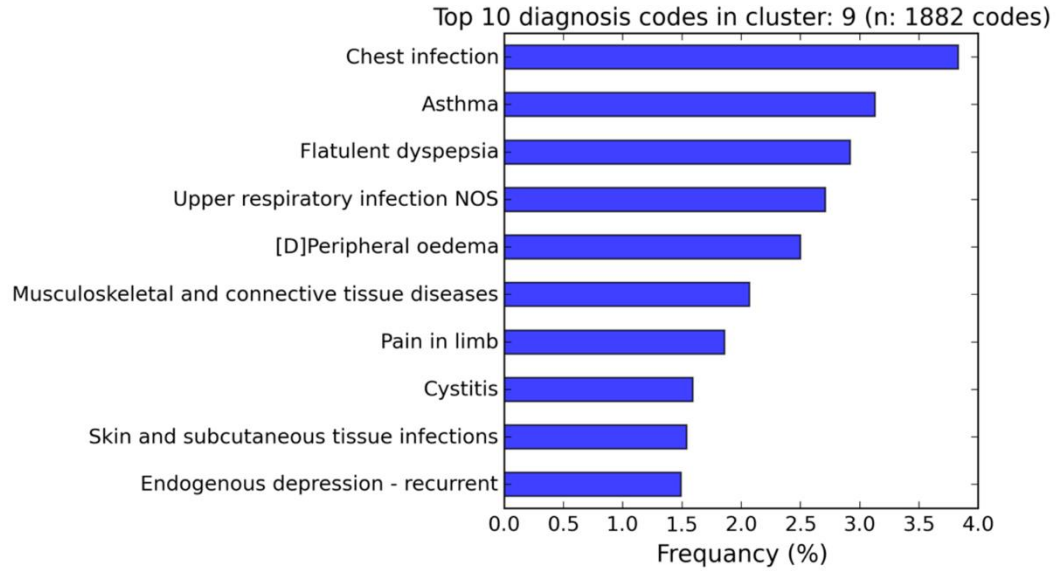

Top 10 diagnosis codes in cluster: 4 (n: 1851 codes)
Chest infection
Upper respiratory infection NOS
Asthma
Pain in limb
Musculoskeletal and connective tissue diseases
Flatulent dyspepsia
Skin and subcutaneous tissue infections
[D]Peripheral oedema
Endogenous depression - recurrent
Chronic obstructive pulmonary disease
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Frequancy (%)


Top 10 diagnosis codes in cluster: 5 (n: 1906 codes)
Upper respiratory infection NOS
Pain in limb
Chest infection
Flatulent dyspepsia
Musculoskeletal and connective tissue diseases
Asthma
Skin and subcutaneous tissue infections
Endogenous depression - recurrent
[D]Peripheral oedema
Chronic obstructive pulmonary disease
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Frequancy (%)


Top 10 diagnosis codes in cluster: 6 (n: 2309 codes)
Essential hypertension
BP - hypertensive disease
Type II diabetes mellitus
Diabetes mellitus
Hypertensive disease NOS
Stroke and cerebrovascular accident unspecified
Type 2 diabetes mellitus
Musculoskeletal and connective tissue diseases
Skin and subcutaneous tissue infections
Ischaemic heart disease
0 5 10 15 20 25 30
Frequancy (%)


Top 10 diagnosis codes in cluster: 7 (n: 2046 codes)
Angina pectoris
MI - acute myocardial infarction
Ischaemic heart disease
Type II diabetes mellitus
Diabetes mellitus
Upper respiratory infection NOS
Stroke and cerebrovascular accident unspecified
Chest infection
Cardiac failure
Type 2 diabetes mellitus
0
5
10
15
20
25
30
Frequancy (%)


Top 10 diagnosis codes in cluster: 8 (n: 2225 codes)
Angina pectoris
MI - acute myocardial infarction
Ischaemic heart disease
Stroke and cerebrovascular accident unspecified
Cardiac failure
Congestive heart failure
Cardiovascular system diseases
Transient ischaemic attack
Diabetes mellitus
Type II diabetes mellitus
0
2
4
6
8
10
12
14
16
Frequancy (%)


Top 10 diagnosis codes in cluster: 9 (n: 1882 codes)
Chest infection
Asthma
Flatulent dyspepsia
Upper respiratory infection NOS
[D]Peripheral oedema
Musculoskeletal and connective tissue diseases
Pain in limb
Cystitis
Skin and subcutaneous tissue infections
Endogenous depression - recurrent
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Frequancy (%)

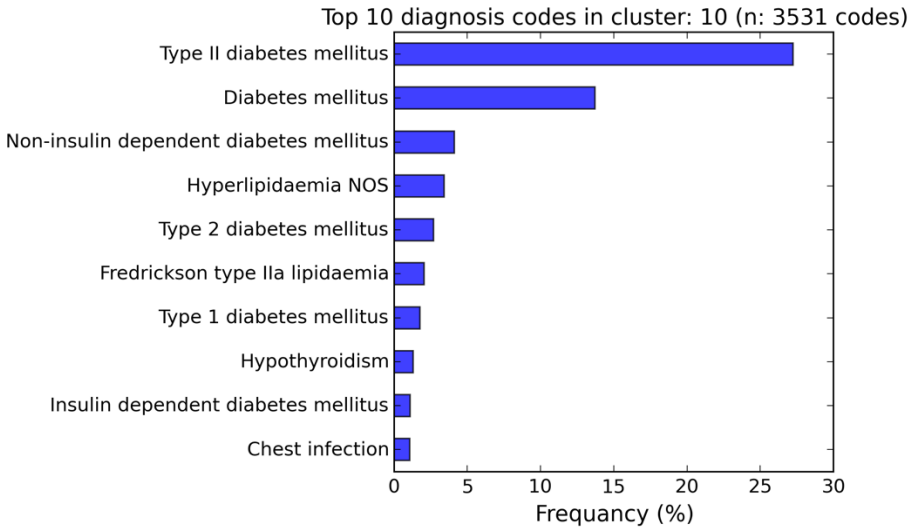


Figure A.2. Cluster analysis of patient records using the k-means algorithm. This shows the top five most frequent diagnosis codes in the 10 clusters.

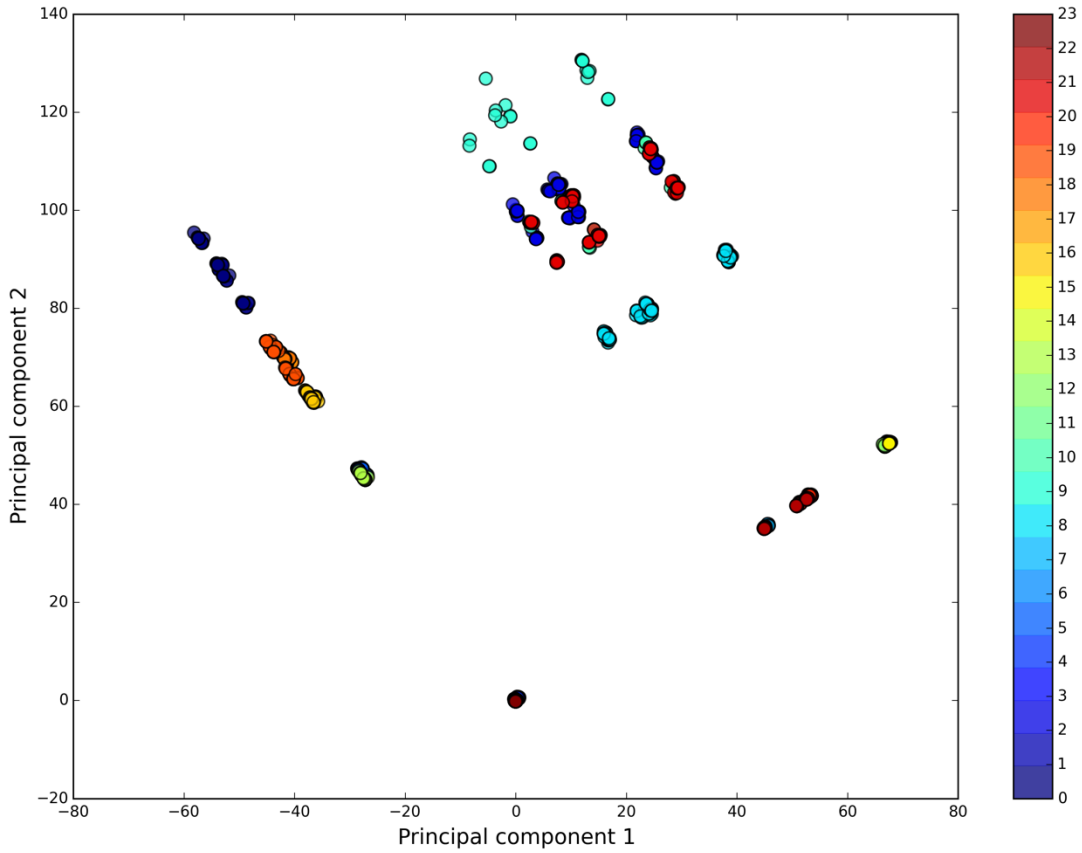
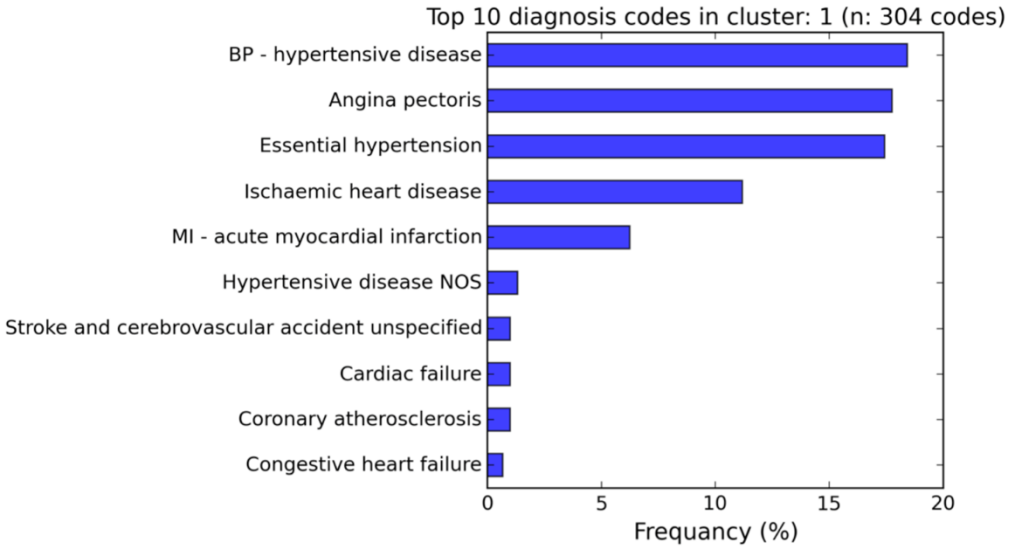
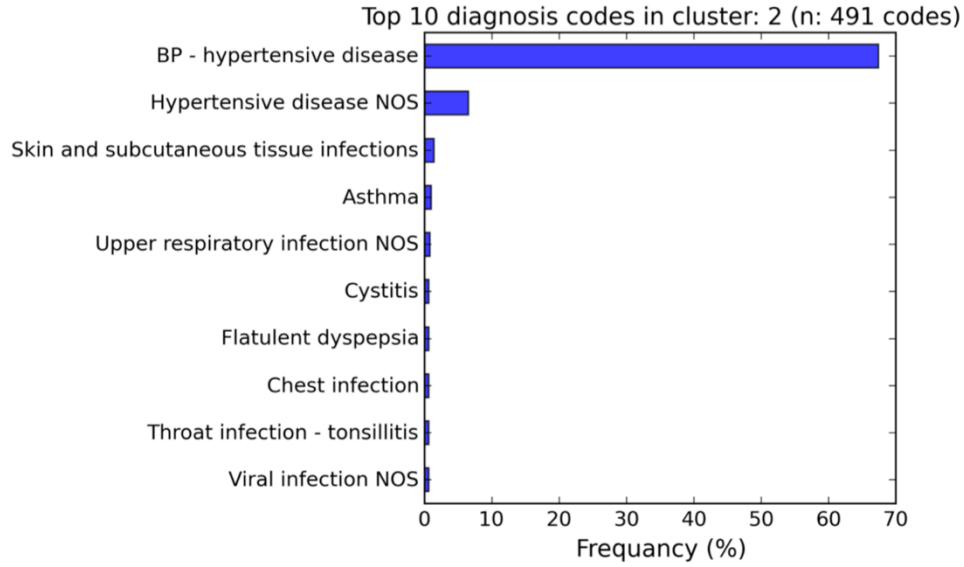
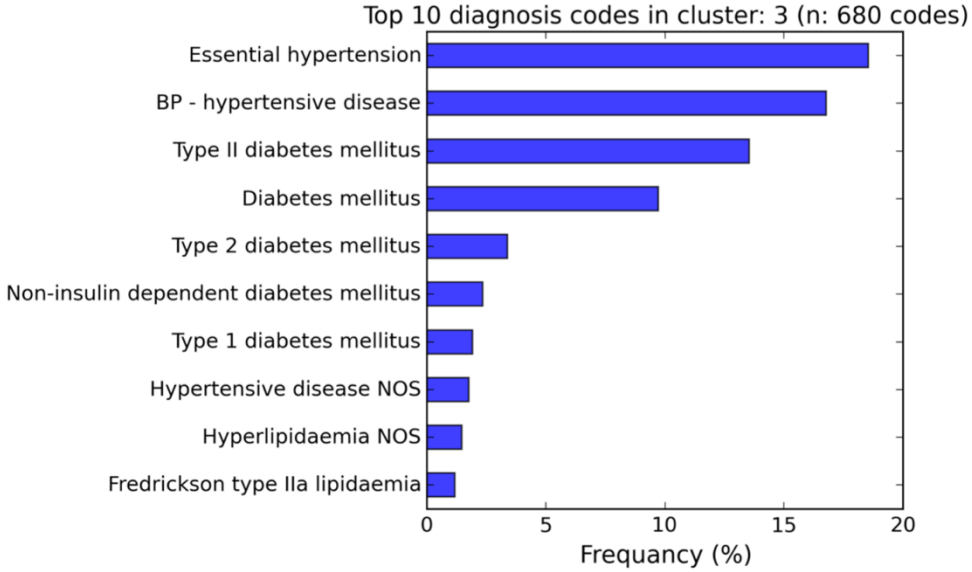
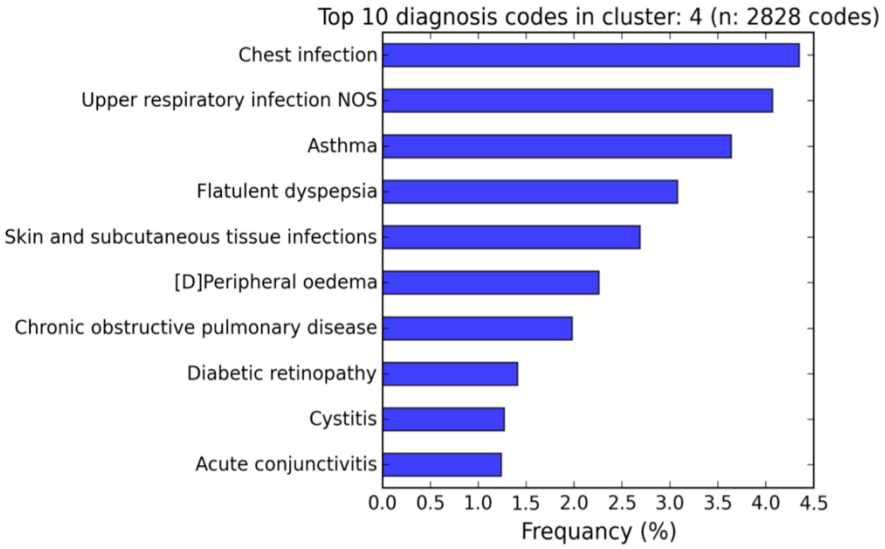


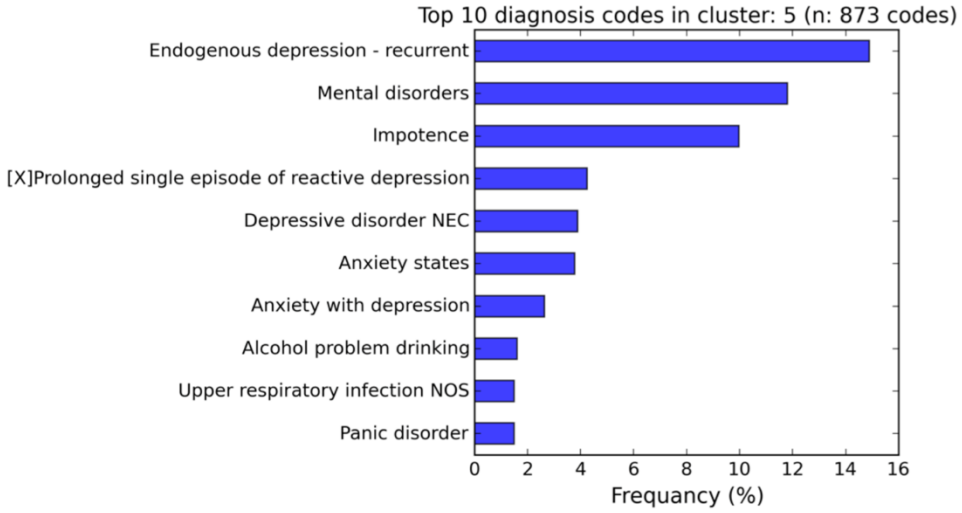
Figure A.3. Clustering analysis of patient records using the Expectation Maximization (EM) clustering algorithm. The PCA representation of patient records was calculated using the Resnik measure with the Maximum approach. 24 clusters were generated by EM algorithm.

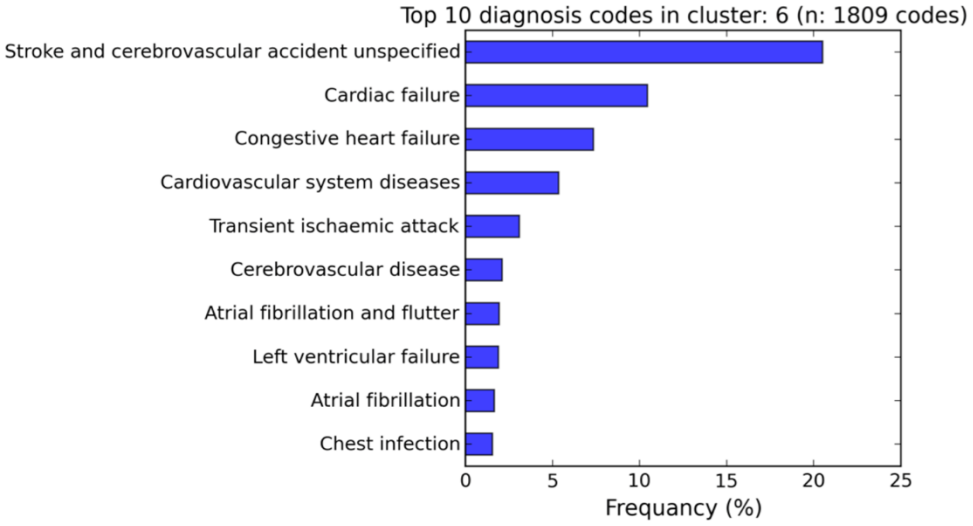

Top 10 diagnosis codes in cluster: 1 (n: 304 codes)
BP - hypertensive disease
Angina pectoris
Essential hypertension
Ischaemic heart disease
MI - acute myocardial infarction
Hypertensive disease NOS
Stroke and cerebrovascular accident unspecified
Cardiac failure
Coronary atherosclerosis
Congestive heart failure
0
5
10
15
20
Frequancy (%)

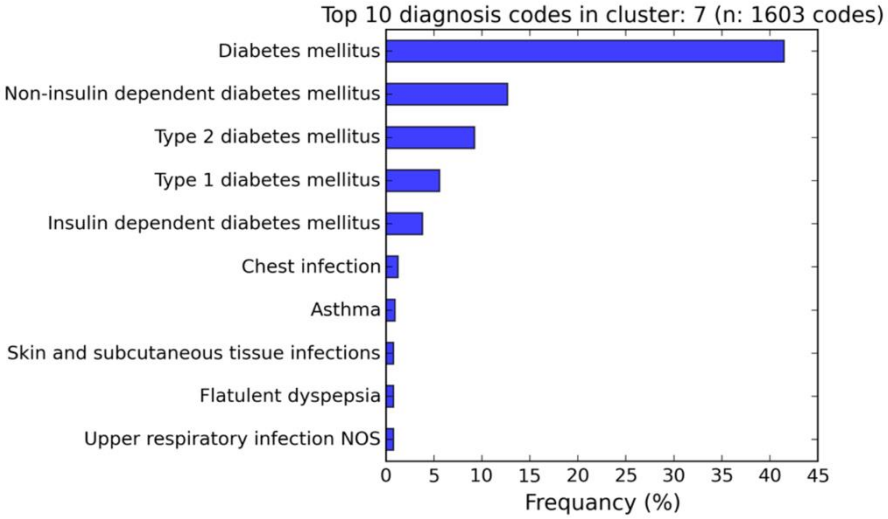

Top 10 diagnosis codes in cluster: 2 (n: 491 codes)
BP - hypertensive disease
Hypertensive disease NOS
Skin and subcutaneous tissue infections
Asthma
Upper respiratory infection NOS
Cystitis
Flatulent dyspepsia
Chest infection
Throat infection - tonsillitis
Viral infection NOS
0
10
20
30
40
50
60
70
Frequancy (%)

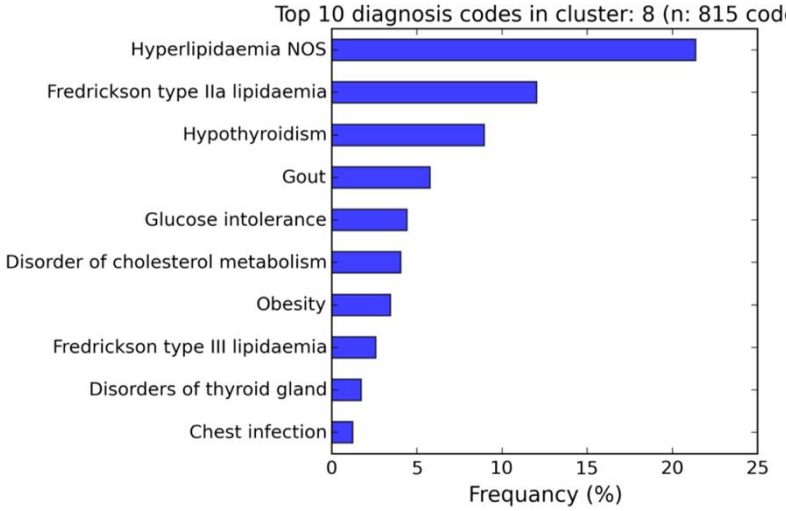

Top 10 diagnosis codes in cluster: 3 (n: 680 codes)
Essential hypertension
BP - hypertensive disease
Type II diabetes mellitus
Diabetes mellitus
Type 2 diabetes mellitus
Non-insulin dependent diabetes mellitus
Type 1 diabetes mellitus
Hypertensive disease NOS
Hyperlipidaemia NOS
Fredrickson type IIa lipidaemia
0
5
10
15
20
Frequancy (%)

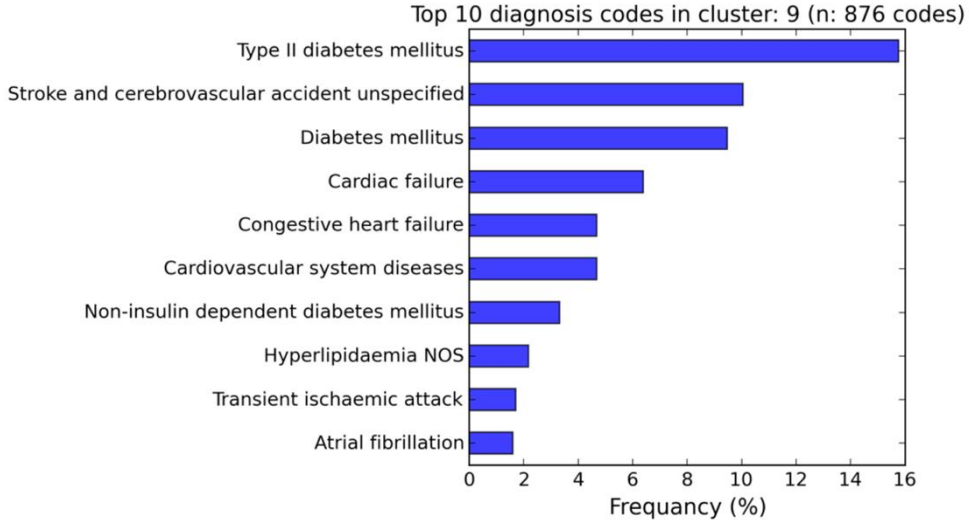

Top 10 diagnosis codes in cluster: 4 (n: 2828 codes)
Chest infection
Upper respiratory infection NOS
Asthma
Flatulent dyspepsia
Skin and subcutaneous tissue infections
[D]Peripheral oedema
Chronic obstructive pulmonary disease
Diabetic retinopathy
Cystitis
Acute conjunctivitis
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5
Frequancy (%)

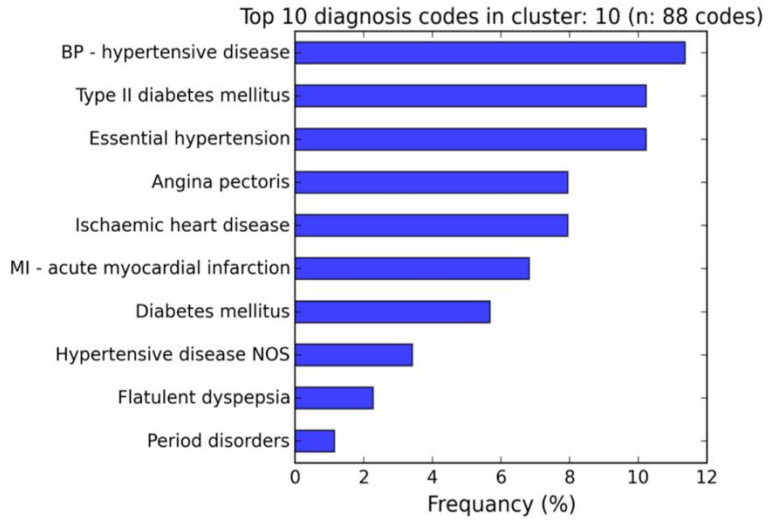

Top 10 diagnosis codes in cluster: 5 (n: 873 codes)
Endogenous depression - recurrent
Mental disorders
Impotence
[X]Prolonged single episode of reactive depression
Depressive disorder NEC
Anxiety states
Anxiety with depression
Alcohol problem drinking
Upper respiratory infection NOS
Panic disorder
0 2 4 6 8 10 12 14 16
Frequancy (%)

