June 2017

Mining of Primary Healthcare Patient Data with Selective Multimorbid Diseases

Annette Megerdichian Azad
The University of Western Ontario

Supervisor
Dr. Michael Bauer
The University of Western Ontario

Graduate Program in Computer Science

A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

Abstract

Despite a large volume of research on the prognosis, diagnosis and overall burden of multimorbidity, very little is known about socio-demographic characteristics of multimorbid patients. This thesis aims to analyze the socio-demographic characteristics of patients with multiple chronic conditions (multimorbidity), focusing on patient groups sharing the same combination of diseases. Several methods were explored to analyze the co-occurrence of multiple chronic diseases as well as the associations between socio-demographics and chronic conditions. These methods include disease pair distributions over gender, age groups and income level quintiles, Multimorbidity Coefficients for measuring the concurrence of disease pairs and triples, and $k-$modes clustering to examine the demographics of patients with the same chronic condition. The experiments suggest that patient income quintile is not associated with multimorbidity rates, although gender and age group may play an important role in prevalence of multimorbidity and diagnosis of certain disease combinations.

Keywords: multimorbidity, pairwise disease association, multimorbidity coefficients, $k-$modes clustering, data mining

Acknowlegements

I am sincerely grateful to my supervisor, Dr. Mike Bauer, for his tremendous amount of support, guidance, wise advice and patience over the course of this research. I was very lucky to learn researching guidelines under his supervision. Mike and his positive attitude towards life were also a critical part of my positive experience at Western.

I would also like to thank Heather Maddocks and Rick Truant for their support and guidance in database access and troubleshooting. I appreciate Heather's availability whenever I had questions related to the database. Rick's support on setting up database access, tips on working with the tables and extracting information were very helpful and much appreciated.

Finally, I would like to extend my thanks to my dear parents, sister and better half for their continuous presence and encouragement during the past two years. They all explored and experienced the ups and downs of the research day by day along my side, and never hesitated to be supportive and willing to help.

Contents

Abstract i

Acknowlegements ii

List of Figures vi

List of Tables xii

List of Algorithms xvi

List of Appendices xvii

1 Introduction 1

1.1 What Happens When Chronic Diseases Co-occur 1

1.1.1 Definition of Chronic Disease 2

Characteristics of chronic conditions 2

1.1.2 Defining Comorbidity and Multimorbidity 4

1.1.3 Morbidity Burden and Patient's Complexity 5

1.1.4 How Do Diseases Co-occur? 7

Disease co-occurrence by chance 7

Disease co-occurrence by selection bias 8

Disease co-occurrence by causal association 8

1.2 Problem Definition 10

1.3 Outline of This Thesis 12

2 Approaches to Measurement and Analysis of Multimorbidity 14

2.1 How to Measure Multimorbidity 14

2.1.1 Total Disease Counts . . . 14
2.1.2 Pairwise Association . . . 15
Odds Ratios . . . 15
Risk Ratios . . . 16
Multimorbidity Coefficients . . . 17
Kappa statistic . . . 19
Concordance statistics measurement methods . . . 20
2.1.3 Clustering of Diseases . . . 22
K-means Clustering . . . 22
Factor Analysis Techniques . . . 24
Latent Class Analysis . . . 25
2.2 Comorbidity and Multimorbidity Indices . . . 25
2.2.1 Definition of a Co/multimorbidity Index . . . 26
2.2.2 Challenges Associated with Co/multimorbidity Indices . . . 28
2.2.3 Examples of Multimorbidity Indices . . . 29
2.3 Impact of Socioeconomic Status on Diagnosis and Treatment of Multimorbidity 33
2.3.1 Socioeconomic Factors and Their Impact on Health . . . 34
2.3.2 Associating Multimorbidity with Socioeconomic Factors . . . 35
2.4 Summary of Multimorbidity Measurement Approaches . . . 36

3 Data Collection 38
3.1 The Database . . . 38
3.2 Data Pre-processing . . . 39
3.2.1 Data Extraction . . . 40
3.2.2 Socioeconomic Score Computation . . . 41
3.2.3 Labeling the Attributes . . . 41

4 Data Analysis 45
4.1 Disease Frequencies and Distributions . . . 46
4.2 Analysis of Two Coexisting Chronic Diseases . . . 51
4.2.1 Total Counts and Pairwise Association of Two Chronic Diseases . . . 51

4.2.2 Disease Combinations Based on Gender 53
4.2.3 Disease Combinations Based on Age Groups 53
4.2.4 Disease Combinations based on Socioeconomic Scores 54
4.3 Clustering of Data for Two Coexisting Diseases 56
4.3.1 K-means vs. K-modes 56
4.3.2 K-modes Adjustments and Implementation 60
4.3.3 Experimental Results 64
4.4 Statistical Analysis of Three Coexisting Chronic Diseases 69

5 Conclusion **74**
5.1 Main Contributions 74
5.2 Strengths, Limitations and Implications of the Research 76
5.3 Future Research 77

Bibliography **79**

Appendices **89**

A Additional Material and Results **90**

B Modified K-modes Clustering Results **109**

Curriculum Vitae **118**

List of Figures

1.1 Conceptual diagram of comorbidity and multimorbidity 5

1.2 Comorbidity Constructs 6

1.3 Major models of comorbidity between disorders A, B and C. A=disorder A; B=disorder B; C=disorder C; RF stands for Risk Factor 9

2.1 Multimorbidity Indices Conceptual Diagram 30

4.1 Average number of diseases 46

4.2 Disease percentages for all age groups and individual diseases. The reason for representing disease trends using percentages instead of absolute counts is the variation in counts of patients in each age groups. In this line graphs, the vertical axis values are the percentages of diagnosis among age groups. 49

4.3 Disease percentages for the three income quintiles of individual diseases. 50

4.4 Disease combination counts for two coexisting diseases. Each circle on the plot represents the relative percentage of instances in the dataset who have the two chronic conditions as named on *x* and *y* coordinates. 52

4.5 Clusters of patients with hypertension and hyperlipidemia. Each pie with color c on block $[x, y]$ represents the patients with socioeconomic score x, age group y who belong to cluster c. The radius of the pie represents the relative number of patients existing in corresponding block in graph. Color shade represents gender. As the clusters suggest, patients belonging to the fourth and fifth age group have been divided from patients of age group 6. Cluster 1 belongs to male and female patients from age group four and five. The other two clusters belong to elder patients and each one holds patients of one gender attribute. The clusters for this disease pair do not separate the patients over their socioeconomic score. 68

4.6 Multimorbidity occurs mostly among middle aged or elder patients, and it affects a greater proportion of these two age groups as the number of coexisting diseases grows. 87.53% of patients diagnosed with two chronic conditions are either middle aged or elder, and this percentage grows with number of coexisting diseases, such that co-occurrence of 8 or more diseases happens only among middle aged and elder patients. 70

A.1 Disease counts for all age groups and all diseases 93

A.2 Disease percentages for all age groups and individual diseases based on gender. According to the charts, some diseases have the same trend of growth among males and females (e.g. hypertension, diabetes, heart failure, kidney disease and stomach problem), whiles some other diseases follow different trends for males and females (i.e. obesity, cancer and urinary problem). Some disease tend to appear with a higher frequency among females (e.g. depression, thyroid, and stomach problem), while some tend to appear more among males (e.g. dementia and liver disease). 94

A.3 Disease correlations of two coexisting diseases for female patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among females population. 98

A.4 Disease correlations of two coexisting diseases for male patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among males population. 99

A.5 Disease correlations of two coexisting diseases among children. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among children's population. There are nine patients in the database under *child* age group who are diagnosed with two diseases. Obesity, bronchitis and colon problem form the majority of disease combinations among children. The disease combination of bronchitis and colon problem is most commonly occurring disease pair affecting 3 children. The family income level of all patients in this age group belongs to high or highest income quintiles. 100

A.6 Disease correlations of two coexisting diseases for adolescent. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among adolescent. Bronchitis is the dominant disease which coexists with many other chronic conditions among adolescent, and it correlates with a wider variety of diseases in adolescent compared to children. The family income level of all patients in this age group belongs to high or highest income quintiles. 101

A.7 Disease correlations of two coexisting diseases for young adults. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among young adults. Coexistence of hypertension with other chronic conditions starts to appear among patients from this age group. Bronchitis is still one of the dominant diseases mostly co-occurring with depression and musculoskeletal problem among young adults. Depression and obesity become the two other dominant diseases coexisting with a variety of chronic conditions. Correlation of thyroid problem with some chronic diseases starts to appear from this age group. 102

A.8 Disease correlations of two coexisting diseases for adult patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among adults population. Bronchitis is still on the list of dominant diseases for this age group, coexisting with eleven other chronic conditions. Musculoskeletal problem becomes another dominant disease coexisting with ten chronic conditions. Hypertension correlates with many other chronic conditions in this age group and depression becomes less dominant compared to young adults age group. 103

A.9 Disease correlations of two coexisting diseases for middle age patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among middle aged patients. The variety of disease correlations become relatively larger compared to adults. Hypertension and musculoskeletal problem become two dominant diseases, coexisting with seventeen and eleven other chronic conditions respectively. Prevalence of coexistence of cancer with other conditions is greater compared to adults. 104

A.10 Disease correlations of two coexisting diseases for elder patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among elder population. Hypertension becomes the most prevalent disease and co-occurs with all other chronic conditions. 105

A.11 Disease correlations of two coexisting diseases for patients belonging to moderate income quintile. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pairs. The most dominant disease correlation belongs to the pair of musculoskeletal problem and bronchitis. The pattern of disease correlations among patients with socioeconomic score 3, is different from those who belong to group of socioeconomic score 4 or 5, and the reason is small population of patients belonging to socioeconomic score 3. 106

A.12 Disease correlations of two coexisting diseases for patients belonging to high income quintile. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pairs. 107

A.13 Disease correlations of two coexisting diseases for patients belonging to highest income quintile. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pairs. 108

B.1 Clusters of patients with depression and musculoskeletal problem. The clusters have appeared based on gender split. Every clusters represents the patients belonging to a few consecutive age groups and socioeconomic scores, which are mainly male or female. The second clusters holds the most patients who belong to all socioeconomic groups, are between age group 3 and 6, and are mainly male, while forth cluster specifies middle aged patients who belong to highest income quintile and are mainly female. 109

B.2 Clusters of patients with hypertension and musculoskeletal problem. The data split occurs according to the three attributes age, gender and socioeconomic score. The second cluster refers to female patients who belong to socioeconomic score 4, while the third cluster refers to elder male patients of the same socioeconomic score. 110

B.3 Clusters of patients with hypertension and diabetes. The clusters split the data over age group and gender. Cluster 1 belongs to patients of socioeconomic score 4 and 5, who are either adult or middle aged, while clusters 2 and 3 belong to elder female and male patients respectively. 111

B.4 Clusters of patients with cancer and musculoskeletal problem. Clusters mainly appear over socioeconomic score and gender. Cluster 1 belongs to female patients of 4 age groups who belong to highest income quintile, while clusters 2 and 3 refer to patients with high socioeconomic quintile and split males from females. 112

B.5 Clusters of patients with cancer and depression. This disease pair mainly affects females according to out dataset. Clusters divide patients of different age groups and males of elder age group from elder females. 113

B.6 Clusters of patients with hypertension and depression. Clusters divide the patients over gender and age group. Females of age group 5 and 6 are clustered in clusters 3 and 2 respectively, while most of the male patients who have this disease combination are clustered in cluster 1. 114
B.7 Clusters of patients with hypertension and cancer. Clusters clearly appear over gender and socioeconomic score split. Patients from highest income quintiles are clustered separately from patients from high income quintile. Male and females who belong to socioeconomic score 4, have been divided into two clusters. 115
B.8 Clusters of patients with hyperlipidemia and musculoskeletal problem. Clusters mainly divide the patients by gender and age group. However, this is not a perfect split of patients based on demographic characteristics. 116
B.9 Clusters of patients with bronchitis and musculoskeletal problem. Patients distribution over age groups and socioeconomic scores is similar to patients with depression and musculoskeletal problem. The division of patients into clusters is also similar in these two disease pairs. There are five clusters dividing the data: other than cluster 2 which mainly divides males from females, the other clusters belong to patients of one to three consecutive age groups. 117

List of Tables

1.1 Alternate models of comorbidity between disorders A and B. r_{AB} = the association between risk factor of disorder A and risk factor of disorder B. 10

2.1 All possible observations for two arbitrary disorders occurring in an individual . 15

2.2 Interpretation of Kappa 20

3.1 Missing attributes counts for patients with one or more chronic diseases. 40

3.2 Life stages and its characteristics. 42

3.3 The structure of final data table used for processing, analyzing and clustering . 44

4.1 This table aims to represent the distribution of patients who share minimum one characteristic (i.e. number of coexisting diseases, gender, age group). The first column stands for the number of coexisting diseases within one patient. In our data table, the most number of coexisting diseases are found among four patients who are diagnosed with eleven chronic conditions. Every row in the table represents percentage of patients who have a specific number of diseases. The distributions are represented for male and female sub groups as well as every one of the six age groups separately. 47

4.2 Percentages of patients for all age groups based on gender 53

4.3 Patient distribution percentages among all age groups for patients with two chronic conditions. 54

4.4 Patient counts for each age group based on family income quintiles. Only patients with two chronic condition are listed in this table. 55

4.5 10 most frequently occurring disease combinations derived from CPCSSN database [90]. Disease combinations are sorted by frequency of occurrence among patients in CPCSSN database. The column *Patient Counts in DELPHI* represents the number of individuals with corresponding diseases combination in DELPHI database which was used for this research. 67

4.6 The distribution of patient characteristics for individuals diagnosed with commonly co-occurring disease pairs. 69

4.7 The summary of clustering results for 10 commonly curring disease pars. Every row represents the summary of clustering results for one disease pair. The number of clusters which minimize the within-cluster sum of squares are included in second column. The attributes upon which the clustering has occurred are indicated with check marks. 69

4.8 The list shows most frequently occurring disease combinations derived from DELPHI database. All disease combinations occurring among more than 20 patients are listed in this table. Disease combinations are sorted by frequency of occurrence among patients. The column *Patient Counts* represents the number of individuals with corresponding disease combination in the DELPHI database. 71

4.9 The distribution of patients characteristics for individuals diagnosed with commonly co-occurring chronic disease triples. Abbreviations: M=Male, F=Female, HT=Hypertension, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, DP=Depression, AT=Arthritis, TD=Thyroid, MP=Musculoskeletal Problem. 71

4.10 Multimorbidity coefficients of commonly occurring pair of chronic conditions with a third disease. Abbreviations: HT=Hypertension, OS=Obesity, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, HF=Heart Failure, DP=Depression, AT=Arthritis, SK=Stroke TD=Thyroid, KD=Kidney Disease, OP=Osteoporosis, DT=Dementia, MP=Musculoskeletal Problem, SP=Stomach Problem, CP=Colon Problem, LD=Liver Disease, UP=Urinary Problem. 73

A.1 A list of 20 chronic disease categories and their corresponding ICD-9 (International Classification of Diseases 9th revision) disease codes [90]. 92

A.2 Odds ratios computed for estimating the association between every disease pair. In order to avoid the selection bias, for calculating odds ratios, all existing patient records in DELPHI database have been included, regardless of the conditions they are diagnosed with; This also includes patients who have not been diagnosed with any chronic conditions. Abbreviations: HT=Hypertension, OS=Obesity, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, HF=Heart Failure, DP=Depression, AT=Arthritis, SK=Stroke TD=Thyroid, KD=Kidney Disease, OP=Osteoporosis, DT=Dementia, MP=Musculoskeletal Problem, SP=Stomach Problem, CP=Colon Problem, LD=Liver Disease, UP=Urinary Problem. 95

A.3 95% confidence intervals (CI) for all disease pairs to estimate the precision of the odds ratios. A large CI indicates a low level of precision of the odds ratio, whereas a small CI indicates a higher precision of the odds ratio. The 95% CI can be used to present statistical significance if it does not overlap OR=1. Abbreviations: HT=Hypertension, OS=Obesity, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, HF=Heart Failure, DP=Depression, AT=Arthritis, SK=Stroke TD=Thyroid, KD=Kidney Disease, OP=Osteoporosis, DT=Dementia, MP=Musculoskeletal Problem, SP=Stomach Problem, CP=Colon Problem, LD=Liver Disease, UP=Urinary Problem. The green highlights on some of the cells show a relatively higher association for corresponding disease triples to co-occur. 96

A.4 Multimorbidity coefficients (MC) measure the association of all disease pairs. For avoiding effects of selection bias, the population used for performing the computations of MC contains all patient records in DELPHI database, regardless of the number and type of diseases they are diagnosed with; this also contains patients with no chronic conditions. Abbreviations: HT=Hypertension, OS=Obesity, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, HF=Heart Failure, DP=Depression, AT=Arthritis, SK=Stroke TD=Thyroid, KD=Kidney Disease, OP=Osteoporosis, DT=Dementia, MP=Musculoskeletal Problem, SP=Stomach Problem, CP=Colon Problem, LD=Liver Disease, UP=Urinary Problem. 97

List of Algorithms

1 Modified $k-$modes algorithm for clustering demographic information of patients diagnosed with most commonly occurring pairs of chronic diseases: Part 1 65
2 Modified $k-$modes algorithm: Part 2 66

List of Appendices

Appendix A Additional Material and Results 90
Appendix B Modified K-modes Clustering Results 109

Chapter 1

Introduction

Multimorbidity occurs ten to fifteen years earlier in people living in areas with socioeconomic deprivation [4]. Research also suggests that the number of people with multiple chronic conditions will increase in coming years [3]. Additionally, patients with more than one disease are not only more likely to be diagnosed with new diseases and mental health disorders, they also have higher levels of psychological distress, significant loss of function, higher utilization of overall health-care, poorer mental health and quality of life, and limited physical functioning and limited ability to work [4, 45, 52, 106, 108, 126]. However, not all combinations of chronic conditions have the same effect on functional status of the patient. Specific combinations of chronic conditions may lead to a greater risk of disability for patients [47].

This chapter provides an overview of and basic definition of chronic diseases, comorbidity, multimorbidity, and the possibilities due to which two diseases may co-occur. It also defines the research goals and plans for this thesis, as well as the anticipated strengths and challenges. The outline of this dissertation is presented in the final section of this chapter.

1.1 What Happens When Chronic Diseases Co-occur

This section provides the definition of the basic concepts associated with chronic diseases, their co-occurrences and effects.

1.1.1 Definition of Chronic Disease

A disease is defined as *chronic* if it lasts for a relatively long period of time and is irreversible without any complete recovery [92]. The complete list of chronic diseases can be found at International Classification of Diseases (ICD), which is a clinical cataloging system to classify and code all diseases [81]. The ICD was developed to identify trends and statistics of health. It is a classification standard for clinical and research purposes to identify diseases, disorders, injuries and other related health conditions [93].

The diagnosis of a disease starts by indication of symptoms, risk factors and relevant analysis obtained by physical examinations. The treatment process is a stepwise approach with regular control of patients' laboratory analysis results, symptoms and signs [76]. Currently, most of the medical knowledge is available for a single condition, and it may not be fully applicable and acceptable for patients with multimorbidity[1]. Therefore, during the treatment process of one disease, the efficiency of treatment of other disease(s) might be affected.

Characteristics of chronic conditions

Chronic conditions have nine characteristics: etiology, duration, onset, recurrence/pattern, prognosis, sequelae, diagnosis, severity and prevalence [92].

Etiology: Sometimes it is not easy to define the etiology of a chronic condition. In some cases, the presence of confounding factors may lead to a non-linear relationship between the exposure to a pathogen and the presence of disease. These confounding factors may include exposure to behavioral risk factors, such as smoking, poor nutrition and excessive alcohol consumption, as well as genetic and environmental factors [75, 82].

Duration: Duration of a chronic disease is one of the characteristics of chronic conditions widely used for defining the recovery time of conditions. There are three main time durations

[1]The term "Multimorbidity" is defined in section 1.1.2

when defining chronic conditions: 3 months, 6 months and 12 months. Some studies rely on 3 months duration and reject 12 months when categorizing chronic conditions [94], while others reject 3 months duration for classifying a disease as chronic condition, as they believe that some acute conditions with lengthy recovery times will be classified chronic by mistake [110]. In the majority of studies, the duration of a chronic condition is defined to be at least 6 months, as in [24, 65]. However, it is suggested that no standard duration can be applied to a chronic condition, because the actual duration of a chronic condition is unique. It depends on the condition itself and the person experiencing the condition [22].

Diagnosis: Although the diagnosis can be a defining characteristic of a chronic condition, it is not the only indicative of all perspectives of a chronic condition and basing the definition of chronic conditions on diagnosis might under-represent the prevalence of chronic conditions, as some patients might have conditions that meet the criteria for diagnosis, but are not given a chronic label [111]. However, in this research, the two terms of being diagnosed with a disease and having a chronic condition will be used interchangeably.

Onset: The onset of chronic conditions is insidious and gradual. Patient's age can be one of the relevant aspects to onset of chronic conditions [110, 111].

Recurrence/pattern: The patterns of chronic conditions have wide variation both between and within conditions. Some conditions have a deteriorating course, while some others are episodic [22, 110].

Prognosis: While the main focus of prognosis is placed on the management of the condition and quality of life [30], lack of certainty for cure, incurability and poor prognosis are also concerns discussed in studies regarding chronic conditions [22, 94].

Sequelae: According to "Dorland's Illustrated Medical Dictionary", the term "sequelae" stands for physical or mental consequences that are caused by, or follow the course of a condition [117]. Sequelae affects patients' quality of life by adding physical disabilities, limita-

tions of activity, reliance on medication or technical devices, or an increased need for medical care [55, 94, 110].

Severity: The severity of a chronic condition may change continuously over the course of the condition, and it depends on the stage of the condition [30]. Studies claim that conditions where the severity of a disease, is not keeping the patient from daily functioning, and has a little effect on patients' physical or mental well-being, should not be included in the classification of chronic conditions [94].

Prevalence: According to a study in general practice, a disease is classified as chronic only when its prevalence is relatively high in the population being studied. However, this approach has been disputed by some other studies where the chronic conditions are classified as chronic if their prevalence is relatively low. In this case, the overall prevalence of chronic conditions is the purpose of study [94].

1.1.2 Defining Comorbidity and Multimorbidity

In the medical literature there are several definitions and interpretations for the two terms *comorbidity* and *multimorbidity*. These two terms were used interchangeably by many authors for a long time [44]. Researchers have compared the definitions of these terms in a large number of relevant papers and their results show that the definition of comorbidity and multimorbidity are ambiguous [119]. According to some definitions, coexistence of several diseases is called comorbidity, while some other definitions claim that coexistence of other conditions with respect to an index disease is called comorbidity [119]. In our work we rely on the classical definition for these terms. The term comorbidity was introduced in 1970 by Feinstein and is defined as "co-occurrence of other medical conditions additional to an index disease". In other words, comorbidity refers to any distinct additional entity that has existed or may occur during the clinical course of a patient who has the index disease under study [39]. In this definition, *index disease* refers to the primary disease or disorder under study. The additional clinical entities (diseases or disorders) may occur or exist during the clinical course of index disease.

Figure 1.1 shows the conceptual diagrams of comorbidity and multimorbidity [13].

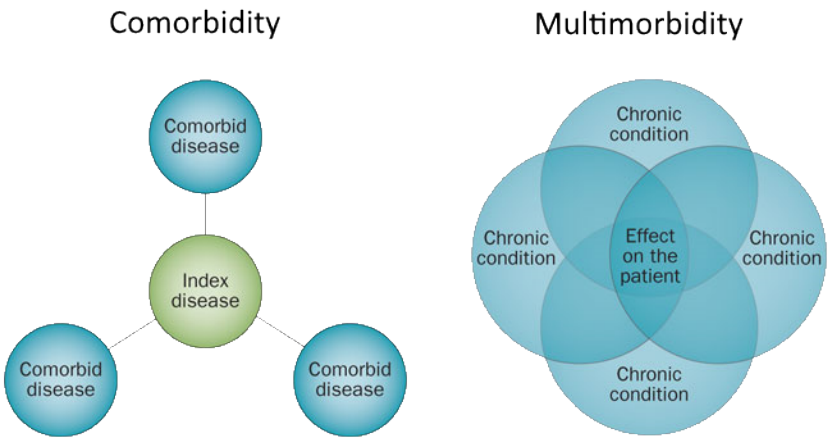


Figure 1.1: Conceptual diagram of comorbidity and multimorbidity

Multimorbidity is a more recently introduced term in chronic disease epidemiology and is defined as "co-occurrence of two or more medical conditions within a person, where one is not necessarily more central than the others" [119].

There are two major considerations in the chronological analysis of comorbidity and multimorbidity: time span and sequence. The former discusses the span of the time across which the patient has been diagnosed with two or more conditions. There has been less interest in conditions which happened at the same point in time than conditions co-occurring across a period of time [12]. The latter addresses the sequence in which comorbidities appear. The order of diseases a patient is being diagnosed with plays a very important role in patient characteristics, although from a cross-sectional perspective two patients may have the two exact same chronic conditions. For instance, a patient diagnosed with hypertension who develops diabetes may be very different from a patient with diabetes who is later diagnosed with hypertension.

1.1.3 Morbidity Burden and Patient's Complexity

Morbidity burden is defined as the overall impact of the different diseases on an individual, taking into account disease severity. *Patient's complexity* is a more broad perspective, considering the morbidity burden while taking into account other health-related attributes. To make these

definitions more explicit, consider the following example:

A 60-year old woman suffering from hypertension, depression and heart failure, who is from an immigrant family living in Canada with limited skills in English, and who is also taking care of a son with multiple sclerosis. Her cardiologist, focusing on her heart failure, would consider her depression and hypertension as comorbidities to the index disease heart failure. Her primary care physician will treat her hypertension, depression and heart failure equally and will describe them as multimorbidity. Her morbidity burden would be shaped by the presence of all chronic conditions the patient is diagnosed with, taking their relative severity into account. Her complexity would be determined considering her health condition, i.e. her chronic diseases, her immigration background, English language skills and fluency and her role as a caretaker for her son (Figure 1.2) [118].

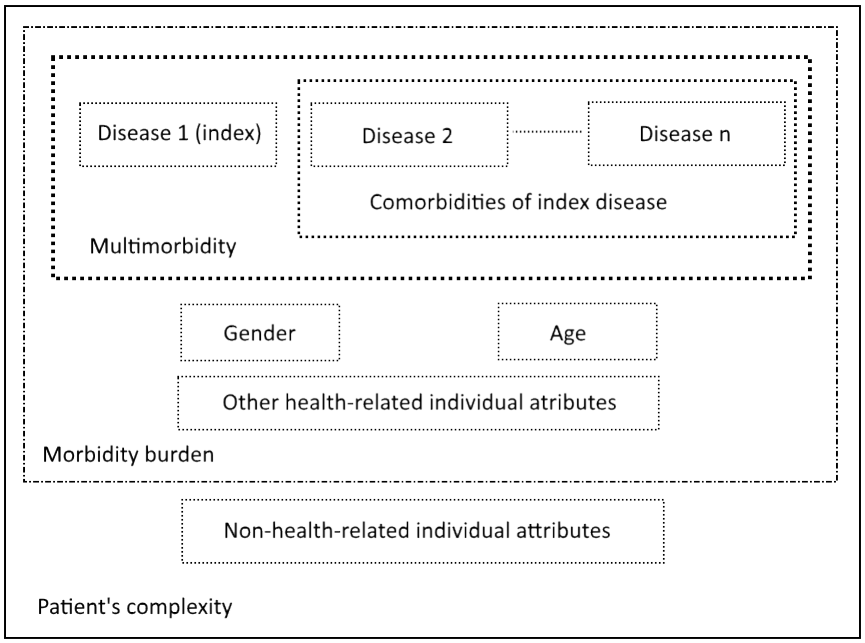


Figure 1.2: Comorbidity Constructs

There are three main areas where comorbidity and multimorbidity constructs are being measured; clinical care research, epidemiology and public health, and health services and financing. In each criteria, certain morbidity constructs are the area of interest for research and computational purposes.

From a clinical research perspective, patient's complexity is relevant to individual's con-

struct of comorbidity, where the index disease gets the highest priority. Patient's complexity is a newly emerging comorbidity construct, which acknowledges the influence of morbidity burden not only by health-related construct, but also by patient behavioural characteristics [99]. In this perspective, although the focus is explicitly on the patient as a whole, the morbidity burden becomes the more reliable construct to be considered. In this context, all conditions of a patient are being given the same privilege.

From an epidemiological and public health perspective, the major focus is on the causes of concurrent diseases. Therefore, both comorbidity and multimorbidity approaches will be of interest in this context.