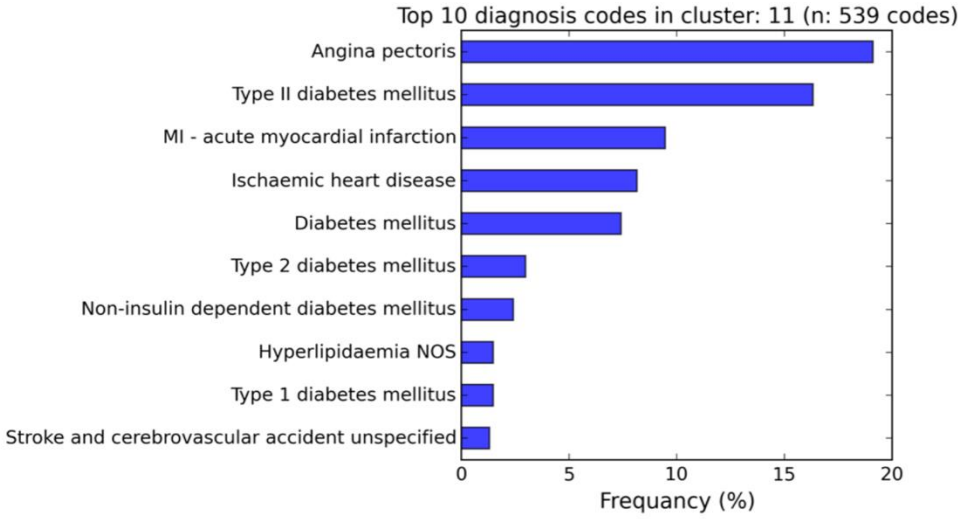

Top 10 diagnosis codes in cluster: 6 (n: 1809 codes)
Stroke and cerebrovascular accident unspecified
Cardiac failure
Congestive heart failure
Cardiovascular system diseases
Transient ischaemic attack
Cerebrovascular disease
Atrial fibrillation and flutter
Left ventricular failure
Atrial fibrillation
Chest infection
0 5 10 15 20 25
Frequancy (%)

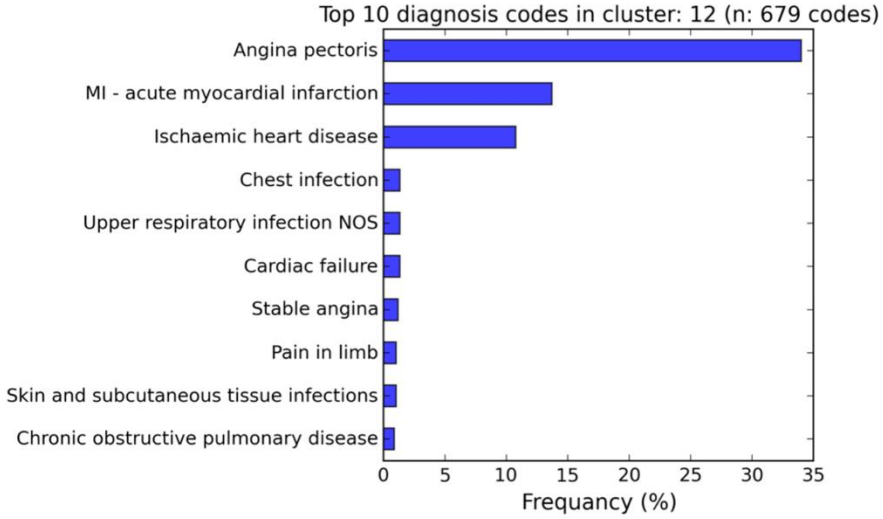

Top 10 diagnosis codes in cluster: 7 (n: 1603 codes)
Diabetes mellitus
Non-insulin dependent diabetes mellitus
Type 2 diabetes mellitus
Type 1 diabetes mellitus
Insulin dependent diabetes mellitus
Chest infection
Asthma
Skin and subcutaneous tissue infections
Flatulent dyspepsia
Upper respiratory infection NOS
0 5 10 15 20 25 30 35 40 45
Frequancy (%)

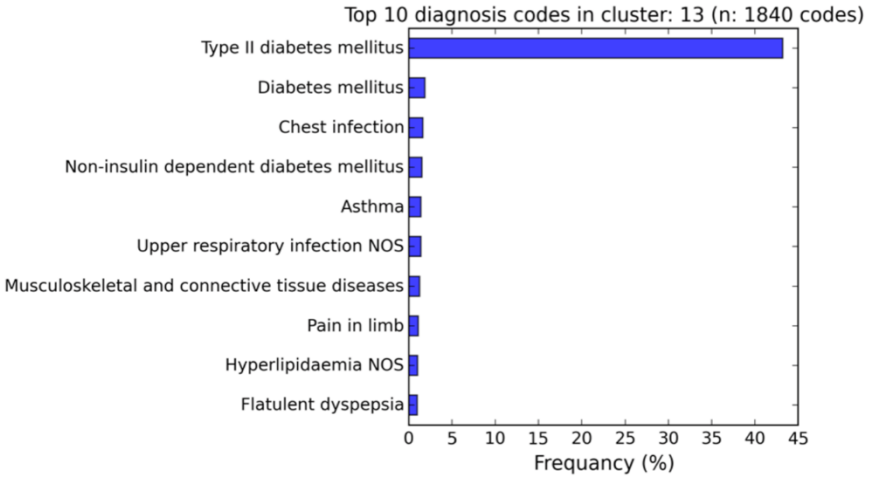

Top 10 diagnosis codes in cluster: 8 (n: 815 codes)
Hyperlipidaemia NOS
Fredrickson type IIa lipidaemia
Hypothyroidism
Gout
Glucose intolerance
Disorder of cholesterol metabolism
Obesity
Fredrickson type III lipidaemia
Disorders of thyroid gland
Chest infection
0 5 10 15 20 25
Frequancy (%)

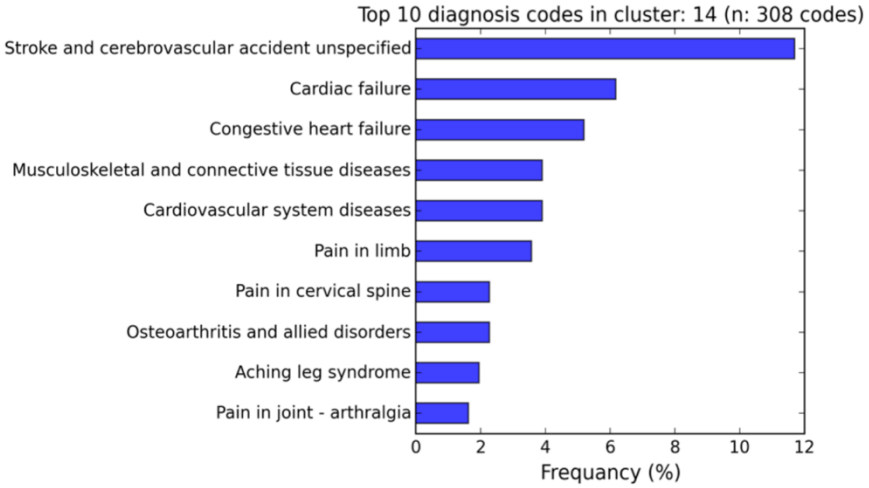

Top 10 diagnosis codes in cluster: 9 (n: 876 codes)
Type II diabetes mellitus
Stroke and cerebrovascular accident unspecified
Diabetes mellitus
Cardiac failure
Congestive heart failure
Cardiovascular system diseases
Non-insulin dependent diabetes mellitus
Hyperlipidaemia NOS
Transient ischaemic attack
Atrial fibrillation
0 2 4 6 8 10 12 14 16
Frequancy (%)

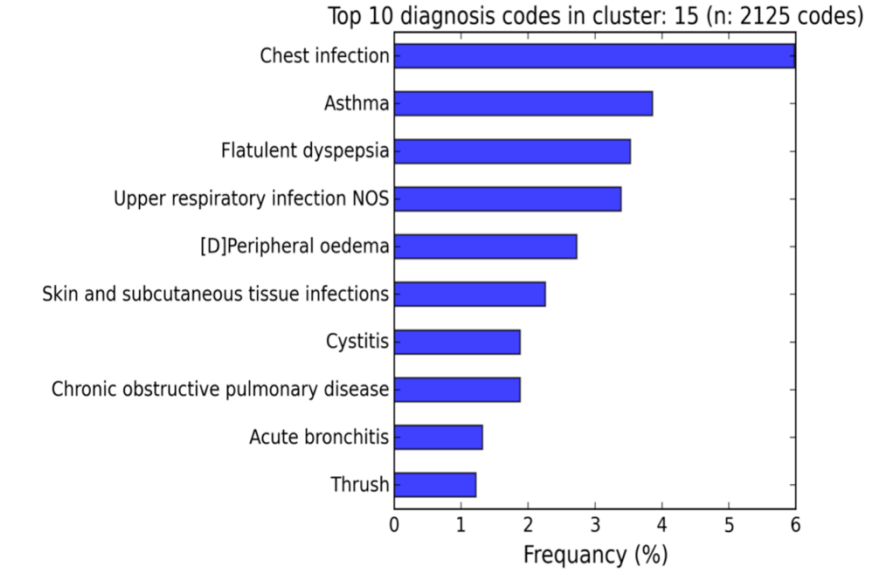

Top 10 diagnosis codes in cluster: 10 (n: 88 codes)
BP - hypertensive disease
Type II diabetes mellitus
Essential hypertension
Angina pectoris
Ischaemic heart disease
MI - acute myocardial infarction
Diabetes mellitus
Hypertensive disease NOS
Flatulent dyspepsia
Period disorders
0
2
4
6
8
10
12
Frequancy (%)

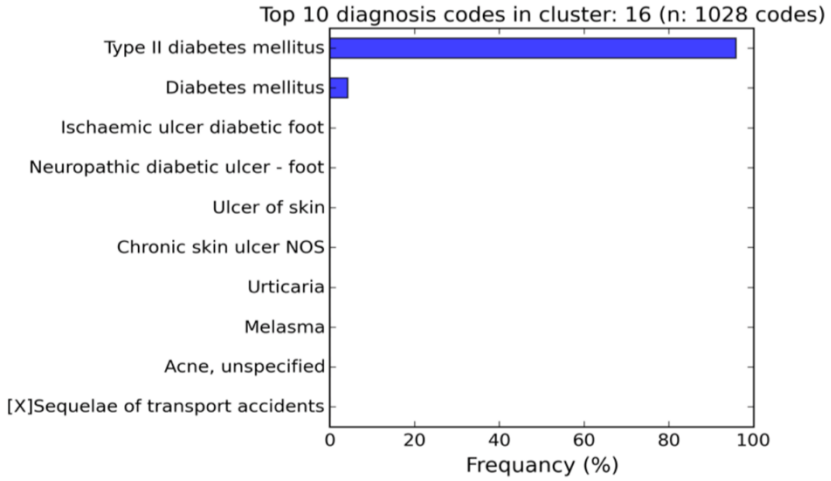

Top 10 diagnosis codes in cluster: 11 (n: 539 codes)
Angina pectoris
Type II diabetes mellitus
MI - acute myocardial infarction
Ischaemic heart disease
Diabetes mellitus
Type 2 diabetes mellitus
Non-insulin dependent diabetes mellitus
Hyperlipidaemia NOS
Type 1 diabetes mellitus
Stroke and cerebrovascular accident unspecified
0
5
10
15
20
Frequancy (%)

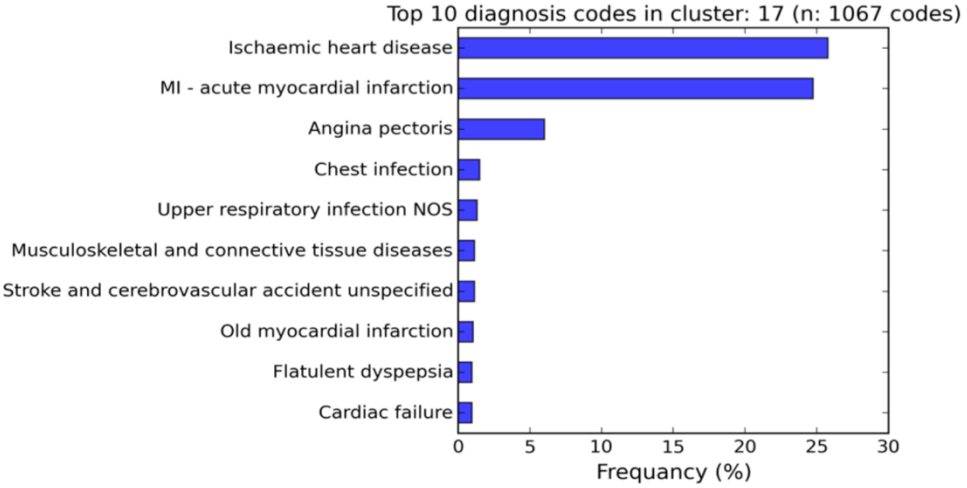

Top 10 diagnosis codes in cluster: 12 (n: 679 codes)
Angina pectoris
MI - acute myocardial infarction
Ischaemic heart disease
Chest infection
Upper respiratory infection NOS
Cardiac failure
Stable angina
Pain in limb
Skin and subcutaneous tissue infections
Chronic obstructive pulmonary disease
0
5
10
15
20
25
30
35
Frequancy (%)

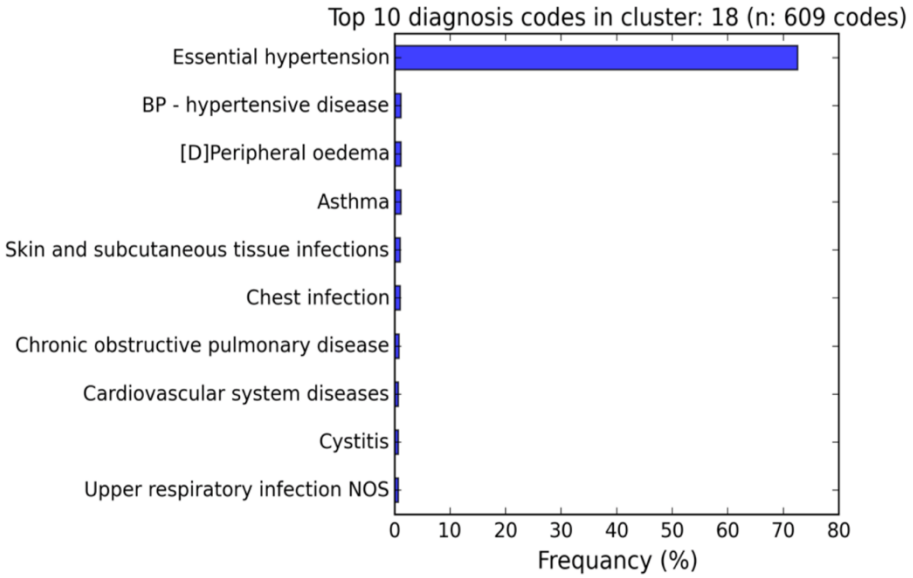

Top 10 diagnosis codes in cluster: 13 (n: 1840 codes)
Type II diabetes mellitus
Diabetes mellitus
Chest infection
Non-insulin dependent diabetes mellitus
Asthma
Upper respiratory infection NOS
Musculoskeletal and connective tissue diseases
Pain in limb
Hyperlipidaemia NOS
Flatulent dyspepsia
0 5 10 15 20 25 30 35 40 45
Frequancy (%)

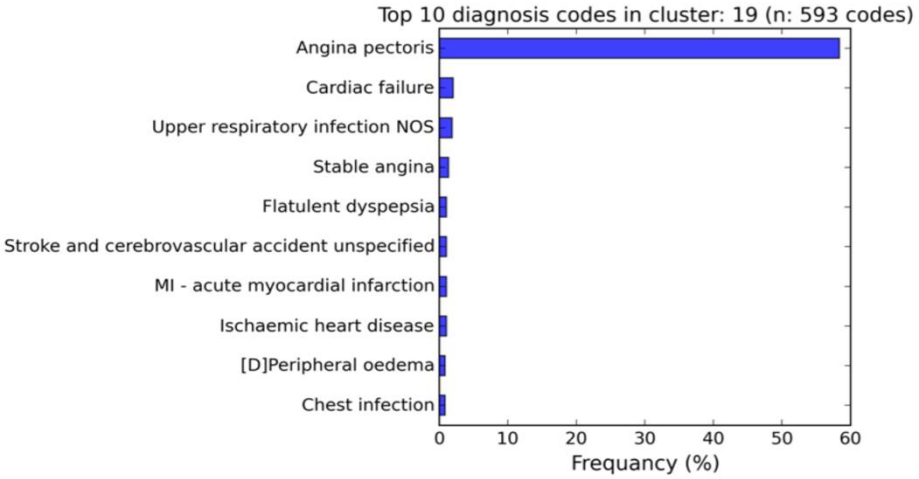

Top 10 diagnosis codes in cluster: 14 (n: 308 codes)
Stroke and cerebrovascular accident unspecified
Cardiac failure
Congestive heart failure
Musculoskeletal and connective tissue diseases
Cardiovascular system diseases
Pain in limb
Pain in cervical spine
Osteoarthritis and allied disorders
Aching leg syndrome
Pain in joint - arthralgia
0 2 4 6 8 10 12
Frequancy (%)

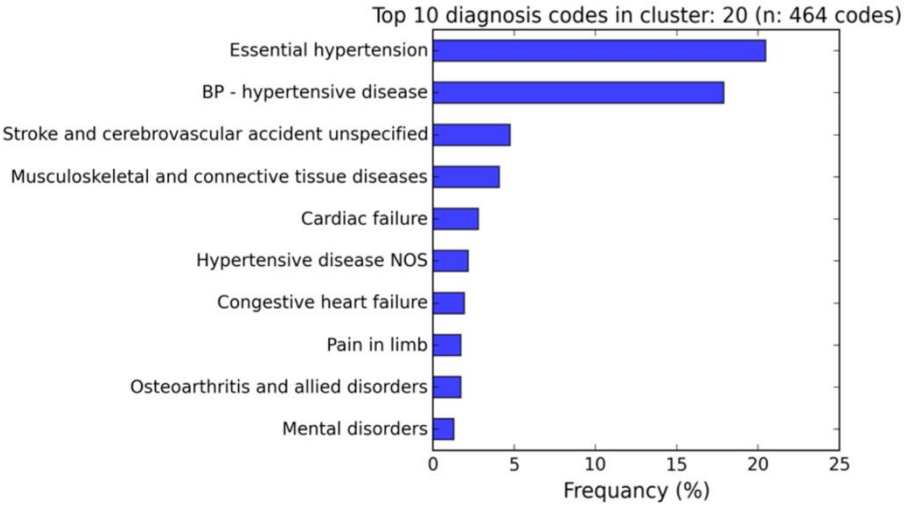

Top 10 diagnosis codes in cluster: 15 (n: 2125 codes)
Chest infection
Asthma
Flatulent dyspepsia
Upper respiratory infection NOS
[D]Peripheral oedema
Skin and subcutaneous tissue infections
Cystitis
Chronic obstructive pulmonary disease
Acute bronchitis
Thrush
0 1 2 3 4 5 6
Frequancy (%)

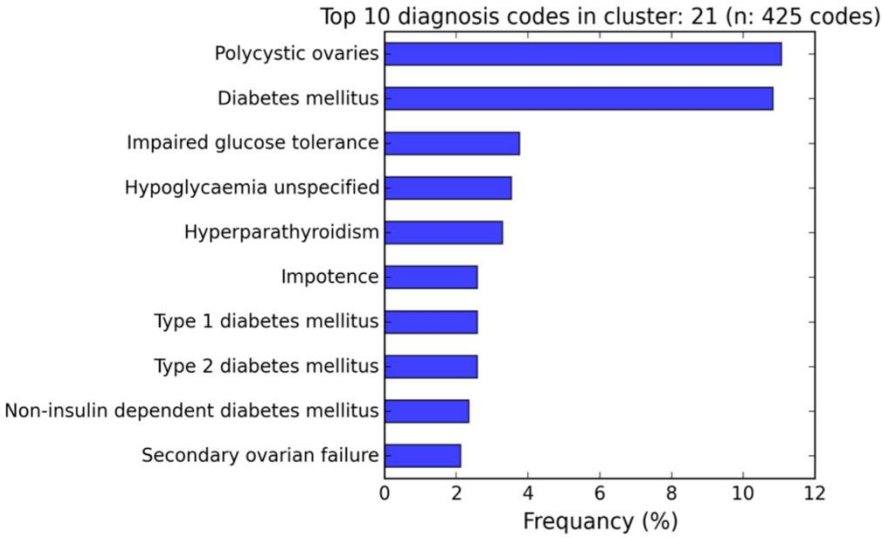

Top 10 diagnosis codes in cluster: 16 (n: 1028 codes)
Type II diabetes mellitus
Diabetes mellitus
Ischaemic ulcer diabetic foot
Neuropathic diabetic ulcer - foot
Ulcer of skin
Chronic skin ulcer NOS
Urticaria
Melasma
Acne, unspecified
[X]Sequelae of transport accidents
0
20
40
60
80
100
Frequancy (%)


Top 10 diagnosis codes in cluster: 17 (n: 1067 codes)
Ischaemic heart disease
MI - acute myocardial infarction
Angina pectoris
Chest infection
Upper respiratory infection NOS
Musculoskeletal and connective tissue diseases
Stroke and cerebrovascular accident unspecified
Old myocardial infarction
Flatulent dyspepsia
Cardiac failure
0
5
10
15
20
25
30
Frequancy (%)


Top 10 diagnosis codes in cluster: 18 (n: 609 codes)
Essential hypertension
BP - hypertensive disease
[D]Peripheral oedema
Asthma
Skin and subcutaneous tissue infections
Chest infection
Chronic obstructive pulmonary disease
Cardiovascular system diseases
Cystitis
Upper respiratory infection NOS
0
10
20
30
40
50
60
70
80
Frequancy (%)


Top 10 diagnosis codes in cluster: 19 (n: 593 codes)
Angina pectoris
Cardiac failure
Upper respiratory infection NOS
Stable angina
Flatulent dyspepsia
Stroke and cerebrovascular accident unspecified
MI - acute myocardial infarction
Ischaemic heart disease
[D]Peripheral oedema
Chest infection
0
10
20
30
40
50
60
Frequancy (%)


Top 10 diagnosis codes in cluster: 20 (n: 464 codes)
Essential hypertension
BP - hypertensive disease
Stroke and cerebrovascular accident unspecified
Musculoskeletal and connective tissue diseases
Cardiac failure
Hypertensive disease NOS
Congestive heart failure
Pain in limb
Osteoarthritis and allied disorders
Mental disorders
0
5
10
15
20
25
Frequancy (%)


Top 10 diagnosis codes in cluster: 21 (n: 425 codes)
Polycystic ovaries
Diabetes mellitus
Impaired glucose tolerance
Hypoglycaemia unspecified
Hyperparathyroidism
Impotence
Type 1 diabetes mellitus
Type 2 diabetes mellitus
Non-insulin dependent diabetes mellitus
Secondary ovarian failure
0
2
4
6
8
10
12
Frequancy (%)

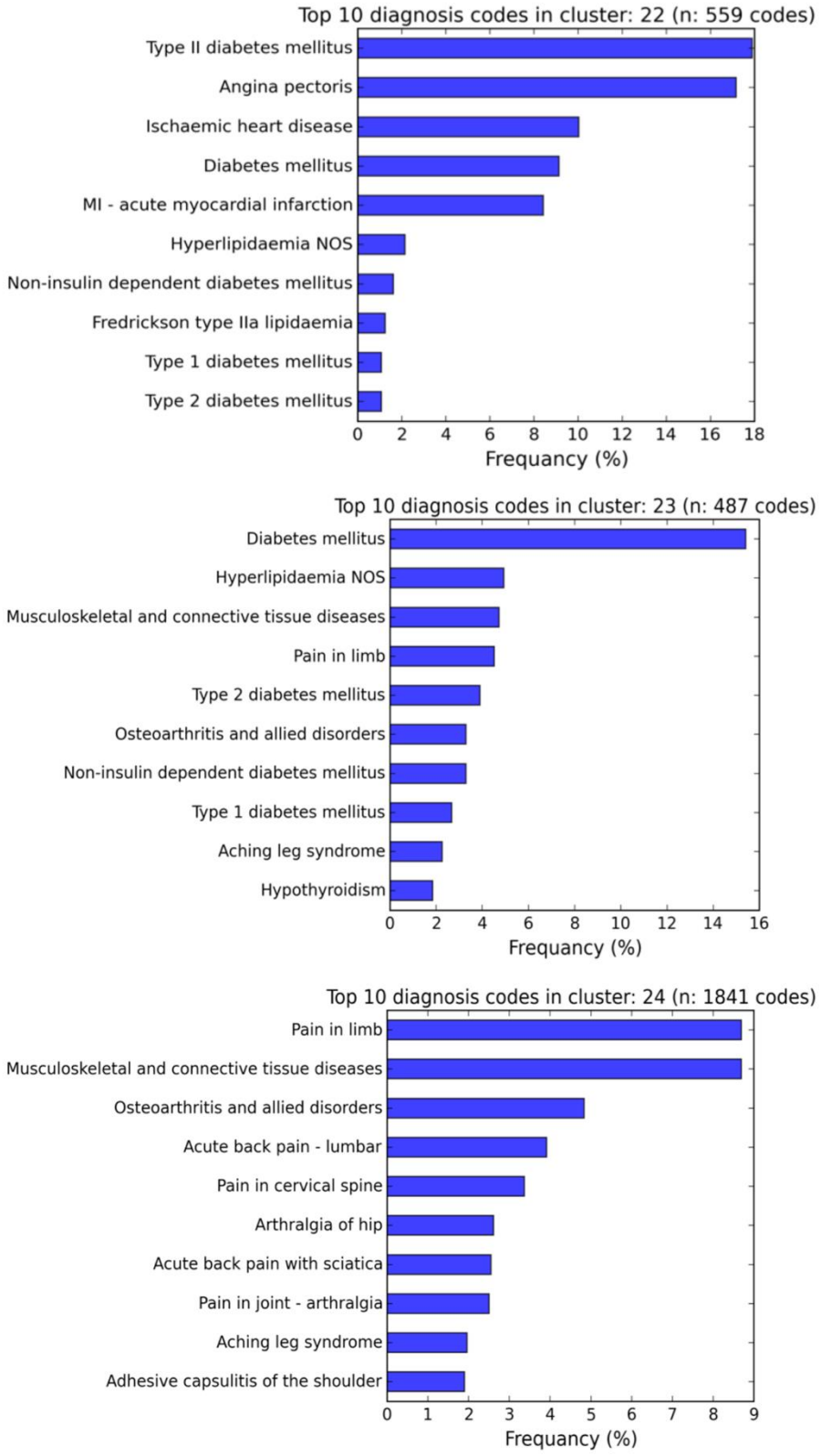


Figure A.4. Cluster analysis of patient records using the Expectation Maximization (EM) algorithm. This shows the top five most frequent diagnosis codes in the 24 clusters.

Appendix B: Mapping analysis on CPRD data

B.1. Distribution of Read code chapters in the data.

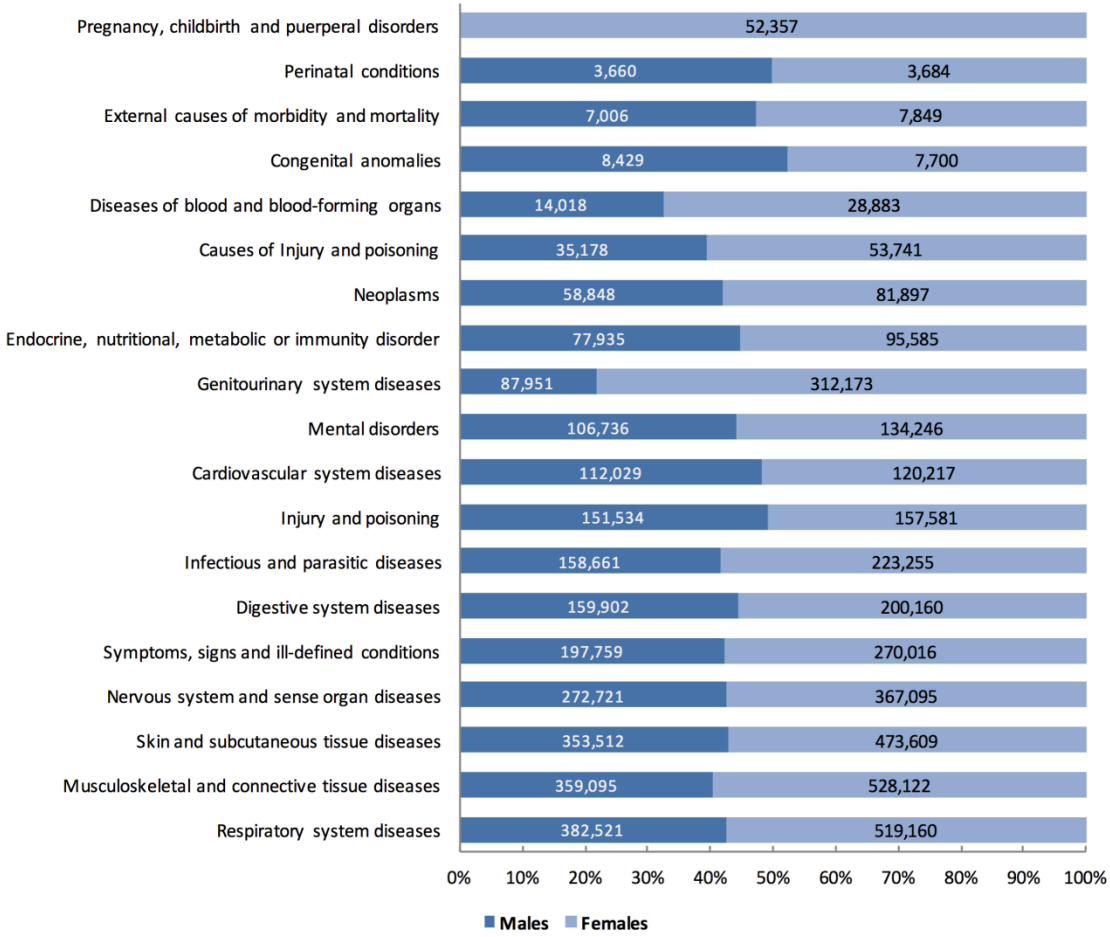


Figure B. 1. The distribution of Read code chapters in the data. This shows the number of patients who have been diagnosed with any of the Read code chapters.

B.2. The distribution of Read chapters based on the 4 ages of male and female patients

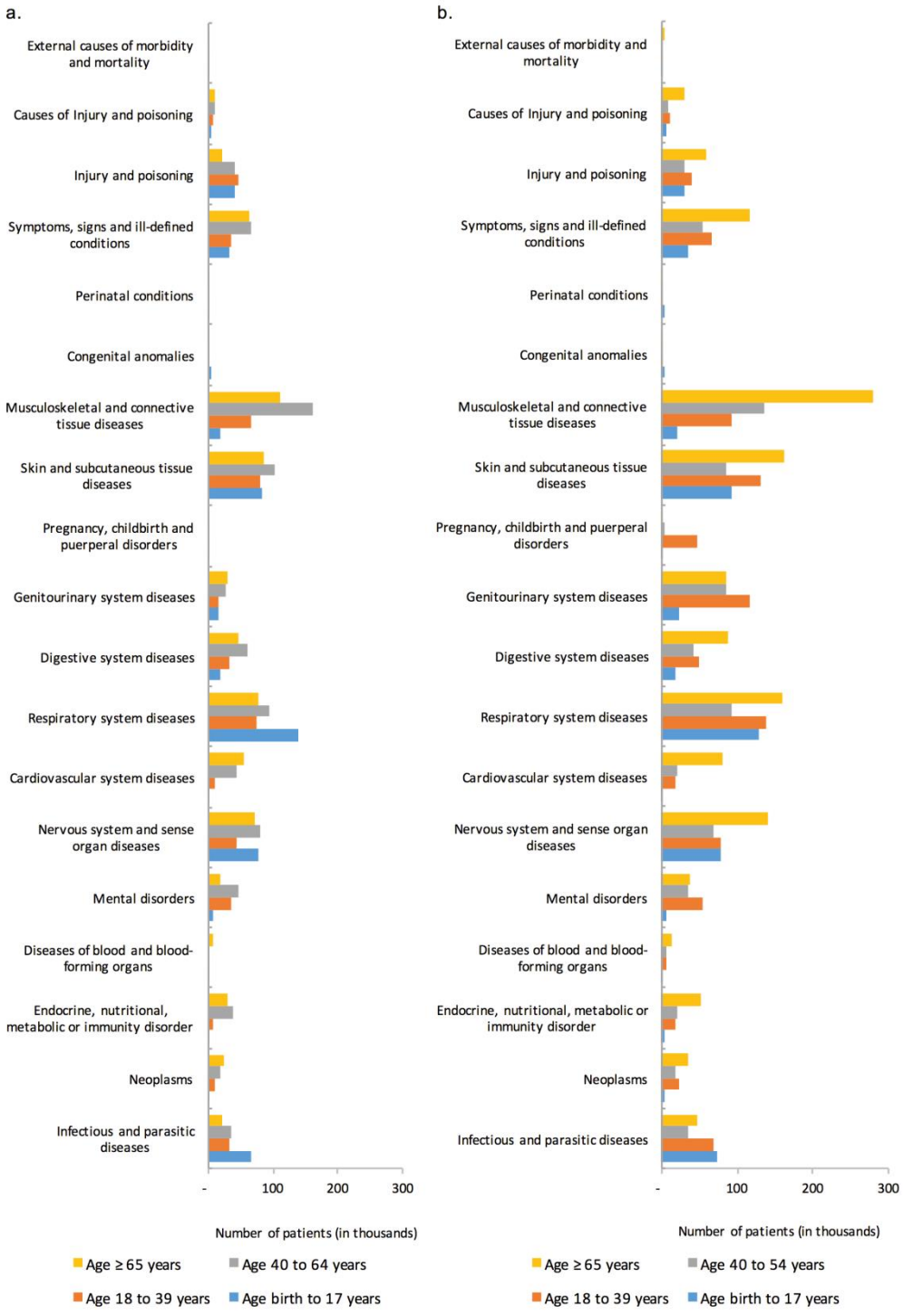


Figure B. 2. The distribution of Read code chapters based on the four ages of male and female patients. The figure shows the number of (a) male and (b) female patients who have been diagnosed with any of the Read code chapters.

Appendix C: Falls in the very elderly

C.1. Falls code in Read codes system

Table C.1. List of Read codes to diagnose accidental falls. These codes are taken from the Read code system provided by UK Terminology Centre in the Health & Social Care Information Centre (HSCIC).

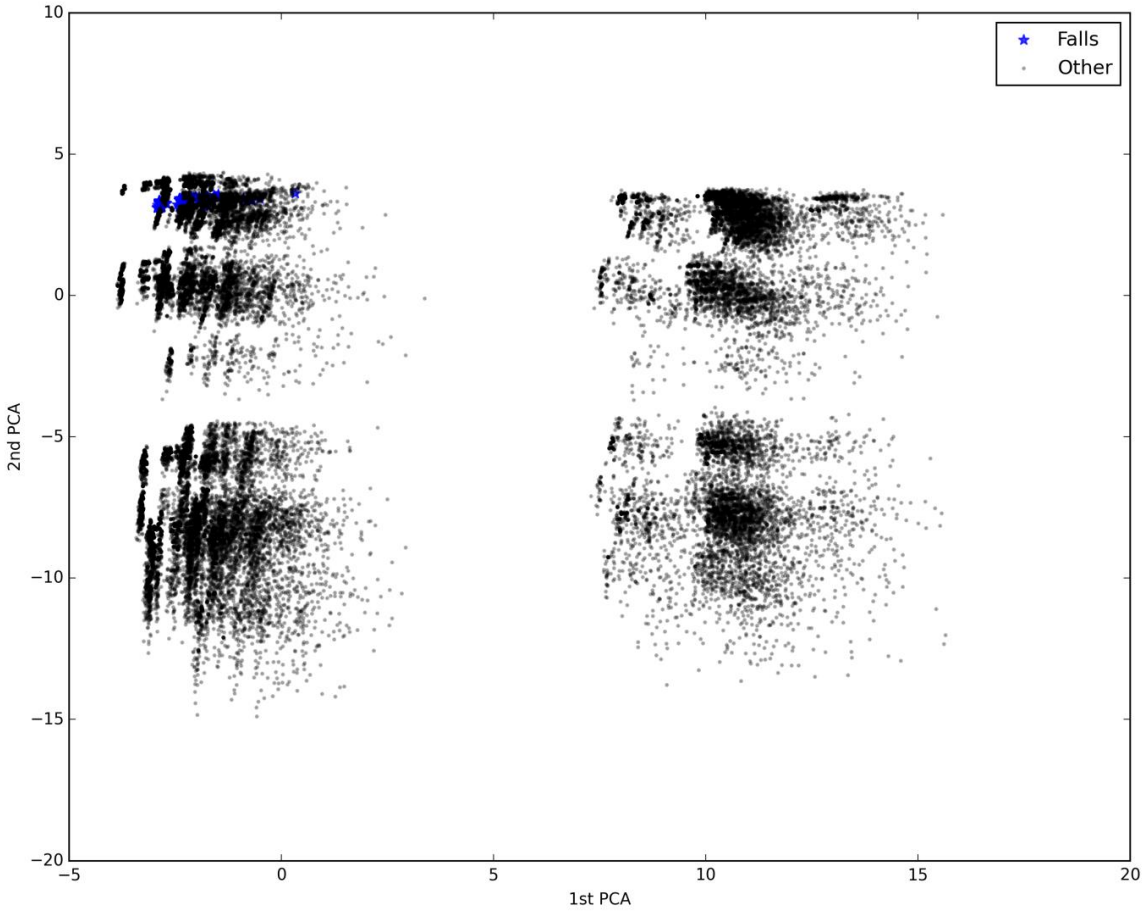
Read codes	Title
TC...	Accidental falls
TC0..	Fall on or from stairs or steps
TC00.	Fall on or from escalator
TC000	Fall on escalator
TC001	Fall from escalator
TC00z	Fall on or from escalator NOS
TC01.	Fall on or from stairs
TC010	Fall on stairs
TC011	Fall from stairs
TC01z	Fall on or from stairs NOS
TC02.	Fall on or from steps
TC020	Fall on steps
TC021	Fall from steps
TC02z	Fall on or from steps NOS
TC0z.	Fall on or from stairs or steps NOS
TC5..	Fall on same level from slipping, tripping or stumbling
TC6..	Fall on same level from collision, pushing or shoving, by or with other person
TC7..	Fracture, cause unspecified
TCy..	Other falls
TCyz.	Other accidental fall NOS
TCz..	Accidental falls NOS
TC1..	Fall on or from ladders or scaffolding
TC10.	Fall from ladder
TC11.	Fall from scaffolding
TC1z.	Fall from ladder or scaffolding NOS
TC2..	Fall from or out of building or other structure
TC21.	Fall from bridge
TC22.	Fall from building
TC20.	Fall from balcony
TC24.	Fall from tower
TC25.	Fall from turret
TC28.|	Fall from window
TC29.	Fall through roof

TC23.	Fall from flagpole
TC26.	Fall from viaduct
TC27.	Fall from wall
TC2z.	Fall from or out of building or other structure NOS
TC40.	Fall from playground equipment
TC4y0	Fall from embankment
TC4y1	Fall from haystack
TC4y2	Fall from stationary vehicle
TC3..	Fall into hole or other opening in surface
TC30.	Accident caused by diving or jumping into water
TC300	Hit against bottom when diving into shallow water
TC301	Hit against bottom when jumping into shallow water
TC302	Hit wall of swimming pool
TC303	Hit board of swimming pool
TC304	Accident caused by hitting water surface
TC305	Accident caused by fall into swimming pool
TC30z	Accident caused by diving or jumping into water NOS
TC31.	Fall into well
TC32.	Fall into manhole
TC320	Accidental fall into manhole, unspecified
TC321	Accidental fall into storm drain
TC32z	Accidental fall into manhole NOS
TC3y.	Fall into other hole or other opening in surface
TC3y0	Fall into cavity, unspecified
TC3y1	Fall into dock
TC3y2	Fall into hole
TC3y3	Fall into pit
TC3y4	Fall into quarry
TC3y5	Fall into shaft
TC3y6	Fall into tank
TC3yz	Fall into other hole, unspecified
TC3z.	Fall into hole NOS
TC4y.	Other fall from one level to another
TC4y3	Fall from tree
TC4yz	Other fall from one level to another NOS
TC4z.	Fall from one level to another NOS
TC41.	Fall from cliff
TC42.	Fall from chair or bed
TC420	Fall from chair
TC421	Fall from bed
TC42z	Fall from chair or bed NOS
TC50.	Fall on same level from slipping
TC5z.	Fall on same level from slipping, tripping or stumbling NOS

TC51.	Fall on same level from tripping
TC52.	Fall on same level from stumbling
TC53.	Fall on moving sidewalk
TC6y.	Fall on same level from other pushing, shoving or collision, with or by other person
TC6y0	Fall on same level from collision with other person, unspecified
TC6y1	Fall on same level from pushing by other person, unspecified
TC6y2	Fall on same level from shoving by other person, unspecified
TC6yz	Other fall on same level from pushing, shoving or collision, with or by other person, NOS
TC6z.	Fall on same level from pushing, shoving or collision, with or by other person
TC60.	Fall on same level from sports contact
TC600	Fall from tackle in sport
TC60y	Other fall in sport
TC60z	Fall on same level from sports contact NOS
TCy0.	Fall from bump against object

C.2. Clusters analysis

A.



B.

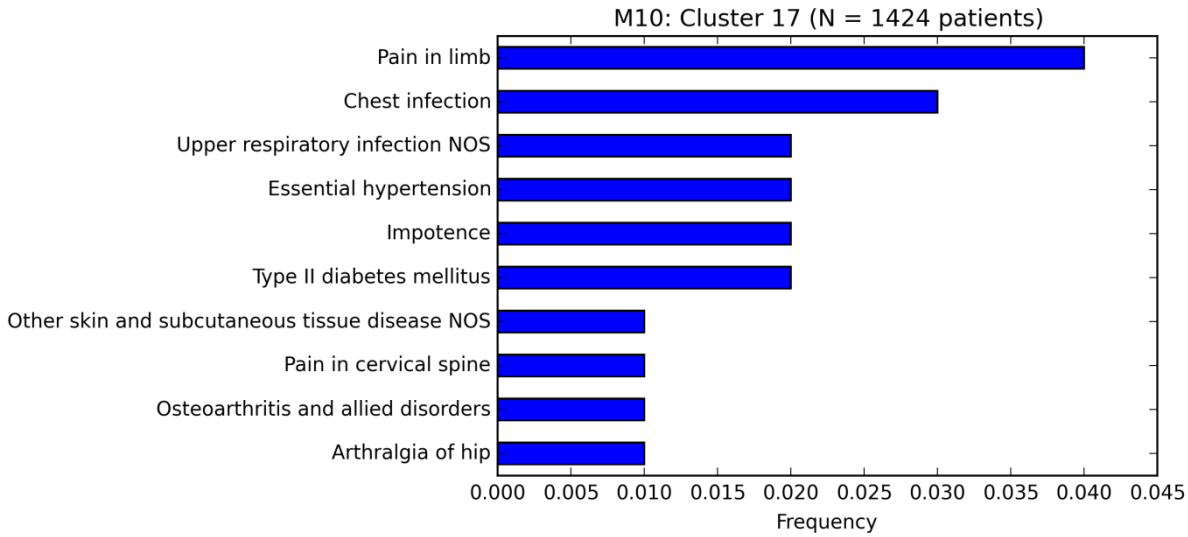
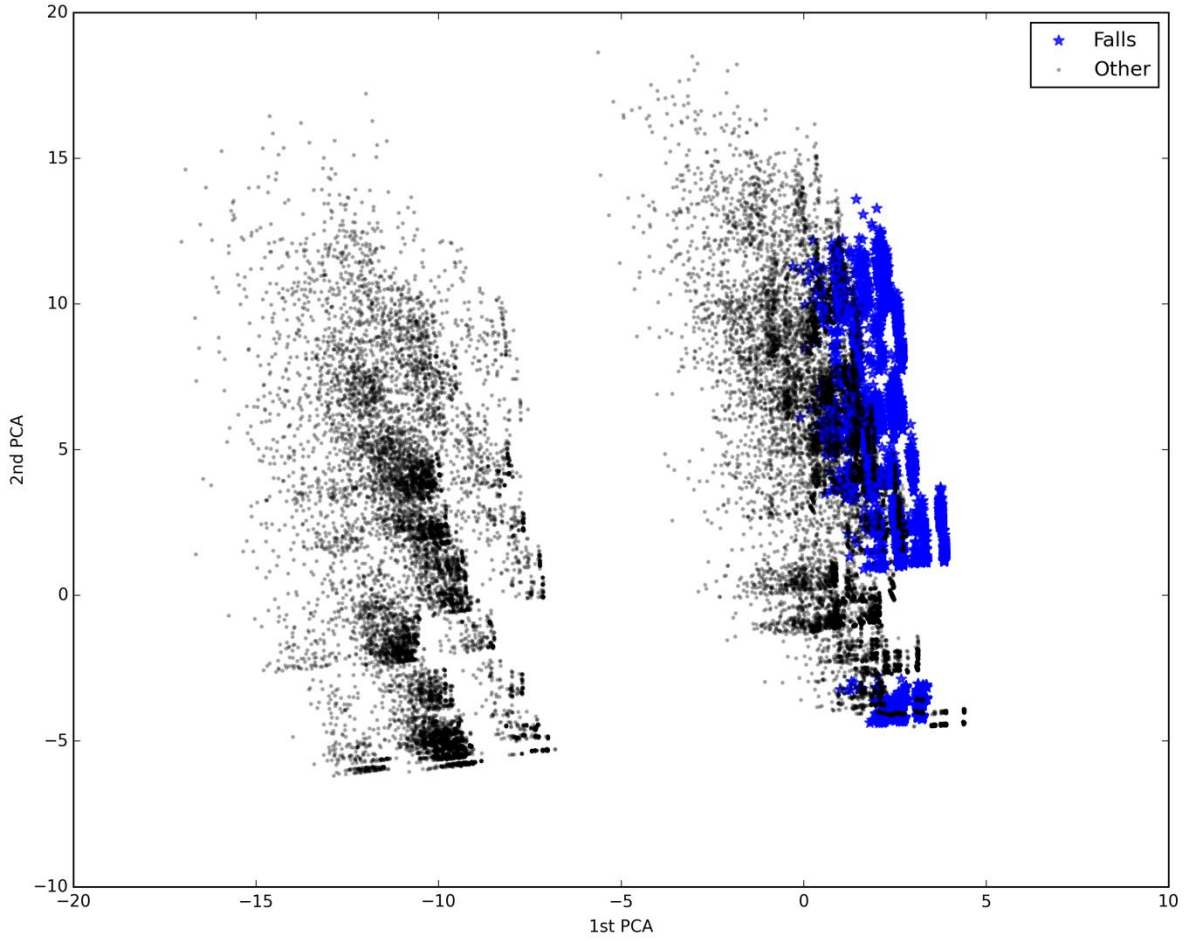


Figure C.1. Clusters analysis for men patients aged between 65 to 69 years (n=75,733) based on semantic similarity. A. Clusters enriched of falls (in blue/ stars). B. The top ten diseases appear in each cluster enriched of falls.

A.



B.

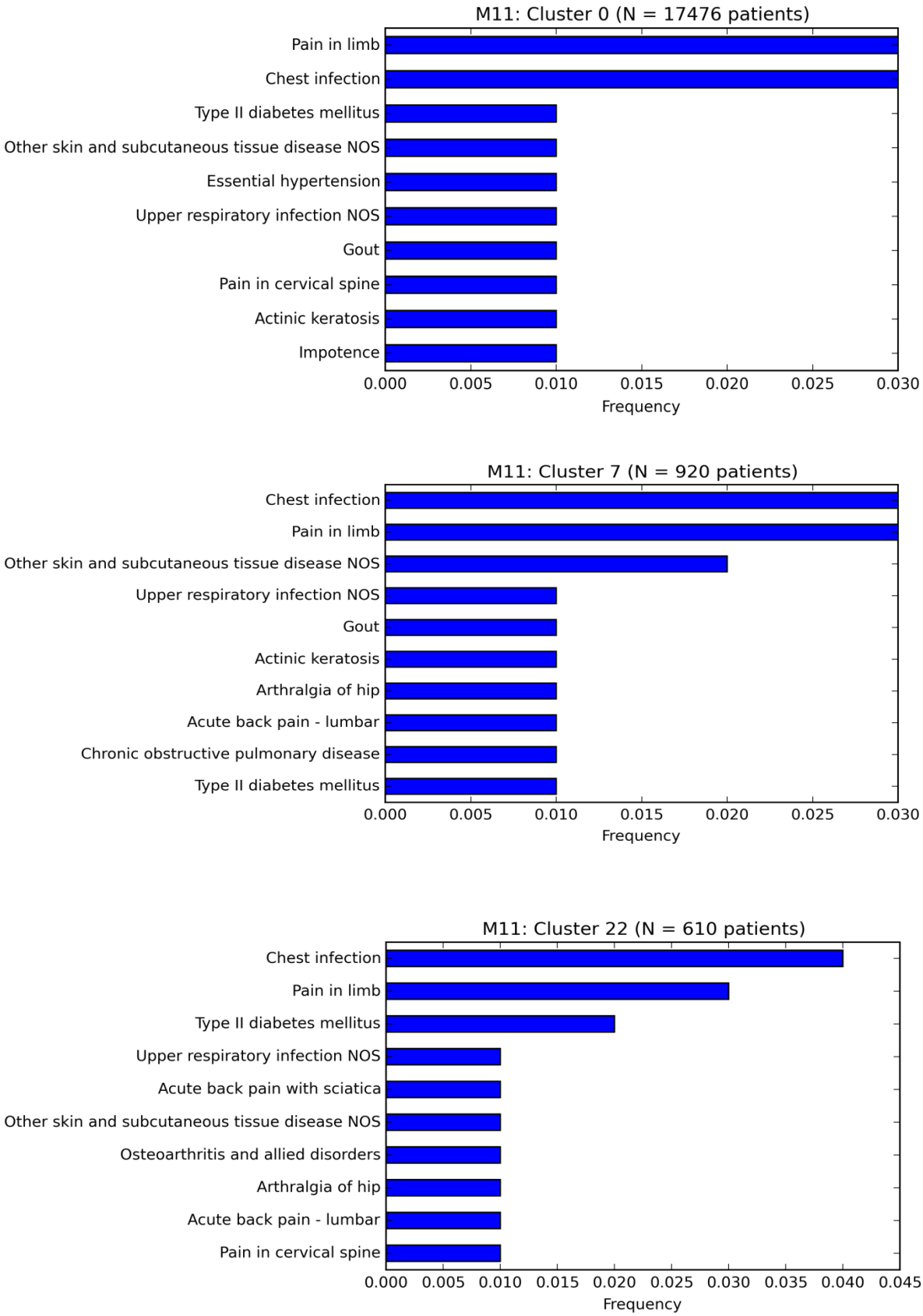
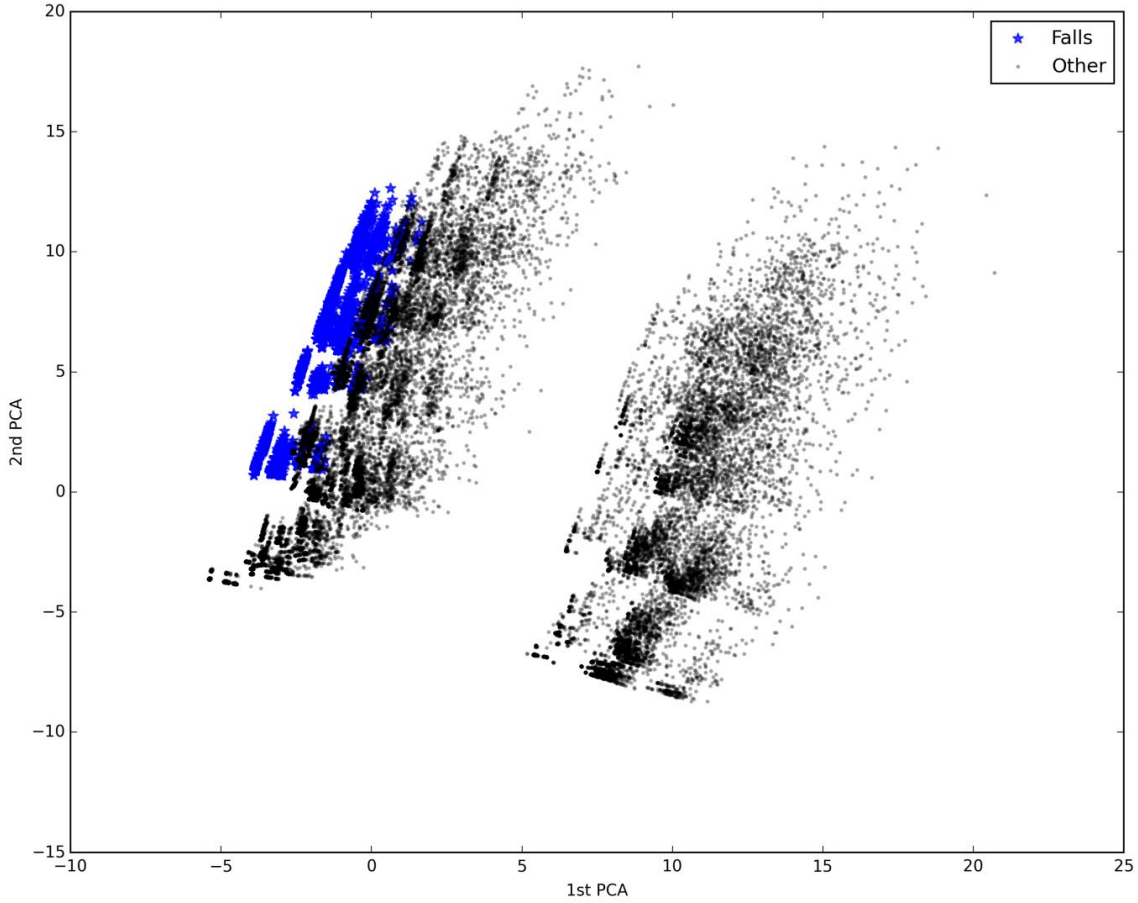


Figure C.2. Clusters analysis for men patients aged between 70 to 74 years (n=59,795) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (B) The top ten diseases appear in each cluster enriched of falls.