From a health services and financing perspective, estimation of costs for patient treatment will not be calculated as sum of the costs for every single condition the patient has. Depending on a patient's condition and outcomes of interactions among co-occurring diseases, this cost can be either greater or less than the sum of separate costs. Therefore, patient complexity plays an important role in cost estimation in this area [118].

1.1.4 How Do Diseases Co-occur?

Diseases can be observed to co-occur due to three main reasons: chance, selection bias, or one or more types of causal association.

Disease co-occurrence by chance

When diseases co-occur by chance, there is no etiological association between them. If we assume $x\%$ of individuals in a population are affected with disease A and $y\%$ of them are affected with disease B, then $(x/100) \cdot (y/100)$ of individual will have both x and y. There are two assumptions made in this context: probability can be assigned to any disease, and the probabilities of disease occurrences are independent in an individual who is suffering from more than one disease. This factor of measurement only applies to diseases that have no associations between their risk factors. In other words the $r_{AB} = 0$ where r_{AB} is the risk factor of A and risk factor of B. Another name for referring to disease co-occurrence by chance, is *coincidental comorbidity*, or *random comorbidity*.

Disease co-occurrence by selection bias

Assume that a hospital wants to conduct a study where the goal is to estimate the association between the occurrence of two diseases D_1 and D_2. For doing so, a control-case study is conducted [2]. In the study, hospitalized patients are considered as cases if they have D_2 and controls if they have D_3. The prevalence of disease associations then will be calculated by comparing the prevalence between D_1 with cases D_2 and with controls D_3. The results show that although there is an association between D_1-D_2 and D_1-D_3 in hospitalized individuals, the association were null in source population.

As a conclusion, two diseases that are independent in the general population may become "seriously associated" in hospital-based case-control studies, and the reason is higher hospitalization probability for patients with more than one disease - even if the diseases have no association [7, 107]. In other words, clinical samples are not random samples from all people in the general population. Individuals in clinical databases have more severe and greater numbers of symptoms than the ones in general population [97]. Although, this bias can be avoided by using community samples rather than clinical databases [118].

Disease co-occurrence by causal association

According to Neale and Kendler [87], comorbidity can be categorized into thirteen models. Table 1.1 provides a brief description for each of these models. One of the categories among the thirteen models is "co-occurring by chance", which was covered in previous sub-section. The twelve remaining categories will be described in this subsection.

In *Alternate Forms (AF)*, the co-occurring disorders are alternate manifestation of a single liability. In *Random Multiformity (RM)* model, the risk factors for disorders are not associated, but each one of them can increase the probability of the other disorder to increase. *Extreme Multiformity (EM)* is a specific form of Random Multiformity, where the effect of risk factors of each disorder is extremely high; if the individual has one of the disorders, probability of having the other disorder is almost equal to one. In multiformity models the hypothesis is the idea

[2] Control-case study is a study which compares patients who have a disease or outcome of interest (cases) with patients who do not have the disease or outcome (controls), and looks back retrospectively to compare how frequently the exposure to a risk factor is present in each group to determine the relationship between the risk factor and the disease [77].

that comorbidity between two disorders occurs because being affected by one disorder directly increases the risk for having the other disorder [87, 97]. In *Three Independence Disorders (TD)* the presence of the diagnostic features of each disorder is actually due to its specific risk factors. *Correlated Liabilities (CL)* refers to a group of disorders where the risk factors for each disorder are associated while *Reciprocal Causation (RC)* refers to a group of disorders where one of the disorders may cause the others. The difference between *Reciprocal Causation (RC)* and *Correlated Liabilities (CL)* models is that, in RC models genetic and environmental risk factors are united into a common latent phenotype before causation, whereas in CL model, no such combination between the genetic and environmental risk factors has occurred before causation [97]. Figure 1.3 shows a simple diagram of main causal association models described by Neale and Kendler:

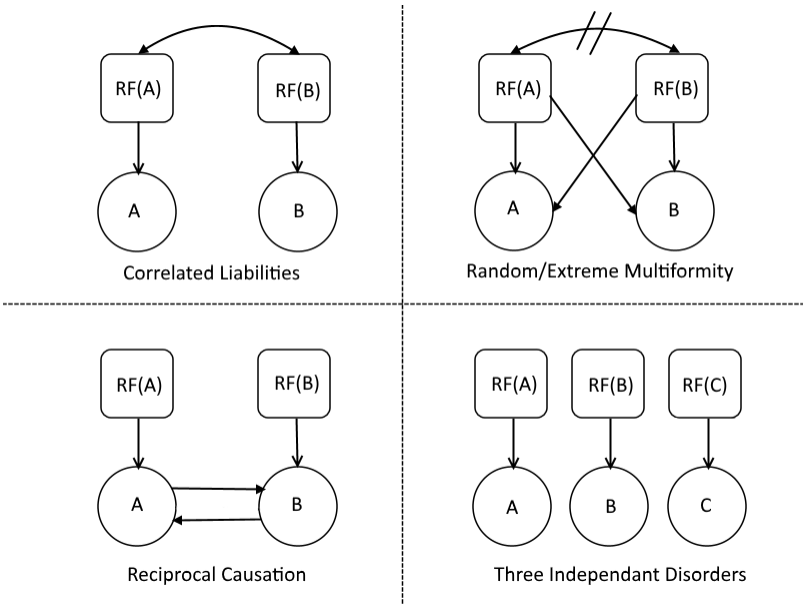


Figure 1.3: Major models of comorbidity between disorders A, B and C. A=disorder A; B=disorder B; C=disorder C; RF stands for Risk Factor

According to Neale and Kendler's comorbidity models, existence of a certain disease or condition, does not necessarily cause the other. That is, "association does not necessarily imply causation" [20].

Name	Description
Alternate Forms (AF)	The disorders are comorbid because they are alternate manifestation of a single liability. In this category individuals have disorder A with probability $p(A)$, disorder B with probability $p(B)$ and both disorders A and B with probability $p(A)p(B)$
Random Multiformity (RM)	Individuals who have disorder A, have the risk of disorder B increased by p_A. Individuals having disorder B, have the risk of disorder A increased by p_B.
Random Multiformity of A (RMA)	Submodel of RM where $p_B = 0$.
Random Multiformity of B (RMB)	Submodel of RM where $p_A = 0$.
Extreme Multiformity (EM)	Individuals who have disorder A, will have disorder B and individuals who have disorder B, will have disorder A.
Extreme Multiformity of A (EMA)	Submodel of EM in which only individuals who have disorder A, will have disorder B and not the other way around.
Extreme Multiformity of B (EMB)	Submodel of EM in which only individuals who have disorder B, will have disorder A and not the other way around.
Three Independent Disorders (TD)	The comorbid disorders are separate from either disorder occurring alone.
Correlated Liabilities (CL)	Two disorders are comorbid because their risk factors are associated. The relationship between risk factors of disorder A and risk factors of disorder B is greater than zero ($r_{AB} > 0$).
Reciprocal Causation (RC)	Comorbidity between two disorders occur because two disorders cause each other. In other words, disorder A and disorder B cause each other in a feedback loop.
A Causes B (ACB)	Submodel of RC where disorder A cases disorder B but not the other way around.
B Causes A (BCA)	Submodel of RC where disorder B cases disorder A but not the other way around.

Table 1.1: Alternate models of comorbidity between disorders A and B. r_{AB} = the association between risk factor of disorder A and risk factor of disorder B.

1.2 Problem Definition

Given disease combinations in patients, one question that arises is what might be similar among people who are sharing same diseases? Are they of the same gender? Do they belong to the same age group? Are they from the same ethnic group? Are they coming from a similar socio-

demographic background? What similarities or differences can be found in their life style and quality of life?

To the most part research is focused on identifying causes and effects of a single disease, and as there is very little known about differences and similarities among patients who have more than one disease. Although fully realizing this goal still has a long way to go, in this research we are taking a step forward by exploring and identifying methods that may be used to shed light on socio-demographic characteristics of patients with multiple chronic conditions. This includes investigating various statistical and clustering methods which can efficiently be applied for examining patients' socio-demographic characteristics in terms of their similarities and differences, as well as approaches to visualize the data of multimorbid patients and their relationships.

Starting from identifying patients in our database with multimorbidity, we also need to identify socio-demographic information associated with those patients. Besides performing statistical measurements such as total disease counts and pairwise disease associations, we will apply clustering algorithms to identify the similarities among patients who share same chronic diseases and determine whether or not the factors we have extracted as socio-demographic information are related to multimorbidity in terms of affecting multimorbidity.

There are three anticipated challenges for this research: (1) Limited availability of socioeconomic variables for patients data; (2) Limited generalizability of the research results to national population; and (3) Limited patient data over a long time span. The first anticipated challenge means that we must contend with incomplete/missing variables for patient data which limits the project scope because it limits the availability of correct/accurate data to work with. Collected data becomes smaller in size with selection of complete and free of error instances to increase the level of accuracy of the research, and on the other hand the availability of sociodemographic information, such as patient ethnicity, languages they speak, income level, education level, etc., is very limited. Specifically, in the DELPHI database[3], which is the source of data used in this research, these variables were absent for almost all patients existing in the database. For this research, patient age, gender and residential codes will be used for analysis purposes; the lack of all other socio-demographic variables represents a significant

[3]The definition and detailed information about DELPHI database can be found in Chapter 3.

limitation on the scope of the project. The second anticipated challenge impacts the generalizability of research outcomes to a national population, as the DELPHI database does not contain data from health care providers nationally. The third challenge means that the work must contend with a lack of data for each patient over time. Information such as when a patient was first diagnosed with a certain disease, what is the order of diseases the patient got diagnosed with over time, how long was the time period between the diagnosis of diseases, what was the patient's socioeconomic status when the diseases appeared for first time, would allow us to do more in-depth analyses. This would also allow the research to identify potential factors that might impact diseases and which ones are being affected due to diagnosis of multiple chronic illnesses.

The initial basis for this research is data on frequently occurring combinations of multiple chronic diseases from a different research project [90], where disease combination counts have been derived from a larger data source (the CPCSSN database[4]). Having access to this information derived from a greater population size ensures the reliability of results to some extent, and allows us to do further research on patient characteristics of those diagnosed with commonly occurring disease combinations.

1.3 Outline of This Thesis

In this research we will elaborate on measurement and analysis approaches of multimorbidity. To do so, we review the available methods for measuring and analyzing the burden of multimorbidity and its relationship with a patient's socioeconomic status in life.

Chapter 2 consists of three main sections. The first section describes how to measure multimorbidity by applying existing statistical and clustering methods. It mainly focuses on most frequently used pairwise association measurement methods and commonly used clustering methods for clustering chronic diseases; describing how each one is calculated, and the advantages and weaknesses of them. The second section of Chapter 2 is dedicated to illustration of multimorbidity indices (which are scoring tools for existing diseases within a person along

[4]The definition and detailed information about CPCSSN database can be found in Chapter 3. This data source contains a total of 600265 electronic patient records as of data extraction period (September 30, 2013).

with their severity degree) and challenges associated with them. This section concludes with a brief introduction to four most commonly used indices developed for measuring outcomes like subsequent risk of mortality, prediction of hospital re-admissions after getting a prescription and health-related quality of life. The final section of Chapter 2 focuses on the influence of socio-economic factors on the burden of multimorbidity among patients, and its association with a patient's age, gender, race and ethnicity.

Chapter 3 introduces the database used for analysis in this research. This Chapter describes when and how the database was created, how often it gets updated, the number of records it holds, the features and the coding strategies used for de-identification of existing records. The Chapter also includes the description of the data extraction process, including identification of missing values, data cleaning and strategies applied for dealing with missing data. It concludes with the representation of a data table created from the database after data extraction and pre-processing steps.

Chapter 4 describes the data analysis. It presents the results of statistical and clustering techniques used for analyzing the relationships between socioeconomic status and two/three coexisting diseases among patients sharing the same chronic diseases. The first section of this Chapter focuses on analytical results of statistical multimorbidity measurement methods explained in Chapter 2 for two coexisting chronic diseases. The second section describes the clustering method used for clustering patients with two coexisting diseases. The clustering algorithm was performed separately on all patients who share the same pair of diseases and for most commonly occurring disease pairs. Some modification have been made to the algorithm in order to adjust it for the data set, as well as the outputs and results of the algorithm, which is also described in this section. The third and final section of this Chapter describes the results for analysis of three coexisting diseases. As the counts for combination of three coexisting diseases were far less than counts of two coexisting diseases, not all measurement methods applied on diseases pairs were applicable for three coexisting diseases. Therefore, this section only includes statistical distribution results.

The final Chapter summarizes and concludes the research and its final results. Research strengths, limitations and recommendations for future investigations in the area of multimorbidity are also represented in this Chapter.

Chapter 2

Approaches to Measurement and Analysis of Multimorbidity

2.1 How to Measure Multimorbidity

Determining the prevalence of co-occurring diseases in certain population groups has always been a research need. The most straightforward approach for measuring multimorbidity is counting the co-occurring diseases among individuals, which is discussed in the first part of this section. More complex approaches try to find some kind of association among diseases. The second part of this section, is mainly focused on measuring the association of disease pairs. The methods described in this part are odds ratios, risk ratios, multimorbidity coefficients, Kappa statistics and concordance statistics.

2.1.1 Total Disease Counts

The most straightforward way to measure multimorbidity is to count the number of diseases for each individual in a group and then calculate the average number of diseases per person by age. This approach can be used in research to determine the prevalence of multimorbidity in a certain age group. For instance,Van den Akker *et al.*[120] have used this method to determine the prevalence of multimorbidity in the elderly for the Dutch population. Due to the exponen-

tial nature[1] and discrete values derived from disease counts, any regression method works for calculating the total burden of multimorbidity. Lappenschaar [76] found that there was approximately one disease at age 40 and five at age 80 in the Netherlands. However, this method is very coarse and does not provide much information about the distribution of diseases and their association with factors other than age. Therefore, in order to gain more insight, other metrics are used.

2.1.2 Pairwise Association

When diseases are to be compared in pairs, there are several ways to express their association. Some of the measurement methods for pairwise association are described in the following.

Odds Ratios

One of the most widely used methods is using odds ratios to figure out the association between two disorders. Odds ratios are commonly used in case control studies, however, they can also be used in cross-sectional and cohort studies [102]. Their popularity is due to ease of calculation and good estimation of relative risks [5].

		Disorder 1 (D_1)		
		Presence	Absence	Total
Disorder 2 (D_2)	Present	a	b	$a + b = R_1$
	Absent	c	d	$c + d = R_2$
	Total	$a + c = C_1$	$b + d = C_2$	$a + b + c + d = N$

Table 2.1: All possible observations for two arbitrary disorders occurring in an individual

For disorder D_1 and disorder D_2 the odds ratios is a quantitative measurement method defined as the odds of being exposed to disorder D_2 if one has D_1, divided by the odds of being diagnosed with D_2 if one does not have D_1. Table 2.1 shows all possible combinations of two arbitrary disorders D_1 and D_2, where every cell shows the prevalence of occurring (or not

[1]The results of disease counts from several researches show that the number of co-occurring chronic diseases within an individual drops very fast, as the number of diseases increase. That is the reason why the term exponential nature for chronic disease count is being used. Due to the exponential nature of co-occurring diseases, it is very challenging to figure out the association between diseases among patients diagnosed with more than three or four diseases, as the available data for them is negligible, regardless of the collected database.

occurring) of the corresponding disorders. According to Table 2.1, the odds of being exposed to D_2 if one has D_1 is $\frac{a}{c}$, and the odds of being exposed to D_2 if one does not have D_1 is $\frac{b}{d}$. Therefore, the odds ratios (OR) of D_1 and D_2 is calculated as follows:

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc} \tag{2.1}$$

This means that if a patient is exposed to D_2, he is OR times more likely to have D_1 than if he is not exposed to D_2. Odds ratios are used to determine whether a particular exposure is a risk factor for a particular outcome [102]. Many major studies on psychiatric comorbidity, such as *National Comorbidity Survey of United States* [70], the *National Psychiatric Morbidity Survey of Great Britain* [62], and *The Netherlands Mental Health Survey* [8], are using odds ratios for quantifying comorbidity. The association of narcolepsy (a neurological disorder) with many psychiatric disorders has also been discovered using odds ratios [34, 46].

Odds ratios can also be modified to accommodate other patient demographics, such as their gender, age, education level and other factors, using logistic regression. An example of using odds ratios with additional patient demographics can be found in a study identifying multimorbidity patterns in elderly performed in Stockholm [80].

Risk Ratios

Risk ratios (also called as rate ratios or relative risks) is another popular method for quantifying association among two disorders, which calculates the ratio of the risk of occurrence of a disease among exposed group of people to that among the unexposed [11]. Risk ratios are not symmetric, therefore they can be defined for both diseases. According to Table 2.1, the relative risk of D_2 associated with D_1 (RR_{D1}), is defined as the risk of occurrence of D_2 among individuals who are exposed to D_1, divided by the risk of occurrence of D_2 among individuals who are not exposed to D_1. So, the risk ratio will be calculated as following:

$$RR_{D1} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{a(b+d)}{b(a+c)} = \frac{aC_2}{bC_1} \tag{2.2}$$

Similarly, the risk ratio of D_1 associated with D_2 will be defined as:

$$RR_{D2} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{a(c+d)}{c(a+b)} = \frac{aR_2}{cR_1}$$

Risk ratios are used in assessing the likelihood that an association represents a causal relationship [11]. Risk ratios are widely used in psychiatric epidemiology [5].

Odds ratios are very similar to risk ratios, particularly in less prevalent diseases. According to Table 2.1, if a disease is rare, it will occur among very few individuals in a population, therefore, $b + d \approx d$ and $a + c \approx c$. This results that, $OR \approx RR$ in rare disease cases.

Multimorbidity Coefficients

Another commonly used method for measuring pairwise association is the use of a multimorbidity coefficient (also known as cluster coefficient or CC), which is defined as the division of observed rate of co-morbidity (multimorbidity) by the rate which is expected under the null-hypothesis of no substantive association between the separate disorders [5]. According to Table 2.1, the observed rate of comorbidity/multimorbidity, is $\frac{a}{N}$, and the expected rate of association is $\frac{a+c}{N} \cdot \frac{a+b}{N}$. Therefore the multimorbidity coefficients (MC) will be calculated as follows:

$$MC = \frac{a/N}{[(a+c)/N] \cdot [(a+b)/N]} = \frac{aN}{(a+c) \cdot (a+b)} = \frac{aN}{C_1 R_1} \tag{2.3}$$

Odds and risk ratios can calculate the overall association among two diseases, but they cannot separate non-random comorbidity (disease co-occurrences by causal association) from co-incidental(random) comorbidity, neither can they measure directly the amount of co-occurrence

among pairs [5]. According to the example explained in *Quantifying psychiatric comorbidity* [5], all three measures of *OR*, *RR*, and *MC* are equal to one if the association between disorders is only coincidental, which indicates that the two disorders are not associated and cannot be clustered in the same cluster. However, when there is an association among diseases, *MC* shows a relatively smaller value compared to other two measures. Also, the more the prevalence among two disorders increases, the difference between *MC* and the two other measures grows bigger, and the odds ratios value becomes progressively larger than the risk ratio. The fact that measures of association diverge increasingly in *OR* and *RR* when the prevalence of disorders rises, makes the two measures not appropriate to be used for sicker populations or large populations. Because in sicker (or large) populations, the average number of disorders within a person is higher compared to a relatively healthier population, and as long as the two measures of associations are calculating the prevalence according to the population, they will show a very high prevalence among majority of disease pairs. For more details, refer to the original paper [5].

Odds ratios and risk ratios can only show the association between pairs of disorders, while multimorbidity coefficients can express association among any number of diseases by dividing the actual rate of multimorbidity by expected numbers of cases. Although calculating the expected rate is a tedious process due to the very large number of combinations which should be considered. To be more specific, if there are n disorders in a population, each with a different prevalence (aka $p_1, p_2, ..., p_n$), and the goal is to find the association of any k disorders ($k \leq n$), the number of combinations for expected rate of associations will be the combination of n and k:

$$C(n,k) = \binom{n}{k} = \frac{n!}{k!}(n-k)!$$

And the expected probability (P_e) for each combination will be the product of probabilities of k disorders that occur, multiplied by the product of probabilities of the rest of disorders, if they do not occur. If we arrange the n disorders the way that k disorders expected to occur are

in first k spots, the expected probability of an individual combination will be:

$$P_e = \prod_{i=1}^{k} p_i * \prod_{j=k+1}^{n} (1 - p_j)$$

As it is shown above [120], it is possible, but laborious to calculate the association between many diseases using multimorbidity coefficients.

Kappa statistic

Kappa is a statistical measure primarily used for assessing the degree of interobserver[2] agreement in the form of binary ratings (in our case, presence or absence of diseases), which can be used for finding pairwise disease association. The advantage of statistical methods over other pairwise association methods is that they can be adjusted to measure expected coincidental (random) co-morbidity [89]. Kappa is defined based on the difference between actual observed association and expected association by chance [121]. It is calculated as follows:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{2.4}$$

where, according to Table 2.1, $p_0 = (a + d)/N$ which relates to the observed proportion of association of diseases, and $p_e = (C_1R_1 + C_2R_2)/N^2$, which relates to the expected proportion of association by chance[3]. Kappa is standardized to lie between -1 and 1. If $p_0 = 1$, its value will be equal to one[4], which is interpreted as complete association among diseases. A zero

[2]Interobserver: occurring between individuals performing the same and especially a visual task [85]. Studies that measure the level of agreement between two (or more) observers should also consider that an agreement (or disagreement) can simply occur by chance [121] among observers.

[3]As mentioned above, p_e is the probability of both diseases occurring together. To calculate it, having the prevalence of occurrence (and non-occurrence) of each disease, p_e will be calculated as the sum of probabilities of both diseases occurring together, and probability of neither of diseases occurring. According to Table 2.1 the probability of occurrence of D_1 and D_2 is $(a + c)/N$ and $(c + d)/N$ respectively. Similarly, the probability of non-occurrence of D_1 and D_2 is $(a + b)/N$ and $(c + d)/N$ respectively. Therefore p_e will be calculated as follows:

$$p_e = \frac{(a+c)}{N}.\frac{(a+b)}{N} + \frac{(b+d)}{N}.\frac{(c+d)}{N} = \frac{C_1R_1 + C_2R2}{N^2}$$

[4]The value of p_0 will be equal to 1, if and only if $a + d = N$. This implies that $b + c = 0$, and as long as b and c can be zero or positive, they both have to be zero, to meet the requirements for p_0 to be 1.

value for Kappa, implies that $p_0 = p_e$, which means that the observed association is equal to the expected association by chance. In other words, the co-occurrence among diseases is exactly what would be expected by chance. A negative value for Kappa implies that there might not be any association among disorders because the observed association rate is less than expected association by chance [121].

Kappa	Association level
<0	Less than chance
0.01–0.20	Slight association
0.21–0.40	Fair association
0.41–0.60	Moderate association
0.61–0.80	Substantial association
0.81–0.99	Almost perfect association

Table 2.2: Interpretation of Kappa

For judging any value between 0 (association based on chance) and 1(perfect association), Table 2.2 can be used, which is a commonly cited scale that divides the interval into five groups and assigns an interpretation to each group.

The main drawback with the Kappa coefficient is that it strongly depends on the prevalence of diseases. In some cases the value for actual association of disease pairs and the value for Kappa coefficient lead to paradoxical results[5]. The Kappa coefficient is not a reliable measurement method for observations where either the two conditions have very high (or very low) prevalence rates, or when the conditions have very different prevalence rates (one very high and the other very low). However, Feinstein and Cicchetti have developed methods to resolve the paradoxical behavior of Kappa [19, 40].

Concordance statistics measurement methods

For finding the association between two disorders, these methods compare every possible pair of individuals in a medical data-set; for a population of N persons, it will be $N(N-1)/2$ com-

[5]Assume that among 100 patients, 84 of them have disorders D_1 and D_2, 6 of them have D_1 but do not have D_2, 9 have D_2 but not D_1, and 1 has neither of the disorders. In this case the actual association among disorders (p_0) is $85/100$ which is considered high, whereas $\kappa = 0.04$, which according to Table 2.2 is interpreted as slight association.

parisons. In order to calculate concordance statistics, concordant pairs and discordant pairs need to be calculated. For doing so, assume that p_1 and p_2 are a pair of arbitrary individuals (patients) in a study population. This pair is concordant if diseases D_1 and D_2 are both found exclusively in p_1 or p_2 (but not both). The number of concordant pairs in population is calculated as $P = ad$ (a and d are values specified in Table 2.1). A discordant pair is a pair of patients who both have only one, but not the same disease. The number of discordant pairs is calculated as $Q = bc$ (b and c are variables from Table 2.1). The remaining pairs in the population, which are not concordant or discordant, are called ties, which refers to pairs where p_1 and p_2 both have at least one concurrent disorder, i.e. they both have either D_1, or D_2, or both disorders. T_c, T_r represent the number of ties in column variables only, and row variables only, respectively.

Somers' D, *Kendall's Tau-b* and *Gamma* are concordance statistics methods which can be used for measurement of pairwise disease association.

$$Somers'D = \frac{P - Q}{min(W_r, W_c)} \tag{2.5}$$

$$Kendall'sTaub = \frac{P - Q}{\sqrt{W_r W_c}} \tag{2.6}$$

$$Gamma = \frac{P - Q}{P + Q} \tag{2.7}$$

where $W_r = P+Q+T_r$ and $W_c = P+Q+T_c$. The Gamma statistic ignores the tied pairs in calculating the association which overestimates the strength of association when many tied pairs are present in the study population [10]. In a study performed to identify co-morbidity patterns, the four methods of Kappa coefficient, Somers' D, Kendall's Tau-b and Gamma have been compared, and the study concluded that the Somers' D and Kendall's Tau-b have the best performance among other measurement methods for identifying non-random co-morbidity [89].

2.1.3 Clustering of Diseases

Clustering techniques are common approaches used to study aspects of multimorbidity, particularly to link psychiatric phenotypes to clinical measurements [76]. Some of the methods for clustering individuals or variables in health-care projects are described in this section.

K-means Clustering

Clustering methods aim to partition data points into clusters such that data points in the same cluster are more similar to each other than data points in different clusters according to some defined criteria. The *K−means clustering* algorithm (first introduced by MacQueen in 1967) is a vector quantization method which aims to partition data points based on Euclidean distance. For a data set $(x_1, x_2, ..., x_n)$, where each data point x_i for $(1 \leq i \leq n)$ is a multi-dimensional vector, $k-$means aims to partition n points into $k, (k \leq n)$ clusters or sets, in which each data point belongs to the cluster with the nearest mean distance. The objective of $k-$means is to minimize within cluster sum of squares defined as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2 \tag{2.8}$$

Where J calculates within-cluster sum of squares for all clusters, and $\|x_i^{(j)} - c_j\|$ provides the distance between a data point and it's cluster's centroid.

The algorithm is composed of the following steps:

1. Initialization: Define and place k data points in data space.
2. Assigning data points to clusters.
3. Recalculation of positions of the K centroids.
4. Repeat steps 2 and 3 until the number of data points moving between clusters falls bellow a predefined threshold.

Two commonly used methods for initializing k points are Forgy method and random partition method [54]. The Forgy method randomly picks k points from data set and uses them as

initial centroids[6]. In contrast to Forgy method, the random partition method starts from step 2 and assigns a cluster to each data point. Once the initial assignment is over, the centroids for clusters are defined by computing the means of data vectors in each cluster.

After initializing the centroids, the algorithm proceeds to step 2, where every data point gets assigned to its nearest cluster. To accomplish this, Euclidean distance between every data point and all centroids is computed, and the data point is assigned to a cluster by the smallest Euclidean distance. In step 3, the mean vector for every cluster is calculated and the centroids of clusters are updated to the new mean vector. The algorithm iterates between steps 2 and 3, unless the number of data points which have been moved between clusters is smaller than a predefined threshold or is zero. However, there is no guarantee that the algorithm will converge to a global optimum[7], the results depend on the initial centroids and how they are being defined. On the bright side, practically the algorithm usually converges very fast despite being computationally difficult (NP-hard)[8], therefore it is possible to run the algorithm multiple times with different starting points.

The main drawback of $k-$means is its cluster model. The concept of this method is based on spherical clusters, such that the assignment of the nearest cluster is the correct assignment, and clusters separating factor is their mean value (vector) converging towards the cluster center. This limitation makes the selection of the optimal number of cluster (k) difficult for any given dataset. Therefore, an inappropriate initialization of a data point as a centroid, might lead to poor/inappropriate data space divisions.

One of the other limitations of the $K-$means clustering algorithm, is the nature of distance based computations. Due to this, the algorithm is applicable only on numerical data types and cannot be used for categorical data types. Different approaches have been introduced to overcome this limitation: One of the traditional solutions is to convert categorical values into

[6]In mathematics and physics, centroid refers to the arithmetic mean or average position of all the points in the space. A centroid in clustering algorithms, is a candidate data point in a cluster, which can be used as a measure of cluster location. Therefore, every cluster carries one centroid as its representative. In $k-$means algorithm, the data vectors of centroid c in cluster C for $1 \leq C \leq k$, are created by computing the means of all data points in that cluster.