A.



B.

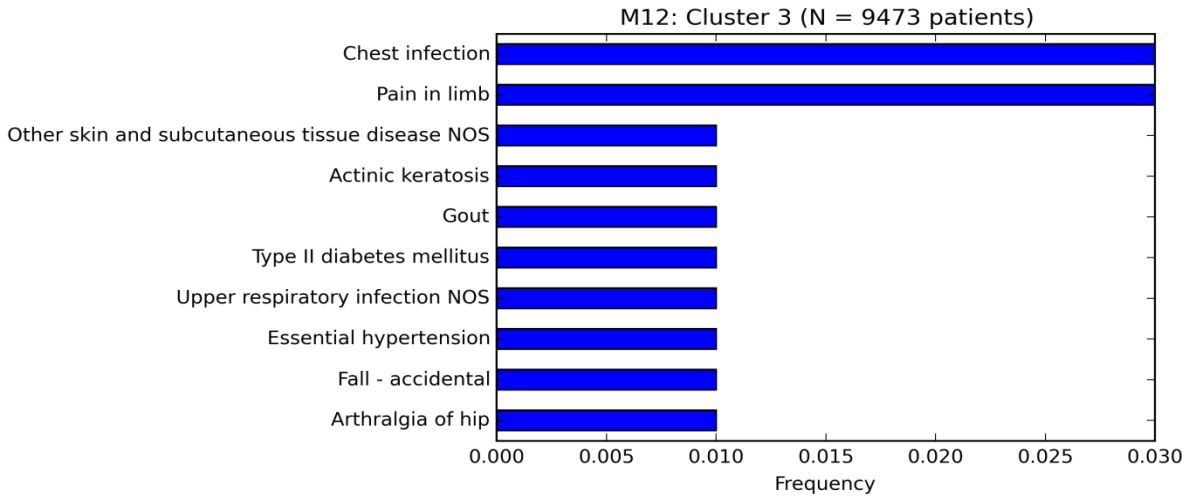
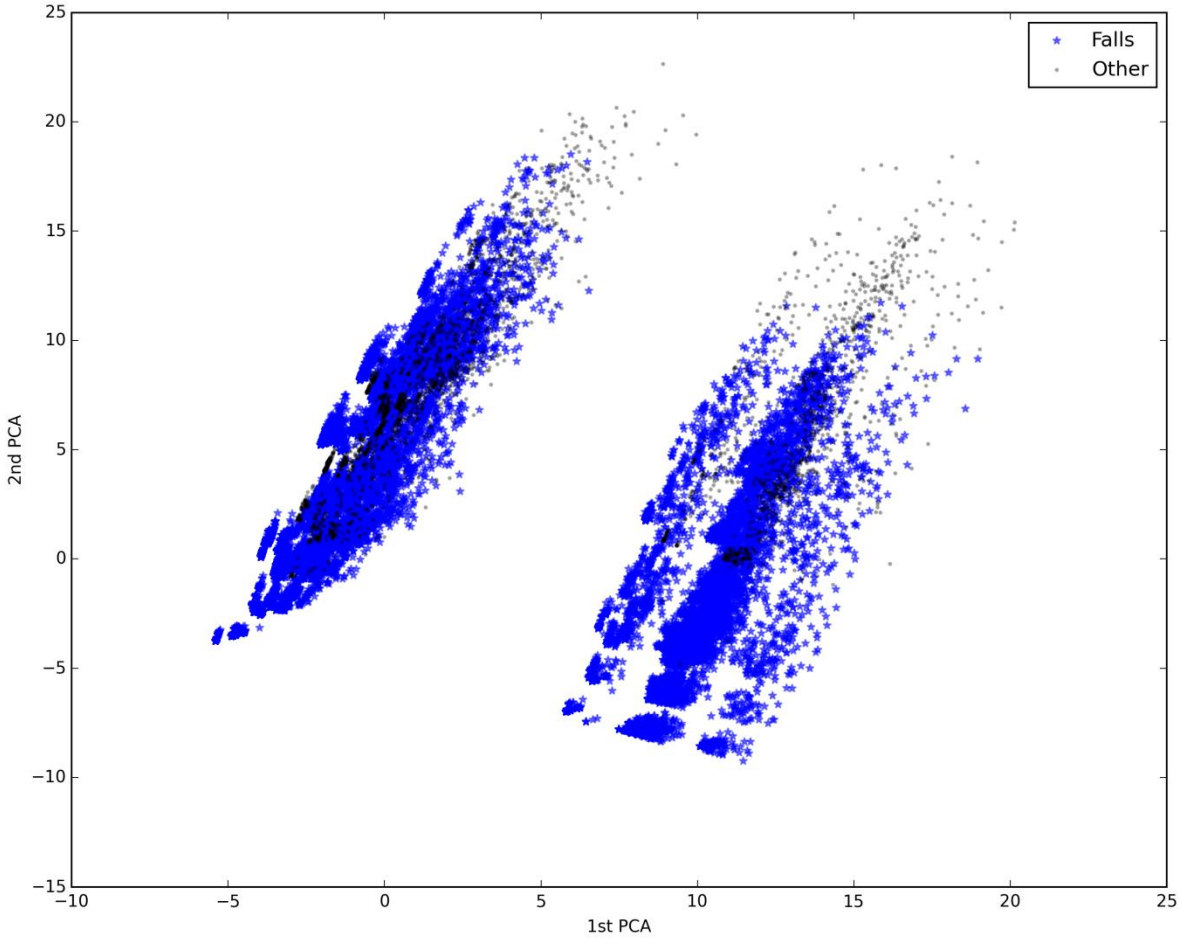
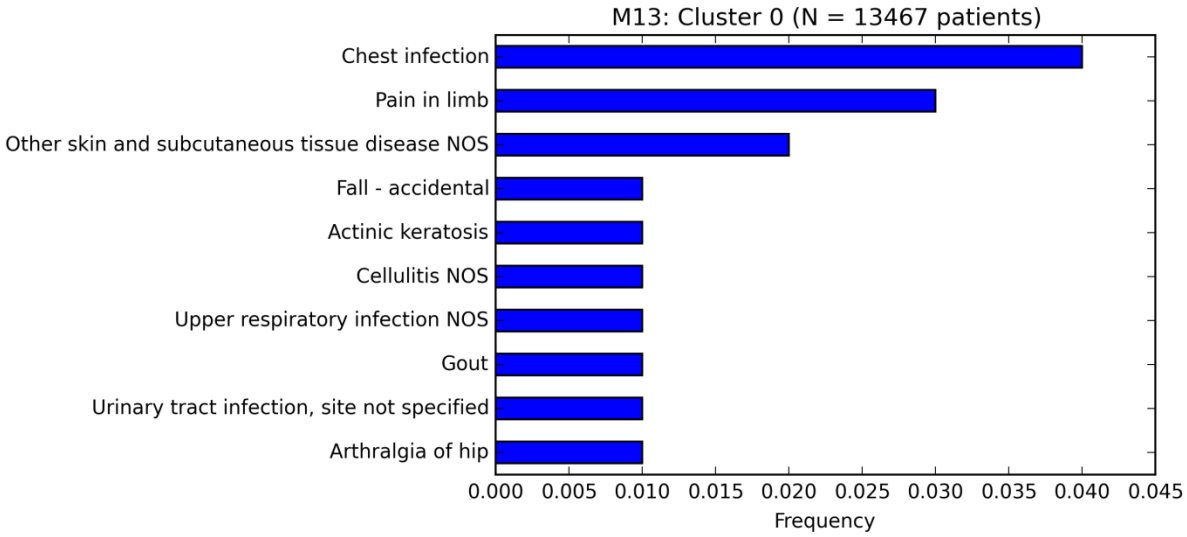


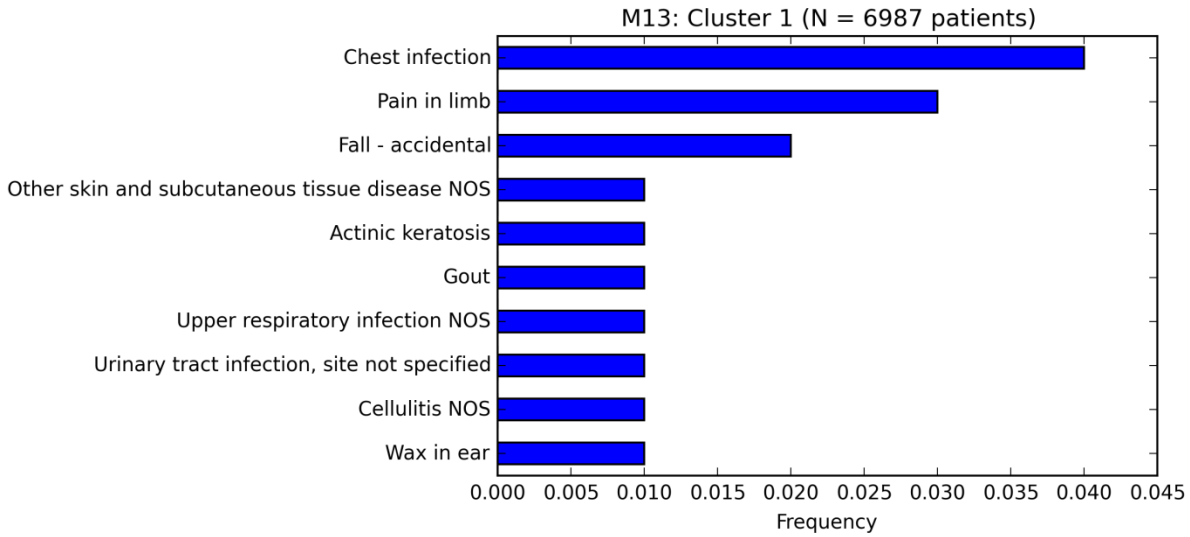
Figure C.3. Clusters analysis for men patients aged between 75 to 79 years (n=50,942) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (B) The top ten diseases appear in each cluster enriched of falls.

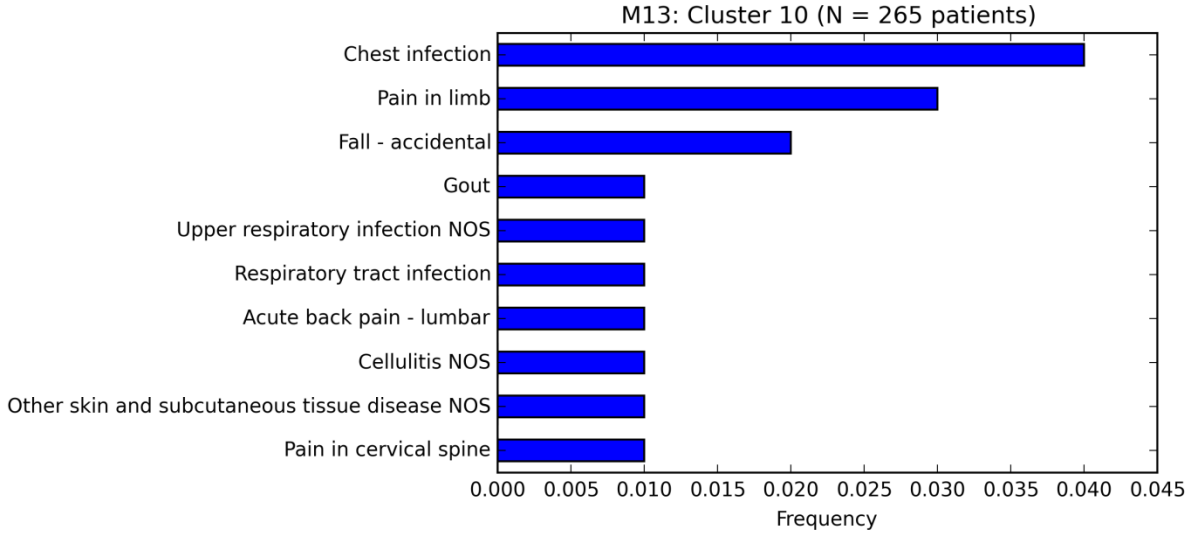
A.

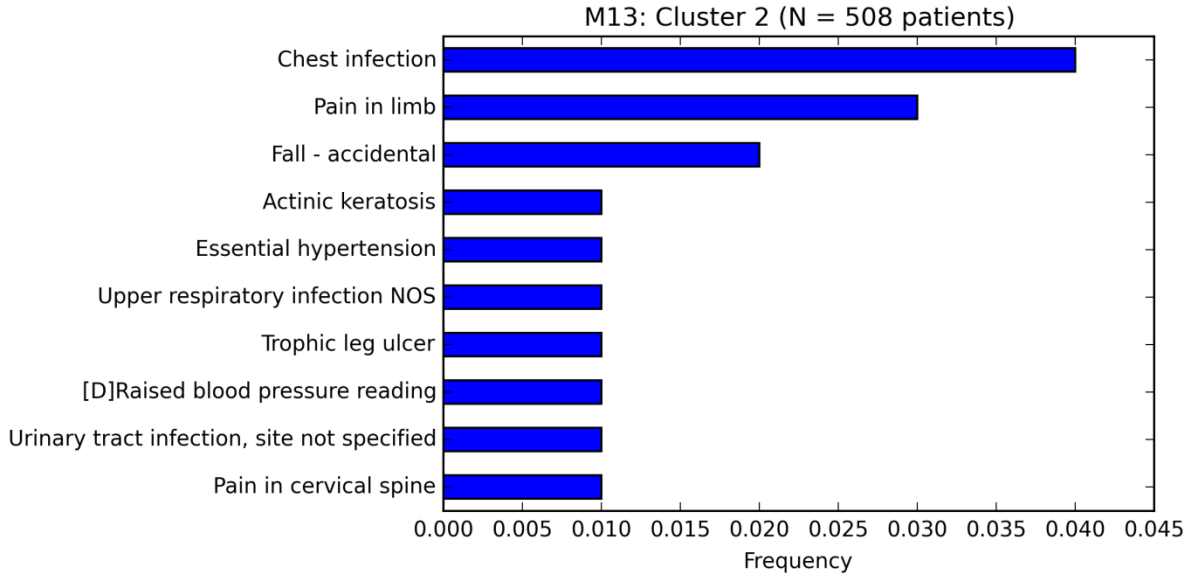

Falls
Other
2nd PCA
1st PCA
25
20
15
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
20
25

B.


M13: Cluster 0 (N = 13467 patients)
Chest infection
Pain in limb
Other skin and subcutaneous tissue disease NOS
Fall - accidental
Actinic keratosis
Cellulitis NOS
Upper respiratory infection NOS
Gout
Urinary tract infection, site not specified
Arthralgia of hip
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0.045
Frequency


M13: Cluster 1 (N = 6987 patients)
Chest infection
Pain in limb
Fall - accidental
Other skin and subcutaneous tissue disease NOS
Actinic keratosis
Gout
Upper respiratory infection NOS
Urinary tract infection, site not specified
Cellulitis NOS
Wax in ear
0.000 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045
Frequency


M13: Cluster 10 (N = 265 patients)
Chest infection
Pain in limb
Fall - accidental
Gout
Upper respiratory infection NOS
Respiratory tract infection
Acute back pain - lumbar
Cellulitis NOS
Other skin and subcutaneous tissue disease NOS
Pain in cervical spine
0.000 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045
Frequency


M13: Cluster 2 (N = 508 patients)
Chest infection
Pain in limb
Fall - accidental
Actinic keratosis
Essential hypertension
Upper respiratory infection NOS
Trophic leg ulcer
[D]Raised blood pressure reading
Urinary tract infection, site not specified
Pain in cervical spine
0.000 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045
Frequency

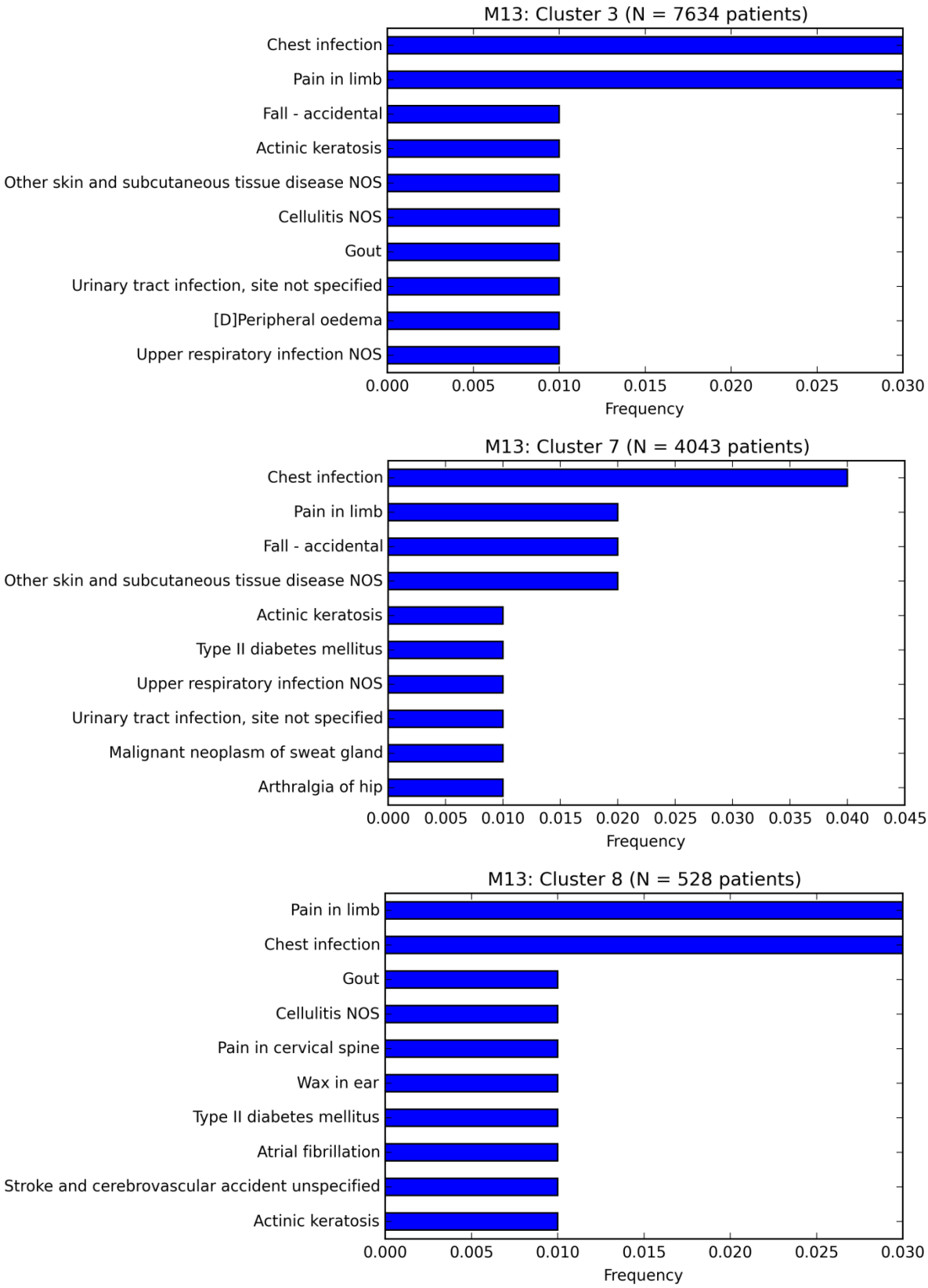
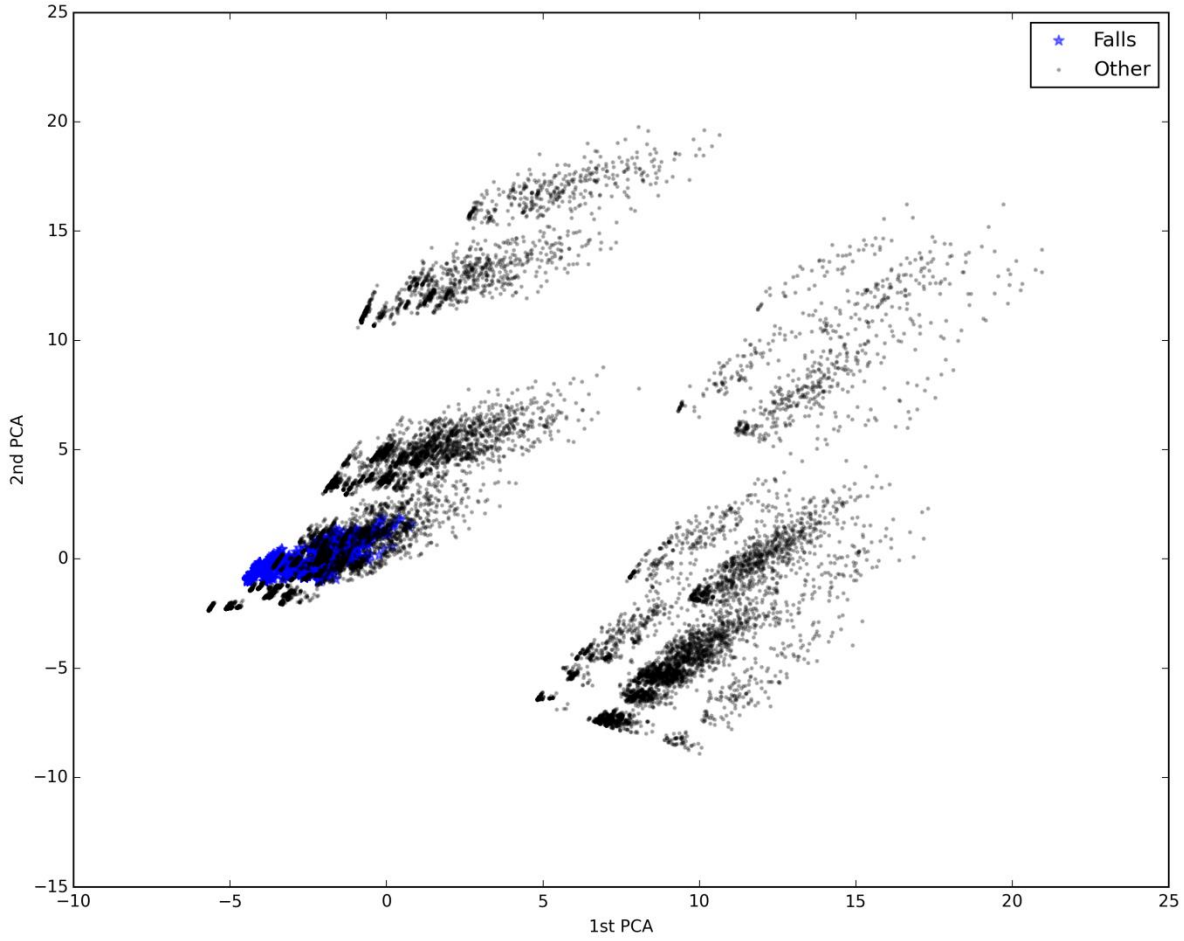
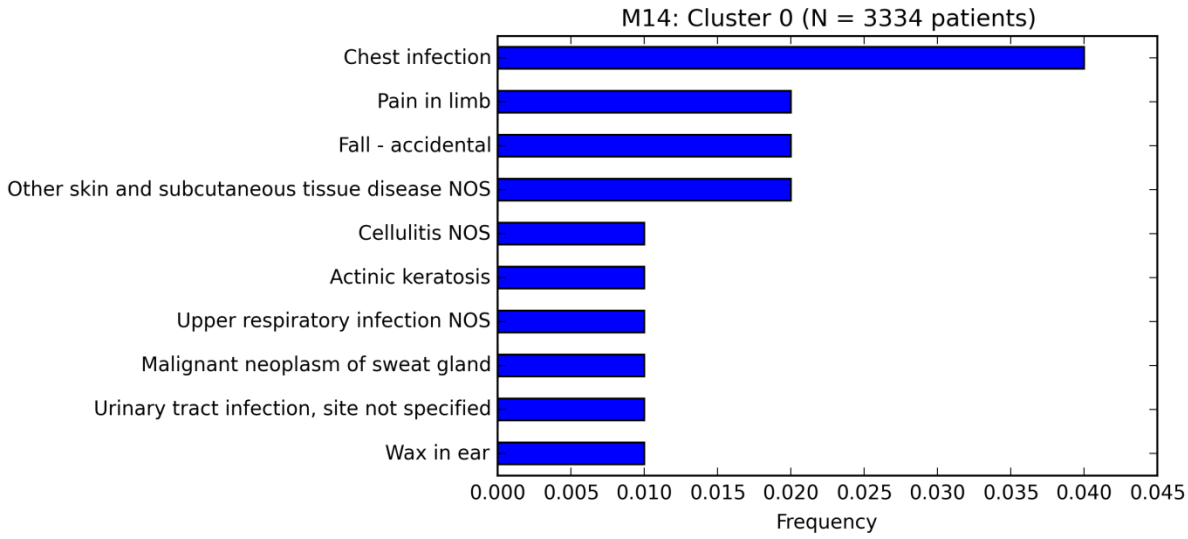


Figure C.4. Clusters analysis for men patients aged between 80 to 84 years (n=36,730) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (B) The top ten diseases appear in each cluster enriched of falls.

A.


Falls
Other
2nd PCA
1st PCA
25
20
15
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
20
25

B.