[7]The global optimum is reached when the assignments of data points to clusters no longer change.

[8]Finding an optimal solution for $k-$means clustering problem in d dimensions is NP-hard in $d-$dimensional space even for 2 clusters, and also for k clusters even in $2-$dimensional space [2, 79]. In worst case, $k-$means can be solved in $O(n^{dk+1})$ [60].

numeric values, which does not necessarily produce meaningful results [58]. Another approach is the $k-$modes algorithm, which is a variation of $k-$means for clustering data sets of categorical type. This approach is explained in detail in Chapter 4 section 4.1. The $K-$prototypes algorithm is another approach which integrates $k-$means and $k-$modes and enables clustering of data sets with mixed data attributes (numeric and categorical) [57].

Factor Analysis Techniques

Factor analysis techniques have been widely used in behavioural sciences. The basic logic of factor analysis was first developed by Charles Spearman in 1904 [109]. Factor analysis is a multivariate statistical method which aims to identify the factors or dimensions that underlies the relations among a set of observed variables [73]. In factor analysis, an observed score for person i measured at time j is X_{ij} which is made up of two components:

$$X_{ij} = T_i + E_{ij}$$

T_i is the true component and E_{ij} is the error component of X_{ij}. The error component is random and unrelated to true component. Therefore, the variances in observed scores can be computed as following:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

where σ_T^2 and σ_E^2 are non-overlapping parts and σ_T^2 is the true variance which consists of common variance and specific variance. The common variance is shared among a set of indicators and it appears due to effects of underlying factors. The specific variance is made up of unshared variance of indicators and measurement error E_{ij}. Factor analysis methods aim to identify a smaller number of factors that explain most of the common variance [73].

Factor analysis methods in general are useful for identifying groups of items that are strongly correlated. There are two broad categories of factor analysis, *Exploratory Factor Analysis (EFA)* and *Confirmatory Factor Analysis (CFA)*. Exploratory factor analysis uses latent variables to discover the correspondence of large sets of variables in a data set. In EFA, it is assumed that any observed variable (indicator) can be associated with any number of factors (*a.k.a.* latent variables). EFAs have been commonly used in research to identify clusters of

chronic diseases (variables of data set) [56, 63, 72, 95]. Confirmatory factor analysis uses the same approach as EFA, and hypothesizes the number of factors in the model before starting the analysis. One of the applications of CFA in research is the work of Johnson and Wolinsky [63], where CFA was used to explore a model for multiple diseases, disability, functional limitations, and perceived health amongst elderly.

Latent Class Analysis

The main goal in latent class analysis (LCA) is to find a reduced set of dimensions that explain the relationship between variables. This method is similar to factor analysis methods (EFA and CFA) in terms of discovering the structure of variable sets and different from these methods in assuming that the latent variables are categorical. A latent class analysis defines membership levels for latent classes and every set of attributes belongs to each latent class to some degree. A complete overview of LCA can be found at the research of Schüz *et al*.[101], which aims to use LCA to measure the subjective well being in later adulthood, and concludes that health conditions in older adults can be divided into four groups: no disease, cardiovascular disease, joint problems, and multiple illnesses.

2.2 Comorbidity and Multimorbidity Indices

In this section we define and discuss a research tool to organize, summarize, compare and measure comorbidity and multimorbidity. This research tool is called an “index”. In most of these indices, a group of diseases are selected and weighted, either according to their individual severity [15], or their influence on other factors, such as mortality [18], quality of life [52], or hospital stays [116]. There is a lack of agreement on the number and type of diseases to be included in an index. Therefore, there is a large heterogeneity on number of diseases in a certain index [31]. Some indices consider only a small subset of chronic diseases [83], and some include up to 185 different diseases [64].

According to the definition of comorbidity and multimorbidity, the indices are designed based on the research needs. In indices for multimorbidity, no index disease is defined or used. Whereas, in comorbidity indices, defining an index disease is obligatory. Although, depending

on the focus of the research (whether it is on measuring the total burden of disease, or the burden of comorbid disease in addition to the condition of interest), some of the indices can be used in both types of research. In the second case, the index disease is being treated/weighted the same way as the other diseases, if the method is being used for multimorbidity [31].

Although there is no gold standard for the measurement of comorbidity or multimorbidity, and the existing indices are too coarse to be replaced with clinical judgment in clinical decision making, the numerous measurement methods and indices exist which can be used in health research [4, 33, 53].

The main focus of this section is to focus on some basics of co/multimorbidity indices, such as what they are, how are they being measured, when it is best to use them, and how to measure an index's validity, reliability, feasibility and generalizability [53]. For simplification purposes, we refer to *comorbidity* and *multimorbidity* indices as *co/multimorbidity indices*.

2.2.1 Definition of a Co/multimorbidity Index

A co/multimorbidity index is a measurement tool which reduces all existing diseases and their severity into a numeric score. This score can then be used to compare patients having co/multimorbidities [53]. In every index, all morbidities that are being included in calculation are given a weight. For some indices the same disease is given multiple weights based on its severity. The final score for diseases is then calculated. Some indices assume that the impact of diseases is additive [78], so their final score is the addition of weights of all existing diseases within the patient which are included in the index disease list. Some other indices determine the weight of single most severe index as the prognosis [66]. Therefore, a patient can get two widely different scores from two different morbidity measurement indices due to three main reasons: list of available diseases for an index, different weights associated to the same disease in different indices, different scoring system among indices. For co-morbidity indices the score can vary for the same patient and same index, only by changing the disease of interest. These indices are being used wherever it is needed to classify patients into groups with similar risks [53].

An index is often designed based on researchers' specific needs. The population for design-

ing and testing the index is usually a specific group of patients being hospitalized because of a certain condition. Therefore the development of a co/multimorbidity measure is influenced by population and outcome used, and as a result, the weights assigned to diseases for an index might be different if the population of the study was different [37].

Co/multimorbidity indices are predictive indices [53] which contain three main components: items (morbidities), severity scales for items (weights), and a scoring system. Any co/multimorbidity index should be assessed with following principles: content validity, face validity, reliability, predict ability, feasibility, and generalizability. In following part of this section, these terms are defined briefly.

Content validity is a term used for assessing the validity and completeness of items, as well as the relevance of content of items to evaluate what they claim to measure. Content validity examines whether the items and their severity scales appropriately describe the co/multimorbidity phenomenon and whether any irrelevant areas are included [53].

Face validity refers to the sensibility of the index and whether it can be used in general clinical studies. The method used for the generation of items (judgmental, statistical, etc.) and the weights assignment method (equal or weighted) are examples of two features for face validity assessment. Most of the indices have not accounted for the validity of the scoring system when it was developed and tested. Therefore, in a hypothetical index A, a score x for hypertension, does not necessarily have the same impact as score x for asthma. In most of the indices, the scoring system is made up based on assumptions. Furthermore, it has not yet been proven that any index is superior to another one, or if any of them are more appropriate to be used [53].

Almost all of the co/multimorbidity indices have been developed based on a specific population, in a specific time and place. *Reliability* refers to the extent to which the results of the indices will be similar, by changing the population, the time and the place of the experiment [42]. Due to the diversity of indices (i.e. list of items they each accept as input, system of weighing every morbidity, scoring system differences, and the way they each deal with common confusing situations such as unstated diagnosis when a person is taking a specific medication, or uncertainty in diagnosis), it is almost always essential to define a unique set of rules for every index to measure their reliability [53].

Co/multimorbidity indices should be simple, easy to understand and use. Therefore the *feasibility* is defined and it refers to the simplicity, ease of use, availability, training requirements, and the time of administration of an index [53].

All the indices have been designed and tested on a specific group of patients (population). It is important to decide whether the results gained from the study group when designing and testing the index, are generalizable to other groups or not. Therefore, *generalizability* is an important feature that should be examined when designing a new index.

2.2.2 Challenges Associated with Co/multimorbidity Indices

Almost all comorbidity and multimorbidity indices use ICD(International Classification of Diseases) to create their list of items, and the different levels of disease abstractions in every index further increases the variation and number of conditions, thus, the complexity of indices increases when it comes to result comparisons. As an example "heart disease" can be categorized as one condition referring to all problems and conditions related to heart disease (e.g. [122]), or it can be categorized in a few different items (conditions), such as hypertensive heart disease, ischemic heart disease, and rheumatic heart disease(e.g. [48]). Besides all the differences in types of listing of the conditions, there are a few studies where the list of the items used in the index is not included in the review (e.g. [64] or [78]). These studies are practically inefficient either for being applied to different population groups, or further research for modifications and development.

As mentioned earlier in this section, the heterogeneity in weighting methods among different indices makes it difficult to compare the results. The origin of this heterogeneity is the different approaches the indices take for weighing the items. In some studies the weights are used to analyze the effects of multimorbidity on objective outcomes, such as costs [116], hospitalization [38], and mortality [18]. In some indices, morbidities are weighted according to predefined criteria such as prescribed medicine [122] or clinical parameters [88]. In another group of studies, weights are applied according to subjective outcomes, such as depression [27], and health related quality of life [15]. Future work on improving co/multimorbidity indices would be to find a way to standardize weights for different indices, so that they would

be easily comparable [33].

The data source used for designing most of indices is extracted from self-reports, and the studies show that there is a poor agreement on data obtained from self-reports and medical records [25, 67, 115]. Also, the population considered for designing and testing indices is biased mostly to older people because those are the ones with higher risk for morbidities. Therefore, there might be some validity issues when using the indices for general population [33].

All these challenges aside, the question becomes whether it is possible to have an established index which can fit to general population is various studies. Multimorbidity in reality is more complicated than simply addition of a number of diseases; there are so many other factors affecting the phenomenon such as social, physiological, emotional and cultural factors, as well as living condition, socioeconomic status and individual preferences in treatment strategies. All these parameters are required to be considered in assessment of multimorbidity, and it is not very straightforward to aggregate all these factors into a single number. Although, in studies where morbidities are indicated with their severity, the goal of incorporation of various factors has been partially achieved. Moreover, in recent studies, a few general questions are added to the interviews to measure the extent of social, physical, and mental impairments, and the results can be included in indices [33].

As mentioned earlier in this section, there are numerous indices available for measuring multimorbidity and comorbidity. In the next subsection, we describe some of the most frequently used indices for multimorbidity. More details on comorbidity indices can be found in research of Vincent and colleagues in Amsterdam [31] and the research of Stephen Hall at Queen's Cancer Research institute in Canada [53].

2.2.3 Examples of Multimorbidity Indices

According to a systematic literature review study upon measurement methods of multiple chronic diseases (e.g. [33]) performed in 2011, 39 multimorbidity indices have been identified which analyze the association among multiple chronic diseases. Data sources for assessment of indices are extracted from self reports, physician reports, medical records and administrative data. Most of the indices are using hospitalized patients data from health maintenance organi-

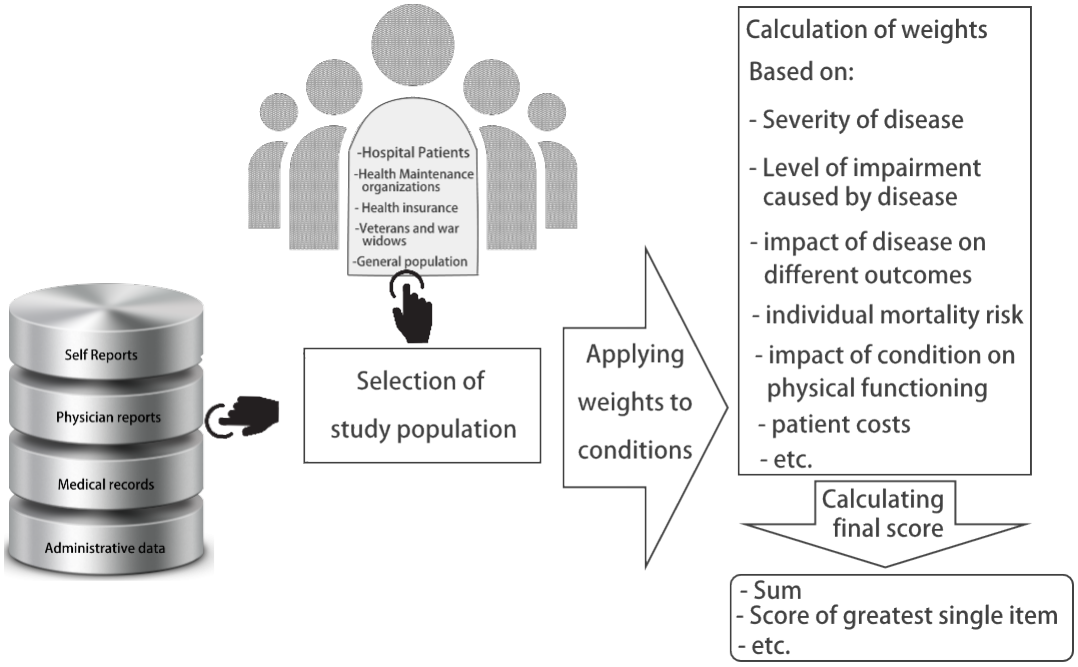


Figure 2.1: Multimorbidity Indices Conceptual Diagram

zations(e.g. [6]) and health insurance centers(e.g. [88]), as study population. However, veterans and war widows have also been studied as population source in some indices (e.g. [15] or [38]). Some of the well-known and frequently used indices which have used medical records, physician reports and administrative data as data sources, are introduced below briefly.

Charlson Index is the best known and most frequently applied method for classifying comorbid conditions (diseases) that might alter the risk of mortality [18]. The index calculates the risk of mortality based on number and seriousness of diseases co-occurring within an individual. The data source for testing the index's functionality and predictability is observed from medical records of 685 patients who were treated for primary breast cancer. There are 18 chronic diseases being studied in this index: They are coded as 1 or 0, based on being or not being present in an individual, respectively. The severity is coded as an integer number between 1 and 5, based on how severe the disease is. The unadjusted and adjusted relative risks (risk ratios) of mortality are calculated. Unadjusted relative risks calculate the mortality risks of patients having a certain condition, ignoring the other conditions that may coexist within the individual. For a specific disease, unadjusted relative risk is calculated as the proportion

of patients with the condition who died, divided by the proportion of patients without the condition who died. The adjusted relative risks take into account the coexisting conditions that may occur within the individual, as well as the conditions severity. A patients subsequent risk of mortality is calculated by combining the adjusted relative risks and patients' age, which is used to predict the survival rate. The statistical difference between the expected and predicted mortality rates is examined by chi-square test [18].

Medication-Based Disease Burden Index (MDBI) is a convenient method for quantifying disease burden and predicting health outcomes using medication records and data on chronic drug therapy [48]. The results from the analysis of MDBI consist of a table of 20 chronic diseases, and drug classes corresponding to each one of them. There are weights associated to each disease. If patient's medication contains any of the agents listed in MDBI, the corresponding weight will be assigned to patient, and the total score for the patient is the sum of all weights associated to him. The results of MDBI tested on patients show that the scores are significantly related to number of medications being taken at a time, and patients' age, as well as a decent predictor of death and planned or unplanned re-admissions during 12 week period [48]. More details on the procedure performed for analysis methods used in the study are described elsewhere (i.e. [48])

Index of Co-existing Diseases (ICED) is another famous index for multiple chronic disease measurement. Potential candidates for the study, were patients older than 18 years of age, who were discharged after having total hip replacement over the period of eighteen months. The data was collected from four hospitals, and 356 patients were selected randomly among potential candidates [50]. The main goal of examinations in ICED is to measure the impact of baseline preoperative co-existing diseases on postoperative complications, which was proven to be significantly high [50]. More details on the disease association measurement methods and data analysis procedure can be found at the main paper explaining the index (i.e. [50]).

Nottingham Health Profile (NHP) is an estimation for health-related quality of life (HRQL), where the presence of different symptoms (including physical, emotional and social indicators)

and lack of ability in various basic activities (i.e. handicap and health-related problems in different life areas) is documented [59]. NHP was applied to patients of only 76 years of age, in order to estimate their mortality risks. It is evaluating the outcome in two separate parts. The first part displays the degree of discomfort or distress using six dimensions: lack of energy, pain, emotional reactions, sleep, social isolation and physical mobility. The overall score for this part of NHP is 6, and the higher the score, the greater severity the problems. The second part of NHP contains seven polar questions (i.e. yes/no questions) regarding to occupation, social and personal lifestyle, interests, holidays and hobbies [51]. The details about statistical analysis can be found in the main paper (i.e. [59]). A modification of NHP has been used to analyze HRQL of five hundred sixty five 76 years old citizens of Sweden (i.e. [51]). The results indicate that older people accept being diagnosed with certain number of diseases and having some disability as natural consequences of being old. Therefore, the health related quality of life changes for elderly (in this study for a sub-sample of 76 years olds), such as they consider themselves to have good HRQL even when they are diagnosed with certain conditions such as hypertension, cancer, or lung disease.

Incalzi Index is an index used to predict the mortality of geriatric elderly patients with acute medical illness by measuring the interaction between comorbidity and age of patients. The test has been performed on 370 patients between 70 and 90 years of age who were consecutively admitted in hospital in an 18 month period. The index consists of two sub indices: a comorbidity index and an age-comorbidity index. The comorbidity index is based on a scoring system which quantifies the prognostic weight of individual diseases, while the age-comorbidity index increases mortality risk by accounting an age-related risk factor. In order to assign a weight to each disease in the comorbidity index, active illnesses of each one of the 370 patients was listed, and, according to relative risk of death for each person, the scoring system was developed.[61] In the age-comorbidity index, an additional score of 0,2,3 and 4 points are added to patients aged 66-75, 76-85, 86-95 years and over 95 years, respectively. Kappa statistical measure is used to assess the reliability in computing the index of comorbidity[23]. More details on complete list of analysis methods used to develop the index can be found in the main paper (i.e. [61]).

The indices introduced above are more commonly used measurement methods for predicting health related quality of life or mortality risks. A full reference for all existing multimorbidity indices can be found in [33]. Unlike most of the existing multimorbidity indices which use self reports as their data source, the data source of all indices introduced here are based on medical records, and physician reports, which is assumed to be more reliable. The outcome of Charlson index, Incalzi index and MDBI is mortality, while the goal of ICED is prospective complications and NHP focuses on health related quality of life.

2.3 Impact of Socioeconomic Status on Diagnosis and Treatment of Multimorbidity

As mentioned in previous chapters, various aspects can play an important role on onset of multimorbidity, and socioeconomic status[9] of the individual is one of them. Individuals with low socioeconomic status are more likely to have diseases (chronic and acute), be physically and mentally impaired and face partial or full loss of function [1, 98]. The influence of socioeconomic status of health-related quality of life is assumed to begin in the prenatal environment and continues throughout individuals life [14, Chapter 4]. For instance, childhood socioeconomic differences have been shown to play a significantly important role in mortality gaps later in life between blacks and whites [123]. The research of Dr. Barnett and colleagues (i.e. [4]) reveal that multimorbidity occurs ten to fifteen years earlier among people with low socioeconomic status. However, health and socioeconomic status (mainly wealth) have dual relationships in the way that it is not only poor socioeconomic status which affects health; but poor health conditions also affect socioeconomic status, as well. If someone is facing health related problems which causes physical or mental distress, they often cannot be as productive in their workplace as they would be if they did not have health issues. A self reported study in

[9]Socioeconomic status (SES) is an economic and sociological combined total measure of a person's work experience and their or their family's economic and social position in relation to others, which can be based on income level, education, and occupation. When analyzing a family's socioeconomic status, household income, earner's education and occupation are examined. The SES index is a composite of often equally weighted, standardized components of parent's education, family income and household items. The terms high, middle, and low SES refer to ranges of the weighted SES composite index distribution [69].

1999 shows that those with fair or poor self-reported health had less wealth growth compared to others over the next ten years [105]. The various factors of socioeconomic status, their impact on health of the individual and prevalence of multimorbidity are discussed in the following sections of this Chapter.

2.3.1 Socioeconomic Factors and Their Impact on Health

Education, income level, wealth, occupational status, and residential neighbourhood are socioeconomic factors which can affect individual's health behaviors, diagnosis of diseases, and health care services [71, 103].

In general, multimorbidity is less common in individuals with higher socioeconomic status, and they are more likely to get access to treatments and be able to handle and afford the treatment if necessary [49]. For instance, having high level of education (which is one of socioeconomic factors) helps individuals to read and research more accurately about the health conditions that they have or might be exposed to, which will help them to prevent the disease or start and monitor the treatment in early stages of their diagnosis.

Socioeconomic factors have an affect on early diagnosis of most, but not all, diseases. Socioeconomic deprivation does not change the prevalence of diagnosis of some chronic conditions, such as cancer, skin diseases and kidney diseases [29]. However, the prevalence of certain conditions, such as depression, drugs misuse, anxiety, dyspepsia, pain, CHD (coronary heart disease), and diabetes is higher in people who are living in areas with high socioeconomic deprivation [84]. According to the results of research performed by a group of medical researchers in 158 general practices in Germany (i.e.[100]), the impact of education and income levels on multimorbidity is as follows: Patients with medium or high education level, had a mean of 0.3 diagnoses less than individuals with low education level [100], while allergies have proven to be more common among them [29], and patients with higher net income (or bigger household size), had 0.3 diagnoses less on average [100].

2.3.2 Associating Multimorbidity with Socioeconomic Factors

One of the complications in measuring socioeconomic factors is that the same results for an arbitrary socioeconomic factor may possibly signify something different among two different population groups. For instance, when measuring income level, the absolute amount of income cannot be interpreted as high if it is not being translated into same purchasing power for two different communities. As another example, junior year high school students may exhibit different levels of knowledge if they are being trained in different countries, cities, schools, or even communities [68, 125]. Whenever it is possible in studies, it helps to split the population into smaller relatively consistent sub-groups and measure the factors separately.

Many studies have shown that socioeconomic factors (education, income, occupational status, sometimes residential neighbourhood), play an important role in health conditions of individuals. However, due to the reasons explained here, it is not straightforward to come up with a standardized and uniform pattern of factors influencing health outcomes[10].

Due to the dual relationship of socioeconomic status and health impairment, it is challenging to decide the lack of which one has caused the impairment of the other. People living in poor socioeconomic status are more likely to have impaired health conditions, and not being able to afford to pay off the health costs makes it even worse in the way that they do not get the required treatments to overcome the condition with which they are diagnosed. For instance, low income patients are at increased risk not only for developing asthma but also for reduced access to appropriate care [86]. On the other hand, not being in good health can have a negative impact on an individual's career through missed workdays and in some cases work loss [35].

Race and ethnicity are other dominant factors which should be taken into account when predicting the impact of socioeconomic status on health care. For instance, there is a higher prevalence of hypertension among black men, and they experience significantly lower life expectancy compared to whites. Also, Hispanics and black men are more likely to have diabetes. This ethnic and racial differences are not explained by socioeconomic status [28].

Levels of stress (it could be the work related stress of personal life related stress), differ-

[10]There are many population subgroups within larger population groups which may be affected by different factors. For instance, Hispanics are known to have lower age-adjusted mortality rates despite higher poverty, compared to whites. As another example, among socioeconomically same groups of people, blacks are more likely to have hypertension, and blacks and Hispanics are more likely to have diabetes [28].

ences in health care access, and behavioural risk factors (such as smoking, overeating, lack of exercise, and excessive alcohol consumption) are other factors that need to be considered for more precise results when studying health impairment factors [14].

2.4 Summary of Multimorbidity Measurement Approaches

Various aspects, causes, effects and measurement and computational approaches for understanding multimorbidity have been analyzed throughout this chapter. Multimorbidity has been defined as the coexistence of multiple diseases within the same individual. Co-occurrence of diseases can happen due to chance, or it can be due to causal association. In studies conducted for measuring disease association rates or prevalence, selection bias occurs when clinical samples are used to conduct the study, and the reason is the relatively sicker source of population being used for analysis purposes. In order to overcome this effect, the population of the study should be selected from non-clinical samples.

Some approaches for measuring multimorbidity are: total disease counts, pairwise association methods, disease clustering approaches and disease networks. The most common methods used for pairwise disease association are, odds ratios, risk ratios, multimorbidity coefficients, kappa statistics and concordance statistics measurement methods. The advantage of statistical measures compared to non-statistical methods (odds ratio, risk ratio and multimorbidity coefficients), is that they can be adjusted to measure expected coincidental rate of multimorbidity. Among the three non-statistical methods, multimorbidity coefficients was proven to have linear behavior when the prevalence of disorders rise, while the other two measures have an exponential raise as the prevalence of multimorbidity rises, which makes them not appropriate to be used for sicker populations.

Using disease networks to represent diseases and their associations is a relatively more recent approach for visualizing the association among diseases. In disease networks, diseases are represented as nodes, and their associations are represented by undirected edges [76], where each edge drawn between a pair of nodes (a pair of diseases), is showing the association between that pair of diseases. The association rate can be calculated using any of the pairwise association measurement methods (i.e. multimorbidity coefficients). However, when the asso-

ciation between two diseases is caused due to a third disease, this approach cannot be used.

The impact of socioeconomic deprivation was another topic discussed in previous chapters. It was discussed that certain diseases have higher prevalence among individuals living in more deprived areas with lower socioeconomic status. However, besides socioeconomic factors, race and ethnicity are many other factors which make differences in prevalence of diseases; The influence of race, ethnicity, age-adjusted mortality rates, levels of stress, feasibility of access to health-care centers and many other factors on individuals health, makes it impossible to conclude a general statement about the importance of particular socioeconomic factors.

This chapter has also introduced multimorbidity indices which are used for measuring multimorbidity. Despite the numerous multimorbidity indices being introduced over time, there is no gold standard for measuring the complex phenomenon of multimorbidity, because the idea of summarizing the complex reality of multimorbidity into a single indicator, which is the general process for a multimorbidity index calculation, makes the further decision making prone to external criticism.

As mentioned in section 2.3, socioeconomic status of the individual can play an important role in diagnosis, prognosis and treatment of chronic diseases. One purpose of the research addressed in this thesis is to analyze the impact of socioeconomic and socio-demographic factors on patients diagnosed with multiple chronic diseases. To accomplish this goal, we have collected the available socio-demographic information of patients from the DELPHI database, and analyzed them using some of the approaches introduced in this chapter. Disease associations, their distribution over gender, age group or socioeconomic score, and the impact of socio-demographic factors on groups of patients sharing the same diseases are analyzed using three main approaches: (1) statistical analysis methods have been used in order find the distribution of individual diseases and disease pairs over gender, age groups and socioeconomic status pf patients, (2) pairwise association methods have been applied for measuring the association of disease pairs in the database and (3) a variation of $k-$means algorithm introduced in sub-section 2.1.3 has been implemented in order to find clusters among patients sharing the same diseases for the most commonly occurring disease pairs.

Chapter 3

Data Collection

3.1 The Database

The DELPHI (Deliver Primary Healthcare Information) project started in 2003 with the goal of creating a primary health care researchable database derived from electronic medical records (EMR)[1] of ten primary health care practices throughout Southwestern Ontario. The project is based in the Centre for Studies in Family Medicine (CSFM) in the Department of Family Medicine, University of Western Ontario. The current version of the database de-identified data from ten practices and includes around 30,000 patients [113].

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) project started in 2008 with the goal of developing a network to collect health-related information from patients with chronic diseases across Canada. It is Canada's first multi-disease surveillance system based on primary care EMR data [112]. Initially, the network began amalgamating data nationally from seven existing academic research networks in Calgary, Edmonton, London, Toronto, Kingston, Montreal and St. John's. Later the network was also developed in Winnipeg , Halifax and Vancouver. CPCSSN database currently holds the data of ten practice based research networks and it contains the medical records of over one million patients. Patients' health information is being extracted from these networks, cleaned and coded into CPCSSN database every three months [9].

[1]The term electronic medical record is used to describe electronic patient's records that are being kept in one location and are accessible by only one provider's site [114]

DELPHI database is one of the several networks included in the CPCSSN project. The database used for study and analysis purposes in this research, is a version of DELPHI database which is coded with CPCSSN standards. The database is designed to hold all medical related information for patients in multiple tables including patients' characteristics (i.e. gender, birth year, postal code, etc.), demographic information, billing information, medication history, family history, vaccine history, allergy intolerance history, physical examinations, laboratory tests and health conditions history. The database holds the medical records of 27670 patients, from which 5720 records have been saved with erroneous diagnostic codes including null, zero or invalid values (codes). Diagnostic codes in the database are recorded using 9th revision of the ICD system [2]. As the purpose of this thesis is to study patients with multimorbidity, the ICD codes have been used to identify chronic diseases, which led to the identification of patients with multimorbidity. Further details about how the data for analysis was collected and cleaned is described in the following section.