M14: Cluster 0 (N = 3334 patients)
Chest infection
Pain in limb
Fall - accidental
Other skin and subcutaneous tissue disease NOS
Cellulitis NOS
Actinic keratosis
Upper respiratory infection NOS
Malignant neoplasm of sweat gland
Urinary tract infection, site not specified
Wax in ear
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0.045
Frequency

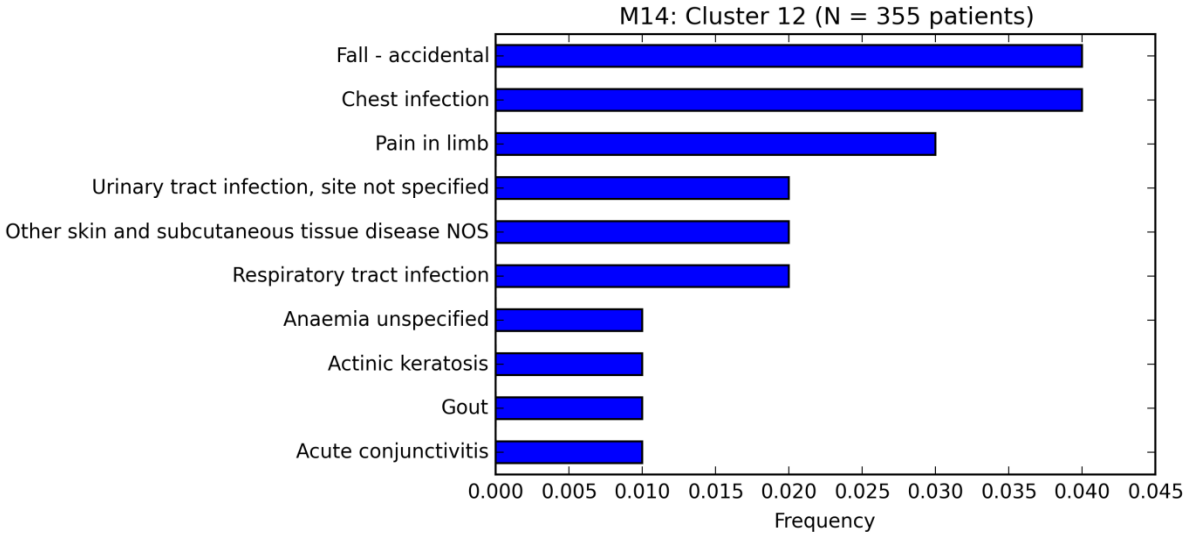
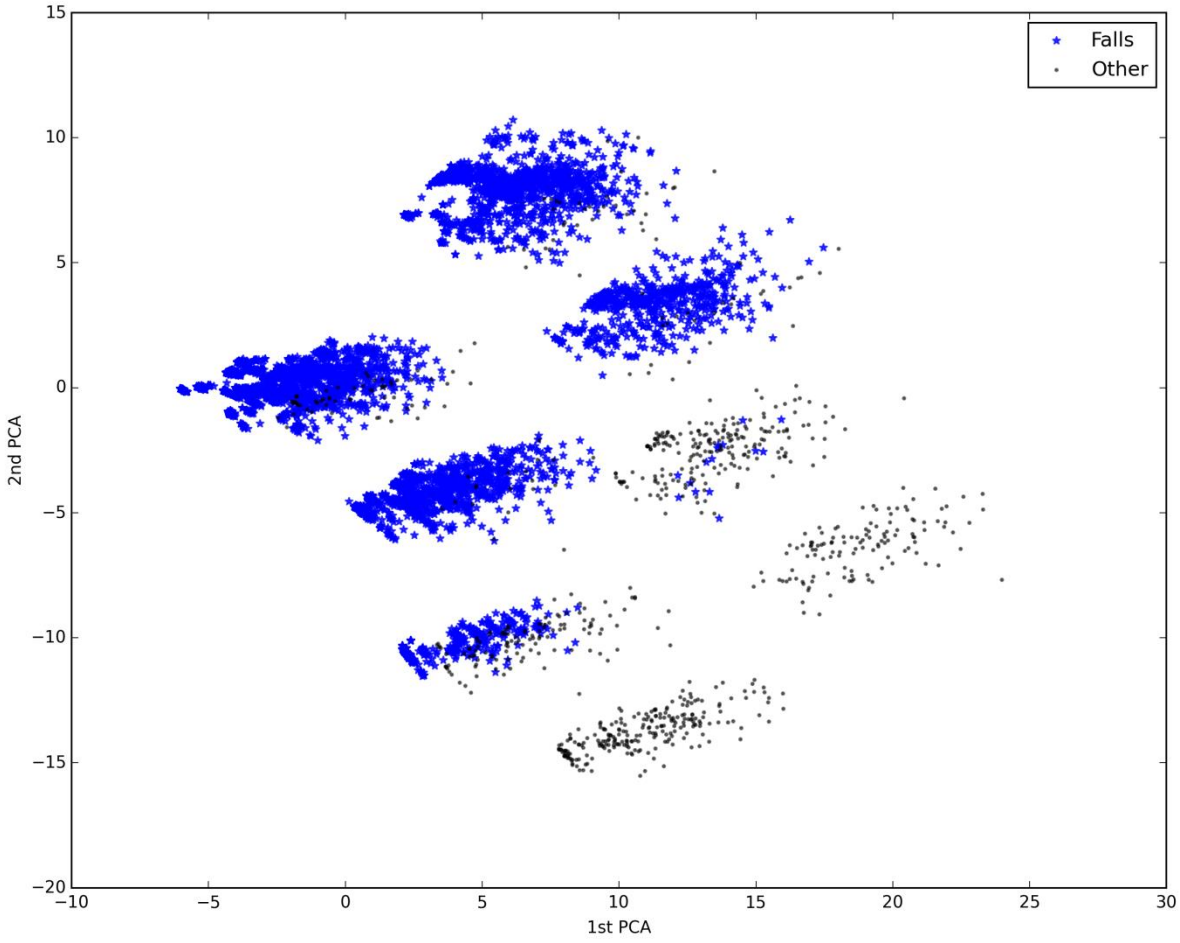
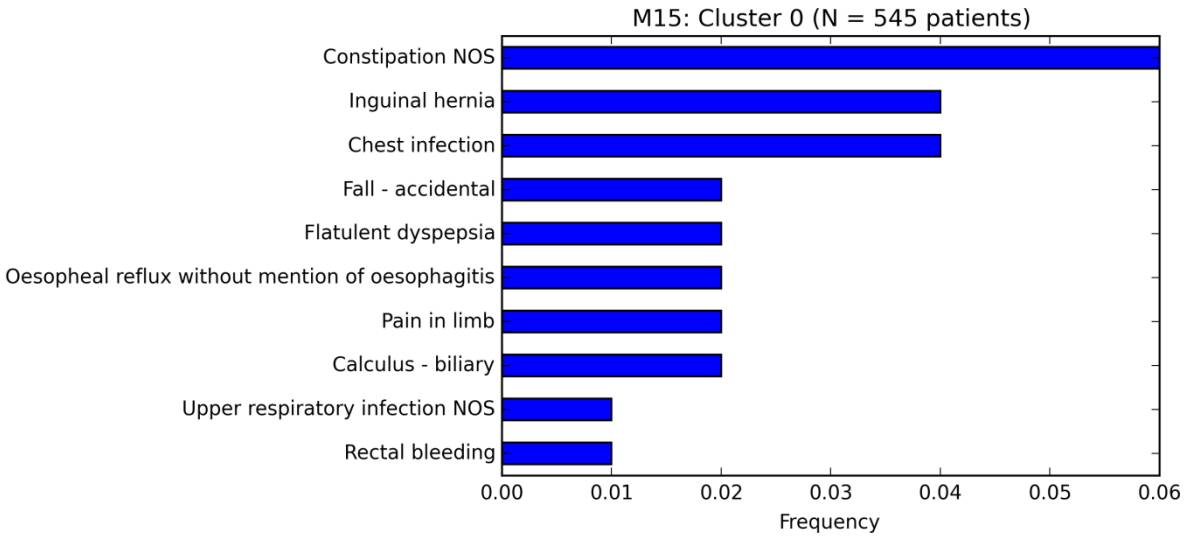


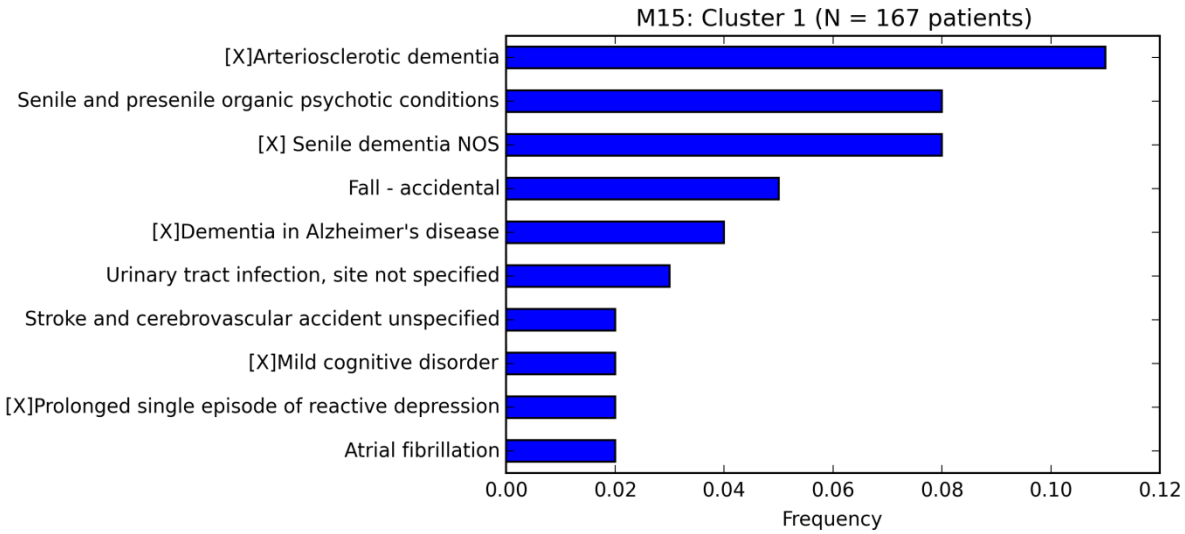
Figure C.5. Clusters analysis for men patients aged between 85 to 89 years (n=21,571) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (B) The top ten diseases appear in each cluster enriched of falls.

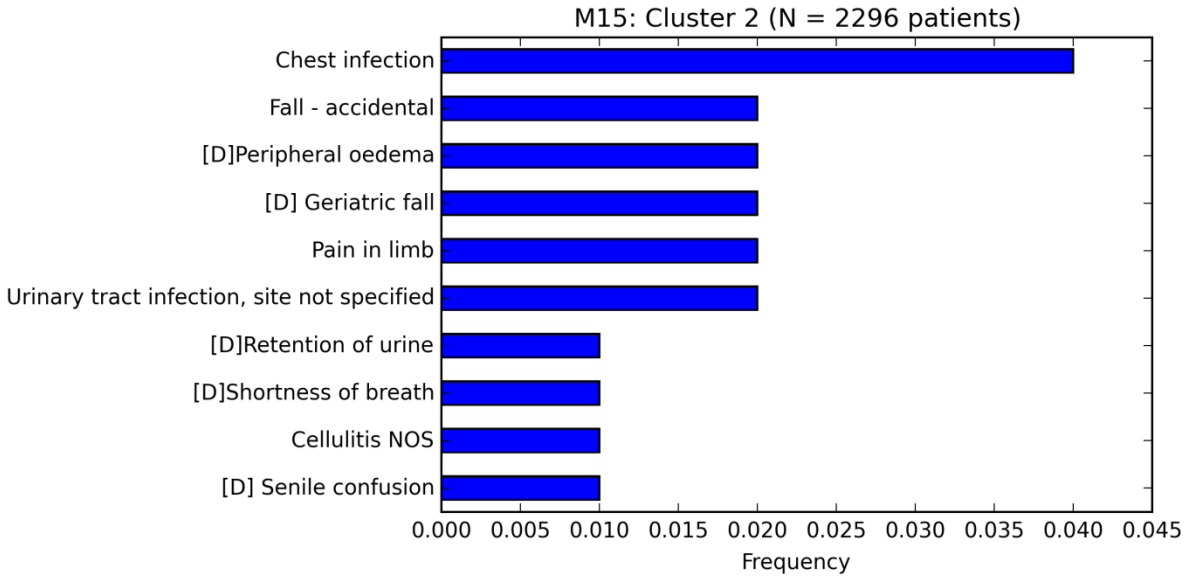
A.

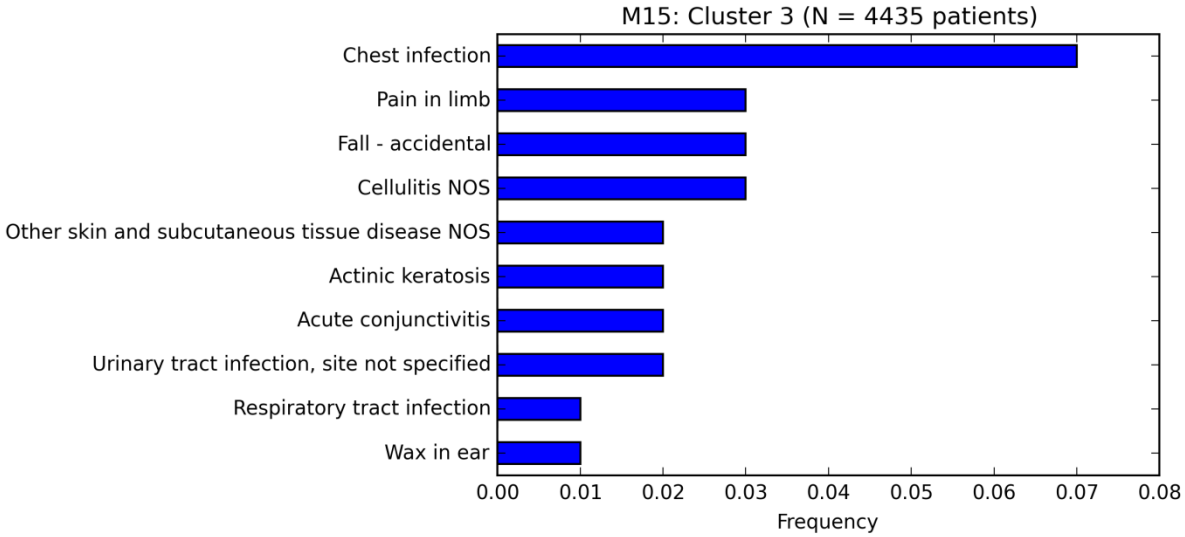


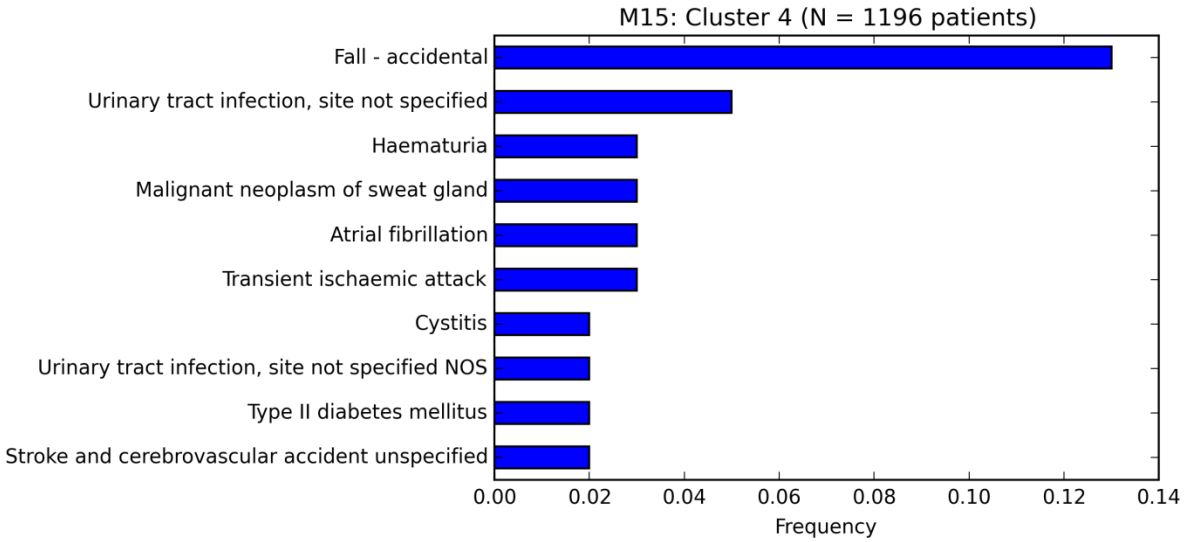
B.

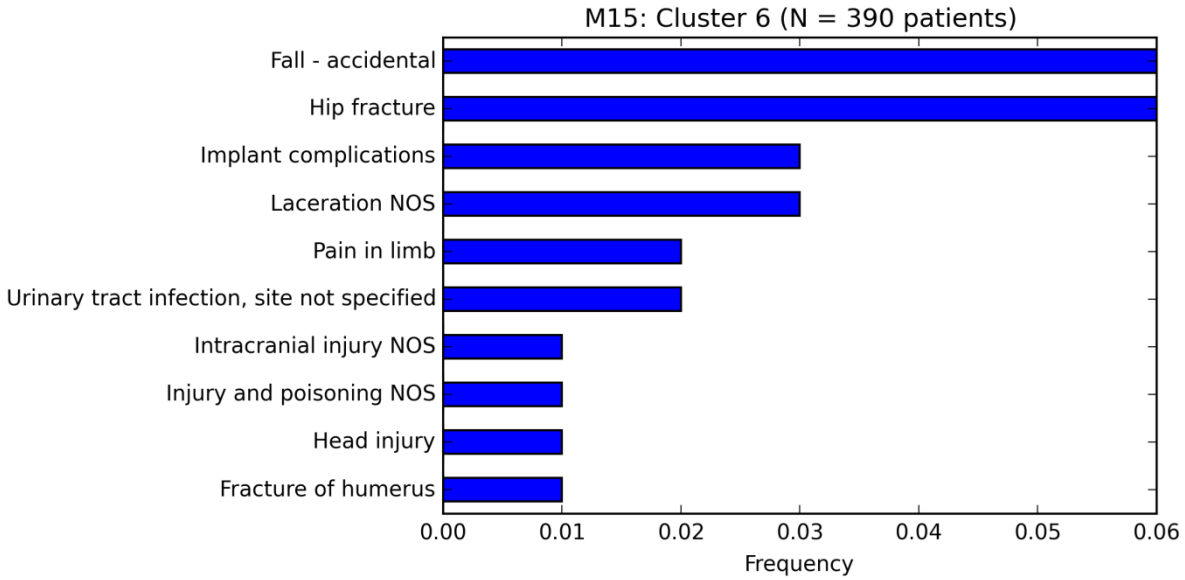

M15: Cluster 0 (N = 545 patients)
Constipation NOS
Inguinal hernia
Chest infection
Fall - accidental
Flatulent dyspepsia
Oesopheal reflux without mention of oesophagitis
Pain in limb
Calculus - biliary
Upper respiratory infection NOS
Rectal bleeding
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Frequency


M15: Cluster 1 (N = 167 patients)
[X]Arteriosclerotic dementia
Senile and presenile organic psychotic conditions
[X] Senile dementia NOS
Fall - accidental
[X]Dementia in Alzheimer's disease
Urinary tract infection, site not specified
Stroke and cerebrovascular accident unspecified
[X]Mild cognitive disorder
[X]Prolonged single episode of reactive depression
Atrial fibrillation
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Frequency


M15: Cluster 2 (N = 2296 patients)
Chest infection
Fall - accidental
[D]Peripheral oedema
[D] Geriatric fall
Pain in limb
Urinary tract infection, site not specified
[D]Retention of urine
[D]Shortness of breath
Cellulitis NOS
[D] Senile confusion
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0.045
Frequency


M15: Cluster 3 (N = 4435 patients)
Chest infection
Pain in limb
Fall - accidental
Cellulitis NOS
Other skin and subcutaneous tissue disease NOS
Actinic keratosis
Acute conjunctivitis
Urinary tract infection, site not specified
Respiratory tract infection
Wax in ear
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
Frequency


M15: Cluster 4 (N = 1196 patients)
Fall - accidental
Urinary tract infection, site not specified
Haematuria
Malignant neoplasm of sweat gland
Atrial fibrillation
Transient ischaemic attack
Cystitis
Urinary tract infection, site not specified NOS
Type II diabetes mellitus
Stroke and cerebrovascular accident unspecified
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
Frequency


M15: Cluster 6 (N = 390 patients)
Fall - accidental
Hip fracture
Implant complications
Laceration NOS
Pain in limb
Urinary tract infection, site not specified
Intracranial injury NOS
Injury and poisoning NOS
Head injury
Fracture of humerus
0.00 0.01 0.02 0.03 0.04 0.05 0.06
Frequency

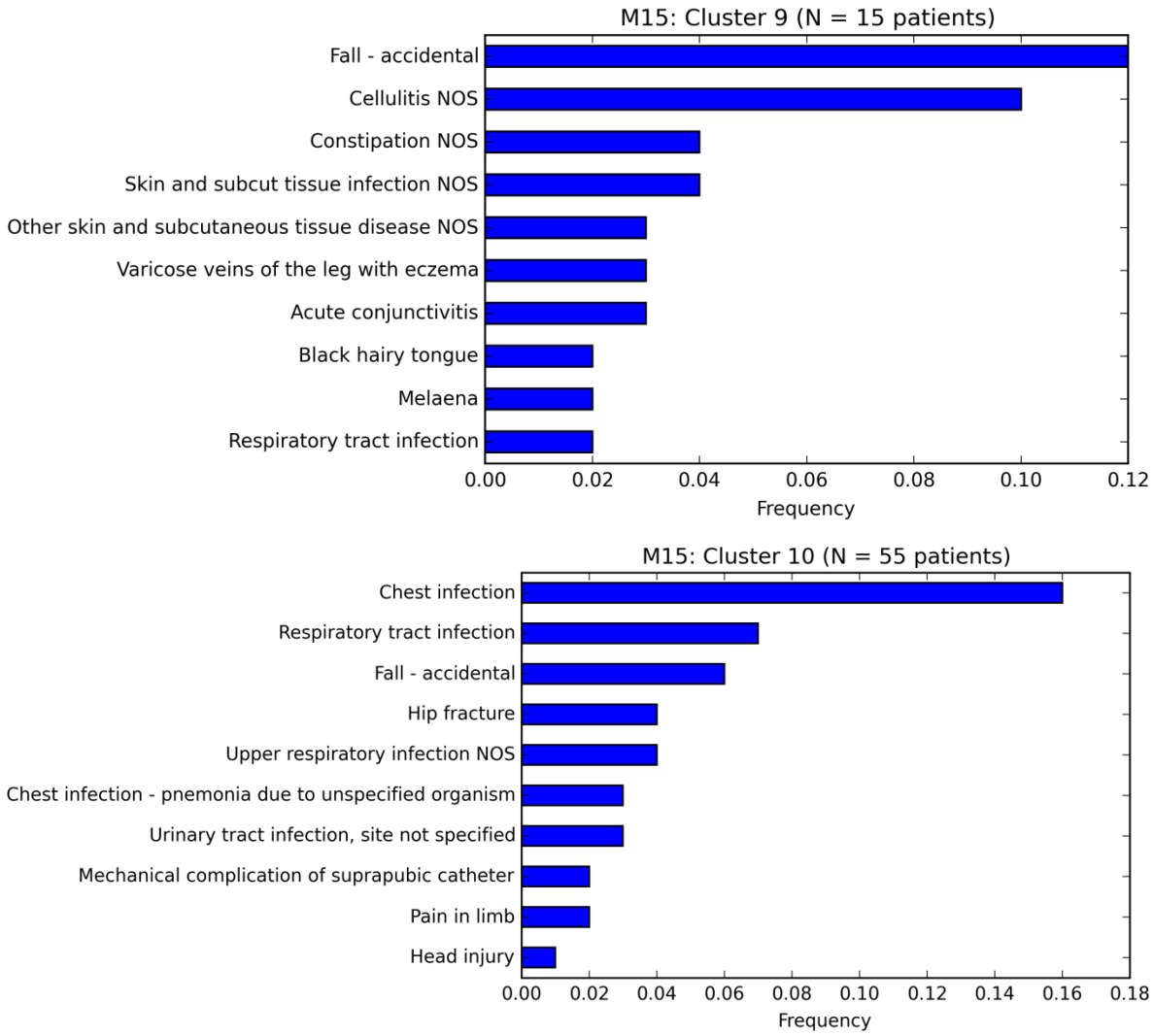
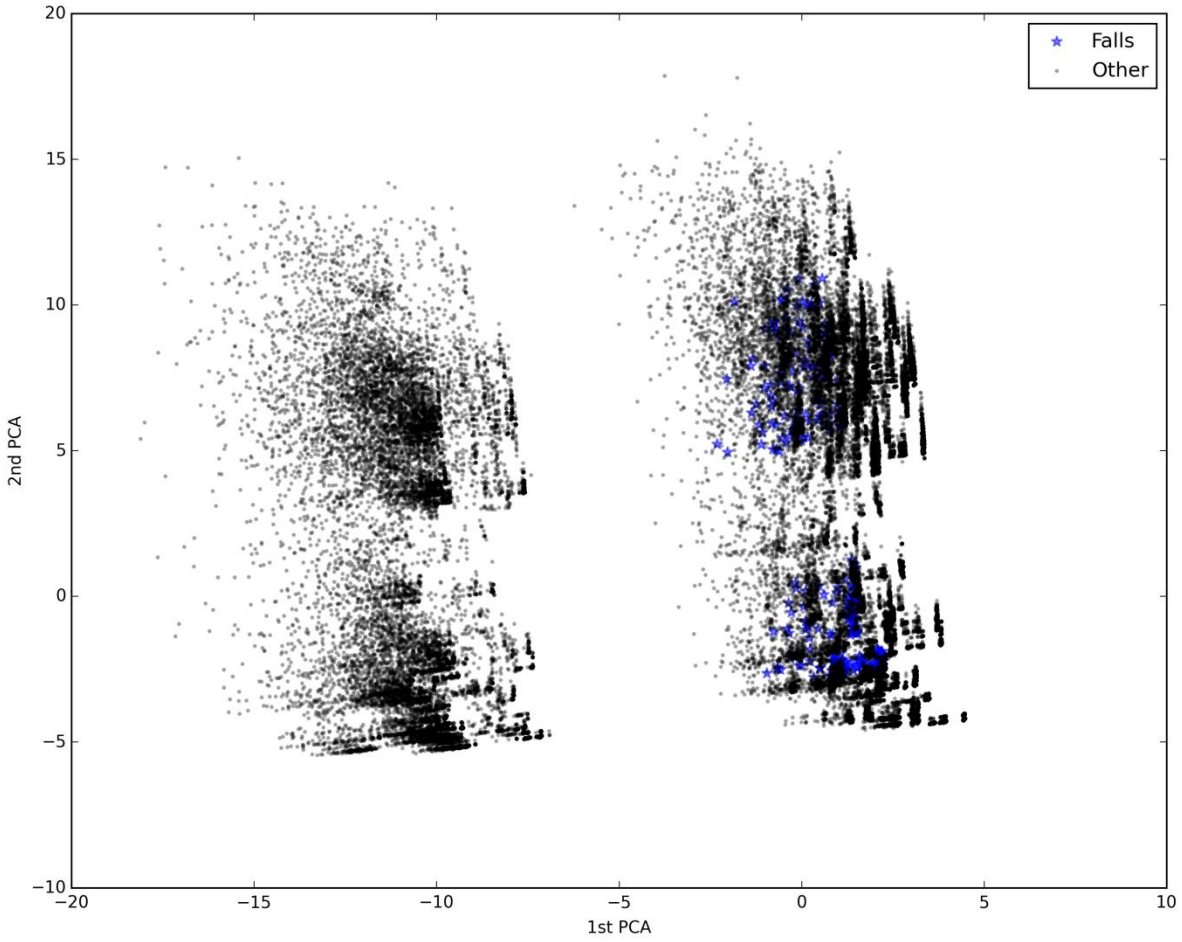


Figure C.6. Clusters analysis for men patients aged above 89 years (n=9,932) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (B) The top ten diseases appear in each cluster enriched of falls.

A.



B.