3.2 Data Pre-processing

In order to identify patients with multimorbidity, a list of 20 chronic disease categories (particularly relevant in clinical and general population in Canada) will be used. This list was created by a nationally funded research project at the Canadian Institute of Health-Centered Innovations, which aims to improve the quality of patient-centered interventions for patients with multimorbidity [43]. The age, gender and postal code attributes have been extracted for all the patients having at least one of the listed chronic conditions. The complete list of the twenty chronic disease categories along with their corresponding ICD-9 disease codes is available in Appendix A, Table A.1.

Detailed description of data extraction process and pre-processing of collected data can be found in the following subsections.

[2]The International Classification of Diseases (ICD) is a clinical cataloging system for coding diseases. The latest version of ICD catalogs is ICD-10, which offers more disease classification options compared to ICD-9. ICD-9 has been used from 1979 to 2015.

3.2.1 Data Extraction

As of the data extraction period for this research, a total of 15462 de-identified patients who have at least one of the twenty chronic diseases listed in Table A.1, have been collected. Among these patients, a total of 8245 have multimorbidity (the rest are diagnosed with only one chronic disease). Patient characteristics include information on patient's gender, masked date of birth (year of birth without mentioning the exact day and month of birth), and masked postal code (only first three digits of residential postal codes (a.k.a. FSA).[3] [4]

Table 3.1 represents the counts of missing or incorrectly filled attributes for patients with minimum of one chronic disease. The invalid instances for 'Gender' and 'Birth Year' in the database, are the ones set to zero, and the invalid or missing values of 'FSA' are either not filled, or filled with an invalid value (an FSA value is considered to be invalid if it is not included in list of postal codes in Canada).

Attribute Name	Missing/invalid attribute counts among patients with at least one disease	Missing/invalid attribute counts among patients with multimorbidity
'Gender'	1	1
'Birth Year'	114	9
'FSA'	1645	835

Table 3.1: Missing attributes counts for patients with one or more chronic diseases.

To perform a reliable analysis, we removed the instances where the value for either one of the attributes is missing or is invalid, and the analysis was performed based on instances which did not have any missing attributes. Therefore, the number of patients with no missing attributes in the database, shrank down to 13697 for patients with at least one chronic condition, and 7395 for patients with multimorbidity.

Besides personal information attributes extracted for the patients from the main database, a socioeconomic score was also computed from external resources for each patient in the

[3]The first three digits of a Canadian postal code is called Forward Sortation Area (FSA). The first character which is a letter, identifies the province or territory (Nunavut and Northwest Territories share the letter *X*). For Ontario and Quebec, the first character further identifies a particular part of the province. The second character, which is a numeric character identifies whether the area is urban or rural. A zero indicates the area is rural, while all other digits indicate urban areas. The third character in combination with the first and second character is a letter which identifies a more precise geographic district [91].

[4]The database also includes a table for holding patients' demographic information, such as their higher education level, languages, ethnicity, occupation, housing status, etc. which could potentially be used to define socioeconomic status for patients, but unfortunately there were only a handful of non-empty instances for these attributes, therefore, it was impractical to include them in extracted attributes.

database. The *'Gender'*, *'Birth Year'*, *'FSA'* and *'Socioeconomic Score'* attributes from the data table used for processing and analyzing. The process of how a socioeconomic score was computed is explained in detail in the next subsection.

3.2.2 Socioeconomic Score Computation

As discussed in Chapter 2, the income level is assumed to be one of the best indicators of socioeconomic status for families. Research has shown that relatively low-income families have lower mortality rates and higher rates in health-related complications compared to relatively high-income individuals [14]. In order to research the impact of income levels on health condition and investigate the relationship between disease patterns and income level of individuals, a median family income was assigned to each individual in the data table. This information was extracted from a separate data source created in 2013, where the median family income was calculated for every FSA in Canada based on census tract-based income maps [16]. The data source uses Google's Fusion Tables mapping tool to map income across the country in 1576 postal areas. The results are visualized into color-coded interactive maps, each color representing the family income level of a certain FSA. There are six colors representing six income ranges, and each FSA is colored based on where the median family income for that FSA belongs to, in income ranges. The map can be zoomed and panned to focus on a specific FSA, and is also click-able. Once and FSA is clicked on, a message pops up showing the population of that specific area along with its median family income value. Median income values have been extracted from this data source, and assigned to all patients in the data table according to individuals' residential code. The data table was then finalized for processing by assigning categorical labels for numeric values of 'Birth Year' and 'Income by FSA'.

3.2.3 Labeling the Attributes

In order to perform efficient analysis and draw patterns out of data, two numeric attributes 'Birth Year' and 'Income by FSA' have transformed into categorical attributes.

For transforming the 'Birth Year' into 'Age Group', the stages of a person's life have been used as categories. To do so, the description of life stages was found in second Chapter of the

book called *It's your future, make it a good one!* [124]. As the book suggests, one's life can be divided up into ten stages according to Table 3.2.

Life Stage	Age Range	Characteristics of Life Stage
Infant	0-2	Dependent, brain developing, learning motor skills and sensory abilities.
Child	3-9	Growing and mastering motor skills and language, learning to play and socialize, continued growth, formal school and organized activities
Adolescent	10-19	Growth spurts, puberty, hormonal changes and strong emotions may rule decisions. Behavioural risks.
Young Adult	20-29	Completing education and beginning career and family. Potential coping and financial pressure.
Adult	30-39	Managing family and career growth. Increasing numbers of couples are starting families in this stage. Continued coping pressures.
Middle Age	40-60	First signs of aging and its effects on lifestyle. Career peak. Children having their nest, grandchildren arrive. Aging parents may require care.
Independent Elder	60 onward	More signs of aging and lifestyle effects. Eligible for government retirement and health-care benefits. Retirement. Some health problems arise, and medications may be required. Care for others.
Vulnerable Elder	60 onward	Beginning of frailty. Cognitive or multiple health problems. Require some assistance. Not able to drive. Possible move to *Assisted Living*.
Dependent Elder	60 onward	Requires daily care. Unable to perform all personal functions. Possible move to a *nursing home*.
End of life (up to six months)	60 onward	Diagnosed with terminal condition(s) or end stage disease(s). May require *hospice* care or *nursing home* care.

Table 3.2: Life stages and its characteristics.

There are only two infant patients in our data table, therefore the two groups of 'Infant' and 'Child' have been merged together and labeled as 'Child'. The four final stages of life have also been treated the same way and labeled as 'Elder' in the database, as there were no further information about the elderly on how they live or what kind of age related problems they might be experiencing. The rest of age group categories have been labeled as they appear in the table.

With sensitivity analysis [32] we ensured that the available data for every chronic disease in every age group is consistent. The analysis can be summarized in following steps:

1. For every one of the six age groups $G^*_{k,d} = [i, j]$ $(1 \leq k \leq 6,\ 1 \leq d \leq 20,\ i < j)$ where k represents the age group, d represents the chronic disease code, i and j represent the minimum and maximum ages in G_k, calculate the counts of patients diagnosed with every one of the twenty chronic diseases.

2. Compute step 1 for the following age intervals:

   $G^{(1)}_{k,d} = [i-1, j-1]$, $G^{(2)}_{k,d} = [i-2, j-2]$, $G^{(3)}_{k,d} = [i+1, j+1]$, $G^{(4)}_{k,d} = [i+2, j+2]$.

3. Compare the results of diseases counts between G^*_k and every one of the intervals in step 2. If counts of patients in $0.9 \times G^*_{k,d} \leq G^{(l)}_{k,d} \leq 1.1 \times G^*_{k,d}$ $(1 \leq l \leq 4)$ for all four age intervals, the data for disease d and age group k is consistent.

The sensitivity analysis showed the data is consistent for all diseases among all age groups.

The numeric 'Income by FSA' attribute has also been labeled to form a categorical attribute. To do so, income quintiles were used from Census Statistics Canada [17]. As the statistics suggest, the population of individuals in Census program has been broken down into five equal sized groups (quintiles), from lowest to highest income and the individuals' income was labeled based on the quintile it lies under. In our data table, there were no individuals lying under lowest or second lowest quintiles.

Based on our considerations, the final data table after pre-processing and data cleaning stages is as shown in Table 3.3

Data Attribute Name	Attribute Type	Values	Description	
Hypertension	Boolean	{0, 1}	1 if diagnosed, else 0	
Obesity	Boolean	{0, 1}	1 if diagnosed, else 0	
Diabetes	Boolean	{0, 1}	1 if diagnosed, else 0	
Bronchitis	Boolean	{0, 1}	1 if diagnosed, else 0	
Hyperlipidemia	Boolean	{0, 1}	1 if diagnosed, else 0	
Cancer	Boolean	{0, 1}	1 if diagnosed, else 0	
Cardiovascular Disease	Boolean	{0, 1}	1 if diagnosed, else 0	
Heart Failure	Boolean	{0, 1}	1 if diagnosed, else 0	
Depression	Boolean	{0, 1}	1 if diagnosed, else 0	
Arthritis	Boolean	{0, 1}	1 if diagnosed, else 0	
Stroke	Boolean	{0, 1}	1 if diagnosed, else 0	
Thyroid	Boolean	{0, 1}	1 if diagnosed, else 0	
Kidney Disease	Boolean	{0, 1}	1 if diagnosed, else 0	
Osteoporosis	Boolean	{0, 1}	1 if diagnosed, else 0	
Dementia	Boolean	{0, 1}	1 if diagnosed, else 0	
Musculoskeletal Problem	Boolean	{0, 1}	1 if diagnosed, else 0	
Stomach Problem	Boolean	{0, 1}	1 if diagnosed, else 0	
Colon Problem	Boolean	{0, 1}	1 if diagnosed, else 0	
Liver Disease	Boolean	{0, 1}	1 if diagnosed, else 0	
Urinary Problem	Boolean	{0, 1}	1 if diagnosed, else 0	
Gender	Categorical	$\{Male, Female\}$		
Age Group	Ordinal	$\{1, 2, 3, 4, 5, 6\}$	1: Child	0-9 years
			2: Adolescent	10-19 years
			3: Young Adult	20-29 years
			4: Adult	30-39 years
			5: Middle Age	40-60 years
			6: Elder	60 onward
Socioeconomic Score	Categorical	$\{3, 4, 5\}$	3: Third income quintile	
			4: Fourth income quintile	
			5: Higest income quintile	

Table 3.3: The structure of final data table used for processing, analyzing and clustering

Chapter 4

Data Analysis

This chapter presents different approaches taken to analyze the patient data with the aim of finding similarities in demographic characteristics among patients with multimorbidity and specifically those with the same combination of diseases.[1] As it is shown in Table 4.1, the total number of patients with multimorbidity decreases as the number of coexisting diseases grows, and every group of patients with the same number of diseases breaks down into even smaller groups for each combination of diseases. Therefore, the available data for analysis of subgroups of patients with same combination of diseases is scarce for more than two coexisting diseases. Due to this reason, the main cluster analysis in this research has been performed only on data of most commonly occurring disease combinations among patients with two chronic diseases.

All statistical and mathematical analysis approaches taken to analyze this data set are divided into three main parts and explained in following three sections. Section 4.1 explains the overall analysis of the database, it represents patient distribution over gender, age groups and socioeconomic score without focusing on number of coexisting diseases. Sections 4.2 and 4.3 represent the analysis performed on a subgroup of patients diagnosed with two chronic diseases: Section 4.2 includes the different statistical approaches taken to analyze data for two coexisting diseases, while Section 4.3 provides detailed description of the clustering approach for clustering demographic information of patients with specific combinations of two chronic

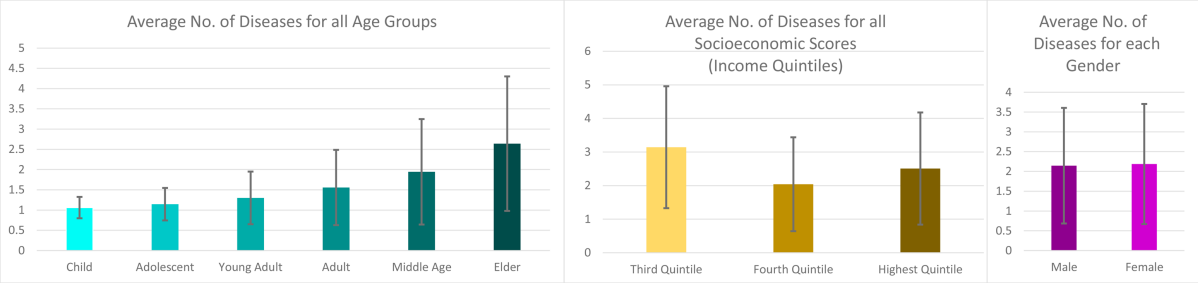
[1] According to Table 3.3 the available patient demographics in this research is *gender*, *age group* and *socioeconomic score*.

diseases. Finally, Section 4.4 includes the analytical results of patients diagnosed with three chronic diseases.

4.1 Disease Frequencies and Distributions

This section provides detailed analysis about the prevalence of diseases among patients. Disease counts and distributions have been computed for all age groups and socioeconomic scores separately and the results are represented in form of charts, tables, graphs and other statistical measurements.

Table 4.1 and Figure 4.1 show that the average number of coexisting diseases grows with age. This evidence shows that prevalence of multimorbidity increases by age. According to the Table 4.1, elder and middle aged groups are the most dominant age groups in this dataset in terms of frequencies, forming 46.4 and 36.1 percent of patients respectively. The distribution of male and female patients is 44.77 and 55.23 percent respectively, and regardless of number of coexisting diseases, the counts for female patients is always greater than male patients. However, this does not imply that females are generally sicker than males, because the limited sources of data forming the DELPHI database. As it is shown in Figure 4.1 the average number of diseases per patient among males and females is 2.15 and 2.19 respectively, with standard deviations of 1.46 and 1.52, respectively.



AVG	1.062	1.146	1.301	1.558	1.9454	2.640	3.143	2.040	2.508	2.145	2.187
SD	0.2638	0.4009	0.6505	0.9295	1.3023	1.6608	1.8157	1.3983	1.6696	1.4604	1.5174
SE	0.0198	0.0203	0.0240	0.0282	0.0185	0.0208	0.1772	0.0139	0.0282	0.0186	0.0174

Figure 4.1: Average number of diseases

Table 4.1 also provides information regarding to percentage of patients according to the number of diseases they are diagnosed with. As shown in the table, the total counts of patients

No. of Diseases	Total Patients(%)	Male(%)	Female(%)	Child(%)	Adolescent(%)	Young Adult(%)	Adult(%)	Middle Age(%)	Elder(%)
1	46.02	45.15	54.85	2.67	5.41	9.03	11.33	40.23	31.33
2	23.24	44.96	55.04	0.28	1.35	3.99	6.85	37.45	50.08
3	14.06	44.50	55.50	0.05	0.36	1.51	5.50	33.18	59.87
4	7.96	44.31	55.69	0	0	0.46	3.30	26.97	69.27
5	4.67	43.82	56.18	0	0	0.63	1.88	27.07	70.42
6	2.34	44.86	55.14	0	0	0.31	0.62	24.30	74.77
7	1.12	38.96	61.04	0	0	0	1.30	16.88	81.82
8	0.34	44.68	55.32	0	0	0	0	17.02	82.97
9	0.18	40	60	0	0	0	0	12	88
10	0.04	0	100	0	0	0	0	0	100
11	0.03	0	100	0	0	0	0	50	50
Total(%)	100	44.77	55.23	1.30	2.85	5.37	7.97	36.08	46.44

Table 4.1: This table aims to represent the distribution of patients who share minimum one characteristic (i.e. number of coexisting diseases, gender, age group). The first column stands for the number of coexisting diseases within one patient. In our data table, the most number of coexisting diseases are found among four patients who are diagnosed with eleven chronic conditions. Every row in the table represents percentage of patients who have a specific number of diseases. The distributions are represented for male and female sub groups as well as every one of the six age groups separately.

decreases exponentially as the number of coexisting diseases grows. This makes the analysis of patient data with larger numbers of coexisting diseases more challenging or sometimes impossible due to lack of data. Because of this, the main focus in this research is on statistical analysis and clustering of patients with two coexisting diseases. However, some statistical analyses have been done on patients diagnosed with one, two or three diseases.

As presented in Figure 4.2, increasing age can be associated with increasing prevalence of some chronic diseases. Some conditions tend to appear more with increases in age (e.g. hypertension, diabetes, hyperlipidemia, cancer, cardiovascular disease, hearth failure, arthritis, osteoporosis and dementia) [104], while other conditions tend to occur less with growth of age (such as bronchitis and obesity), and some other conditions tend to appear more among certain groups and less in other age groups (an example is depression which occurs the most among young adults and adults, and decreases in older individuals). Hypertension and hyperlipidemia are the two most commonly occurring diseases, appearing 46% and 34.5% of the elderly, respectively. Liver disease is the least occurring disease among all age groups and, at its peak, appears 0.4% of middle aged group. There are two charts included in Appendix A illustrating the disease distributions over all age groups in more detail: Figure A.1 represents the absolute counts of diseases for every age group in a greater detail, and Figure A.2 represents disease percentages for all age groups for males and females separately.

Patient distribution and their basic characteristics have also been analyzed for the three socioeconomic scores separately: The majority (73.66%) of patients in this data set belong to fourth income quintile, almost about 25.57% of the patients belong to highest income quintile and the rest belong to third income quintile.[2] As shown in Figure 4.1, the average number of diseases for patients with moderate income level equals 3.14 with a standard deviation of 1.82, which is greater when compared to the other two groups. This can imply that patients with relatively lower income levels are more likely to be diagnosed with multiple diseases when compared to individuals with higher income levels. However, exploring this hypothesis is not in the scope of this research.

Figure 4.3 is shows the percentages of patients diagnosed with each of the twenty chronic

[2]Third, fourth and highest income quintiles represent families with moderate, high and highest income levels, respectively.

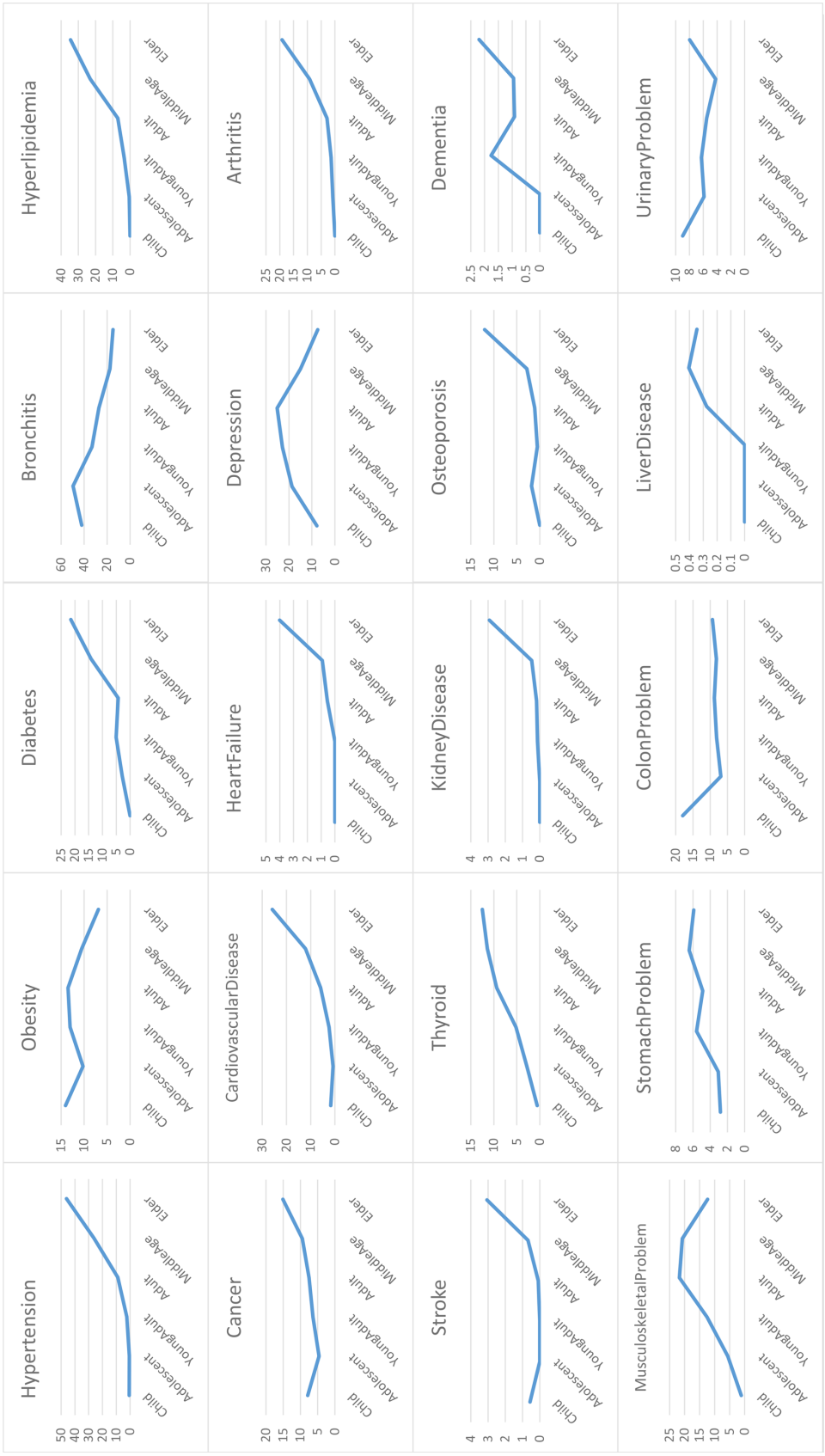


Figure 4.2: Disease percentages for all age groups and individual diseases. The reason for representing disease trends using percentages instead of absolute counts is the variation in counts of patients in each age groups. In this line graphs, the vertical axis values are the percentages of diagnosis among age groups.

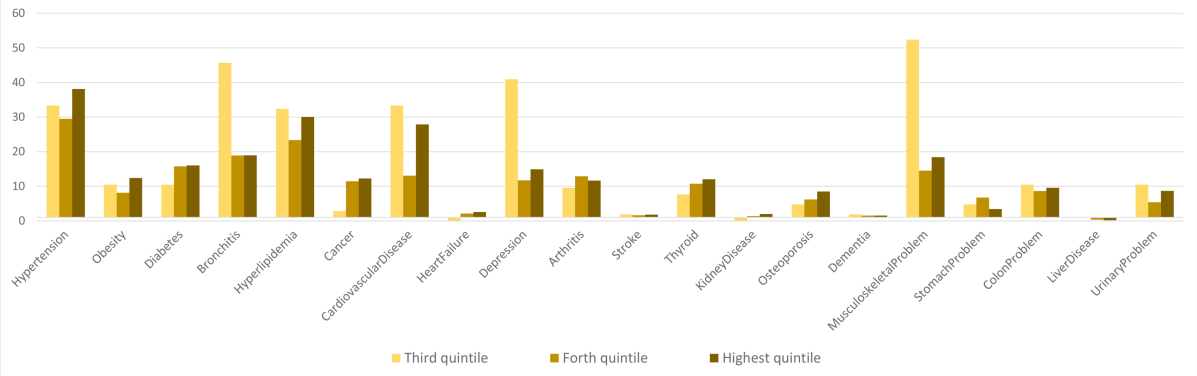


Figure 4.3: Disease percentages for the three income quintiles of individual diseases.

diseases for each income quintile. If the diagnosis of chronic diseases was not associated with socioeconomic status, the heights of the three bars for every disease were expected to be in the same range. As shown in the figure, bar heights are in the same range only for some of the diseases. Some diseases are shown to appear more frequently among individuals with moderate income level (examples in this data are bronchitis, cardiovascular disease, depression and musculoskeletal problems). However, caution must be used in drawing any conclusions about the relationship between socioeconomic status and the prevalence of diseases mentioned above for two reasons: (1) the factors included for measuring socioeconomic status are limited to the area the patient lives in and an estimation of their income level, as well as the lack of many other important parameters such as race, ethnicity, level of education, occupation, smoking and drinking habits, etc. (2) available instances for every socioeconomic group are not of equal (or almost equal) size (there are only 105 individuals from third quintile of income in this database).

However, there has been other research which has specifically focused on the impact of socioeconomic status on diagnosis, prognosis, and prevalence of a certain chronic disease. For example Alexis Ferré *et al.*have researched and confirmed that socioeconomic status influences both the prevalence (higher in low categories) and rate of diagnosis (lower in high categories) of bronchitis [41]. Similarly, Alexander M. Clark and colleagues have researched the relationship between socioeconomic status and cardiovascular disease by associating lower socioeconomic status to higher prevalence of cardiovascular disease [21]. Other related research analyzed the influence of socioeconomic status on depression, obesity and diabetes [36], which suggests

poor socioeconomic status affects physical and mental health and functioning of young adults and the effects persist across the life time. Some evidence of this is also illustrated in Figure 4.3.

On the other hand, there is research showing no socioeconomic differences is evident for prevalence of some chronic conditions such as cancer, kidney disease and skin disease [29]. A detailed survey analyzing the racial and ethnic influences on health can be found in [26].

In this section, we have analyzed the prevalence of chronic diseases individually among patients belonging to different age groups, gender categories and socioeconomic status, without concentrating on coexisting diseases in patients. The following section provides similar data analysis for individuals with two coexisting diseases.

4.2 Analysis of Two Coexisting Chronic Diseases

In order to analyze prevalence of multimorbidity, combinations between two diseases have been analyzed with statistical measurement methods introduced in Chapter 2 subsections 2.1.1 and 2.1.2.

4.2.1 Total Counts and Pairwise Association of Two Chronic Diseases

As it was discussed in Chapter 2, disease counts is the most preliminary way to determine the prevalence of multimorbidity. According to Table 4.1, 43.06% of patients with multimorbidity are diagnosed with two chronic conditions, 87.53% of which are elder or middle aged.

Figure 4.4 shows the total counts of diseases as they appear among patients with two chronic conditions. Each circle in the plot represents the existence of corresponding disease combinations. The diameter of each circle refers to the relative percentage of patients diagnosed with the corresponding disease combination. As shown in Figure 4.4, hypertension is the most commonly occurring disease among patients with multimorbidity: It appears among 31.17% of patients diagnosed with two conditions, and 47.96% of patients with multimorbidity. Hypertension and hyperlipidemia is the most occurring disease combination which affects 225 patients in the database. Musculoskeletal problems is the other relatively dominant disease commonly co-occurring with other conditions.

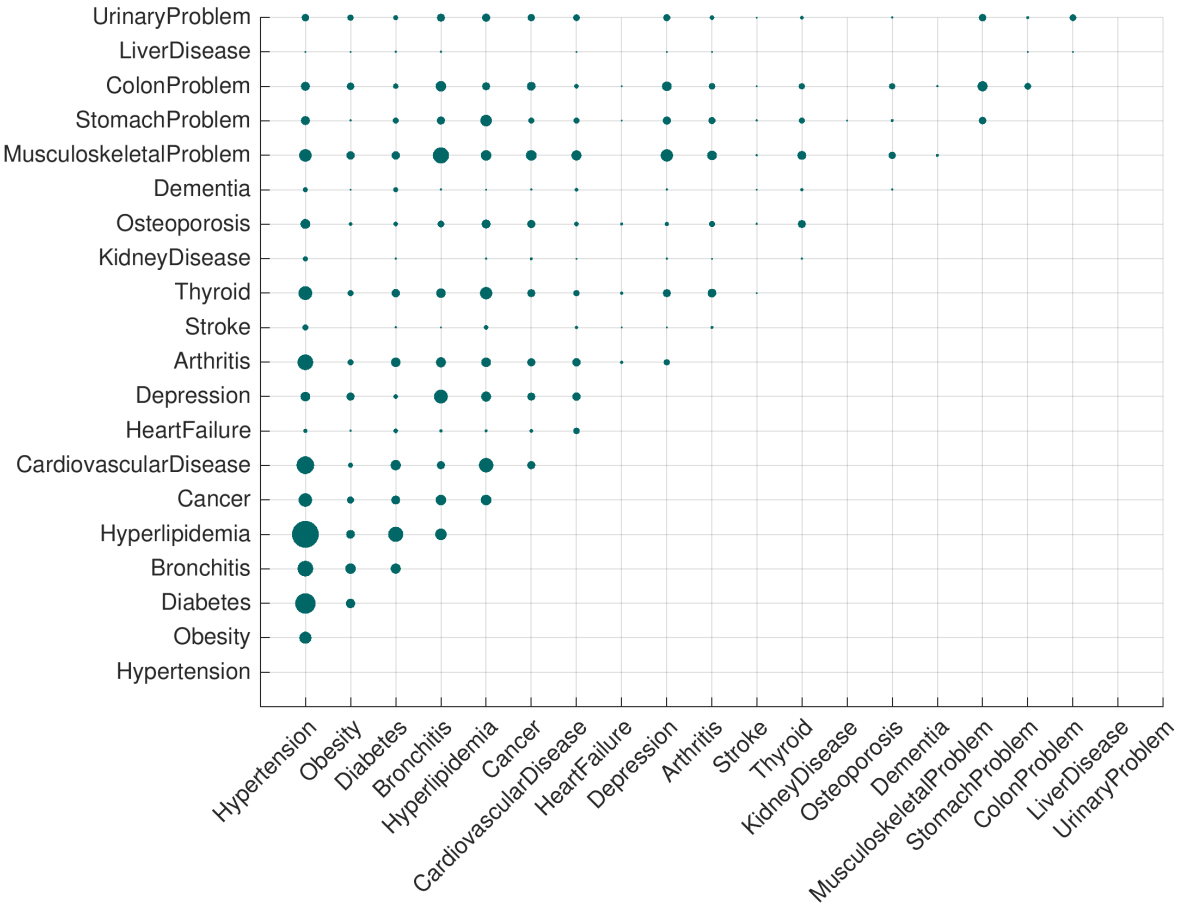


Figure 4.4: Disease combination counts for two coexisting diseases. Each circle on the plot represents the relative percentage of instances in the dataset who have the two chronic conditions as named on x and y coordinates.