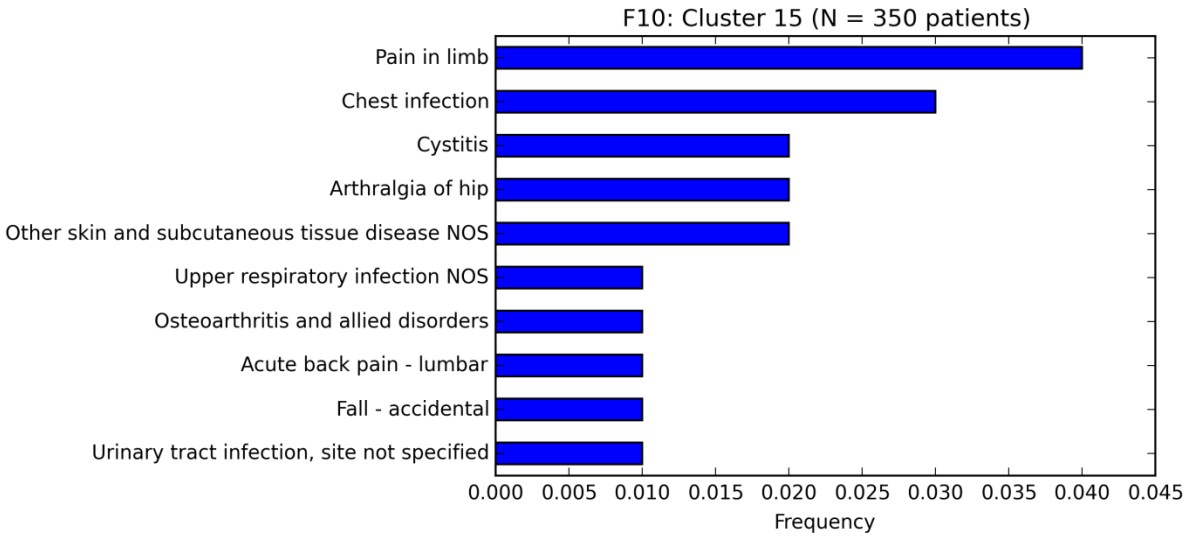
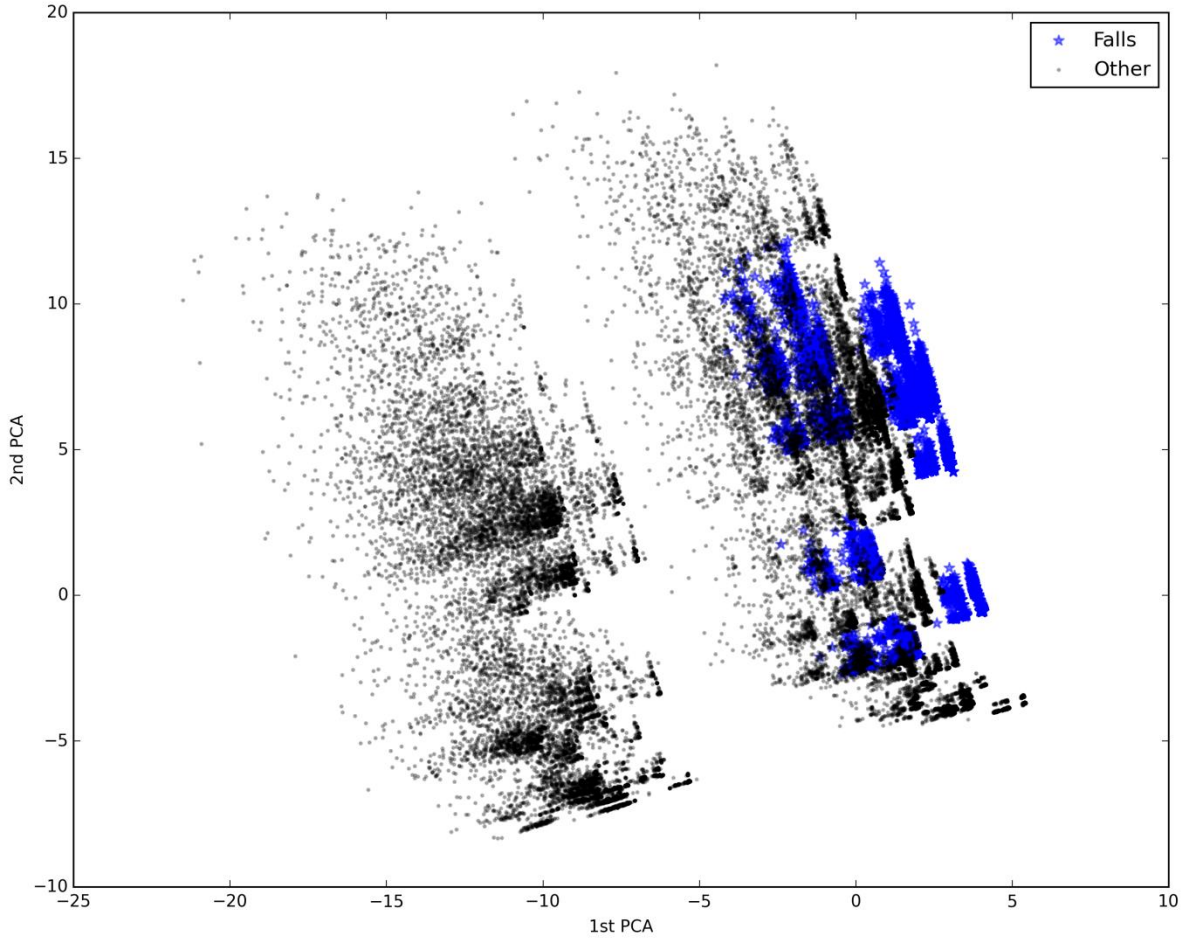
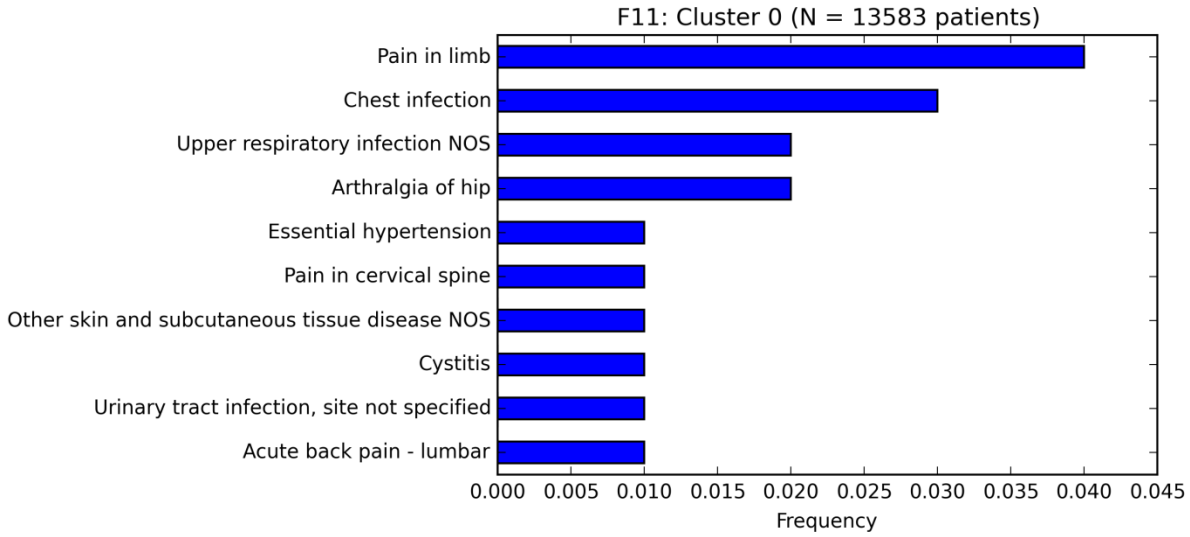


Figure C. 7. Clusters analysis for women patients aged between 65 to 69 years (n=85,381) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (A) The top ten diseases appear in each cluster enriched of falls.

A.


Falls
Other
2nd PCA
1st PCA
20
15
10
5
0
−5
−10
−25
−20
−15
−10
−5
0
5
10

B.


F11: Cluster 0 (N = 13583 patients)
Pain in limb
Chest infection
Upper respiratory infection NOS
Arthralgia of hip
Essential hypertension
Pain in cervical spine
Other skin and subcutaneous tissue disease NOS
Cystitis
Urinary tract infection, site not specified
Acute back pain - lumbar
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0.045
Frequency

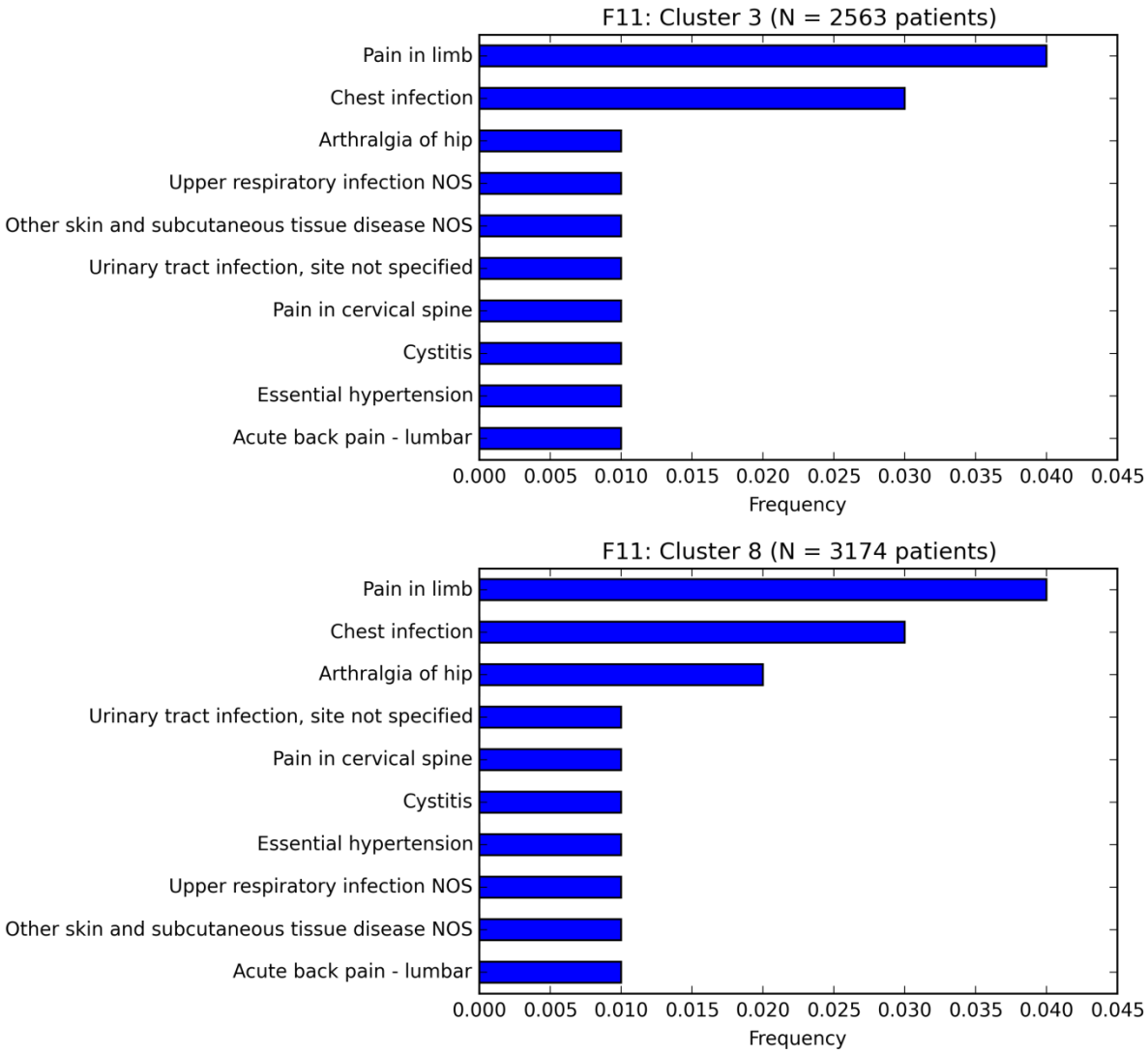
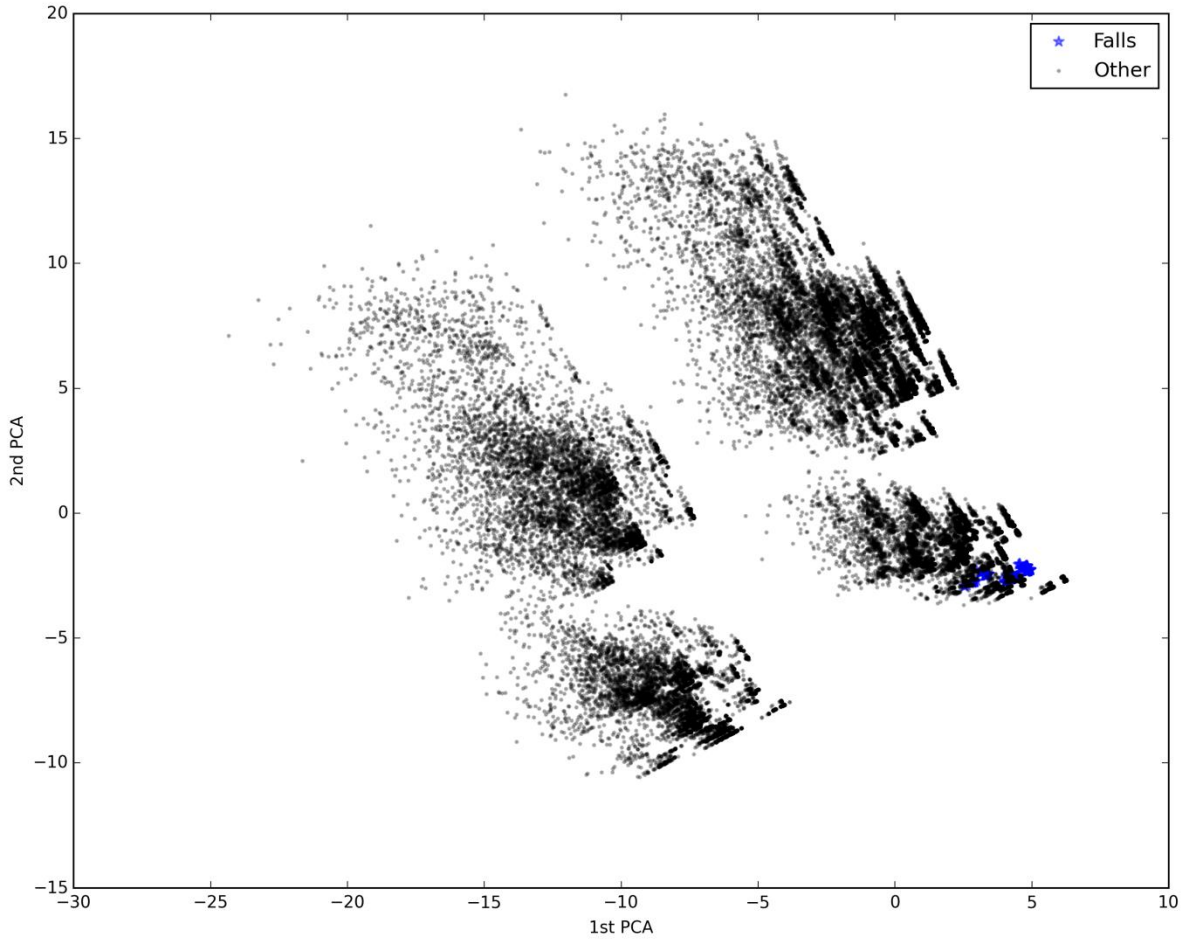


Figure C. 8. Clusters analysis for women patients aged between 70 to 74 years (n=69,938) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (B) The top ten diseases appear in each cluster enriched of falls.

A.



B.

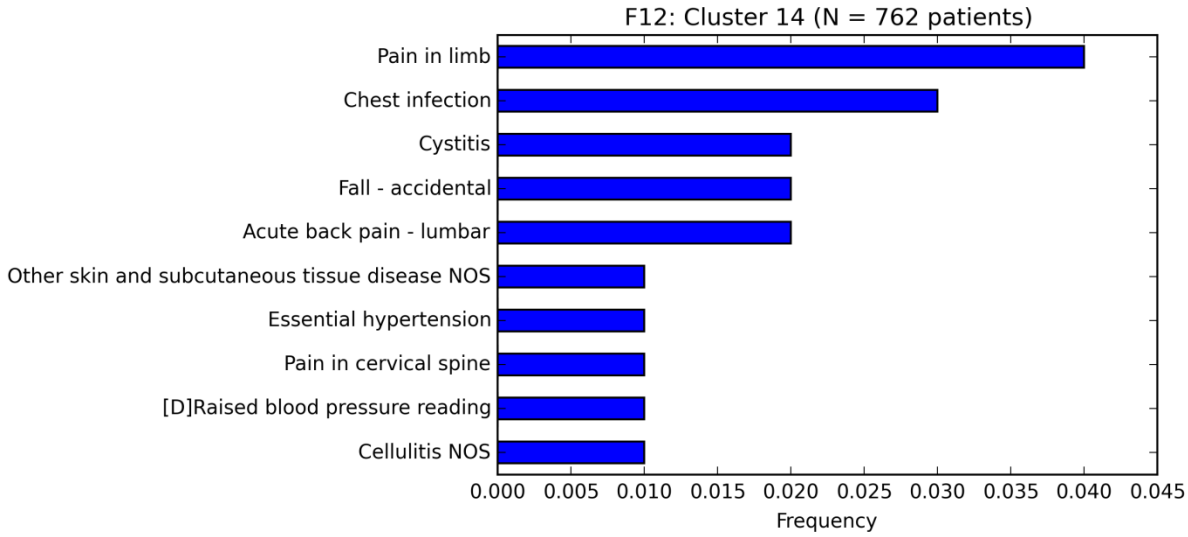
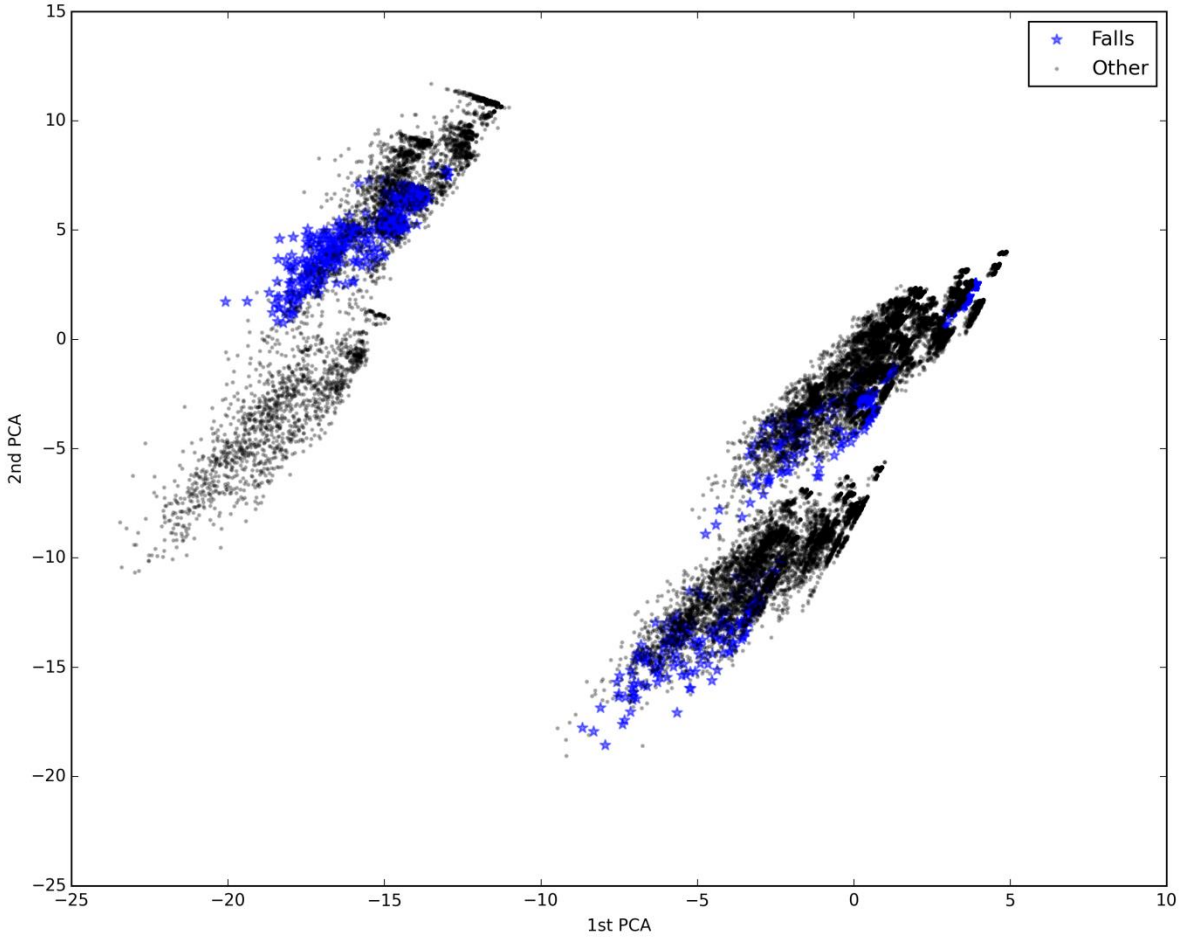
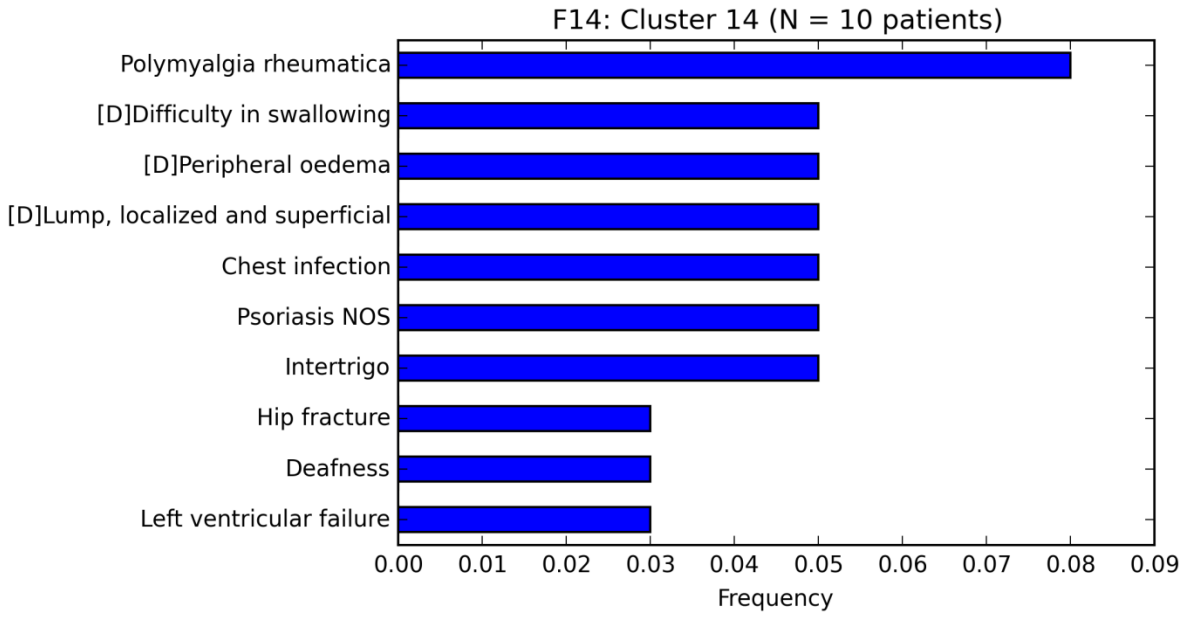


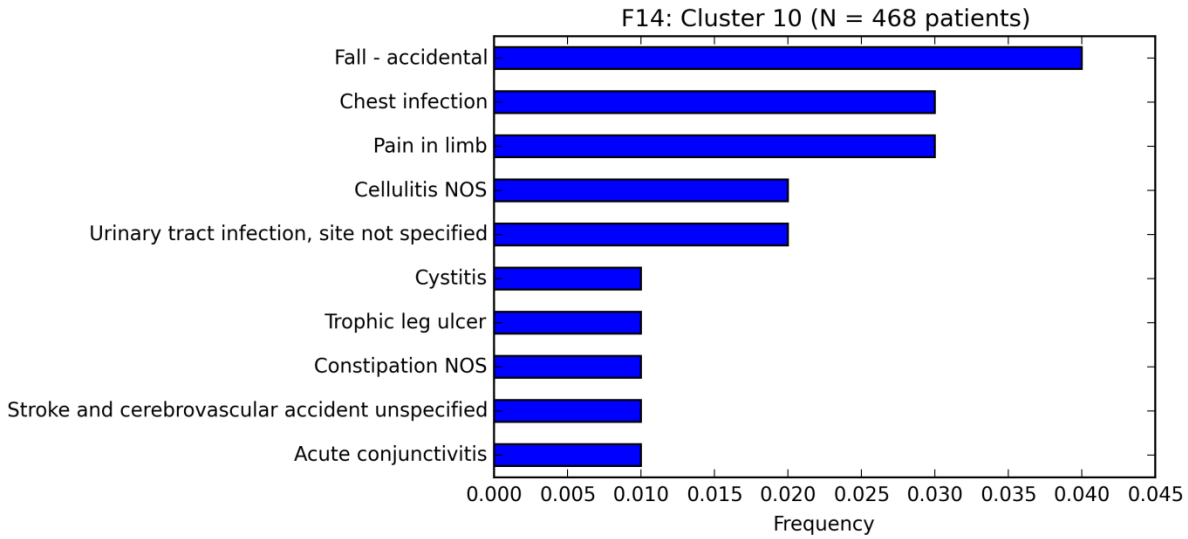
Figure C. 9. Clusters analysis for women patients aged between 75 to 79 years (n=62,849) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (B) The top ten diseases appear in each cluster enriched of falls.