Disease associations have also been measured according to pairwise association measurement methods introduced in Chapter 2[3]. Odds ratios, multimorbidity coefficients, and Kappa statistics [4] suggest the highest rate of pairwise association refers to heart failure and kidney disease. According to odds ratios ($OR_{HeartFailure,KidneyDisease} = 10.927$), the chance of having kidney disease if one has heart failure is 10.927, and with 95% confidence the true odds ratio lies in the rage of 6.97 and 17.11. Similarly, Kidney disease and hypertension, cardiovascular disease and heart failure, are defined as other highly associated disease pairs. A complete list

[3]In order to avoid the inaccuracy of results due to selection bias, all patient records available in DELPHI database have been included in measuring pairwise association between diseases; this also includes patients who are not diagnosed with any chronic conditions.

[4]Computed using equations 2.1, 2.3, 2.4.

of odds ratios for all disease pairs is available in Appendix A, Table A.2, and the corresponding 95% confidence intervals are presented in Table A.3. Multimorbidity coefficients of all pairwise disease combinations is also included in Appendix A, Table A.4: greater multimorbidity coefficients refer to higher pairwise association levels. Some of the highly associated pairs have also appeared commonly in disease pair counts (Figure 4.4), such as hypertension and hyperlipidemia, hypertension and diabetes.

4.2.2 Disease Combinations Based on Gender

The distribution of male and female patients with two chronic conditions is 44.96% and 55.04%, respectively. Table 4.2 shows percentages of patients who have two chronic diseases in every age group based on gender. As the distributions suggest, multimorbidity in younger age groups, according to the data in the DELPHI database, appears to be slightly more prevalent among females compared to males.

	Child	Adolescent	Young Adult	Adult	Middle Age	Elder	Total
Male (%)	0.42	1.12	2.87	6.43	36.55	52.62	44.96
Female (%)	0.17	1.54	4.91	7.19	38.18	48	55.04

Table 4.2: Percentages of patients for all age groups based on gender

According to Figures A.3 and A.4 in Appendix A, the disease pair hypertension and hyperlipidemia is most commonly appearing among both males and females. However, some diseases might be more gender specific. For instance, hypertension and hyperlipidemia tend to co-occur with a second disease more often among males than females, while depression, thyroid problems and osteoporosis are the three diseases which co-occurs with a second disease relatively more often among females compared to males.

4.2.3 Disease Combinations Based on Age Groups

As mentioned earlier in Section 4.1, the prevalence of some chronic conditions increase in older population. Therefore, it is expected that multimorbidity is also occurring with a higher prevalence and more variety among older age groups. We have analyzed the validity of this hypothesis by performing statistical analysis on patients with multimorbidity among all age

groups. As shown in Table 4.3, the prevalence of having two chronic conditions grows by age, such that it is extremely rare among children, and mainly appears among middle age and elder groups. The table provides the percentage of patients with two chronic conditions among all age groups.

	Child	Adolescent	Young Adult	Adult	Middle Age	Elder
2 coexisting diseases (%)	0.28	1.35	3.90	6.75	37.17	50.02

Table 4.3: Patient distribution percentages among all age groups for patients with two chronic conditions.

The variation of disease combinations also grows by age. Almost all possible combinations of two chronic conditions can be found among elderly. The distribution of disease combinations among every age group is presented in six figures in Appendix A (Figures A.5 to A.10). Based on the results in these Figures, disease combinations appear with three main trends: (1) The occurrence between some disease pairs is more dominant among younger age groups and decreases among older age group. For instance, the occurrence of bronchitis and other diseases is dominant among younger age groups and diminishes among older age groups. (2) Some disease pairs tend to appear more with growth of age, including hypertension and other diseases, which start to appear from age group 3 (young adults) and grows with age and becomes the most dominant disease among elderly. (3) Some disease combinations seem to be age specific, which means they seem to peak among certain age groups. An example is the combination between depression and other diseases: this first appears among adolescents, increases in young adults, and starts to decrease in later stages of life, such that among elders the combination of depression with another chronic condition is relatively rare.

4.2.4 Disease Combinations based on Socioeconomic Scores

As Table 4.4 suggests, patients with two chronic conditions mainly belong to families in high income quintile and form 74.96% percent of the population of patients with two chronic conditions.

The distribution of the prevalence of disease pairs belonging to each one of the socioeconomic groups is illustrated in Figures A.11, A.12 and A.13 in Appendix A. As shown in those

	Child	Adolescent	YoungAdult	Adult	MiddleAge	Elder	Total
Moderate Income	0	0	3	3	9	2	17
High Income	6	35	86	159	920	1180	2386
Highest Income	3	8	38	56	263	412	780

Table 4.4: Patient counts for each age group based on family income quintiles. Only patients with two chronic condition are listed in this table.

Figures, the distributions of disease pairs among patients belonging to the moderate family income quintile is very different from those belonging to high or highest family income quintiles, while the high and highest income quintiles suggest a similar distribution among disease pairs. However, these results do not imply that disease combinations have a different pattern among patients with two chronic conditions who belong to moderate family income level, compared to those belonging to high or highest family income levels. The reason for different disease patterns might be related to insufficient population of patients belonging to the moderate family income level in this database. Disease pair combinations among two subgroups of patients from high and highest income levels are very similar to the combinations for all patients with two chronic conditions (Figure 4.4). This observation suggests that belonging to different income levels does not necessarily imply a visible change in multimorbidity patterns for patients with two chronic conditions.

In analyzing disease pair combinations for gender and age groups we were able to see some patterns which may suggest the presence of clusters. For socioeconomic scores, it is harder to see whether it is a parameter upon which disease pair combinations can be clustered.[5] In the next section we introduce a clustering algorithm for clustering characteristics of patients diagnosed with the same two chronic conditions and explore the presence or absence of possible clusters.

[5]It is important to note that there are major assumptions made for calculating the socioeconomic score in this research, and the socioeconomic score basically takes into account patients FSA and their potential family income level. Therefore, the results implied from disease combinations among patients belonging to different socioeconomic levels will possibly not be applicable for the national population of Canada.

4.3 Clustering of Data for Two Coexisting Diseases

This section is devoted to introducing the clustering algorithm used for clustering most commonly appearing disease pairs, as well as associated experimental results. The algorithm used for this purpose is a variation of $k-$modes adjusted to fit with the data table used in this research. In this regard, $k-$modes algorithm is explained in detail; it is compared with the original $k-$means in terms of similarities and differences. Following the introduction of the algorithm and its adjustments, the experimental results are presented in the final sub-section of this section.

4.3.1 K-means vs. K-modes

As mentioned in section 2.1.3, one of the limitations of the $k-$means algorithm is its distance based computational nature, which allows the algorithm to fit only for numerical data sets. To overcome this limitation, researchers have suggested different approaches[6], one of which is the $k-$modes algorithm. This approach is a variation of the $k-$means algorithm used for clustering data containing only categorical data types by making three main modifications to $k-$means: (1) using a simple matching dissimilarity measurement method for categorical data types, (2) computing modes instead of mean to define cluster centers, and (3) using a frequency-based method to update centers of clusters.[7] The following paragraphs represent the notations and description of the $k-$modes algorithm in detail.

Assume that D is a set of n elements noted as $D = \{X_1, ..., X_n\}$, where each element is represented by a set of m attributes noted as $A_1, ..., A_m$, and each attribute contains a set of values described as domain of values or $Dom(A_i)$ for $1 \leq i \leq m$. As noted earlier, the $k-$modes

[6]One of these approaches was presented by Ralambondrainy in 1995 [96], where categorical attributes were converted into binary attributes by using 0 or 1 to show whether the category is absent or present, and the binary values were treated as numeric attributes in the same $k-$means algorithm. In this approach, every categorical attribute with m possible categories(values), will need to be divided into m binary attributes, each indicating whether or not the value i for $1 \leq i \leq m$ is present. Therefore, it must handle a very large set of attributes for databases where there are large sets of categorical attributes, because every value of each categorical attribute will need to become an independent binary attribute. Due to this, the computational and space complexity of the algorithm can inevitably increase in some cases. Also, clusters means lose their meaning as characteristics of clusters in this approach because of the presence of binary values which will be real numbers 0.0 and 1.0.

[7]To discuss the clustering algorithms in this chapter, an assumption has been made that all data types can be converted into either numerical or categorical data types.

algorithm is valid for clustering data-sets containing categorical data types only. Therefore, in this notation $Dom(A_i)$ is a finite and unordered set. An instance in data-set D is represented by X_j for $1 \leq j \leq n$, which is an $m-$dimensional data vector represented as $X_j = [x_{j,1}, x_{j,2}, ..., x_{j,m}]$ where $x_{j,i} \in A_i$, and all attributes of X_j are of type categorical. In this notation, it is assumed that all data instances have exactly m attributes, all attributes are categorical, and there are no missing values. Consequently, $X_i = X_j$ if $x_{i,k} = x_{j,k}$ for $1 \leq k \leq m$ and $1 \leq i, j \leq n$. However the equation $X_i = X_j$ does not imply that the two instances X_i and X_j are pointing to the same object in real word, it means that the two separate instances have equal values for all attributes $A_1, ..., A_m$ [58].[8]

The main steps of the $k-$modes algorithm are same as the $k-$means. Although, the following modifications have been made to $k-$means to make it computable for categorical attributes. Here are the main computational changes for clustering a categorical data-set using $k-$modes:

Data points' dissimilarity measure: In $k-$means the Euclidean distance is used to define the dissimilarity between two data points, while it is not generally applicable to categorical data types. Therefore, in $k-$modes, the dissimilarity measure is the total number of mismatches of all attributes for two instances. In other words, the dissimilarities between two data points X and Y is the *Hamming distance* of the two instances, which is calculated as follows:

$$\eta(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \tag{4.1}$$

where

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \tag{4.2}$$

Using modes instead of means: As noted in definition of $k-$means, after assignment of all data points to cluster groups based on the distance measure, all clusters' centroids are updated

[8] In our database, $X_i = X_j$ if gender, age group and socioeconomic score are equal in two patients who have the same diseases.

by computing the mean vector for every one of the clusters. Due to the nature of categorical attributes, computing *means* is not useful for this attribute type. Therefore, a different but similar measure has been used in $k-$modes - to find and place a data point in a cluster (e.g. q) in a way that minimizes the overall Hamming distance between all data points and q. Assuming that S is a set of n categorical instances defined as $S = \{s_1, s_2, ...s_n\}$, a mode of S is a vector $q = \{q_1, q_2, ..., q_m\}$ that minimizes the sum of hamming distances between all instances of S and q. In other words, q is the mode of S if it minimizes the following equation:

$$d(S, q) = \sum_{i=1}^{n} \eta(s_i, q) \tag{4.3}$$

According to the following theorem, the computational metric to minimize equation 4.3 is proven to be *mode* of the cluster. The proof of the theorem can be found in Appendix A.

Theorem1. *The function d(S,q) is minimised if and only if* $f_r(A_j = q_j|S) \geq f_r(A_j = c_{k,j}|S)$ *for* $q_j \neq c_{k,j}$ *for all* $j = 1, ..., m$, *where* $f_r(A_j = c_{k,j}|S) = \frac{n_{c_{k,j}}}{n}$ *stands for the relative frequency of* $k-$*th category* $c_{k,j}$ *in attribute* A_j *from set* S.

From Theorem 1 it is implied that the mode of a set S is not unique, and it is not necessarily an element of S.

Using frequency-based methods for updating centers of clusters: Cluster representatives in both the $k-$means and $k-$modes algorithms are vectors containing the same attributes as all data points in data set. The cluster representative vector may or may not be one of the data points in data set, and they are used to compute the distance between the data points and clusters in order to assign every data point to the nearest cluster. In contrast to the $k-$means algorithm where cluster representatives are *mean* vectors for every cluster (also called centroids), in the $k-$modes algorithm the *mode* vector is the cluster representative, and it is created by calculation of modes for all attributes in every cluster.

$K-$**modes algorithm steps**: By introducing the main differences between $k-$means and $k-$modes, the main steps of the latter algorithm are provided below:

1. Initialization: Select k initial modes, one for representing each cluster

2. Data point assignments: Every one of the data points will be allocated to a cluster whose mode is nearest to it based on equation 4.1. Update the mode for the cluster every time a new instance is assigned to it.

3. Reassigning instances: Once all data points have been allocated to clusters, retest the Hamming distances between every data point in all the clusters. If a data point is nearer to a different cluster mode according to equation 4.1, reallocate the instance to that cluster, then update the mode for both clusters.

4. Repeat step 3 until the number of data points moving between clusters is zero.

Similar to the $k-$means algorithm, the results of the $k-$modes algorithm are dependent on the initial modes and the order of objects in the dataset [58]. There are two main methods for selecting initial modes: (1) Randomly selecting k instances from the data set and use them as k modes. (2) Distribute most frequently occurring categories of attributes equally among initial modes. The second method needs more processing power, and is explained in full detail in subsection 4.3.2

One of the limitations of $k-$modes algorithm is its applicability on data containing only categorical data types. In order to overcome this limitation, a more generalized algorithm named $k-$prototypes has been introduced by Huang in 1997 [57], which integrates $k-$means and $k-$modes to allow data clustering for databases containing mixed numerical and categorical attributes.[9] Detailed description of the algorithm can be found in [57].

The attribute types of all features in the data table used in this research are categorical. Therefore, the $k-$modes algorithm is well suited to be used for clustering the data. The algorithm has been adjusted for the database and implemented in order to find clusters among patients sharing the same diseases.

[9]The clustering process of the $k-$prototypes algorithm is similar to $k-$means, except that it uses the $k-$modes algorithm approach to update the categorical attribute values in centroids. Assuming s^r is dissimilarity measure for numeric attributes (which was defined by Euclidean distance in $k-$means) and s^c is dissimilarity measure for categorical attributes, defined by counting the matches and mismatches of categorical attributes, the dissimilarity measure for two data instances having mixed categorical and numeric values in the $k-$prototypes algorithm will be $s^r + \gamma s^c$, where γ is a coefficient (weight) to balance the effect of either type of attribute when computing distance between two data points.

4.3.2 K-modes Adjustments and Implementation

In this research the $k-$modes algorithm has been implemented for finding clusters among patients who share same pair of diseases. However, an adjustment has been made to distance measurement method of the algorithm, in order to make the clustering results more reliable. This sub-section explains the alternative measurement method and the implementation of the algorithm in full detail.

Alternative distance measurement method: The socio-demographic attributes of data points in our data are the following:

$$Gender = \{Male, Female\}$$
$$AgeGroup = \{Child, Adolescent, YoungAdult, Adult, MiddleAge, Elder\}$$
$$SocioeconomicScore = \{Moderate, High, Highest\}$$

Even though these attributes are categorical, the categories of the latter two attributes can be ordered from lowest/youngest to highest/oldest. Using Hamming distance to calculate the distance between two data points in this case, will show dissimilarities to some extent, but it will fail to recognize the level of dissimilarity in many cases. For instance, assume that there are three patients with following attributes:

$$p_1 = [Male, Child, Highest] \quad p_2 = [Male, Elder, Moderate] \quad p_3 = [Female, Child, High]$$

According to distance measure in $k-$modes (i.e. equation 4.1), the distance between the two pairs (p_1, p_2) and (p_1, p_3) will be $\eta(p_1, p_2) = \eta(p_1, p_3) = 2$, while the pair (p_1, p_3) appears to be sharing more similar socio-demographic information compared to pair (p_1, p_2). Dissimilarity of data points can be measured more accurately by using the *Manhattan distance*[10][74].

Manhattan distance: For two $m-$dimensional vectors $X = [x_1, ..., x_m]$ and $Y = [y_1, ..., y_m]$, the Manhattan distance is defined as the sum of absolute differences for every attribute in these vectors. More formally,

$$d_\mu(X, Y) = \sum_{j=1}^{m} |x_j - y_j| \tag{4.4}$$

[10]The Manhattan distance is also known as rectilinear distance, L_1 distance, snake distance or city block distance.

More specifically, if we associate the three attributes *Gender*, *AgeGroup*, *Socioeconomic-Score* as attribute 1, 2 and 3 respectively, the Manhattan distance of two patients p_i and p_j in our database can be defined as following:

$$d_\mu(p_i, p_j) = |p_{i,1} - p_{j,1}| + |p_{i,2} - p_{j,2}| + |p_{i,3}, p_{j,3}| \tag{4.5}$$

Methods for generating k initial mode vectors[11]**:** The $k-$modes algorithm can be initialized either by randomly selecting and associating k data points as initial cluster centers, or by generating initial k modes by distributing most frequently occurring categories for each attribute. In random selection of k initial mode vectors, k data points are randomly selected from the main data set and placed as initial cluster representatives. However, the second approach for generating initial mode vectors requires more programming effort, and there is more than a unique method for implementing it. The strategy applied in this research for generating initial mode vectors is explained in the following steps:

1. Calculating frequencies: The frequencies of all categories for all attributes are calculated and stored in a *category array*.

2. Sorting frequencies: The frequencies for every attribute is sorted in descending order such that $f(c_{i,j}) \geq f(c_{i+1,j})$ where $f(c_{i,j})$ is the frequency of category i for attribute j. There is an example of a *category array* holding categories of m (in this case 5) attributes in equation 4.6. For every attribute a_j where $1 \leq j \leq m$, $n(a_j)$ represents the number of categories of a_j. For instance, in this example $n(a_2) = 5$.

[11] A *mode vector* for cluster C is a data vector $v = \{v_1, v_2, ..., v_m\}$ where $v_i \in Dom(a_i)$ represents the one category in attribute a_i with most occurrences among all data points of cluster C.

$$C_{array} = \left\{ \begin{array}{ccccc} c_{1,1} & c_{1,2} & c_{1,3} & c_{1,4} & c_{1,5} \\ c_{2,1} & c_{2,2} & c_{2,3} & c_{2,4} & c_{2,5} \\ & c_{3,2} & c_{3,3} & c_{3,4} & c_{3,5} \\ & c_{4,2} & & c_{4,4} & c_{4,5} \\ & c_{5,2} & & c_{5,4} & \\ & & & c_{6,4} & \end{array} \right\} \tag{4.6}$$

Column j for $1 \leq j \leq 5$ in equation 4.6 represents the frequencies of all categories of attribute j in descending order, and $c_{i,j}$ stands for category i of attribute j.

3. Shifting the *category array*: In order to distribute most frequently occurring categories equally among the k initial modes (in this example $k = 3$), every attribute needs to be shifted circularly in a vertical manner in the *category array* such that $x-$th attribute will have $x - 1$ vertical shifts from bottom to top. With every shift, the category positioned as the last item of the column, will become the first item, and all the other items (categories), will shift down one row. According to the descriptions, the *category array* in equation 4.6 will become as in equation 4.7 after all shifts :

$$C_{array} = \left\{ \begin{array}{ccccc} c_{1,1} & c_{5,2} & c_{2,3} & c_{4,4} & c_{1,5} \\ c_{2,1} & c_{1,2} & c_{3,3} & c_{5,4} & c_{2,5} \\ & c_{2,2} & c_{1,3} & c_{6,4} & c_{3,5} \\ & c_{3,2} & & c_{1,4} & c_{4,5} \\ & c_{4,2} & & c_{2,4} & \\ & & & c_{3,4} & \end{array} \right\} \tag{4.7}$$

As shown in equation 4.7, there are no shifts for a_1, 1 shift for a_2, 2 shifts for a_3, 3 shifts for a_4 and 4 shifts for a_5.

4. Initializing k modes: After shifting the elements of attribute array, the empty spots of

the array will be filled with categories in the same order, for each attribute. For instance, as attribute a_1 has two categories, it has four less categories compared to a_4 which contains the maximum number of categories. Therefore, empty spots in first column (i.e. in attribute 1) in 4.7 will be filled with categories of a_1 with the same order as they appeared after shifting the array. The category array after filling the empty spots will be as following:

$$C_{array} = \begin{Bmatrix} c_{1,1} & c_{5,2} & c_{2,3} & c_{4,4} & c_{1,5} \\ c_{2,1} & c_{1,2} & c_{3,3} & c_{5,4} & c_{2,5} \\ c_{1,1} & c_{2,2} & c_{1,3} & c_{6,4} & c_{3,5} \\ c_{2,1} & c_{3,2} & c_{2,3} & c_{1,4} & c_{4,5} \\ c_{1,1} & c_{4,2} & c_{3,3} & c_{2,4} & c_{1,5} \\ c_{2,1} & c_{5,2} & c_{1,3} & c_{3,4} & c_{2,5} \end{Bmatrix} \tag{4.8}$$

Initial modes can be created once the array is filled. In this example every row of category array in 4.8 represents one initial mode, where the categories have been equally distributed among modes. For $k = 3$, the initial modes according to 4.8 will be:

$$q_1 = [c_{1,1}, c_{5,2}, c_{2,3}, c_{4,4}, c_{1,5}]$$

$$q_2 = [c_{2,1}, c_{1,2}, c_{3,3}, c_{5,4}, c_{2,5}]$$

$$q_3 = [c_{1,1}, c_{2,2}, c_{1,3}, c_{6,4}, c_{3,5}]$$

In order to avoid repetitive modes for $k > n(a_{max})$ where a_{max} stands for the maximum number of categories among all attributes (in this case $n(a_{max}) = n(a_4)$), repeat step 3 after every $n(a_{max})$ generations of modes. For instance, if $k = 14$ and $n(a_{max} = 6)$, step 3 is required to be repeated twice: once after generating 6th initial mode, and once after generating 12th initial mode.

Modified $k-$modes Algorithm: After introducing the modification to distance measurement

method of the $k-$modes algorithm, here we provide the complete pseudocode of the algorithm which is adjusted for creating clusters of patients diagnosed with same pair of chronic diseases in our data table. The purpose of this algorithm is to find clusters over age group, gender and socioeconomic score of the patients who are diagnosed with same two diseases, and analyze whether or not they are also sharing the same or similar demographics. The name (or code) of the two chronic diseases is passed to the algorithm as input parameters, and the initialization step, extracts all patients diagnosed with the same two diseases are from the data table and stored in a temporary memory for further analysis.[12]

The following subsection, provides the experimental results for clustering of commonly occurring disease pairs.

4.3.3 Experimental Results

The number of possible combinations for two diseases to co-occur from a list of 20 diseases equals to $20 \times 19 = 380$. However all of these combinations do not occur among patients in DELPHI database: there are certain combinations of diseases which occur more frequently. In order to perform further research and find the similarities in characteristics of patients with same combination of diseases, a list of most frequently occurring combinations of multiple chronic diseases was collected from a different research [90], where disease combination counts have been derived from the CPCSSN database which is a relatively greater data source compared to the one used in this research. Having access to information derived from a greater population size ensures the reliability of results to some extent. This list contains most frequently occurring disease combinations for two coexisting diseases. These combinations are used for further analysis and clustering of patients in our dataset. Table 4.5 shows the common combinations of diseases for two coexisting conditions according to CPCSSN database.

The individuals with same pair of chronic conditions have been clustered according to the modified version of the $k-$modes algorithm (Algorithm 1). The metric for determining the number of clusters for each pair of disease is minimization of within-cluster sum of squared

[12]It is important to note that for extracting patients with two chronic disease d_1 and d_2, the patients with only two diseases are considered for searching in the main data table. In other words if a patient is diagnosed with three diseases d_1,d_2 and d_3, they will not be included in the list of patients diagnosed with d_1 and d_2.

Algorithm 1 Modified $k-$modes algorithm for clustering demographic information of patients diagnosed with most commonly occurring pairs of chronic diseases: Part 1

procedure MODIFIED-K-MODES($k, Dss1, Dss2, DataSet, InitializeMethod$)
  // k : number of clusters
  // $Dss1$ and $Dss2$: the two diseases existing among all patients in $DataSet$
  // $DataSet$: The data array holding the records of patients having $Dss1$ and $Dss2$.

  // Deciding which method to use for initializing the centroids (initial *mode* vectors)
  if $InitializeMethod$ == "*Random*" **then**
    $centroids \leftarrow$ Select K random patients' from $DataSet$
  else // Generate K initial modes
    $centroids \leftarrow$GENERATECENTROIDS($k, DataSet$)
  end if
  // Continue data point assignments until no data points move between clusters
  repeat
  // assign data points to clusters
    for every data point p_i in $DataSet$ **do**
      Compute Manhattan distance of p_i with all $centroids$
      find j such that $d_\mu(p_i, centroids(j))$ is minimized
      $cluster(p_i) \leftarrow j$
      $centroids(j) \leftarrow$UPDATECLUSTER($j, DataSet$)
    end for
  // Checking for any potential data points that can be switched between clusters
    for every cluster j **do**
      if p_i in cluster j exists such that $d_\mu(p_i, centroids(x)) < d_\mu(p_i, centroids(j))$ **then**
        $cluster(p_i) \leftarrow x$
        $centroids(j) \leftarrow$UPDATECLUSTER($j, DataSet$)
        $centroids(x) \leftarrow$UPDATECLUSTER($x, DataSet$)
      end if
    end for
  until no data points in $DataSet$ switch between clusters (i.e. convergence)
end procedure

errors (computed according to equation 2.8). Both methods of centroid initialization (random selection and creation of centroids based on attribute modes) were tested on all commonly occurring disease pairs: Number of clusters and their distribution was not affected by the initialization method for all disease pairs. Therefore, the results presented in this section stand for both initialization methods.

Figure 4.5 represents clusters appeared for individuals diagnosed with hypertension and hyperlipidemia. In the DELPHI database this disease pair appears most frequently among

Algorithm 2 Modified $k-$modes algorithm: Part 2

function GENERATECENTROIDS($k, DataSet$)
    Calculate and **sort** the frequency of each category for every attribute
    $[g_1, g_2] \leftarrow$ frequencies of $Male$ and $Female$ categories in $gender$ attribute
    $[a_1, a_2, a_3, a_4, a_5, a_6] \leftarrow$ frequencies of age group categories
    $[s_1, s_2, s_3] \leftarrow$ frequencies of socioeconomic score categories
    $C_{array} \leftarrow$Sort each column of C_{array} based on frequency in decreasing order
    $C_{array} \leftarrow$Apply 1 and 2 circular shifts to column 2 and 3 in C_{array} respectively
    for $i = 1 : k$ **do** // Generate $i-$ centroid and store in $i - th$ row of $centroids$
        if $\text{mod}(i, 6) == 0$ **then**
            $C_{array} \leftarrow$Apply 1 and 2 circular shifts to column 2 and 3 in C_{array} respectively
        end if
        $centroids(i) \leftarrow C_{array}[i, :]$
    end for
    return $centroids$
end function

function UPDATECLUSTER($j, DataSet$) // update mode vector of cluster j
    Count frequency of every category in all attributes of data points which belong to cluster j
    $[g_1, g_2] \leftarrow$ frequencies of $Male$ and $Female$ categories in $gender$ attribute
    $[a_1, a_2, a_3, a_4, a_5, a_6] \leftarrow$ frequencies of age group categories
    $[s_1, s_2, s_3] \leftarrow$ frequencies of socioeconomic score categories
    $g^* \leftarrow \max(g_1, g_2)$, $a^* \leftarrow \max(a_1, a_2, a_3, a_4, a_5, a_6)$, $s^* \leftarrow \max(s_1, s_2, s_3)$
    $centroids(j) \leftarrow [g^*, a^*, s^*]$
    return $centroids(j)$
end function

patients diagnosed with two chronic illnesses. As the figure suggests, the two dimensions of the plot represent socioeconomic score and age group of individuals. Clusters are distinguished with different colors: the lighter luminance of a color represents the male individuals, and the darker luminance stands for the female individuals in a cluster. Each pie on every block of the graph represents the cluster appeared for corresponding socioeconomic score and age group. The radius of every pie shows the relative number of instances (patients) belonging to that cluster. Every pie is partitioned into two slices, showing the proportion of male and female patients in the cluster. Two clusters overlap if they appear in the same block and correspond to the same gender attribute. For instance, in Figure 4.5 clusters 2 and 3 would overlap if they both belonged to whether male or female patients. The legend of the graph shows every cluster's corresponding colors for both male and female patients. For each cluster, one of the

	Diseases Combination	Patient Counts in DELPHI
1	Depression and Musculoskeletal Problem	49
2	Hypertension and Musculoskeletal Problem	50
3	Hypertension and Hyperlipidemia	225
4	Hypertension and Diabetes	132
5	Cancer and Musculoskeletal Problem	37
6	Cancer and Depression	20
7	Hypertension and Depression	30
8	Hypertension and Cancer	58
9	Hyperlipidemia and Musculoskeletal Problem	36
10	Bronchitis and Musculoskeletal Problem	82

Table 4.5: 10 most frequently occurring disease combinations derived from CPCSSN database [90]. Disease combinations are sorted by frequency of occurrence among patients in CPCSSN database. The column *Patient Counts in DELPHI* represents the number of individuals with corresponding diseases combination in DELPHI database which was used for this research.

squares in the legend is marked with a cross: this refers to the gender attribute of the centroid of corresponding cluster. For instance, in Figure 4.5, the gender attribute of centroids for clusters 1 and 3 is male and for cluster 2 is female. Finally, the pie on the right side of the plot represents patients distributions for every cluster. The distribution of weights on all attributes is equal for measuring the distance between two data points.