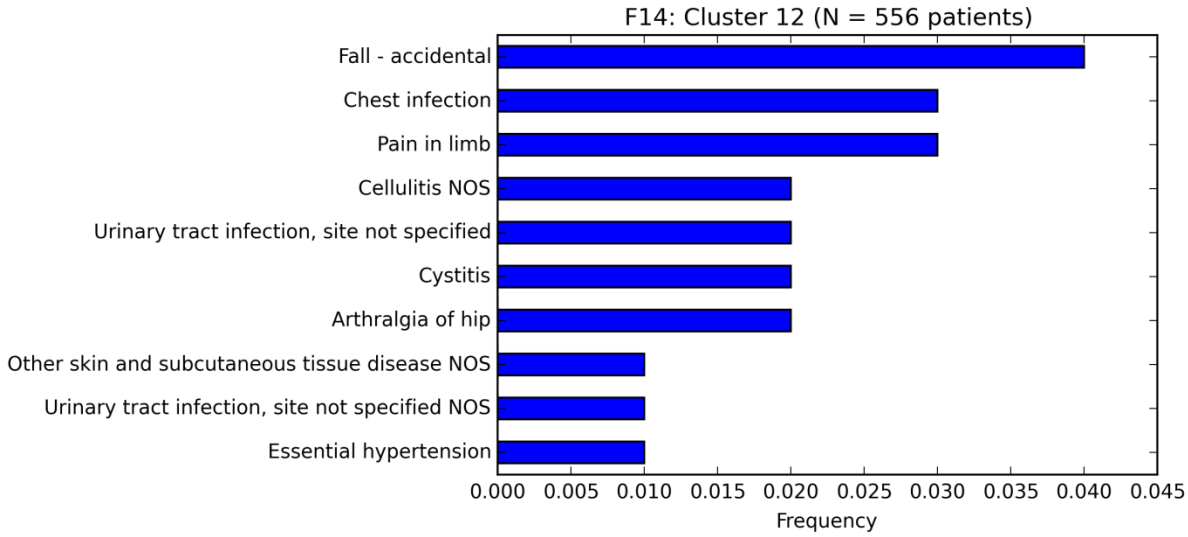
A.

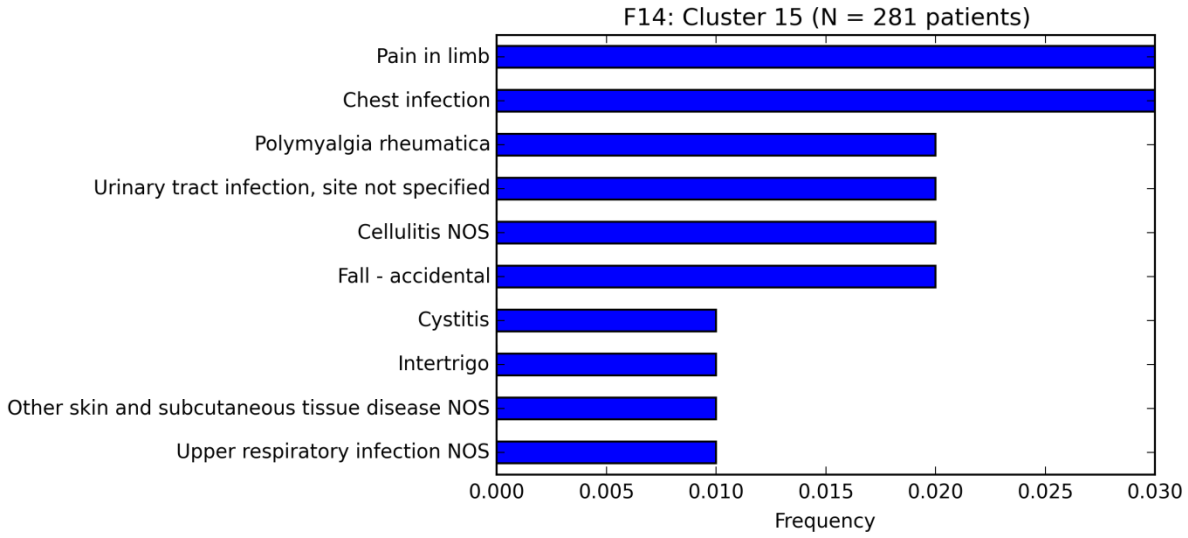

Falls
Other
2nd PCA
1st PCA
15
10
5
0
−5
−10
−15
−20
−25
−25
−20
−15
−10
−5
0
5
10

B.


F14: Cluster 14 (N = 10 patients)
Polymyalgia rheumatica
[D]Difficulty in swallowing
[D]Peripheral oedema
[D]Lump, localized and superficial
Chest infection
Psoriasis NOS
Intertrigo
Hip fracture
Deafness
Left ventricular failure
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
Frequency


F14: Cluster 10 (N = 468 patients)
Fall - accidental
Chest infection
Pain in limb
Cellulitis NOS
Urinary tract infection, site not specified
Cystitis
Trophic leg ulcer
Constipation NOS
Stroke and cerebrovascular accident unspecified
Acute conjunctivitis
0.000 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045
Frequency


F14: Cluster 12 (N = 556 patients)
Fall - accidental
Chest infection
Pain in limb
Cellulitis NOS
Urinary tract infection, site not specified
Cystitis
Arthralgia of hip
Other skin and subcutaneous tissue disease NOS
Urinary tract infection, site not specified NOS
Essential hypertension
0.000 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045
Frequency


F14: Cluster 15 (N = 281 patients)
Pain in limb
Chest infection
Polymyalgia rheumatica
Urinary tract infection, site not specified
Cellulitis NOS
Fall - accidental
Cystitis
Intertrigo
Other skin and subcutaneous tissue disease NOS
Upper respiratory infection NOS
0.000 0.005 0.010 0.015 0.020 0.025 0.030
Frequency

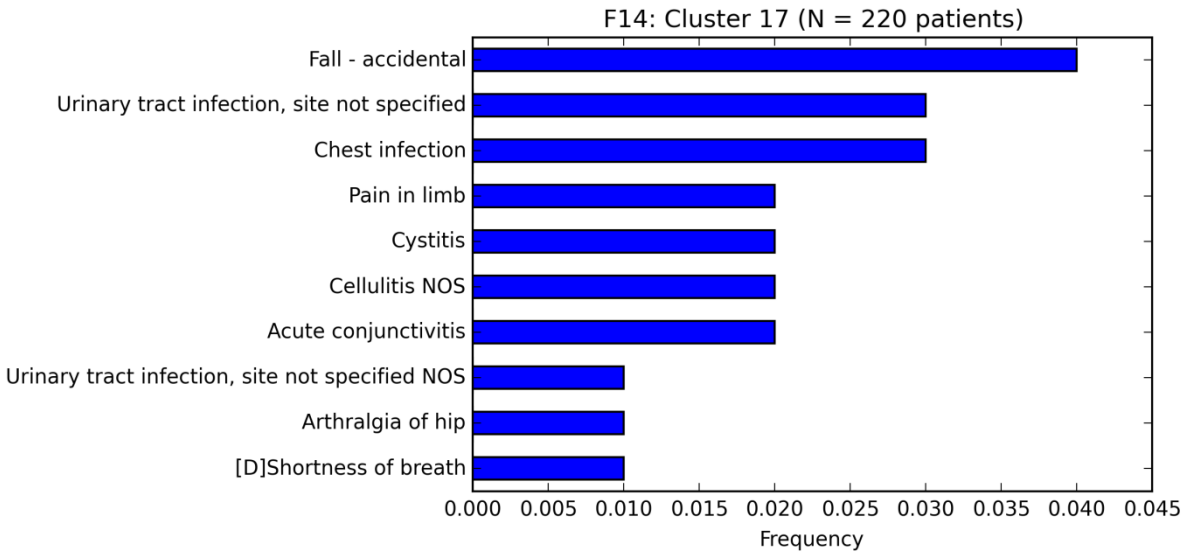
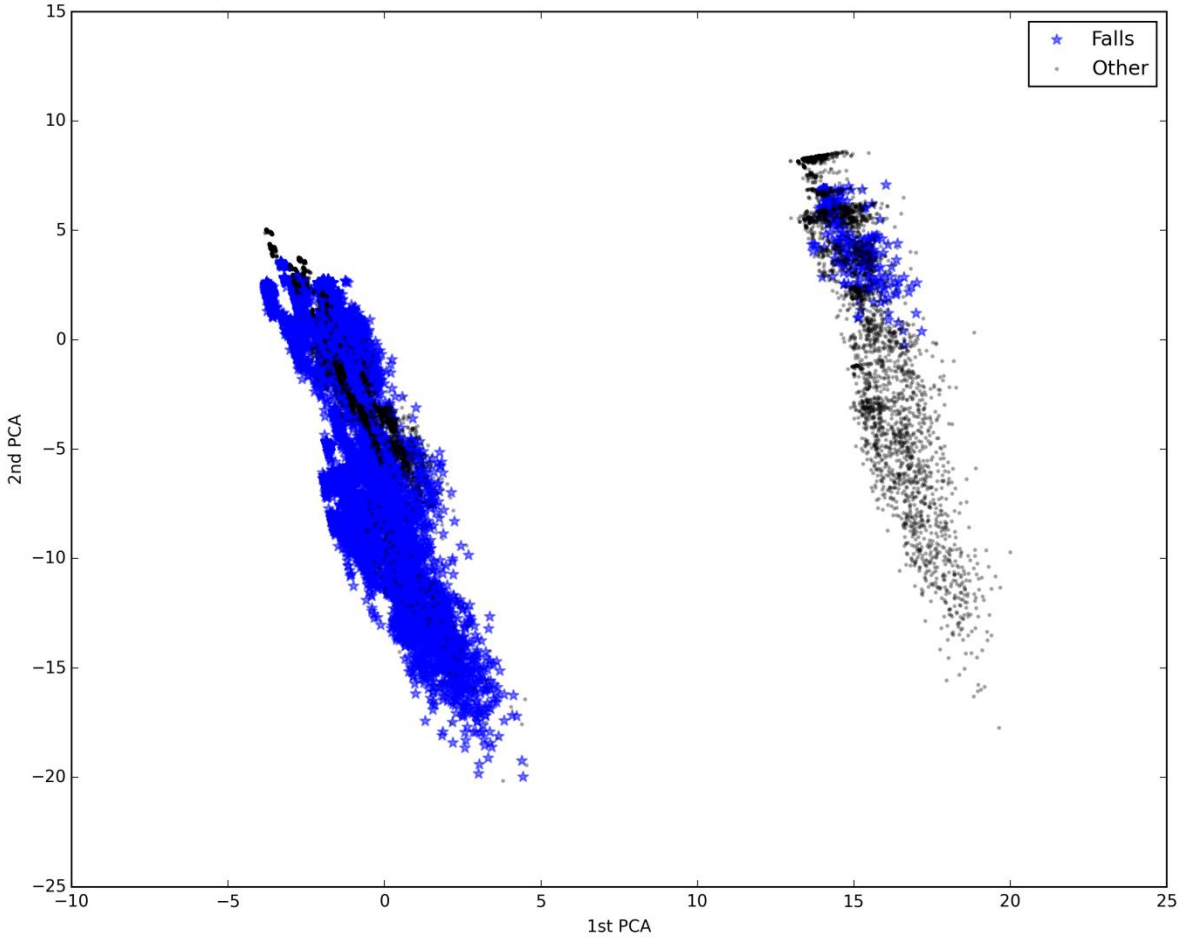
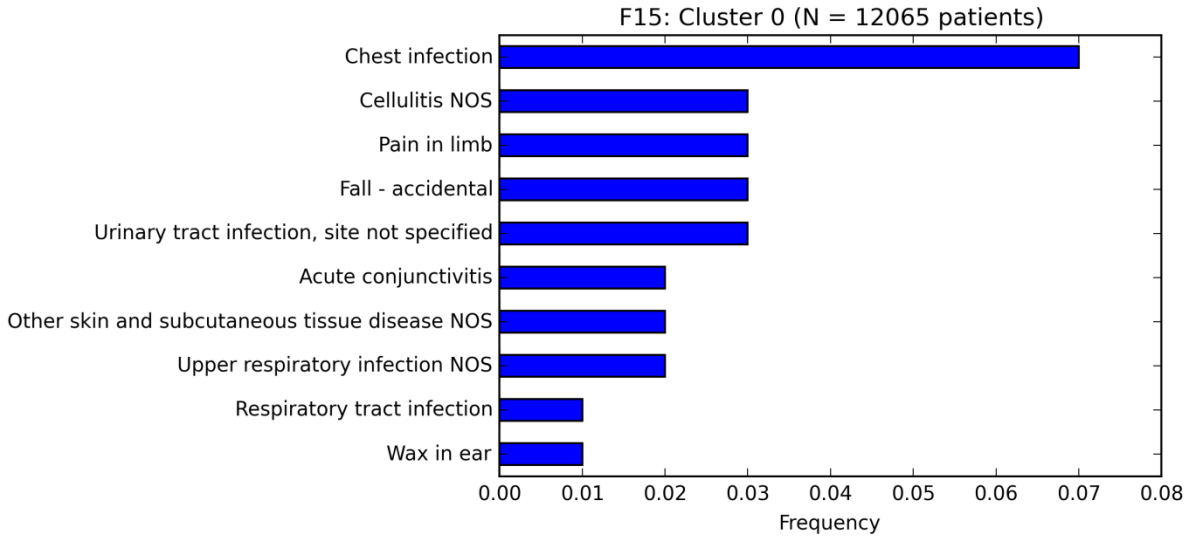


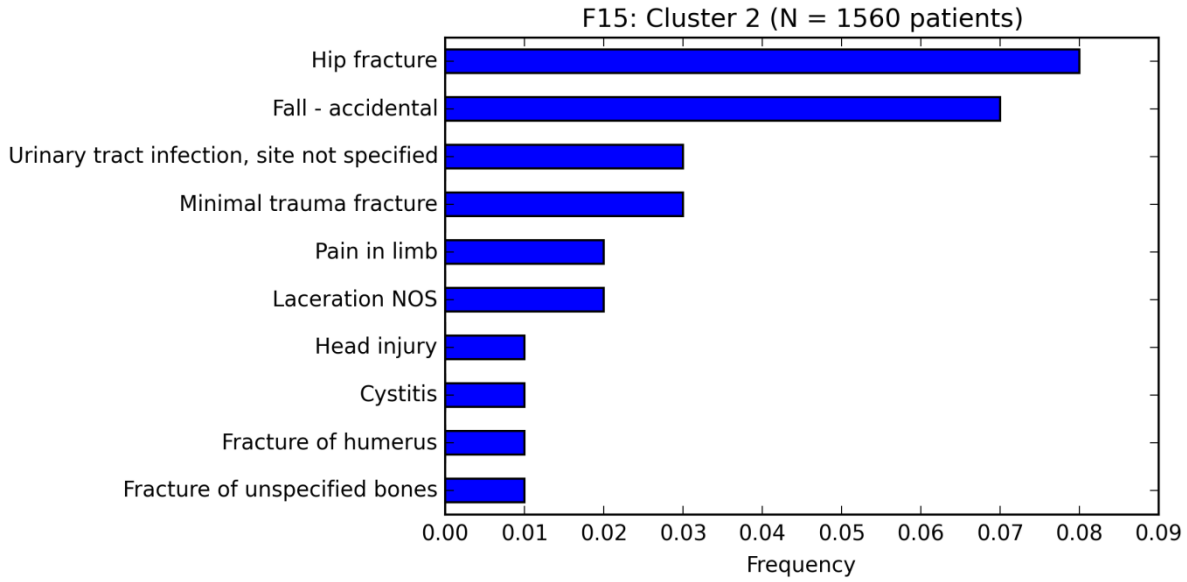
Figure C. 10. Clusters analysis for women patients aged between 85 to 89 years (n=37,647) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (B) The top ten diseases appear in each cluster enriched of falls.

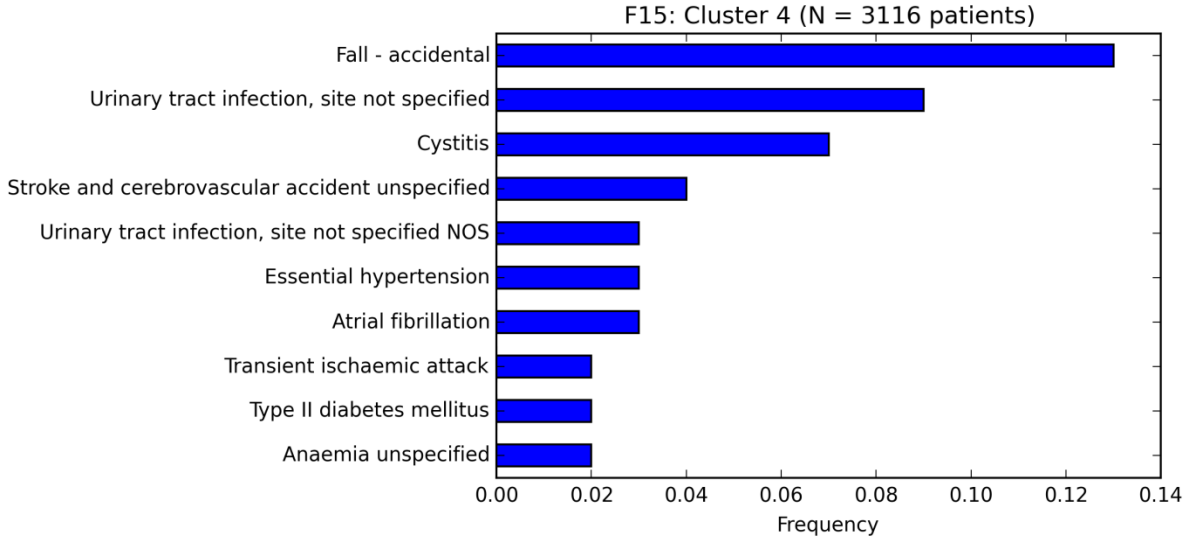
A.

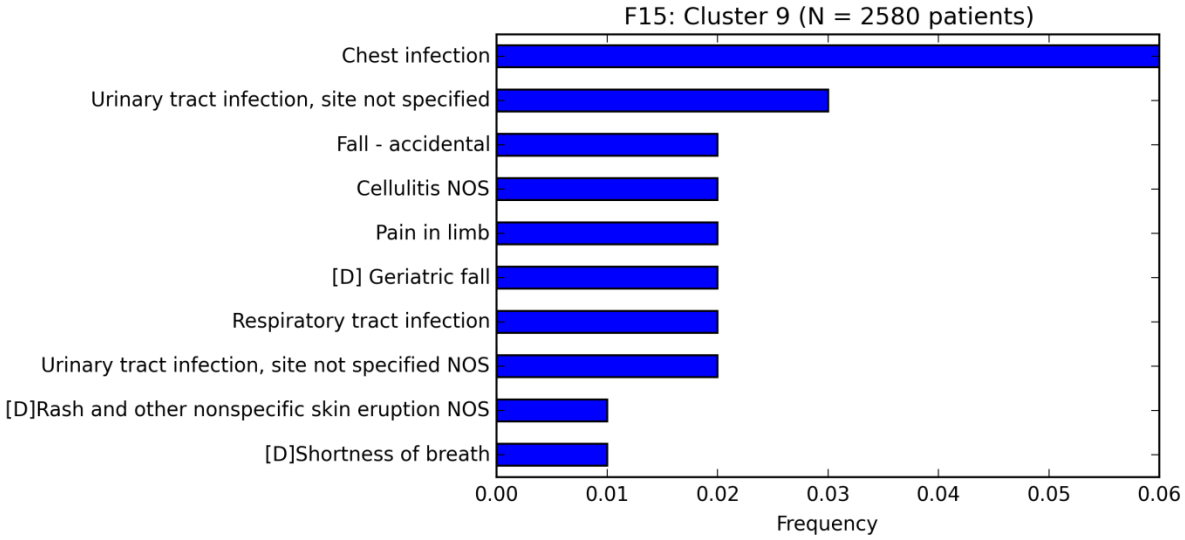


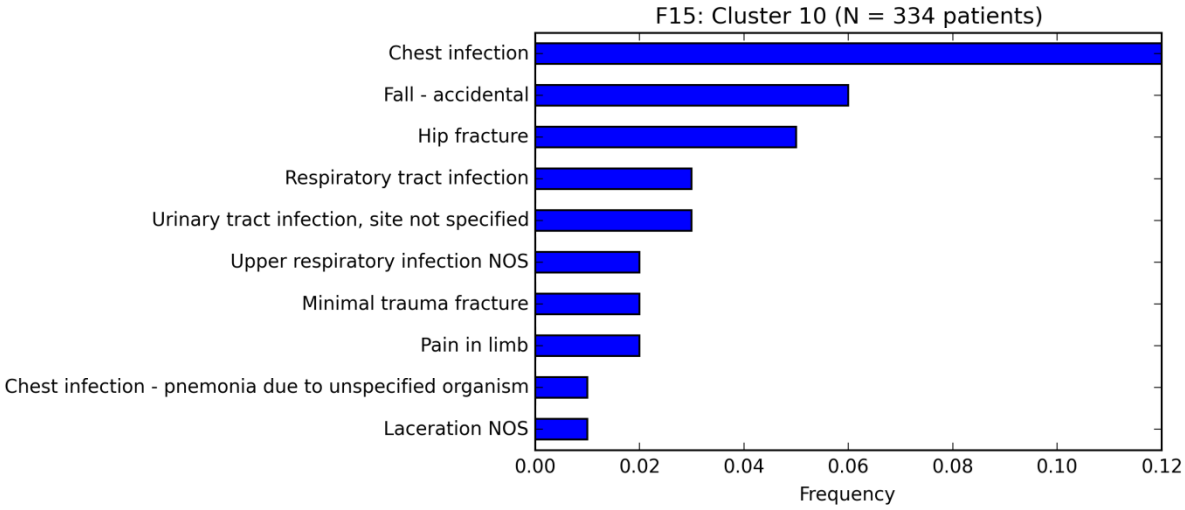
B.

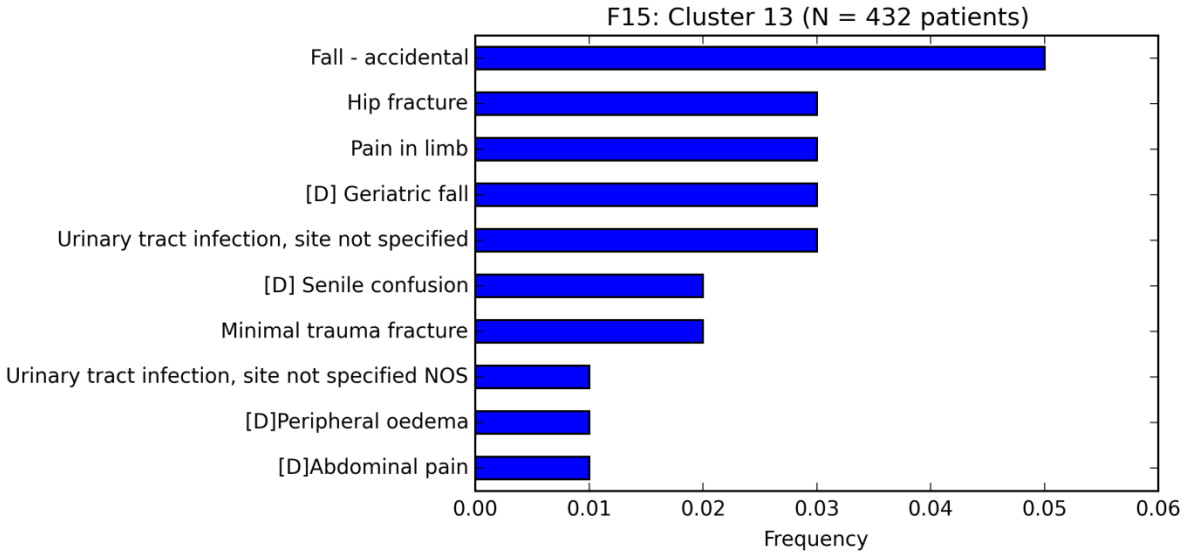

F15: Cluster 0 (N = 12065 patients)
Chest infection
Cellulitis NOS
Pain in limb
Fall - accidental
Urinary tract infection, site not specified
Acute conjunctivitis
Other skin and subcutaneous tissue disease NOS
Upper respiratory infection NOS
Respiratory tract infection
Wax in ear
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
Frequency


F15: Cluster 2 (N = 1560 patients)
Hip fracture
Fall - accidental
Urinary tract infection, site not specified
Minimal trauma fracture
Pain in limb
Laceration NOS
Head injury
Cystitis
Fracture of humerus
Fracture of unspecified bones
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09
Frequency


F15: Cluster 4 (N = 3116 patients)
Fall - accidental
Urinary tract infection, site not specified
Cystitis
Stroke and cerebrovascular accident unspecified
Urinary tract infection, site not specified NOS
Essential hypertension
Atrial fibrillation
Transient ischaemic attack
Type II diabetes mellitus
Anaemia unspecified
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
Frequency


F15: Cluster 9 (N = 2580 patients)
Chest infection
Urinary tract infection, site not specified
Fall - accidental
Cellulitis NOS
Pain in limb
[D] Geriatric fall
Respiratory tract infection
Urinary tract infection, site not specified NOS
[D]Rash and other nonspecific skin eruption NOS
[D]Shortness of breath
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Frequency


F15: Cluster 10 (N = 334 patients)
Chest infection
Fall - accidental
Hip fracture
Respiratory tract infection
Urinary tract infection, site not specified
Upper respiratory infection NOS
Minimal trauma fracture
Pain in limb
Chest infection - pnemonia due to unspecified organism
Laceration NOS
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Frequency


F15: Cluster 13 (N = 432 patients)
Fall - accidental
Hip fracture
Pain in limb
[D] Geriatric fall
Urinary tract infection, site not specified
[D] Senile confusion
Minimal trauma fracture
Urinary tract infection, site not specified NOS
[D]Peripheral oedema
[D]Abdominal pain
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Frequency

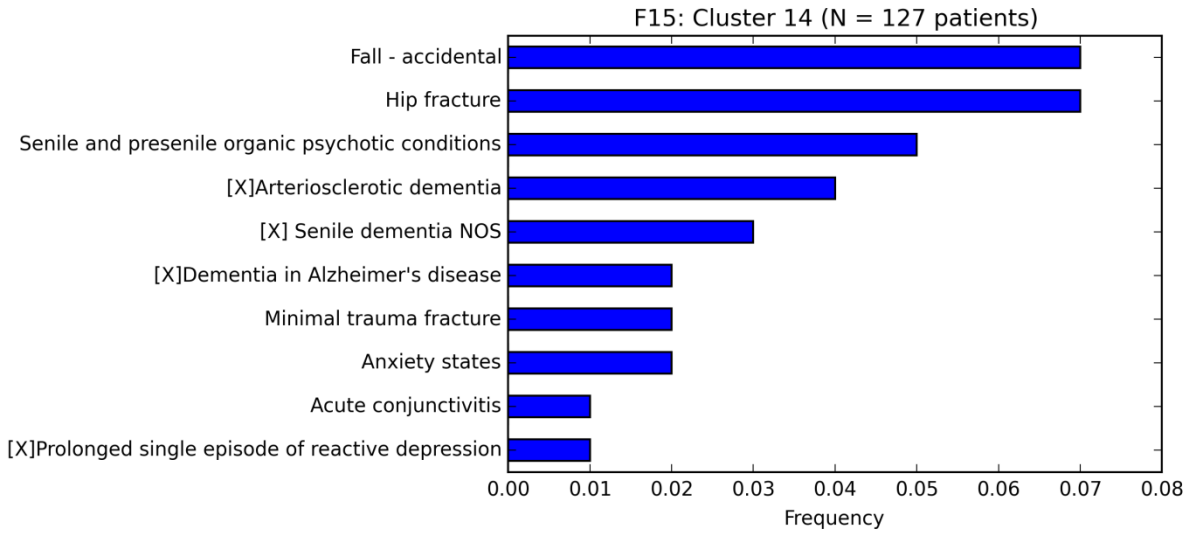


Figure C. 11. Clusters analysis for women patients aged above 89 years (n=25,649) based on semantic similarity. (A) Clusters enriched of falls (in blue/ stars). (B) The top ten diseases appear in each cluster enriched of falls.

C.3. Results

Table C. 2. Significant associated diseases with falls in men elderly population level.

Disease	RR	*Φ*-correlation	95% LCI	95% UCI	P value
Minor head injury	6.92	0.03	5.27	9.09	<0.0001
Head injury	6.71	0.03	5.22	8.63	<0.0001
Intracranial injury NOS	6.05	0.03	4.64	7.89	<0.0001
Closed fracture pelvis, single pubic ramus	5.14	0.01	2.01	13.17	<0.0001
Fracture of unspecified bones	4.85	0.02	3.57	6.58	<0.0001
Fracture of humerus	4.85	0.02	3.37	6.98	<0.0001
Minimal trauma fracture	4.85	0.02	3.29	7.15	<0.0001
Hip fracture	4.83	0.04	4.01	5.81	<0.0001
Closed fracture of radius (alone), unspecified	4.55	0.01	2.43	8.54	<0.0001
[D] Senile confusion	4.17	0.03	3.36	5.17	<0.0001
Laceration NOS	3.86	0.03	3.13	4.76	<0.0001
Leg bruise	3.84	0.01	2.36	6.25	<0.0001
[X]Arteriosclerotic dementia	3.72	0.03	3.05	4.54	<0.0001
Haematoma with intact skin	3.70	0.02	2.81	4.89	<0.0001
Postural hypotension	3.59	0.03	3.06	4.21	<0.0001
[X] Senile dementia NOS	3.45	0.02	2.69	4.42	<0.0001
Pressure sore	3.38	0.02	2.66	4.29	<0.0001
Senile and presenile organic psychotic conditions	3.31	0.02	2.56	4.28	<0.0001
[D]Restlessness and agitation	3.31	0.01	1.98	5.53	<0.0001
Injury and poisoning NOS	3.19	0.01	2.02	5.04	<0.0001
Post-traumatic wound infection NEC	3.15	0.01	2.23	4.45	<0.0001
[D]Collapse	3.14	0.02	2.49	3.96	<0.0001
Implant complications	2.92	0.01	2.05	4.17	<0.0001
Hyponatraemia	2.92	0.02	2.29	3.73	<0.0001
Osteoporosis	2.77	0.02	2.14	3.59	<0.0001
Urinary tract infection, site not specified	2.75	0.04	2.48	3.04	<0.0001
[D]Difficulty in swallowing	2.68	0.01	1.94	3.69	<0.0001
Trophic leg ulcer	2.61	0.02	2.15	3.18	<0.0001
Chest infection - pnemonia due to unspecified organism	2.59	0.02	2.14	3.12	<0.0001
Swelling of calf	2.57	0.02	2.12	3.12	<0.0001
Chest infection - unspecified bronchopneumonia	2.50	0.01	1.79	3.49	<0.0001
[D]Abnormal loss of weight	2.49	0.01	1.95	3.18	<0.0001
Wasp sting	2.47	0.01	1.91	3.20	<0.0001
Fracture of lower end of radius	2.47	0.00	0.89	6.81	<0.0001
Urinary tract infection, site not specified NOS	2.41	0.02	2.02	2.88	<0.0001
Anaemia unspecified	2.34	0.02	2.03	2.71	<0.0001
[D]Dependent oedema	2.27	0.01	1.47	3.49	<0.0001
[D]Retention of urine	2.24	0.02	1.87	2.68	<0.0001

Congestive heart failure	2.21	0.01	1.77	2.77	<0.0001
Acute lower respiratory tract infection	2.13	0.01	1.70	2.68	<0.0001
DVT - Deep vein thrombosis	2.07	0.01	1.60	2.66	<0.0001
Stroke due to cerebral arterial occlusion	2.05	0.01	1.40	3.00	<0.0001
Stroke and cerebrovascular accident unspecified	1.99	0.01	1.62	2.45	<0.0001
Cardiac failure	1.96	0.01	1.52	2.54	<0.0001
Chronic renal failure	1.94	0.01	1.33	2.84	0.0003
Cellulitis NOS	1.93	0.02	1.72	2.16	<0.0001
Constipation - functional	1.91	0.01	1.21	3.04	0.004
Constipation NOS	1.90	0.02	1.64	2.20	<0.0001
Bowel obstruction	1.89	0.01	1.16	3.07	0.007
[D]Insomnia NOS	1.84	0.01	1.39	2.43	<0.0001
Microcytic hypochromic anaemia	1.83	0.01	1.37	2.44	<0.0001
[X]Dementia in Alzheimer's disease	1.80	0.00	1.07	3.02	0.02
Cystitis	1.80	0.01	1.48	2.18	<0.0001
Gastrointestinal haemorrhage unspecified	1.74	0.00	1.02	2.98	0.03
[D]Uncertain diagnosis	1.72	0.01	1.32	2.25	<0.0001
Respiratory tract infection	1.72	0.02	1.51	1.96	<0.0001
Skin and subcutaneous tissue infections	1.68	0.01	1.44	1.95	<0.0001
[D]Dizziness	1.64	0.01	1.31	2.06	<0.0001
Atrial fibrillation	1.63	0.01	1.40	1.89	<0.0001
Microcytic - hypochromic anaemia	1.61	0.01	1.33	1.95	<0.0001
[D]Shortness of breath	1.61	0.01	1.32	1.95	<0.0001
Left ventricular failure	1.61	0.01	1.19	2.17	0.001
Transient ischaemic attack	1.59	0.01	1.28	1.99	<0.0001
Acute non-ST segment elevation myocardial infarction	1.56	0.01	1.18	2.06	0.001
Vomiting of blood	1.44	0.00	0.83	2.51	0.2
Irritable hip	1.43	0.01	1.13	1.80	0.002
Hypothyroidism	1.41	0.01	1.10	1.82	0.005
Vitamin B12 deficiency	1.41	0.00	1.04	1.89	0.02
Aching leg syndrome	1.40	0.01	1.20	1.63	<0.0001
Intertrigo	1.38	0.00	1.04	1.83	0.02
Acute conjunctivitis	1.37	0.01	1.19	1.58	<0.0001
Acute back pain - lumbar	1.34	0.01	1.18	1.52	<0.0001
[D]Insomnia - symptom	1.33	0.00	0.94	1.87	0.09
Cellulitis of eyelids	1.31	0.01	1.07	1.60	0.006
Squamous cell carcinoma of skin	1.31	0.00	0.88	1.95	0.2
Bleeding PR	1.30	0.00	1.01	1.68	0.03
Haematuria	1.30	0.01	1.10	1.54	0.002
Adverse reaction to aspirin	1.30	0.00	0.69	2.44	0.4
Chest infection	1.28	0.01	1.19	1.36	<0.0001
MI - acute myocardial infarction	1.27	0.00	0.92	1.76	0.1
Other non-infective inflammatory gastroenteritis and colitis	1.26	0.00	0.89	1.79	0.2

Calculus - biliary	1.25	0.00	0.89	1.77	0.2
Skin irritation	1.25	0.00	1.03	1.52	0.02
Pain in limb	1.24	0.01	1.16	1.34	<0.0001
[D]Light-headedness	1.24	0.00	0.91	1.68	0.2
Atrial fibrillation and flutter	1.24	0.00	0.95	1.60	0.1
[M]Squamous cell carcinoma NOS	1.23	0.00	0.83	1.82	0.3
Arthralgia of hip	1.21	0.01	1.06	1.39	0.005
Cataract	1.20	0.00	1.00	1.44	0.04
Oesophagitis	1.20	0.00	0.82	1.74	0.3
Malignant neoplasm of prostate	1.17	0.00	0.95	1.44	0.1
[D]Rash and other nonspecific skin eruption NOS	1.13	0.00	0.89	1.44	0.3
Open wound of leg	1.13	0.00	0.50	2.55	0.9
Acute exacerbation of chronic obstructive airways disease	1.13	0.00	0.93	1.37	0.2
Duodenal ulcer - (DU)	1.10	0.00	0.60	2.01	0.7
Wax in ear	1.10	0.00	0.95	1.28	0.2
Chronic obstructive pulmonary disease	1.09	0.00	0.90	1.33	0.3
Atopic dermatitis/eczema	1.09	0.00	0.90	1.31	0.4
[D]Abdominal pain	1.08	0.00	0.87	1.34	0.5
Varicose veins of the leg with eczema	1.05	0.00	0.85	1.30	0.6
Pain in joint - arthralgia	1.05	0.00	0.84	1.32	0.6
Deafness	1.03	0.00	0.85	1.23	0.8
Varicose veins of the legs	1.02	0.00	0.71	1.47	0.9
Diverticulosis	1.01	0.00	0.79	1.29	0.9
[D]Vertigo NOS	0.97	0.00	0.73	1.30	0.9
[D]Microalbuminuria	0.96	0.00	0.69	1.35	0.8
Gastritis unspecified	0.94	0.00	0.63	1.40	0.8
Other skin and subcutaneous tissue disease NOS	0.94	0.00	0.83	1.06	0.3
Frank haematuria	0.93	0.00	0.54	1.62	0.8
Dermatitis NOS	0.93	0.00	0.71	1.23	0.6
Shingles	0.92	0.00	0.74	1.16	0.5
Discoid eczema	0.92	0.00	0.72	1.18	0.5
Upper respiratory infection NOS	0.91	0.00	0.81	1.03	0.2
Pain in cervical spine	0.91	0.00	0.79	1.05	0.2
Actinic keratosis	0.90	0.00	0.78	1.04	0.2
Ischaemic heart disease	0.89	0.00	0.69	1.16	0.4
Dermatophytosis including tinea or ringworm	0.89	0.00	0.70	1.15	0.4
Acute bronchitis	0.89	0.00	0.71	1.12	0.3
Right inguinal hernia	0.89	0.00	0.57	1.39	0.6
[D]Epistaxis	0.88	0.00	0.62	1.26	0.5
Parasternal hernia	0.88	0.00	0.66	1.17	0.4
Rectal bleeding	0.86	0.00	0.63	1.18	0.4
Asthma	0.86	0.00	0.64	1.16	0.3
Gout	0.83	0.00	0.71	0.97	0.02

[D]Groin pain	0.83	0.00	0.64	1.07	0.2
Polymyalgia rheumatica	0.83	0.00	0.60	1.15	0.3
Haemorrhoids	0.82	0.00	0.62	1.09	0.2
Inguinal hernia	0.81	0.00	0.64	1.03	0.1
Osteoarthritis and allied disorders	0.81	0.00	0.67	0.99	0.04
Oesopheal reflux without mention of oesophagitis	0.81	0.00	0.63	1.05	0.1
Skin lesion	0.81	0.00	0.64	1.02	0.08
Angina pectoris	0.81	0.00	0.57	1.14	0.2
Leg cramps	0.81	0.00	0.64	1.02	0.07
Skin and subcut tissue infection NOS	0.79	0.00	0.55	1.15	0.2
Prostatism	0.79	0.00	0.64	0.97	0.03
Acute back pain with sciatica	0.78	0.00	0.63	0.97	0.03
Malignant neoplasm of urinary bladder	0.78	0.00	0.45	1.35	0.4
[D]Rash and other nonspecific skin eruption	0.78	0.00	0.54	1.12	0.2
Diabetes mellitus	0.77	0.00	0.54	1.12	0.2
Aortic aneurysm	0.77	0.00	0.50	1.19	0.2
Malignant neoplasm of sweat gland	0.75	-0.01	0.61	0.91	0.004
Acid reflux	0.74	0.00	0.52	1.04	0.08
Diverticula of intestine	0.74	0.00	0.49	1.10	0.1
Flatulent dyspepsia	0.73	-0.01	0.60	0.89	0.002
[D]Peripheral oedema	0.72	-0.01	0.61	0.86	0.0003
[D]Cough	0.69	0.00	0.47	1.00	0.05
Type II diabetes mellitus	0.67	-0.01	0.56	0.81	<0.0001
[D]Raised blood pressure reading	0.67	-0.01	0.52	0.87	0.002
Otitis externa NOS	0.65	0.00	0.46	0.92	0.01
Essential hypertension	0.62	-0.01	0.51	0.75	<0.0001
Infective otitis externa	0.62	-0.01	0.50	0.77	<0.0001
Seborrhoeic keratosis	0.62	-0.01	0.48	0.79	<0.0001
Seborrhoeic wart	0.53	-0.01	0.39	0.72	<0.0001
BP - hypertensive disease	0.52	-0.01	0.33	0.81	0.003
Sinusitis	0.45	-0.01	0.33	0.62	<0.0001
Plantar fasciitis	0.40	-0.01	0.27	0.59	<0.0001
[D]Renal colic	0.00	0.00	-1.00	-1.00	0.04
Malignant neoplasm of women breast	0.00	0.00	-1.00	-1.00	0.9

Table C. 3. Significant associated diseases with falls in women elderly population level.

Disease	RR	*Φ*-correlation	95% LCI	95% UCI	P value
Minor head injury	6.92	0.03	5.27	9.09	<0.0001
Minor head injury	6.72	0.02	4.16	10.85	<0.0001
Head injury	6.71	0.03	5.22	8.63	<0.0001
Intracranial injury NOS	6.05	0.03	4.64	7.89	<0.0001
Closed fracture pelvis, single pubic ramus	5.14	0.01	2.01	13.17	<0.0001
Fracture of unspecified bones	4.85	0.02	3.57	6.58	<0.0001
Fracture of humerus	4.85	0.02	3.37	6.98	<0.0001
Minimal trauma fracture	4.85	0.02	3.29	7.15	<0.0001
Hip fracture	4.83	0.04	4.01	5.81	<0.0001
Closed fracture of radius (alone), unspecified	4.55	0.01	2.43	8.54	<0.0001
[D] Senile confusion	4.17	0.03	3.36	5.17	<0.0001
Laceration NOS	3.86	0.03	3.13	4.76	<0.0001
Leg bruise	3.84	0.01	2.36	6.25	<0.0001
[X]Arteriosclerotic dementia	3.72	0.03	3.05	4.54	<0.0001
Haematoma with intact skin	3.70	0.02	2.81	4.89	<0.0001
Postural hypotension	3.59	0.03	3.06	4.21	<0.0001
[X] Senile dementia NOS	3.45	0.02	2.69	4.42	<0.0001
Pressure sore	3.38	0.02	2.66	4.29	<0.0001
Senile and presenile organic psychotic conditions	3.31	0.02	2.56	4.28	<0.0001
[D]Restlessness and agitation	3.31	0.01	1.98	5.53	<0.0001
Injury and poisoning NOS	3.19	0.01	2.02	5.04	<0.0001
Post-traumatic wound infection NEC	3.15	0.01	2.23	4.45	<0.0001
[D]Collapse	3.14	0.02	2.49	3.96	<0.0001
Implant complications	2.92	0.01	2.05	4.17	<0.0001
Hyponatraemia	2.92	0.02	2.29	3.73	<0.0001
Osteoporosis	2.77	0.02	2.14	3.59	<0.0001
Urinary tract infection, site not specified	2.75	0.04	2.48	3.04	<0.0001
[D]Difficulty in swallowing	2.68	0.01	1.94	3.69	<0.0001
Trophic leg ulcer	2.61	0.02	2.15	3.18	<0.0001
Chest infection - pnemonia due to unspecified organism	2.59	0.02	2.14	3.12	<0.0001
Swelling of calf	2.57	0.02	2.12	3.12	<0.0001
Chest infection - unspecified bronchopneumonia	2.50	0.01	1.79	3.49	<0.0001
[D]Abnormal loss of weight	2.49	0.01	1.95	3.18	<0.0001
Wasp sting	2.47	0.01	1.91	3.20	<0.0001
Fracture of lower end of radius	2.47	0.00	0.89	6.81	0.06
Urinary tract infection, site not specified NOS	2.41	0.02	2.02	2.88	<0.0001
Anaemia unspecified	2.34	0.02	2.03	2.71	<0.0001

[D]Dependent oedema	2.27	0.01	1.47	3.49	<0.0001
[D]Retention of urine	2.24	0.02	1.87	2.68	<0.0001
Congestive heart failure	2.21	0.01	1.77	2.77	<0.0001
Acute lower respiratory tract infection	2.13	0.01	1.70	2.68	<0.0001
DVT - Deep vein thrombosis	2.07	0.01	1.60	2.66	<0.0001
Stroke due to cerebral arterial occlusion	2.05	0.01	1.40	3.00	<0.0001
Stroke and cerebrovascular accident unspecified	1.99	0.01	1.62	2.45	<0.0001
Cardiac failure	1.96	0.01	1.52	2.54	<0.0001
Chronic renal failure	1.94	0.01	1.33	2.84	0.0003
Cellulitis NOS	1.93	0.02	1.72	2.16	<0.0001
Constipation - functional	1.91	0.01	1.21	3.04	0.004
Constipation NOS	1.90	0.02	1.64	2.20	<0.0001
Bowel obstruction	1.89	0.01	1.16	3.07	0.007
[D]Insomnia NOS	1.84	0.01	1.39	2.43	<0.0001
Microcytic hypochromic anaemia	1.83	0.01	1.37	2.44	<0.0001
[X]Dementia in Alzheimer's disease	1.80	0.00	1.07	3.02	0.02
Cystitis	1.80	0.01	1.48	2.18	<0.0001
Gastrointestinal haemorrhage unspecified	1.74	0.00	1.02	2.98	0.03
[D]Uncertain diagnosis	1.72	0.01	1.32	2.25	3.28E-05
Respiratory tract infection	1.72	0.02	1.51	1.96	<0.0001
Skin and subcutaneous tissue infections	1.68	0.01	1.44	1.95	<0.0001
[D]Dizziness	1.64	0.01	1.31	2.06	<0.0001
Atrial fibrillation	1.63	0.01	1.40	1.89	<0.0001
Microcytic - hypochromic anaemia	1.61	0.01	1.33	1.95	<0.0001
[D]Shortness of breath	1.61	0.01	1.32	1.95	<0.0001
Left ventricular failure	1.61	0.01	1.19	2.17	0.001
Transient ischaemic attack	1.59	0.01	1.28	1.99	<0.0001
Acute non-ST segment elevation myocardial infarction	1.56	0.01	1.18	2.06	0.001
Vomiting of blood	1.44	0.00	0.83	2.51	0.2
Irritable hip	1.43	0.01	1.13	1.80	0.002
Hypothyroidism	1.41	0.01	1.10	1.82	0.005
Vitamin B12 deficiency	1.41	0.00	1.04	1.89	0.02
Aching leg syndrome	1.40	0.01	1.20	1.63	<0.0001
Intertrigo	1.38	0.00	1.04	1.83	0.02
Acute conjunctivitis	1.37	0.01	1.19	1.58	<0.0001
Acute back pain - lumbar	1.34	0.01	1.18	1.52	<0.0001
[D]Insomnia - symptom	1.33	0.00	0.94	1.87	0.09
Cellulitis of eyelids	1.31	0.01	1.07	1.60	0.006
Squamous cell carcinoma of skin	1.31	0.00	0.88	1.95	0.2
Bleeding PR	1.30	0.00	1.01	1.68	0.03
Haematuria	1.30	0.01	1.10	1.54	0.002

Adverse reaction to aspirin	1.30	0.00	0.69	2.44	0.4
Chest infection	1.28	0.01	1.19	1.36	<0.0001
MI - acute myocardial infarction	1.27	0.00	0.92	1.76	0.1
Other non-infective inflammatory gastroenteritis and colitis	1.26	0.00	0.89	1.79	0.2
Calculus - biliary	1.25	0.00	0.89	1.77	0.2
Skin irritation	1.25	0.00	1.03	1.52	0.02
Pain in limb	1.24	0.01	1.16	1.34	<0.0001
[D]Light-headedness	1.24	0.00	0.91	1.68	0.2
Atrial fibrillation and flutter	1.24	0.00	0.95	1.60	0.1
[M]Squamous cell carcinoma NOS	1.23	0.00	0.83	1.82	0.3
Arthralgia of hip	1.21	0.01	1.06	1.39	0.005
Cataract	1.20	0.00	1.00	1.44	0.04
Oesophagitis	1.20	0.00	0.82	1.74	0.3
Malignant neoplasm of prostate	1.17	0.00	0.95	1.44	0.1
[D]Rash and other nonspecific skin eruption NOS	1.13	0.00	0.89	1.44	0.3
Open wound of leg	1.13	0.00	0.50	2.55	0.8
Acute exacerbation of chronic obstructive airways disease	1.13	0.00	0.93	1.37	0.2
Duodenal ulcer - (DU)	1.10	0.00	0.60	2.01	0.7
Wax in ear	1.10	0.00	0.95	1.28	0.2
Chronic obstructive pulmonary disease	1.09	0.00	0.90	1.33	0.3
Atopic dermatitis/eczema	1.09	0.00	0.90	1.31	0.4
[D]Abdominal pain	1.08	0.00	0.87	1.34	0.5
Varicose veins of the leg with eczema	1.05	0.00	0.85	1.30	0.6
Pain in joint - arthralgia	1.05	0.00	0.84	1.32	0.6
Deafness	1.03	0.00	0.85	1.23	0.8
Varicose veins of the legs	1.02	0.00	0.71	1.47	0.9
Diverticulosis	1.01	0.00	0.79	1.29	0.9
[D]Vertigo NOS	0.97	0.00	0.73	1.30	0.9
[D]Microalbuminuria	0.96	0.00	0.69	1.35	0.8
Gastritis unspecified	0.94	0.00	0.63	1.40	0.8
Other skin and subcutaneous tissue disease NOS	0.94	0.00	0.83	1.06	0.3
Frank haematuria	0.93	0.00	0.54	1.62	0.8
Dermatitis NOS	0.93	0.00	0.71	1.23	0.6
Shingles	0.92	0.00	0.74	1.16	0.5
Discoid eczema	0.92	0.00	0.72	1.18	0.5
Upper respiratory infection NOS	0.91	0.00	0.81	1.03	0.2
Pain in cervical spine	0.91	0.00	0.79	1.05	0.2
Actinic keratosis	0.90	0.00	0.78	1.04	0.2
Ischaemic heart disease	0.89	0.00	0.69	1.16	0.4
Dermatophytosis including tinea or	0.89	0.00	0.70	1.15	0.4

ringworm					
Acute bronchitis	0.89	0.00	0.71	1.12	0.3
Right inguinal hernia	0.89	0.00	0.57	1.39	0.6
[D]Epistaxis	0.88	0.00	0.62	1.26	0.5
Parasternal hernia	0.88	0.00	0.66	1.17	0.4
Rectal bleeding	0.86	0.00	0.63	1.18	0.4
Asthma	0.86	0.00	0.64	1.16	0.3
Gout	0.83	0.00	0.71	0.97	0.02
[D]Groin pain	0.83	0.00	0.64	1.07	0.2
Polymyalgia rheumatica	0.83	0.00	0.60	1.15	0.3
Haemorrhoids	0.82	0.00	0.62	1.09	0.2
Inguinal hernia	0.81	0.00	0.64	1.03	0.09
Osteoarthritis and allied disorders	0.81	0.00	0.67	0.99	0.04
Oesopheal reflux without mention of oesophagitis	0.81	0.00	0.63	1.05	0.1
Skin lesion	0.81	0.00	0.64	1.02	0.08
Angina pectoris	0.81	0.00	0.57	1.14	0.2
Leg cramps	0.81	0.00	0.64	1.02	0.07
Skin and subcut tissue infection NOS	0.79	0.00	0.55	1.15	0.2
Prostatism	0.79	0.00	0.64	0.97	0.03
Acute back pain with sciatica	0.78	0.00	0.63	0.97	0.03
Malignant neoplasm of urinary bladder	0.78	0.00	0.45	1.35	0.4
[D]Rash and other nonspecific skin eruption	0.78	0.00	0.54	1.12	0.2
Diabetes mellitus	0.77	0.00	0.54	1.12	0.2
Aortic aneurysm	0.77	0.00	0.50	1.19	0.2
Malignant neoplasm of sweat gland	0.75	-0.01	0.61	0.91	0.004
Acid reflux	0.74	0.00	0.52	1.04	0.08
Diverticula of intestine	0.74	0.00	0.49	1.10	0.1
Flatulent dyspepsia	0.73	-0.01	0.60	0.89	0.002
[D]Peripheral oedema	0.72	-0.01	0.61	0.86	0.0003
[D]Cough	0.69	0.00	0.47	1.00	0.05
Type II diabetes mellitus	0.67	-0.01	0.56	0.81	<0.0001
[D]Raised blood pressure reading	0.67	-0.01	0.52	0.87	0.002
Otitis externa NOS	0.65	0.00	0.46	0.92	0.01
Essential hypertension	0.62	-0.01	0.51	0.75	<0.0001
Infective otitis externa	0.62	-0.01	0.50	0.77	<0.0001
Seborrhoeic keratosis	0.62	-0.01	0.48	0.79	<0.0001
Seborrhoeic wart	0.53	-0.01	0.39	0.72	<0.0001
BP - hypertensive disease	0.52	-0.01	0.33	0.81	0.003
Sinusitis	0.45	-0.01	0.33	0.62	<0.0001
Plantar fasciitis	0.40	-0.01	0.27	0.59	<0.0001
[D]Renal colic	0.00	0.00	-1.00	-1.00	0.03
Malignant neoplasm of women breast	0.00	0.00	-1.00	-1.00	0.9

C.4. The distribution of comorbidity measures

A number of 160 diseases have appeared significantly with falls in the resulted clusters (p-value less than 0.05 for all diseases codes). We have tested the relationship between these diseases and falls, using RR and Φ-correlation. The distributions of RR and Φ-correlation values found in the dataset are shown in Figure C. 12 A and B. Figure C. 12 C shows that these measures have a positive correlation.

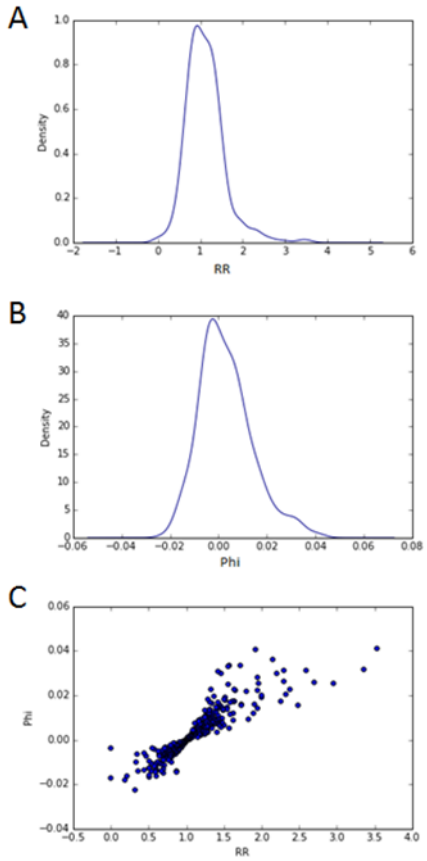


Figure C. 12. Data characteristics of comorbidity measures**.** (A) Distribution of relative risk (RR) between all diseases and falls across patients age groups. (B) Distribution of Φ-correlation between all diseases and falls across patients age groups. (C) Scatterplot