As the clustering results for hypertension and hyperlipidemia suggest, partitioning the data over three clusters minimizes within-cluster sum of squared errors. The clusters do not have any overlap and they partition the data over gender and age group. According to Figure 4.5, the population of the elder age group with hypertension and hyperlipidemia is almost twice as much as middle aged group: therefore the clusters split the data over the two main age groups (middle age and elder) and then divide males and females in elder age group.

Table 4.6 summarizes the distribution of patients diagnosed with one of the commonly occurring chronic disease pairs over all age group and socioeconomic categories. As the table suggests, there are two common disease pairs widely spread among almost all age group (excluding child age group) and all socioeconomic groups. The rest of the disease pairs mainly occur among older age groups and socioeconomic scores 4 and 5. Musculoskeletal problem and hypertension appear among 5 of this disease pair as one of the two diseases.

Clustering results for the other nine commonly occurring disease pairs (listed in Table 4.5)

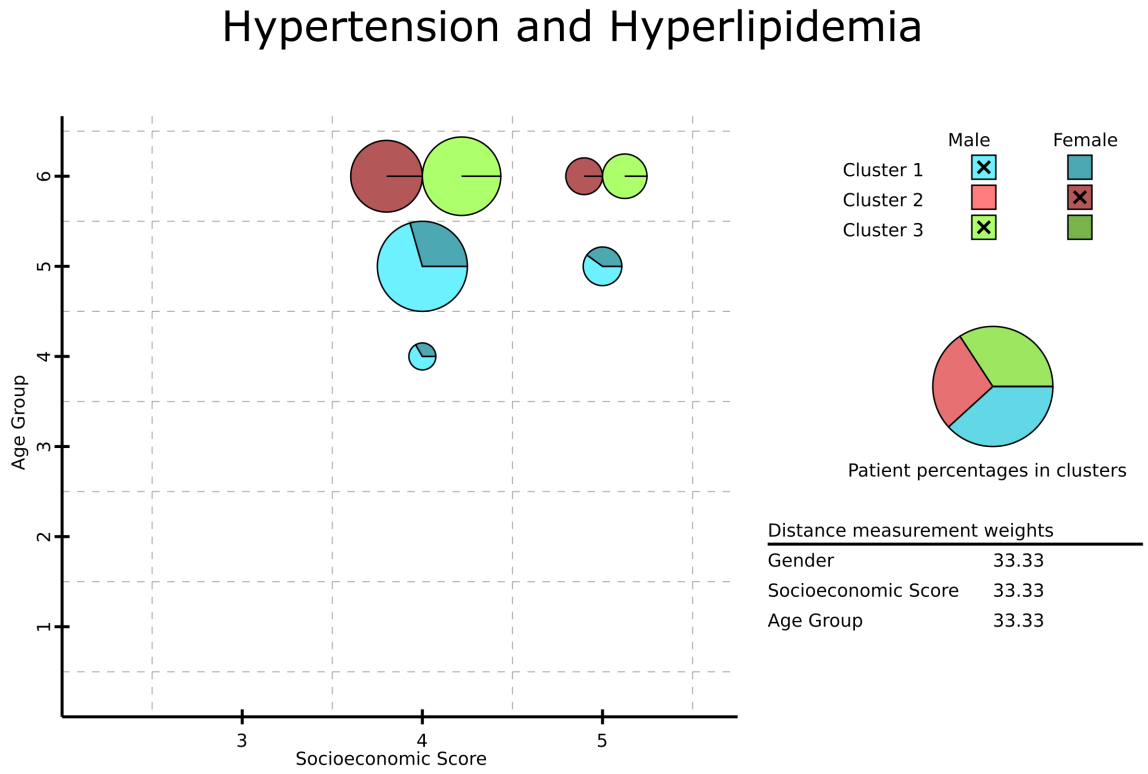


Figure 4.5: Clusters of patients with hypertension and hyperlipidemia. Each pie with color c on block $[x, y]$ represents the patients with socioeconomic score x, age group y who belong to cluster c. The radius of the pie represents the relative number of patients existing in corresponding block in graph. Color shade represents gender. As the clusters suggest, patients belonging to the fourth and fifth age group have been divided from patients of age group 6. Cluster 1 belongs to male and female patients from age group four and five. The other two clusters belong to elder patients and each one holds patients of one gender attribute. The clusters for this disease pair do not separate the patients over their socioeconomic score.

are included in Appendix B Figures B.1 to B.9. According to this results, the number of clusters dividing patients in each disease pair is ranging between 3 and 5. For all ten disease pairs, some of the clustered are divided based on gender and most of them are also divided based on age group. Patients partitioning over socioeconomic score has also happened among 4 disease pairs, and interestingly musculoskeletal problem appeared among 3 of these 4 pairs. The summary of the results of clusters is presented in Table 4.7. There were no clear patterns observed in relation to patients characteristics and the attributes the data was clustered over.

Diseases Combination	Age Group					Socioeconomic Score		
	2	3	4	5	6	3	4	5
{Depression,Musculoskeletal Problem}	✓	✓	✓	✓	✓	✓	✓	✓
{Hypertension, Musculoskeletal Problem}	-	-	✓	✓	✓	-	✓	✓
{Hypertension, Hyperlipidemia}	-	-	✓	✓	✓	-	✓	✓
{Hypertension, Diabetes}	-	✓	✓	✓	✓	-	✓	✓
{Cancer, Musculoskeletal Problem}	-	✓	✓	✓	✓	-	✓	✓
{Cancer, Depression}	-	✓	✓	✓	✓	-	✓	✓
{Hypertension, Depression}	-	-	✓	✓	✓	-	✓	✓
{Hypertension, Cancer}	-	-	✓	✓	✓	-	✓	✓
{Hyperlipidemia, Musculoskeletal Problem}	-	-	✓	✓	✓	✓	✓	✓
{Bronchitis, Musculoskeletal Problem}	✓	✓	✓	✓	✓	✓	✓	✓

Table 4.6: The distribution of patient characteristics for individuals diagnosed with commonly co-occurring disease pairs.

Diseases Combination	Clusters	Clustering results rely on:		
		Gender	Age Group	Socioeconomic Score
{Depression, Musculoskeletal Problem}	5	✓		
{Hypertension, Musculoskeletal Problem}	3	✓	✓	✓
{Hypertension, Hyperlipidemia}	3	✓	✓	
{Hypertension, Diabetes}	4	✓	✓	
{Cancer, Musculoskeletal Problem}	3	✓		✓
{Cancer, Depression}	4	✓	✓	
{Hypertension, Depression}	3	✓	✓	
{Hypertension, Cancer}	3	✓		✓
{Hyperlipidemia, Musculoskeletal Problem}	3	✓	✓	✓
{Bronchitis, Musculoskeletal Problem}	5	✓	✓	

Table 4.7: The summary of clustering results for 10 commonly curring disease pars. Every row represents the summary of clustering results for one disease pair. The number of clusters which minimize the within-cluster sum of squares are included in second column. The attributes upon which the clustering has occurred are indicated with check marks.

4.4 Statistical Analysis of Three Coexisting Chronic Diseases

The chart in Figure 4.6 shows the percentage of middle aged or elder patients with any number of chronic conditions. As the chart suggests, the distribution of patients with multimorbidity focuses more on older population as the number of coexisting chronic conditions grow. According to the chart, 91.53 percent of the patients who are defined as multimorbid by having three chronic conditions are elder or middle aged. On the other hand, the absolute patient counts

with multimorbidity decreases exponentially as the number of coexisting diseases per patient grows. In the database of this research, there are 1926 patients in total from all age groups who have three coexisting diseases. The limited number of instances who have many coexisting chronic conditions, restricts the opportunity of further research. However, this section provides some insight on demographic characteristics of patients with three chronic illnesses.

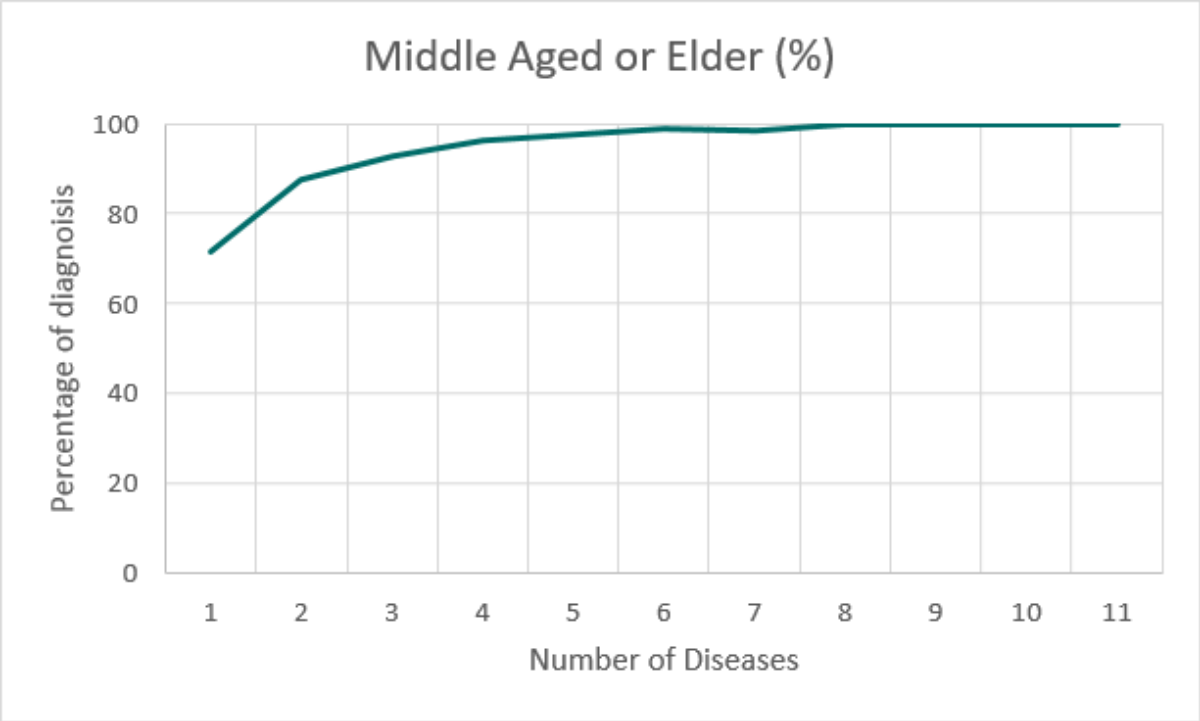


Figure 4.6: Multimorbidity occurs mostly among middle aged or elder patients, and it affects a greater proportion of these two age groups as the number of coexisting diseases grows. 87.53% of patients diagnosed with two chronic conditions are either middle aged or elder, and this percentage grows with number of coexisting diseases, such that co-occurrence of 8 or more diseases happens only among middle aged and elder patients.

The distribution of patients with three chronic diseases among males and females is 45.43% and 54.57%, respectively. The variety of combinations for three co-occurring conditions among patients is 493 based on the data in the DELPHI database. The most commonly occurring disease triples are listed in Table 4.8 and sorted based on patient counts.

The characteristics of patients diagnosed with common disease triples is represented in Table 4.9. As the table suggests some of the disease triples (for instance 1st and 3rd combination) occur among a greater variety of age groups and all socioeconomic score categories, whereas the other combinations are more specific to older age groups and higher socioeconomic scores.

	Disease Combination	Patient Counts
1	Hypertension, Diabetes and Hyperlipidemia	116
2	Hypertension, Hyperlipidemia and Cardiovascular Disease	75
3	Bronchitis, Depression and Musculoskeletal Problem	59
4	Hypertension, Diabetes and Cardiovascular Disease	28
5	Hypertension, Hyperlipidemia and Arthritis	24
6	Hypertension, Diabetes and Thyroid	22
7	Hypertension, Hyperlipidemia and Cancer	21

Table 4.8: The list shows most frequently occurring disease combinations derived from DELPHI database. All disease combinations occurring among more than 20 patients are listed in this table. Disease combinations are sorted by frequency of occurrence among patients. The column *Patient Counts* represents the number of individuals with corresponding disease combination in the DELPHI database.

Some of the combinations are more prevalent among males (such as 2nd combination), some are more prevalent among female population (i.e. 6th disease combination), and there are combinations which are not specific to males or females, and occur among both genders almost equally (3rd combination is an example). Almost 61.21% of the patients with one of these disease combinations are male and the rest are female.

	Diseases Combination	Age Group				Socioeconomic Score			Gender Counts	
		3	4	5	6	3	4	5	M	F
1	{HT, DB, HL}	✓	✓	✓	✓	✓	✓	✓	71	45
2	{HT, HL, CD}	-	✓	✓	✓	-	✓	✓	54	21
3	{BC, DP, MP}	✓	✓	✓	✓	✓	✓	✓	31	28
4	{HT, DB, CD}	-	-	✓	✓	-	✓	✓	17	11
5	{HT, HL, AT}	-	-	✓	✓	-	✓	✓	10	14
6	{HT, DB, TD}	-	-	✓	✓	-	✓	✓	4	18
7	{HT, HL, CC}	-	-	✓	✓	-	✓	✓	16	5

Table 4.9: The distribution of patients characteristics for individuals diagnosed with commonly co-occurring chronic disease triples. Abbreviations: M=Male, F=Female, HT=Hypertension, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, DP=Depression, AT=Arthritis, TD=Thyroid, MP=Musculoskeletal Problem.

As the *Risk Ratios* suggest, patients are 4.9 times more likely to develop a third chronic condition if they are diagnosed with any one of the commonly occurring disease pairs. On the other hand, every one of the commonly occurring disease pairs is a subset of at least one of the commonly co-occurring disease triples. In order to measure the association between commonly occurring disease pairs and commonly occurring disease triples, we have calculated

the multimorbidity coefficients for commonly occurring disease pairs and all twenty chronic conditions to find out how likely it is for a patient to be diagnosed with any of the twenty chronic conditions if they are already diagnosed with any one of the common disease pairs. The results are presented in Table 4.10. According to the results, multimorbidity coefficients show a high association between three of the common disease triples and disease pairs. For instance, patients with hypertension and hyperlipidemia (one of the common disease pairs) are most likely to be diagnosed with diabetes compared to all other diseases, and these three diseases are one of the commonly occurring combinations among patients with three chronic illnesses. The same association can be found among depression, musculoskeletal problem and bronchitis. However, not all of the commonly occurring disease triples were shown to be highly associated according to multimorbidity coefficients, and not all highly associated diseases according to multimorbidity coefficients are among common combinations of three co-occurring diseases. The rates of multimorbidity coefficients show the likelihood of diagnosis with a third chronic disease for a patient with any one of the disease pairs to be diagnosed with a third chronic condition: the higher the rate, the more is the prevalence of the corresponding disease triples to occur.

According to disease association rates computed by multimorbidity coefficients, the common disease triples among patients in DELPHI database which do not appear to be highly associated in the table, have appeared either due to coincidental multimorbidity, or lack of enough data for examination, or any other medical related reasons which multimorbidity coefficients are not taking into account.

The reason for choosing multimorbidity coefficients over the other pairwise association measurement methods for measuring disease associations, is its ability to separate coincidental multimorbidity (random multimorbidity) from disease co-occurrences by causal association (non-random multimorbidity). The other pairwise disease association methods (such as odds ratios or risk ratios) do not separate between random and non-random multimorbidity.

Disease Pair	Multimorbidity Coefficient between disease pairs and a third chronic disease																			
	HT	OS	DB	BC	HL	CC	CD	HF	DP	AT	SK	TD	KD	OP	DT	MP	SP	CP	LD	UP
{HT, DB}	0	0.84	0	0.46	0.93	0.55	0.63	0.54	0.42	0.54	0.63	0.49	0.58	0.37	0.58	0.37	0.61	0.41	1.32	0.49
{HT, HL}	0	0.70	0.96	0.50	0	0.57	0.79	0.49	0.50	0.62	0.74	0.56	0.52	0.56	0.51	0.48	0.48	0.46	0.44	0.67
{HT, CC}	0	0.47	0.68	0.41	0.68	0	0.54	0.47	0.27	0.68	0.68	0.56	0.51	0.76	0.76	0.46	0.76	0.53	0.79	0.54
{HT, DP}	0	0.65	0.56	0.68	0.64	0.29	0.83	0.39	0	0.36	0.27	0.52	0.39	0.48	0.28	0.63	0.59	0.50	1.11	0.43
{HT, MP}	0	0.57	0.50	0.71	0.63	0.49	0.70	0.32	0.64	0.66	0.36	0.55	0.33	0.51	0.31	0	0.36	0.67	0.27	0.42
{BC, MP}	0.50	0.38	0.34	0	0.51	0.38	0.66	0.34	1.44	0.52	0.34	0.55	0.19	0.37	0.07	0	0.39	0.68	0	0.48
{HL, MP}	0.61	0.52	0.49	0.71	0	0.49	0.67	0.32	0.67	0.63	0.36	0.64	0.23	0.44	0.22	0	0.52	0.60	0	0.48
{CC, DP}	0.37	0.51	0.28	0.65	0.50	0	0.55	0.32	0	0.39	0	0.53	0.19	0.69	0.59	0.85	0.97	0.73	0	0.47
{CC, MP}	0.49	0.39	0.41	0.53	0.49	0	0.42	0.30	0.66	0.61	0.70	0.64	0.58	0.73	0.33	0	0.38	0.78	0	0.53
{DP, MP}	0.43	0.49	0.29	1.40	0.47	0.46	0.58	0.18	0	0.43	0.19	0.48	0.13	0.49	0.20	0	0.36	0.73	0	0.52

Table 4.10: Multimorbidity coefficients of commonly occurring pair of chronic conditions with a third disease. Abbreviations: HT=Hypertension, OS=Obesity, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, HF=Heart Failure, DP=Depression, AT=Arthritis, SK=Stroke TD=Thyroid, KD=Kidney Disease, OP=Osteoporosis, DT=Dementia, MP=Musculoskeletal Problem, SP=Stomach Problem, CP=Colon Problem, LD=Liver Disease, UP=Urinary Problem.

Chapter 5

Conclusion

The main goal of the research underlying this thesis was to investigate, explore and identify different approaches for analyzing socio-demographic characteristics of patients with multiple chronic conditions, with a specific focus on patients sharing same chronic disease combinations. This chapter provides an overview on approaches taken to analyze the burden of multimorbidity according to patients demographic characteristics. It summarizes potential similarities and differences identified among patients diagnosed with same number of diseases focusing on patient groups sharing the same combination of diseases, and concludes by providing strengths and limitations of the research, as well as potential areas for future work.

5.1 Main Contributions

In this research, we demonstrated the applicability of analytical and clustering methods to data about patients with multiple chronic diseases. The dataset used for analysis purposes contains 13697 records of patients with one or more chronic conditions. Socio-demographic information extracted for every patient are gender, age group, and socioeconomic score. The methods explored include disease counts, pairwise association measurement approaches (e.g. odds ratios, risk ratios and multimorbidity coefficients), and clustering algorithms ($k-$modes was the primary algorithm used for analysis in this research). Besides exploring analytical methods, we have also suggested different visualization approaches for representing the data of patients with multimorbidity, which includes representation of patients socio-demographic

characteristics and their relationships.

In order to explore relationships between certain chronic conditions and patient demographic characteristics, prevalence and distribution of diseases were analyzed for every chronic disease separately over gender, six age groups and three socioeconomic scores among patients diagnosed with one or more chronic conditions. The results suggest that age is important in the prevalence of occurrences of multimorbidity: Older age groups had more occurrences of multiple chronic conditions compared to younger groups. Some diseases appeared to be age-associated, whereas some others did not.

Pairwise co-occurrence rates of chronic diseases was computed using *multimorbidity coefficients (MC)*, and it was shown that certain combinations (disease pairs) are very likely to co-occur compared to others. Disease pair counts showed that some of the highly associated disease pairs according to MC were frequently occurring among patients; an example of this is the pair hypertension and hyperlipidemia. Disease pair distributions were also computed for males and females, all age groups and all socioeconomic score levels separately. The results suggest that some diseases occur more often with a second disease among males (e.g. hypertension) and some among females (e.g. thyroid). The variation of disease combinations grows by age, and for every age group there are certain disease combinations which tend to occur more compared to other age groups. Disease pair distribution results for patients from socioeconomic score 3 were less varying compared to scores 4 and 5, and the latter two scores suggested a similar distribution to total disease pair distributions, with no visible dissimilarities.

To further explore patient characteristics with same diseases, we clustered patients with certain disease pair combinations. The observations suggest that gender plays a role in dividing patients into clusters in every disease pair, and age group in most of them, while socioeconomic score plays a role in division of clusters mostly among pairs where musculoskeletal problems are one of the two diseases.

Finally, patients' socio-demographic characteristics were analyzed for individuals diagnosed with three chronic conditions. The overall analysis of this group showed that diagnosis of three chronic conditions mainly appears among middle aged and elder population. According to disease counts, the most frequently co-occurring diseases are hypertension, diabetes and hyperlipidemia. Multimorbidity coefficients measured the concurrence between commonly co-

occurring disease pairs and all eighteen other diseases, and it was shown that some of the highly associated triples (according to MC) were also commonly occurring in patients of this dataset (according to disease counts), were a subset of commonly co-occurring disease pairs.

One of the secondary contributions of this research was to identify socio-demographic score for patients from a different data source and tie it to each patient in main database given the lack of these attributes in original data-set. This approach can be applied to similar data-sets (which are greater in size) to investigate the burden of multimorbidity, although the socioeconomic score (as measured) did not appear to be directly associated with co-occurrence of multiple chronic condition according to our dataset due to following reasons: (1) high level of abstraction in defining socioeconomic score; (2)lack of availability of socio-demographic attributes; (3) lack of data.

5.2 Strengths, Limitations and Implications of the Research

Despite numerous past and ongoing research about the prognosis, diagnosis and overall burden of multimorbidity, there is very little known about characteristics of patients with multimorbidity. This research is among the first of its kind exploring methods for analyzing multimorbidity data, focusing on patient groups sharing the same combination of diseases. There were challenges and limitations during the course of this research, which will be pointed out in the following, along with assumptions, estimations and strategies used to overcome them.

Small dataset: In order to get reliable results from analyzing patient data, the main requirement is having access to data with sufficient numbers of patients with multimorbidity. The DELPHI database contained 7394 patients with multimorbidity at the time of extraction. However, the available amount of data drops dramatically when extracting the information of patients with a certain number and certain combination of diseases. This was a limitation for further exploration of dataset. Due to the small number of available data we were only able to explore the demographic characteristics of patients with certain and most occurring combinations of diseases.

Lack of availability of socioeconomic attributes: It is commonly seen that regardless of the source of medical data, socioeconomic variables usually contain incomplete or miss-

ing attributes that cannot be used in statistical analysis. Although, each patient's gender, age, and residential area were collected from the database, these attributes do not account for socioeconomic factors that impact or are affected by health. As we needed a metric identifying the economic status of patients along side the other variables, we have extracted median family income levels of individuals for all FSAs from a different source and matched it with the residential area of patients in our dataset. The final socioeconomic score was computed by matching the FSA-based median family income to the FSA of individuals in DELPHI database which contains a high level of abstraction and makes the results prone to external criticism.

Complex reality of multimorbidity: As discussed in Chapter 2, the phenomenon of multimorbidity is very complex. There are many aspects that can play a role in patients health outcomes. Patients' educational background, current and past occupations, smoking and drinking habits, distress levels, ethnic background, cultural and environmental effects are some examples of factors that can affect or be affected by the prevalence of multimorbidity. Collecting this information for patients in a database is a tedious and challening process. However, assuming all these variables are accessible, the question become how to aggregate these factors into single attributes/variables and how to summarize them to compute the socioeconomic score? The other question is how to identify whether a socioeconomic factor is impacting health outcomes or being affected by health-related complications? Although the scoring system can enable the opportunity to perform computational research on multimorbidity, the research in this area has still a long way to go to create a reliable scoring system for computing the socioeconomic burden of multimorbidity.

5.3 Future Research

There are many opportunities for improving this research as well as further analysis of multimorbid patient data, some of which are provided in this section.

Usage of a larger dataset: Some of the limitations in this research were due to lack of data for multimorbid patients. For instance, we have overgeneralized the metric used in computation of socioeconomic score due to lack of availability of socioeconomic variables. Furthermore, lack of sufficient records for patients with more than three diseases, restricted the opportunity to

perform further exploration on demographic characteristics of patients with a greater number of diseases. With a larger dataset, it is possible to do further and more precise analysis on characteristics of multimorbid patients.

Applying multimorbidity measurement approaches on time series data: It is also important to have access to information on patients over a duration of time. Analysis of sequential data on patients' diagnosis and socioeconomic status over periods of time can also help understand the complexity of multimorbidity by helping to identify key factors and life changes that may affect the occurrence of multimorbidity and to understand the effects multimorbidity have on quality of life.

Exploring co-occurring disease combinations among patients with same socio-demographic characteristics: Beside exploring the distribution of diseases (or disease pairs) for all age groups, socioeconomic scores, and the gender separately, we have analyzed and clustered socio-demographic characteristics of patients with same number (as well as same combination) of chronic conditions. However, analyzing and clustering the distribution of different disease counts and combinations for all patients belonging to the same gender, age group and socioeconomic status would be another interesting approach in analyzing multimorbid patient data.

Replacing disease counts with pairwise associations measurement methods: In order to define highly associated disease pairs, we have used absolute disease pair counts, and ten of the most commonly occurring pairs were used for clustering. Although, it is also possible to derive highly associated diseases using the pairwise association measurement methods, such as risk ratios, multimorbidity coefficients or Kappa statistics methods. These measurement methods provide a more precise notion on disease associations compared to absolute counts. However, the reason for choosing counts over these methods was insufficient available data for most of the highly associated disease pairs defined by pairwise association methods.

Applying other clustering methods: Our approach for clustering patients with common disease pair using modified version of $k-$modes may benefit from further exploration and comparison with results of other clustering approaches. By using Factor Analysis methods it is also possible to identify the type of disease associations and their corresponding risk factors (the causation models for two diseases and their risk factors are presented in Figure 1.3).

Bibliography

[1] Nancy E Adler, Thomas Boyce, Margaret A Chesney, Sheldon Cohen, Susan Folkman, Robert L Kahn, and S Leonard Syme. Socioeconomic status and health: the challenge of the gradient. *American psychologist*, 49(1):15, 1994.

[2] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009.

[3] Gerard F Anderson. *Chronic care: making the case for ongoing care*. Robert Wood Johnson Foundation, 2010.

[4] Karen Barnett, Stewart W Mercer, Michael Norbury, Graham Watt, Sally Wyke, and Bruce Guthrie. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380(9836):37–43, 2012.

[5] L Batstra, EH Bos, and J Neeleman. Quantifying psychiatric comorbidity. *Social psychiatry and psychiatric epidemiology*, 37(3):105–111, 2002.

[6] Elizabeth A Bayliss, Jennifer L Ellis, and John F Steiner. Subjective assessments of comorbidity correlate with quality of life health outcomes: initial validation of a comorbidity assessment instrument. *Health and Quality of life Outcomes*, 3(1):1, 2005.

[7] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[8] Rob V Bijl, Anneloes Ravelli, and G Van Zessen. Prevalence of psychiatric disorder in the general population: results of the netherlands mental health survey and incidence study (nemesis). *Social psychiatry and psychiatric epidemiology*, 33(12):587–595, 1998.

[9] Richard V Birtwhistle. Canadian primary care sentinel surveillance network a developing resource for family medicine and public health. *Canadian Family Physician*, 57(10):1219–1220, 2011.

[10] Norman Blaikie. *Analyzing quantitative data: From description to explanation*. Sage, 2003.

[11] Ruth Bonita, Robert Beaglehole, and Tord Kjellström. *Basic epidemiology*. World Health Organization, 2006.

[12] Andrew Booth. " brimful of starlite": toward standards for reporting literature searches*. *Journal of the Medical Library Association*, 94(4):421, 2006.

[13] Cynthia M Boyd, MD Martin Fortin, et al. Future of multimorbidity research: how should understanding of multimorbidity inform health system design? *Public Health Reviews*, 32(2):1, 2010.

[14] Rodolfo A Bulatao, Norman B Anderson, et al. *Understanding racial and ethnic differences in health in late life: A research agenda*. National Academies Press, 2004.

[15] Julie E Byles, Catherine D'Este, Lynne Parkinson, Rachel O'Connell, and Carla Treloar. Single index of multimorbidity did not predict multiple outcomes. *Journal of clinical epidemiology*, 58(10):997–1005, 2005.

[16] Patrick Cain. Income by postal code: Mapping Canada's richest and poorest neighbourhoods. http://globalnews.ca/news/370804/income-by-postal-code/, 2013.

[17] Statistics Canada. Average adjusted after-tax income by after-tax income quintiles for all persons, 2006 to 2010. http://www.statcan.gc.ca/pub/75-202-x/2010000/analysis-analyses-eng.htm, 2010.

[18] Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987.

[19] Domenic V Cicchetti and Alvan R Feinstein. High agreement but low kappa: Ii. resolving the paradoxes. *Journal of clinical epidemiology*, 43(6):551–558, 1990.

[20] Thomas Claassen. *Causal discovery and logic*. UB Nijmegen [host], 2013.

[21] Alexander M Clark, Marie DesMeules, Wei Luo, Amanda S Duncan, and Andy Wielgosz. Socioeconomic status and cardiovascular disease: risks and implications for care. *Nature Reviews Cardiology*, 6(11):712–722, 2009.

[22] Aileen Clarke. What is a chronic disease? the effects of a re-definition in hiv and aids. *Social Science & Medicine*, 39(4):591–597, 1994.

[23] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[24] WONCA Classification Committee et al. Ichppc-2-defined (international classification of health problems in primary care, 1983.

[25] William Corser, Alla Sikorskii, Ade Olomu, Manfred Stommel, Camille Proden, and Margaret Holmes-Rovner. Concordance between comorbidity data from patient self-report interviews and medical record documentation. *BMC Health Services Research*, 8(1):1, 2008.

[26] National Research Council, Committee on Population, et al. *Critical perspectives on racial and ethnic differences in health in late life*. National Academies Press, 2004.

[27] HL Crabtree, CS Gray, AJ Hildreth, JE O'Connell, and J Brown. The comorbidity symptom scale: a combined disease inventory and assessment of symptom severity. *Journal of the American Geriatrics Society*, 48(12):1674–1678, 2000.

[28] Eileen M Crimmins, Mark D Hayward, and Teresa E Seeman. Race/ethnicity, socioeconomic status, and health. *Critical perspectives on racial and ethnic differences in health in late life*, pages 310–352, 2004.

[29] Jetty AA Dalstra, Anton E Kunst, Carme Borrell, Elizabeth Breeze, Emmanuelle Cambois, Giuseppe Costa, José JM Geurts, Eero Lahelma, Herman Van Oyen, Niels K Rasmussen, et al. Socioeconomic differences in the prevalence of common chronic diseases: an overview of eight european countries. *International journal of epidemiology*, 34(2):316–326, 2005.

[30] R Davis, E Wagner, and Trish Groves. Patients as partners in managing chronic disease. *Bmj*, 320:526–527, 2000.

[31] Vincent de Groot, Heleen Beckerman, Gustaaf J Lankhorst, and Lex M Bouter. How to measure comorbidity: a critical review of available methods. *Journal of clinical epidemiology*, 56(3):221–229, 2003.

[32] Joseph AC Delaney and John D Seeger. Sensitivity analysis. *Agency for Healthcare Research and Quality (US)*, 2013.

[33] Claudia Diederichs, Klaus Berger, and Dorothee B Bartels. The measurement of multiple chronic diseases-a systematic review on existing multimorbidity indices. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 66(3):301–311, 2011.

[34] Hal A Droogleever Fortuyn, Rolf Fronczek, Mirjan Smitshoek, Sebastiaan Overeem, Martijn Lappenschaar, Joke Kalkman, Willy Renier, Jan Buitelaar, Gert Jan Lammers, and Gijs Bleijenberg. Severe fatigue in narcolepsy with cataplexy. *Journal of sleep research*, 21(2):163–169, 2012.

[35] Benjamin G Druss, Steven C Marcus, Mark Olfson, Terri Tanielian, Lynn Elinson, and Harold Alan Pincus. Comparing the national economic burden of five chronic conditions. *Health Affairs*, 20(6):233–241, 2001.

[36] Susan A Everson, Siobhan C Maty, John W Lynch, and George A Kaplan. Epidemiologic evidence for the relation between socioeconomic status and depression, obesity, and diabetes. *Journal of psychosomatic research*, 53(4):891–895, 2002.

[37] Martine Extermann. Measurement and impact of comorbidity in older cancer patients. *Critical reviews in oncology/hematology*, 35(3):181–200, 2000.

[38] Vincent S Fan, David Au, Patrick Heagerty, Richard A Deyo, Mary B McDonell, and Stephan D Fihn. Validation of case-mix measures derived from self-reports of diagnoses and health. *Journal of clinical epidemiology*, 55(4):371–380, 2002.

[39] Alvan R Feinstein. The pre-therapeutic classification of co-morbidity in chronic disease. *Journal of chronic diseases*, 23(7):455–468, 1970.

[40] Alvan R Feinstein and Domenic V Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549, 1990.

[41] Alexis Ferré, Claire Fuhrman, Mahmoud Zureik, Christos Chouaid, Alain Vergnenègre, Gérard Huchon, Marie-Christine Delmas, and Nicolas Roche. Chronic bronchitis in the general population: influence of age, gender and socio-economic conditions. *Respiratory medicine*, 106(3):467–471, 2012.

[42] Robert H Fletcher, Suzanne W Fletcher, and Grant S Fletcher. *Clinical epidemiology: the essentials*. Lippincott Williams & Wilkins, 2012.

[43] Patient-Centred Innovations for Persons with Multimorbidity. PACE in MM. `http://paceinmm.recherche.usherbrooke.ca/index.php`, 2014.

[44] Martin Fortin, Lise Lapointe, Catherine Hudon, and Alain Vanasse. Multimorbidity is common to family practice: is it commonly researched? *Canadian Family Physician*, 51(2):244–245, 2005.

[45] Martin Fortin, Hassan Soubhi, Catherine Hudon, Elizabeth A Bayliss, and Marjan van den Akker. Multimorbidity's many challenges. *bmj*, 334(7602):1016–1017, 2007.

[46] Ha Droogleever Fortuyn, GA Lappenschaar, JW Furer, and PP Hodiamont. Anxiety and mood disorders in narcolepsy. *aspects of the psychiatric phenotype*, 32:49–56, 2011.

[47] Linda P Fried, Karen Bandeen-Roche, Judith D Kasper, Jack M Guralnik, et al. Association of comorbidity with disability in older women: the womenâĂŹs health and aging study. *Journal of clinical epidemiology*, 52(1):27–37, 1999.

[48] Johnson George, Tam Vuong, Michael J Bailey, David CM Kong, Jennifer L Marriott, and Kay Stewart. Development and validation of the medication-based disease burden index. *Annals of Pharmacotherapy*, 40(4):645–650, 2006.

[49] Dana P Goldman and James P Smith. Can patient self-management help explain the ses health gradient? *Proceedings of the National Academy of Sciences*, 99(16):10929–10934, 2002.

[50] Sheldon Greenfield, Giovanni Apolone, Barbara J McNeil, and Paul D Cleary. The importance of co-existent disease in the occurrence of postoperative complications and one-year recovery in patients undergoing total hip replacement: comorbidity and outcomes after hip replacement. *Medical care*, 31(2):141–154, 1993.

[51] Agneta Grimby and Alvar Svanborg. Morbidity and health-related quality of life among ambulant elderly citizens. *Aging Clinical and Experimental Research*, 9(5):356–364, 1997.

[52] Dianne L Groll, Teresa To, Claire Bombardier, and James G Wright. The development of a comorbidity index with physical function as the outcome. *Journal of clinical epidemiology*, 58(6):595–602, 2005.

[53] Stephen F Hall. A user's guide to selecting a comorbidity index for clinical research. *Journal of clinical epidemiology*, 59(8):849–855, 2006.

[54] Greg Hamerly and Charles Elkan. Alternatives to the k-means algorithm that find better clusterings. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 600–607. ACM, 2002.

[55] Catherine Hoffman, Dorothy Rice, and Hai-Yen Sung. Persons with chronic conditions: their prevalence and costs. *Jama*, 276(18):1473–1479, 1996.

[56] Libby Holden, Paul A Scuffham, Michael F Hilton, Alexander Muspratt, Shu-Kay Ng, and Harvey A Whiteford. Patterns of multimorbidity in working australians. *Population health metrics*, 9(1):15, 2011.

[57] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, pages 21–34. Citeseer, 1997.

[58] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304, 1998.

[59] Soiya M Hunt and James McEwen. The development of a subjective health indicator. *Sociology of health & illness*, 2(3):231–246, 1980.

[60] Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the tenth annual symposium on Computational geometry*, pages 332–339. ACM, 1994.

[61] R Antonelli Incalzi, O Capparella, A Gemma, F Landi, E Bruno, F Di Meo, and P Carbonin. The interaction between age and comorbidity contributes to predicting the mortality of geriatric patients in the acute-care hospital. *Journal of internal medicine*, 242(4):291–298, 1997.

[62] Rachel Jenkins, G Lewis, P Bebbington, T Brugha, M Farrell, B Gill, and H Meltzer. The national psychiatric morbidity surveys of great britain–initial findings from the household survey. *Psychological medicine*, 27(04):775–789, 1997.

[63] Robert J Johnson and Fredric D Wolinsky. The structure of health status among older adults: disease, disability, functional limitation, and perceived health. *Journal of health and social behavior*, pages 105–121, 1993.

[64] UT Kadam, PR Croft, et al. Clinical multimorbidity and physical function in older adults: a record and health status linkage study in general practice. *Family practice*, 24(5):412–419, 2007.

[65] David Kalisch. The 14th biennial health report of the australian institute of health and welfare. Technical report, Australian Institution of Health and Welfare, 06 2014.

[66] Moreson H Kaplan and Alvan R Feinstein. The importance of classifying initial comorbidity in evaluating the outcome of diabetes mellitus. *Journal of chronic diseases*, 27(7):387–404, 1974.

[67] Jeffrey N Katz, Lily C Chang, Oliver Sangha, Anne H Fossel, and David W Bates. Can comorbidity be measured by questionnaire rather than medical record review? *Medical care*, 34(1):73–84, 1996.

[68] Jay S Kaufman, Richard S Cooper, and Daniel L McGee. Socioeconomic status and health in blacks and whites: the problem of residual confounding and the resiliency of race. *Epidemiology*, pages 621–628, 1997.

[69] Grace Kena. National center for education statistics, 2008.

[70] Zhao S Nelson CB Hughes M Eshleman S Wittchen HU Kessler RC, McGonagle KA and Kendler KS. Lifetime and 12-month prevalence of dsm-iii-r psychiatric disorders in the united states: Results from national comorbidity survey. *Arch Gen Psychiatry*, 51(1):8–19, 1994.

[71] Raynard S Kington and Herbert W Nickens. Racial and ethnic differences in health: recent trends, current patterns, future directions. *America becoming: Racial trends and their consequences*, 2:253–310, 2001.

[72] Inge Kirchberger, Christa Meisinger, Margit Heier, Anja-Kerstin Zimmermann, Barbara Thorand, Christine S Autenrieth, Annette Peters, Karl-Heinz Ladwig, and Angela Döring. Patterns of multimorbidity in the aged population. results from the kora-age study. *PloS one*, 7(1):e30556, 2012.

[73] Rex Kline, Y Petscher, and C Schatschneider. Exploratory and confirmatory factor analysis. *Applied quantitative analysis in the social sciences*, pages 171–207, 2013.

[74] Eugene F Krause. *Taxicab geometry: An adventure in non-Euclidean geometry*. Courier Corporation, 2012.

[75] Lewis H Kuller. Relationship between acute and chronic disease epidemiology. *The Yale journal of biology and medicine*, 60(4):363, 1987.

[76] Martijn Lappenschaar. *New network models for the analysis of disease interaction with applications in multimorbidity*. PhD thesis, Radbound University, 2014.

[77] Himmelfarb Health Sciences Library. Study Design 101 case-control study. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htmhttps://himmelfarb.gwu.edu/tutorials/studydesign101/casecontrols.html. Accessed: 2011-10-01.

[78] Bernard S Linn, MARGARET W LINN, and LEE Gurel. Cumulative illness rating scale. *Journal of the American Geriatrics Society*, 16(5):622–626, 1968.

[79] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. In *International Workshop on Algorithms and Computation*, pages 274–285. Springer, 2009.

[80] Alessandra Marengoni, Debora Rizzuto, Hui-Xin Wang, Bengt Winblad, and Laura Fratiglioni. Patterns of chronic multimorbidity in the elderly population. *Journal of the American Geriatrics Society*, 57(2):225–230, 2009.

[81] Scott Wallask Margaret Rouse. Icd-10(international classification of diseases, tenth revision), 2015.

[82] Colin Mathers, Theo Vos, and Chris Stevenson. *The burden of disease and injury in Australia*. Australian Institute of Health and Welfare, 1999.

[83] Daniel McGee, Richard Cooper, Youlian Liao, and Ramon Durazo-Arvizu. Patterns of comorbidity and mortality risk in blacks and whites. *Annals of epidemiology*, 6(5):381–385, 1996.

[84] Gary McLean, Jane Gunn, Sally Wyke, Bruce Guthrie, Graham CM Watt, David N Blane, and Stewart W Mercer. The influence of socioeconomic deprivation on multimorbidity at different ages: a cross-sectional study. *Br J Gen Pract*, 64(624):e440–e447, 2014.

[85] Merriam-Webster. *Merriam-Webster's collegiate dictionary*. Merriam-Webster, 2004.

[86] Jane E Miller. The effects of race/ethnicity and income on early childhood asthma prevalence and health care use. *American Journal of Public Health*, 90(3):428, 2000.

[87] Michael C Neale and Kenneth S Kendler. Models of comorbidity for multifactorial disorders. *American journal of human genetics*, 57(4):935, 1995.

[88] Anne B Newman, Robert M Boudreau, Barbara L Naydeck, Linda F Fried, and Tamara B Harris. A physiologic index of comorbidity: relationship to mortality and disability. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 63(6):603–609, 2008.

[89] Shu Kay Ng, Libby Holden, and Jing Sun. Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics. *Statistics in medicine*, 31(27):3393–3405, 2012.

[90] Kathryn Nicholson, Amanda L Terry, Martin Fortin, Tyler Williamson, Michael Bauer, and Amardeep Thind. Examining the prevalence and patterns of multimorbidity in canadian primary healthcare: a methodologic protocol using a national electronic medical record database. *Journal of Comorbidity*, 5(1):150–161, 2015.

[91] Government of Canada. Forward Sortation Area–Definition. https://www.ic.gc.ca/eic/site/bsf-osb.nsf/eng/br03396.html, 2016.

[92] Julie O'Halloran, Graeme C Miller, and Helena Britt. Defining chronic conditions for primary care with icpc-2. *Family Practice*, 21(4):381–386, 2004.

[93] World Health Organization. International classification of disease (icd). http://www.who.int/classifications/icd/en/. Accessed: 2011-06-29.

[94] Ellen C Perrin, Paul Newacheck, I Barry Pless, Dennis Drotar, Steven L Gortmaker, John Leventhal, James M Perrin, Ruth EK Stein, Deborah K Walker, and Michael Weitzman. Issues involved in the definition and classification of chronic health conditions. *Pediatrics*, 91(4):787–793, 1993.

[95] Alexandra Prados-Torres, Beatriz Poblador-Plou, Amaia Calderón-Larrañaga, Luis Andrés Gimeno-Feliu, Francisca González-Rubio, Antonio Poncel-Falcó, Antoni Sicras-Mainar, and José Tomás Alcalá-Nalvaiz. Multimorbidity patterns in primary care: interactions among chronic diseases using factor analysis. *PloS one*, 7(2):e32190, 2012.

[96] Henri Ralambondrainy. A conceptual version of the k-means algorithm. *Pattern Recognition Letters*, 16(11):1147–1157, 1995.

[97] Soo Hyun Rhee, John K Hewitt, Jeffrey M Lessem, Michael C Stallings, Robin P Corley, and Michael C Neale. The validity of the neale and kendler model-fitting approach in examining the etiology of comorbidity. *Behavior Genetics*, 34(3):251–265, 2004.

[98] Richard G Rogers, Robert A Hummer, and Charles B Nam. *Living and dying in the USA: Behavioral, health, and social differentials of adult mortality*. Elsevier, 1999.

[99] Monika M Safford, Jeroan J Allison, and Catarina I Kiefe. Patient complexity: more than comorbidity. the vector model of complexity. *Journal of General Internal Medicine*, 22(3):382–390, 2007.

[100] Ingmar Schäfer, Heike Hansen, Gerhard Schön, Susanne Höfels, Attila Altiner, Anne Dahlhaus, Jochen Gensichen, Steffi Riedel-Heller, Siegfried Weyerer, Wolfgang A Blank, et al. The influence of age, gender and socio-economic status on multimorbidity patterns in primary care. first results from the multicare cohort study. *BMC health services research*, 12(1):1, 2012.

[101] Benjamin Schüz, Susanne Wurm, Lisa M Warner, and Clemens Tesch-Römer. Health and subjective well-being in later adulthood: Different health states-different needs? *Applied Psychology: Health and Well-Being*, 1(1):23–45, 2009.

[102] Nova Scotia. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*, 19:227, 2010.

[103] Teresa E Seeman and Eileen Crimmins. Social environment effects on health and aging. *Annals of the New York Academy of Sciences*, 954(1):88–117, 2001.

[104] J Smith. Chronic diseases related to aging and health promotion and disease prevention. *Report of the Standing Committee on Health*, 75, 2012.

[105] James P Smith. Healthy bodies and thick wallets: the dual relation between health and economic status. *The journal of economic perspectives: a journal of the American Economic Association*, 13(2):144, 1999.

[106] Susan M Smith and Tom O'Dowd. Chronic diseases: what happens when they come in multiples? *Br J Gen Pract*, 57(537):268–270, 2007.

[107] Jaapjan D Snoep, Alfredo Morabia, Sonia Hernández-Díaz, Miguel A Hernán, and Jan P Vandenbroucke. Commentary: A structural approach to berkson's fallacy and a guide to a history of opinions about it. *International journal of epidemiology*, page dyu026, 2014.

[108] Hassan Soubhi, Martin Fortin, and Catherine Hudon. Perceived conflict in the couple and chronic illness management: preliminary analyses from the quebec health survey. *BMC family practice*, 7(1):1, 2006.

[109] Charles Spearman. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

[110] Ruth EK Stein, Laurie J Bauman, Lauren E Westbrook, Susan M Coupey, and Henry T Ireys. Framework for identifying children who have chronic conditions: the case for a new definition. *The Journal of pediatrics*, 122(3):342–347, 1993.

[111] Ruth EK Stein and Dorothy Jones Jessop. What diagnosis does not tell: the case for a noncategorical approach to chronic illness in childhood. *Social science & medicine*, 29(6):769–778, 1989.

[112] Moira Stewart. CPCSSN(Canadian Primary Care Sentinel Surveillance Network) Project. `http://cpcssn.ca/regional-networks/delphi-deliver-primary-healthcare-information-project/`, 2016. [Canadian Primary Care Sentinel Surveillance Network;].

[113] Moira Stewart. DELPHI(Deliver Primary Healthcare Information) Project. `http://cpcssn.ca/research-resources/cpcssn-data-for-research/`, 2016. [Canadian Primary Care Sentinel Surveillance Network;].

[114] Moira Stewart, Amardeep Thind, Amanda L Terry, Vijaya Chevendra, and J Neil Marshall. Implementing and maintaining a researchable database from electronic medical records: a perspective from an academic family medicine department. *Healthcare Policy*, 5(2):26, 2009.

[115] Stephanie R Susser, Jane McCusker, and Eric Belzile. Comorbidity information in older patients at an emergency visit: self-report vs. administrative data had poor agreement but similar predictive validity. *Journal of clinical epidemiology*, 61(5):511–515, 2008.

[116] Leigh Tooth, Richard Hockey, Julie Byles, and Annette Dobson. Weighted multimorbidity indexes predicted mortality, health service use, and health-related quality of life in older women. *Journal of clinical epidemiology*, 61(2):151–159, 2008.

[117] Derrick Vail. Dorland's illustrated medical dictionary, 1957.

[118] Jose M Valderas, Barbara Starfield, Bonnie Sibbald, Chris Salisbury, and Martin Roland. Defining comorbidity: implications for understanding health and health services. *The Annals of Family Medicine*, 7(4):357–363, 2009.

[119] Marjan van den Akker, Frank Buntinx, and J André Knottnerus. Comorbidity or multimorbidity: what's in a name? a review of literature. *The European Journal of General Practice*, 2(2):65–70, 1996.

[120] Marjan Van den Akker, Frank Buntinx, Job FM Metsemakers, Sjef Roos, and J André Knottnerus. Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. *Journal of clinical epidemiology*, 51(5):367–375, 1998.

[121] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[122] Michael Von Korff, Edward H Wagner, and Kathleen Saunders. A chronic disease score from automated pharmacy data. *Journal of clinical epidemiology*, 45(2):197–203, 1992.

[123] David F Warner and Mark D Hayward. Early-life origins of the race gap in men's mortality. *Journal of health and social behavior*, 47(3):209–226, 2006.

[124] Verne Wheelwright. *It's Your Future... Make It a Good One!* Personal Futures Network, 2010.

[125] David R Williams and Chiquita Collins. Us socioeconomic and racial differences in health: patterns and explanations. *Annual review of sociology*, pages 349–386, 1995.

[126] Jennifer L Wolff, Barbara Starfield, and Gerard Anderson. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. *Archives of internal medicine*, 162(20):2269–2276, 2002.

Appendices

Appendix A

Additional Material and Results

The extra figures and tables belonging to the analysis in this research are provided here. The tables and figures included in this appendix represent the distribution of chronic diseases over socio-demographic attributes with different aspects. Figures A.1 A.2 refer to patients diagnosed with one of more chronic conditions, while Figures A.3 to A.13 represent the characteristics of patients with two chronic conditions.

Theorem 1. *The function d(S,q) is minimised if and only if* $f_r(A_j = q_j|S) \geq f_r(A_j = c_{k,j}|S)$ *for* $q_j \neq c_{k,j}$ *for all* $j = 1, ..., m$, *where* $f_r(A_j = c_{k,j}|S) = \frac{n_{c_{k,j}}}{n}$ *stands for the relative frequency of* $k-th$ *category* $c_{k,j}$ *in attribute* A_j *from set* S.

Proof. Assume that n is the total number of instances and m is the number of attributes for

each instance in set S. According to , we have:

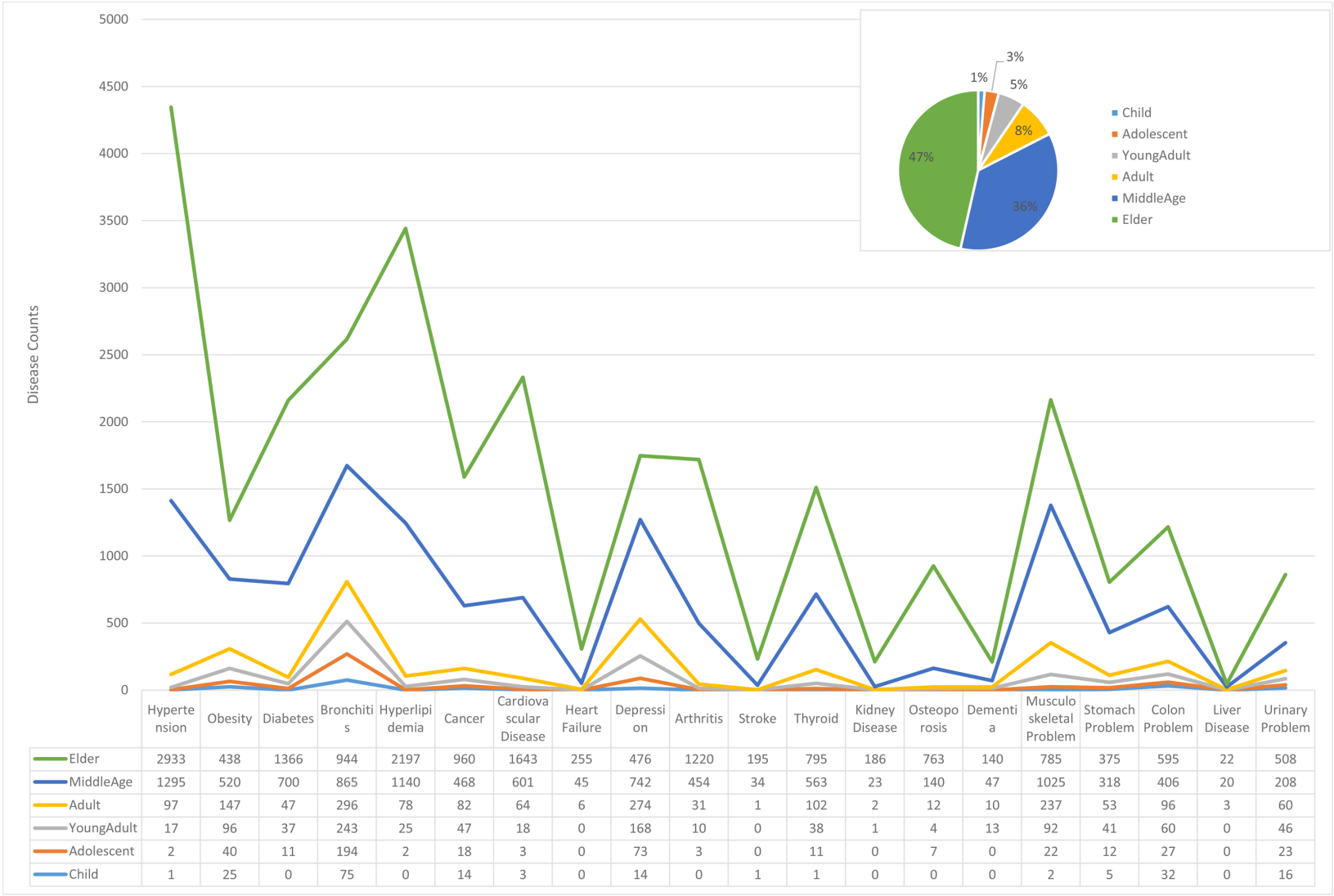
$$
\begin{aligned}
d(S,q) &= \sum_{i=1}^{n} \eta(s_i, q) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \eta(s_{i,j}, q_j) \\
&= \sum_{j=1}^{m} \sum_{i=1}^{n} \eta(s_{i,j}, q_j) \\
&= \sum_{j=1}^{m} [\eta(s_{1,j}, q_j) + \eta(s_{2,j}, q_j) + ... + \eta(s_{n,j}, q_j)] \\
&= \sum_{j=1}^{m} f_r(A_j = q_j|S) \\
&= \sum_{j=1}^{m} (n - n_{q_j}) \\
&= \sum_{j=1}^{m} n(1 - \frac{n_{q_j}}{n}) \\
&= \sum_{j=1}^{m} n(1 - f_r(A_j = q_j|S))
\end{aligned}
$$

As we know $\frac{n_{q_j}}{n}$ is a positive number lying between zero and one, as a results $n(1 - f_r(A_j = q_j)|S) \geq 0$ for $1 \leq j \leq m$. Therefore $d(S,q)$ is minimized if and only if $n(1 - f_r(A_j = q_j)|S)$ is minimal. For achieving this, $f_r(A_j = q_j|S)$ must be maximal[1]. □

[1]Proof of Theorem 1 was derived from [58].

Disease Index	Chronic Disease Category	ICD-9 Codes
1	Hypertension	401–405, 401, 401.1, 401.9, 405, 405.01, 405.09, 405.1, 405.11, 405.19, 405.9, 405.91, 405.99
2	Obesity	278, 278.01
3	Diabetes	250, 250.01, 250.02, 250.03, 250.1, 250.11, 250.12, 250.13, 250.2, 250.21, 250.22, 250.23, 250.3, 250.31, 250.32, 250.33, 250.4, 250.41, 250.42, 250.43, 250.5, 250.51, 250.52, 250.53, 250.6, 250.61, 250.62, 250.63, 250.7, 250.71, 250.72, 250.73, 250.8, 250.81, 250.82, 250.83, 250.9, 250.91, 250.92, 250.93
4	Bronchitis	491, 491.1, 491.2, 491.21, 491.22, 491.8, 491.9, 492, 492.8, 493, 493.01, 493.02, 493.1, 493.11, 493.12, 493.2, 493.21, 493.22, 493.8, 493.81, 493.82, 493.9, 493.91, 493.92, 496
5	Hyperlipidemia	272, 272.1, 272.2, 272.3, 272.4
6	Cancer	140-239, 140-149, 150-159, 160-165, 170-176, 179-189, 190-199, 200-209
7	Cardiovascular Disease	412, 413, 413.1, 413.2, 440-449, 427, 427.3, 427.31, 417.32
8	Heart Failure	428, 394, 394.1, 394.2, 395, 395.1, 395.2, 395.9
9	Depression	296, 296.2, 296.21, 296.22, 296.23, 296.24, 296.25, 296.26, 296.3, 296.31, 296.32, 296.33, 296.34, 296.35, 296.36, 300, 300.01, 300.02, 300.09
10	Arthritis	714, 714.1, 714.2, 714.3, 715, 715.1, 715.2, 715.3, 715.8, 715.9
11	Stroke	434, 434.01, 434.1, 434.11, 433.9, 434.9, 434.91, 435, 435.1, 435.2, 435.3, 435.8, 435.9
12	Thyroid	240-246, 240, 241, 242, 243, 244, 245, 246
13	Kidney Disease	585, 585.1, 585.2, 585.3, 585.4, 585.5, 585.6, 585.9
14	Osteoporosis	733, 733.01, 733.02, 733.03, 733.09
15	Dementia	290, 290.1, 290.11, 290.12, 290.13, 290.2, 290.21, 290.3, 290.4, 294, 294.1, 294.2
16	Musculoskeletal Problem	723, 723.1, 724, 724.1, 724.2, 724.3, 724.4, 724.5, 725, 726, 726.1, 726.2, 726.3, 726.31, 726.32, 726.33, 726.39, 726.4, 726.5, 726.6, 726.61, 726.62, 726.63, 726.64, 726.65, 726.69, 726.7, 726.71, 726.72, 726.73, 726.79, 726.9, 726.91, 727, 727.01, 727.03, 727.04, 727.05, 727.06, 727.09, 727.2, 727.3, 729, 729.1, 729.2, 729.4, 729.5
17	Stomach Problem	530, 530.81, 531, 531.4, 531.41, 531.5, 531.51, 531.6, 531.61, 531.7, 531.71, 531.9, 531.91
18	Colon Problem	555, 555.1, 555.2, 555.9, 556, 556.4, 556.5, 556.6, 556.8, 556.9, 564, 564.1
19	Liver Disease	571, 571.1, 571.2, 571.3, 571.4, 571.41, 571.42, 571.49, 571.5, 571.6, 571.8, 571.9
20	Urinary Problem	593, 593.3, 593.4, 593.5, 593.7, 593.71, 593.72, 593.73, 593.8, 593.82, 593.89, 593.9, 595, 595.1, 595.2, 595.9, 597, 597.8, 597.81, 597.82, 600, 601, 601.1, 601.3, 601.8, 601.9, 602, 602.1, 602.2, 602.3, 602.8, 602.9

Table A.1: A list of 20 chronic disease categories and their corresponding ICD-9 (International Classification of Diseases 9th revision) disease codes [90].



	Hypertension	Obesity	Diabetes	Bronchitis	Hyperlipidemia	Cancer	Cardiovascular Disease	Heart Failure	Depression	Arthritis	Stroke	Thyroid	Kidney Disease	Osteoporosis	Dementia	Musculoskeletal Problem	Stomach Problem	Colon Problem	Liver Disease	Urinary Problem
Elder	2933	438	1366	944	2197	960	1643	255	476	1220	195	795	186	763	140	785	375	595	22	508
MiddleAge	1295	520	700	865	1140	468	601	45	742	454	34	563	23	140	47	1025	318	406	20	208
Adult	97	147	47	296	78	82	64	6	274	31	1	102	2	12	10	237	53	96	3	60
YoungAdult	17	96	37	243	25	47	18	0	168	10	0	38	1	4	13	92	41	60	0	46
Adolescent	2	40	11	194	2	18	3	0	73	3	0	11	0	7	0	22	12	27	0	23
Child	1	25	0	75	0	14	3	0	14	0	1	1	0	0	0	2	5	32	0	16

Figure A.1: Disease counts for all age groups and all diseases

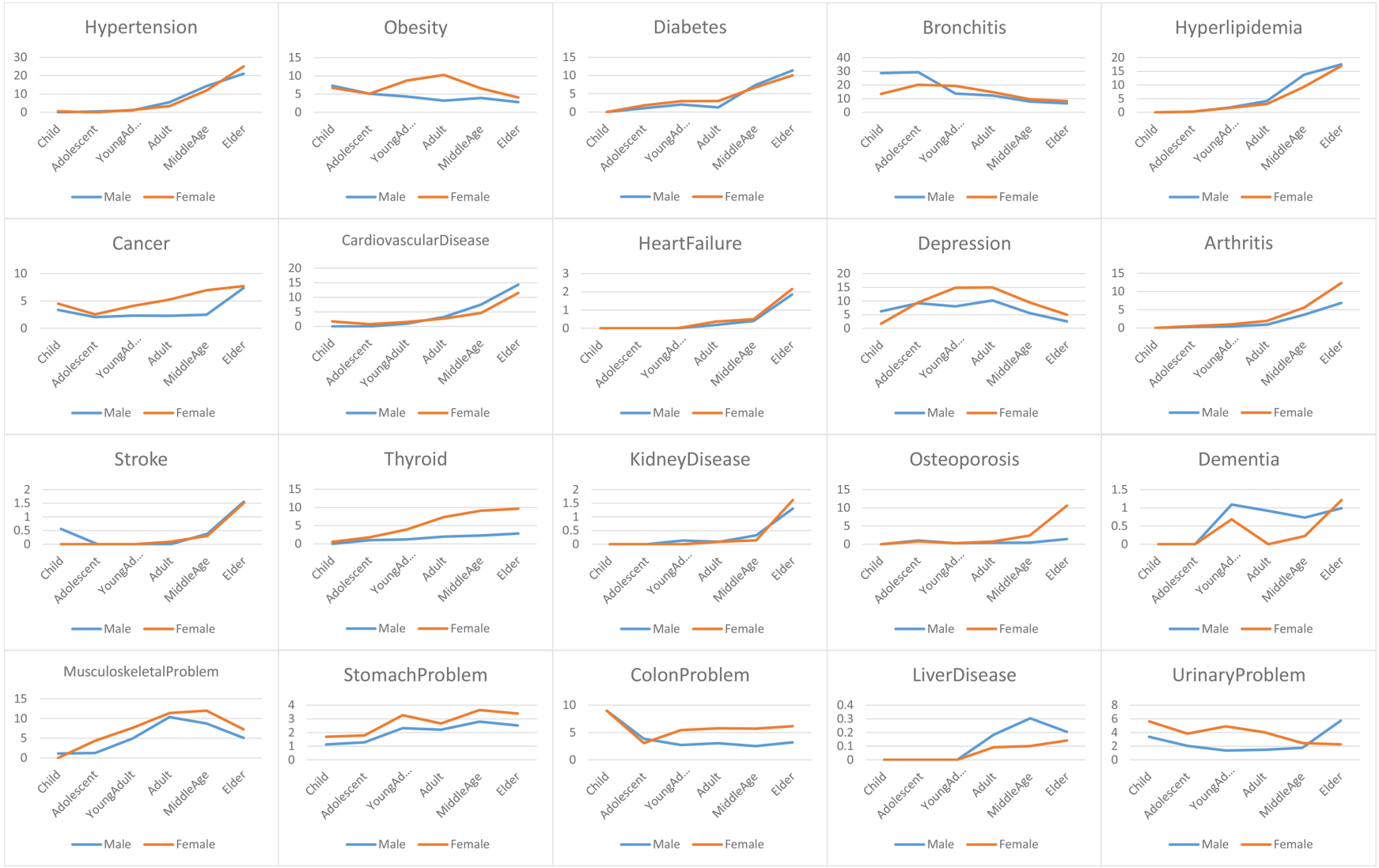


Figure A.2: Disease percentages for all age groups and individual diseases based on gender. According to the charts, some diseases have the same trend of growth among males and females (e.g. hypertension, diabetes, heart failure, kidney disease and stomach problem), whiles some other diseases follow different trends for males and females (i.e. obesity, cancer and urinary problem). Some disease tend to appear with a higher frequency among females (e.g. depression, thyroid, and stomach problem), while some tend to appear more among males (e.g. dementia and liver disease).

Disease		Odds ratios measuring pairwise disease association																		
		OS	DB	BC	HL	CC	CD	HF	DP	AT	SK	TD	KD	OP	DT	MP	SP	CP	LD	UP
1	HT	3.44	5.67	1.79	7.12	2.22	7.58	3.36	1.58	4.00	6.67	2.64	10.60	3.33	2.40	1.85	1.39	1.99	2.49	3.06
2	OS	-	3.73	1.65	3.25	1.55	3.44	2.14	2.13	2.37	1.26	1.79	2.51	1.68	0.87	1.91	0.57	2.09	1.39	2.28
3	DB	-	-	1.31	5.09	1.93	3.53	4.00	1.12	2.48	3.08	1.80	5.44	1.52	2.29	1.32	1.51	1.64	4.48	2.05
4	BC	-	-	-	1.76	1.28	2.83	2.15	2.66	2.02	1.99	1.68	1.96	1.62	1.37	3.13	1.20	1.86	1.11	1.64
5	HL	-	-	-	-	2.05	7.28	3.04	1.87	3.47	4.75	2.74	5.42	3.64	2.11	2.21	2.20	2.12	1.00	4.02
6	CC	-	-	-	-	-	2.18	2.76	1.18	2.35	2.77	1.82	3.86	3.44	1.80	1.71	1.50	2.45	1.09	2.49
7	CD	-	-	-	-	-	-	8.88	2.94	3.88	5.84	2.64	8.51	2.93	3.21	3.20	1.37	2.21	3.28	4.46
8	HF	-	-	-	-	-	-	-	1.07	3.99	4.30	2.25	10.93	3.46	2.49	1.24	1.51	1.82	6.02	2.17
9	DP	-	-	-	-	-	-	-	-	1.40	0.96	1.82	1.69	1.94	1.30	3.38	1.49	2.47	1.73	1.91
10	AT	-	-	-	-	-	-	-	-	-	2.55	2.51	4.45	3.99	2.56	2.75	2.26	2.87	1.37	2.48
11	SK	-	-	-	-	-	-	-	-	-	-	1.37	5.09	4.80	8.39	1.58	1.12	2.61	0	3.12
12	TD	-	-	-	-	-	-	-	-	-	-	-	3.48	3.72	1.52	2.04	1.78	2.16	0.75	1.50
13	KD	-	-	-	-	-	-	-	-	-	-	-	-	5.12	4.95	1.68	2.40	3.96	2.76	5.54
14	OP	-	-	-	-	-	-	-	-	-	-	-	-	-	2.88	2.43	1.92	3.74	1.26	1.97
15	DT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.90	0.92	1.02	2.79	1.62
16	MP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.44	3.03	0.51	2.25
17	SP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.58	1.45	1.42
18	CP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.45	2.65
19	LD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.66
20	UP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table A.2: Odds ratios computed for estimating the association between every disease pair. In order to avoid the selection bias, for calculating odds ratios, all existing patient records in DELPHI database have been included, regardless of the conditions they are diagnosed with; This also includes patients who have not been diagnosed with any chronic conditions. Abbreviations: HT=Hypertension, OS=Obesity, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, HF=Heart Failure, DP=Depression, AT=Arthritis, SK=Stroke TD=Thyroid, KD=Kidney Disease, OP=Osteoporosis, DT=Dementia, MP=Musculoskeletal Problem, SP=Stomach Problem, CP=Colon Problem, LD=Liver Disease, UP=Urinary Problem.

	Confidence intervals of odds ratios presented in Table A.2																		
	OS	DB	BC	HL	CC	CD	HF	DP	AT	SK	TD	KD	OP	DT	MP	SP	CP	LD	UP
HT	[3.1-3.9]	[5.2-6.2]	[1.6-2]	[6.6-7.7]	[2-2.5]	[6.9-8.3]	[2.7-4.2]	[1.4-1.8]	[3.6-4.4]	[5.1-8.7]	[2.4-2.9]	[7.9-14.2]	[2.9-3.8]	[1.8-3.2]	[1.7-2.0]	[1.2-1.6]	[1.7-2.3]	[1.3-4.6]	[2.7-3.5]
OS	-	[3.2-4.3]	[1.4-1.9]	[2.9-3.7]	[1.3-1.9]	[3-3.9]	[1.5-3.1]	[1.8-2.5]	[2-2.8]	[0.7-2.2]	[1.5-2.2]	[1.6-3.9]	[1.3-2.1]	[0.4-1.7]	[1.6-2.2]	[0.4-0.9]	[1.7-2.6]	[0.4-4.5]	[1.8-2.9]
DB	-	-	[1.1-1.5]	[4.6-5.6]	[1.7-2.2]	[3.2-3.9]	[3.1-5.2]	[0.9-1.3]	[2.2-2.8]	[2.2-4.2]	[1.5-2.1]	[4.1-7.3]	[1.2-1.9]	[1.6-3.3]	[1.1-1.5]	[1.2-1.9]	[1.4-2]	[2.4-8.6]	[1.7-2.5]
BC	-	-	-	[1.6-2]	[1.1-1.5]	[2.5-3.2]	[1.6-2.9]	[2.4-3.0]	[1.8-2.3]	[1.4-2.8]	[1.5-1.9]	[1.4-2.8]	[1.3-1.9]	[0.9-2.0]	[2.8-3.5]	[1-1.5]	[1.6-2.2]	[0.4-2.8]	[1.4-2]
HL	-	-	-	-	[1.8-2.3]	[6.6-8]	[2.4-3.9]	[1.7-2.1]	[3.1-3.9]	[3.6-6.2]	[2.4-3.1]	[4.1-7.1]	[3.2-4.2]	[1.5-2.9]	[2-2.5]	[1.9-2.6]	[1.8-2.4]	[0.4-2.5]	[3.5-4.6]
CC	-	-	-	-	-	[1.9-2.5]	[2.0-3.8]	[1-1.4]	[2-2.7]	[1.9-4]	[1.5-2.2]	[2.7-5.4]	[2.9-4.1]	[1.2-2.8]	[1.5-2]	[1.2-1.9]	[2.1-2.9]	[0.3-3.5]	[2-3.1]
CD	-	-	-	-	-	-	[7.1-11.2]	[2.6-3.3]	[3.4-4.4]	[4.4-7.7]	[2.3-3]	[6.5-11.2]	[2.5-3.5]	[2.3-4.4]	[2.9-3.6]	[1.1-1.7]	[1.9-2.6]	[1.7-6.5]	[3.8-5.2]
HF	-	-	-	-	-	-	-	[0.7-1.7]	[3-5.3]	[2.3-8]	[1.6-3.2]	[7-17.1]	[2.4-5]	[1.1-5.7]	[0.9-1.8]	[0.9-2.6]	[1.2-2.8]	[1.9-19.5]	[1.4-3.4]
DP	-	-	-	-	-	-	-	-	[1.2-1.7]	[0.6-1.6]	[1.5-2.2]	[1.1-2.6]	[1.6-2.4]	[0.8-2.1]	[3-3.8]	[1.2-1.9]	[2.1-2.9]	[0.7-4.4]	[1.5-2.4]
AT	-	-	-	-	-	-	-	-	-	[1.8-3.7]	[2.2-2.9]	[3.2-6.1]	[3.4-4.7]	[1.8-3.7]	[2.4-3.1]	[1.8-2.8]	[2.4-3.4]	[0.5-3.8]	[2-3]
SK	-	-	-	-	-	-	-	-	-	-	[0.8-2.2]	[2.6-10]	[3.3-6.9]	[4.8-14.7]	[1.1-2.3]	[0.6-2.3]	[1.7-3.9]	[0-0]	[2-4.9]
TD	-	-	-	-	-	-	-	-	-	-	-	[2.4-5]	[3.1-4.4]	[0.9-2.5]	[1.8-2.4]	[1.4-2.3]	[1.8-2.6]	[0.2-3.1]	[1.2-1.9]
KD	-	-	-	-	-	-	-	-	-	-	-	-	[3.5-7.5]	[2.4-10.2]	[1.1-2.5]	[1.4-4.1]	[2.7-5.7]	[0.4-20.1]	[3.8-8.1]
OP	-	-	-	-	-	-	-	-	-	-	-	-	-	[1.8-4.6]	[2-2.9]	[1.4-2.6]	[3.1-4.5]	[0.3-5.2]	[1.5-2.6]
DT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	[0.5-1.5]	[0.4-2]	[0.5-1.9]	[0.4-20.3]	[0.9-3]
MP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	[1.2-1.8]	[2.6-3.5]	[0.1-2.1]	[1.9-2.7]
SP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	[1.2-2.1]	[0.4-6]	[1-2]
CP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	[0.5-4.7]	[2.1-3.3]
LD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	[0.1-4.8]
UP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table A.3: 95% confidence intervals (CI) for all disease pairs to estimate the precision of the odds ratios. A large CI indicates a low level of precision of the odds ratio, whereas a small CI indicates a higher precision of the odds ratio. The 95% CI can be used to present statistical significance if it does not overlap OR=1. Abbreviations: HT=Hypertension, OS=Obesity, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, HF=Heart Failure, DP=Depression, AT=Arthritis, SK=Stroke TD=Thyroid, KD=Kidney Disease, OP=Osteoporosis, DT=Dementia, MP=Musculoskeletal Problem, SP=Stomach Problem, CP=Colon Problem, LD=Liver Disease, UP=Urinary Problem. The green highlights on some of the cells show a relatively higher association for corresponding disease triples to co-occur.

Disease		Multimorbidity Coefficient measuring pairwise disease association																		
		OS	DB	BC	HL	CC	CD	HF	DP	AT	SK	TD	KD	OP	DT	MP	SP	CP	LD	UP
1	HT	2.32	2.86	1.50	2.92	1.76	3.19	2.38	1.40	2.47	3.38	1.97	4.02	2.31	1.93	1.55	1.29	1.66	1.99	2.20
2	OS	-	2.80	1.51	2.37	1.47	2.62	2.01	1.90	2.06	1.24	1.65	2.32	1.59	0.88	1.71	0.59	1.90	1.36	2.07
3	DB	-	-	1.25	2.90	1.71	2.54	3.13	1.10	2.06	2.59	1.63	3.90	1.44	2.06	1.26	1.43	1.52	3.46	1.84
4	BC	-	-	-	1.52	1.23	2.16	1.91	2.12	1.75	1.80	1.53	1.77	1.50	1.32	2.32	1.17	1.66	1.10	1.52
5	HL	-	-	-	-	1.72	3.40	2.36	1.61	2.41	3.13	2.10	3.37	2.58	1.83	1.79	1.85	1.78	1.00	2.74
6	CC	-	-	-	-	-	1.87	2.45	1.16	2.02	2.47	1.66	3.23	2.82	1.71	1.56	1.44	2.13	1.09	2.20
7	CD	-	-	-	-	-	-	5.05	2.30	2.77	3.99	2.16	4.98	2.39	2.64	2.38	1.31	1.91	2.72	3.20
8	HF	-	-	-	-	-	-	-	1.07	3.25	4.03	2.07	9.18	3.11	2.42	1.21	1.47	1.74	5.64	2.06
9	DP	-	-	-	-	-	-	-	-	1.33	0.96	1.66	1.61	1.78	1.27	2.55	1.42	2.13	1.65	1.76
10	AT	-	-	-	-	-	-	-	-	-	2.28	2.14	3.56	3.11	2.30	2.22	2.03	2.39	1.34	2.17
11	SK	-	-	-	-	-	-	-	-	-	-	1.34	4.76	4.12	7.48	1.50	1.12	2.40	0	2.87
12	TD	-	-	-	-	-	-	-	-	-	-	-	2.99	3.00	1.47	1.79	1.66	1.93	0.76	1.43
13	KD	-	-	-	-	-	-	-	-	-	-	-	-	4.35	4.66	1.58	2.28	3.42	2.72	4.68
14	OP	-	-	-	-	-	-	-	-	-	-	-	-	-	2.66	2.10	1.81	3.08	1.24	1.85
15	DT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.91	0.92	1.01	2.74	1.58
16	MP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.37	2.43	0.53	1.98
17	SP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.51	1.43	1.38
18	CP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.42	2.35
19	LD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.67
20	UP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table A.4: Multimorbidity coefficients (MC) measure the association of all disease pairs. For avoiding effects of selection bias, the population used for performing the computations of MC contains all patient records in DELPHI database, regardless of the number and type of diseases they are diagnosed with; this also contains patients with no chronic conditions. Abbreviations: HT=Hypertension, OS=Obesity, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, HF=Heart Failure, DP=Depression, AT=Arthritis, SK=Stroke TD=Thyroid, KD=Kidney Disease, OP=Osteoporosis, DT=Dementia, MP=Musculoskeletal Problem, SP=Stomach Problem, CP=Colon Problem, LD=Liver Disease, UP=Urinary Problem.

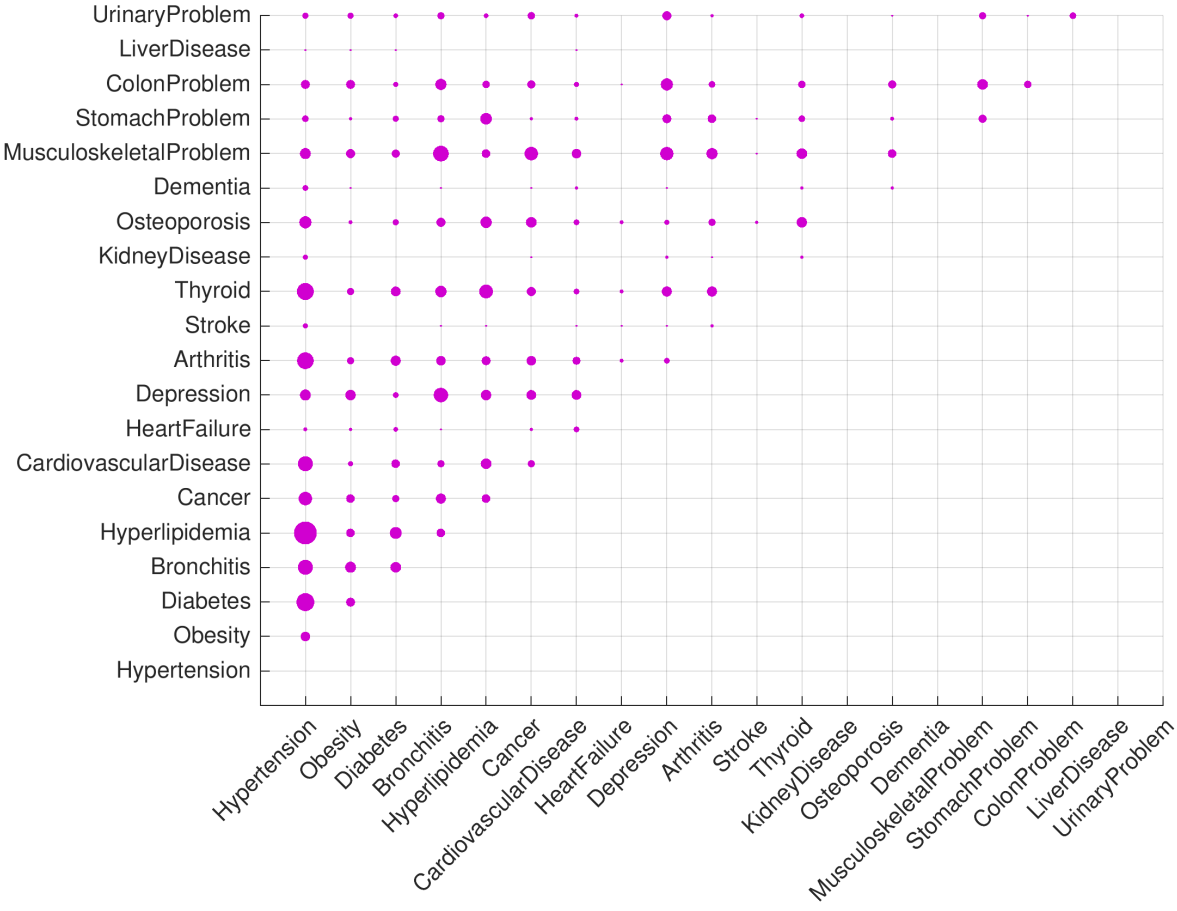


Figure A.3: Disease correlations of two coexisting diseases for female patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among females population.

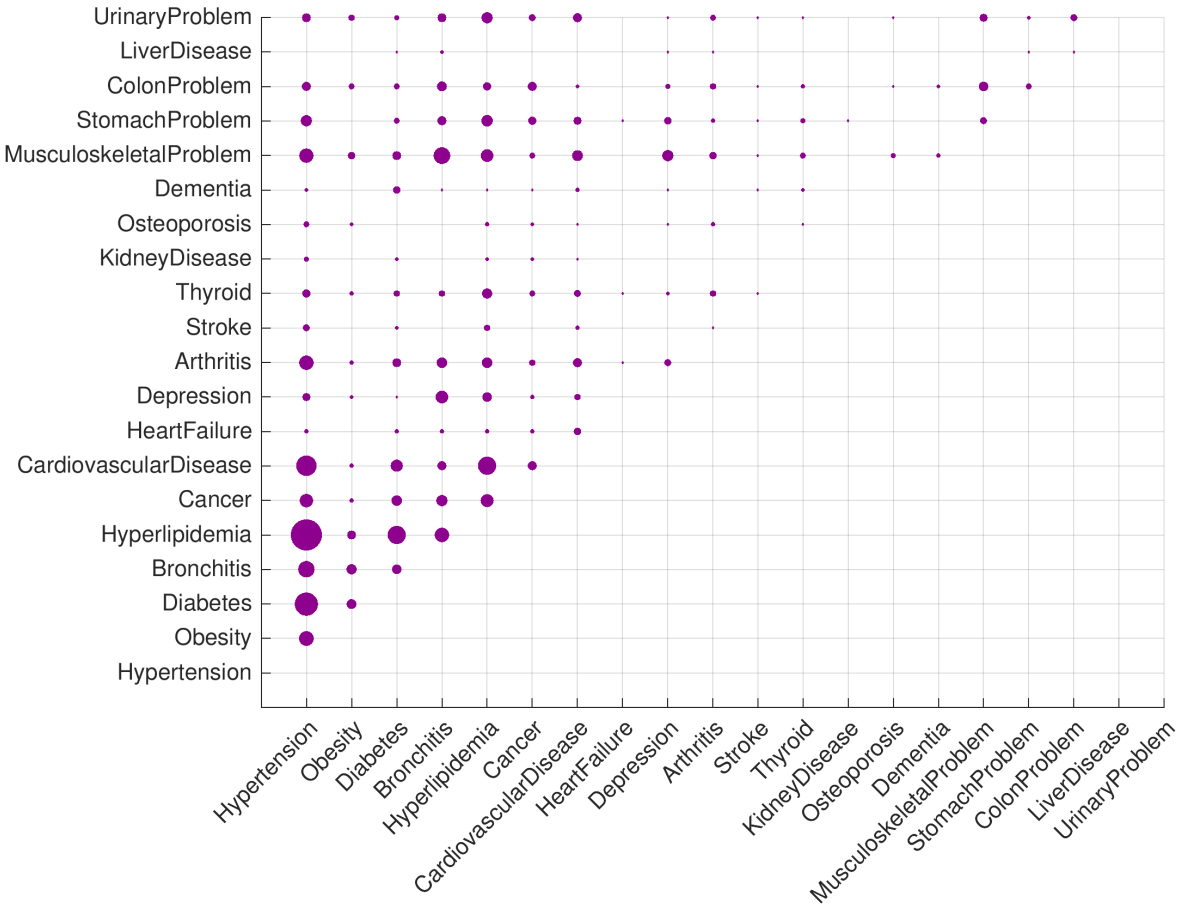


Figure A.4: Disease correlations of two coexisting diseases for male patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among males population.

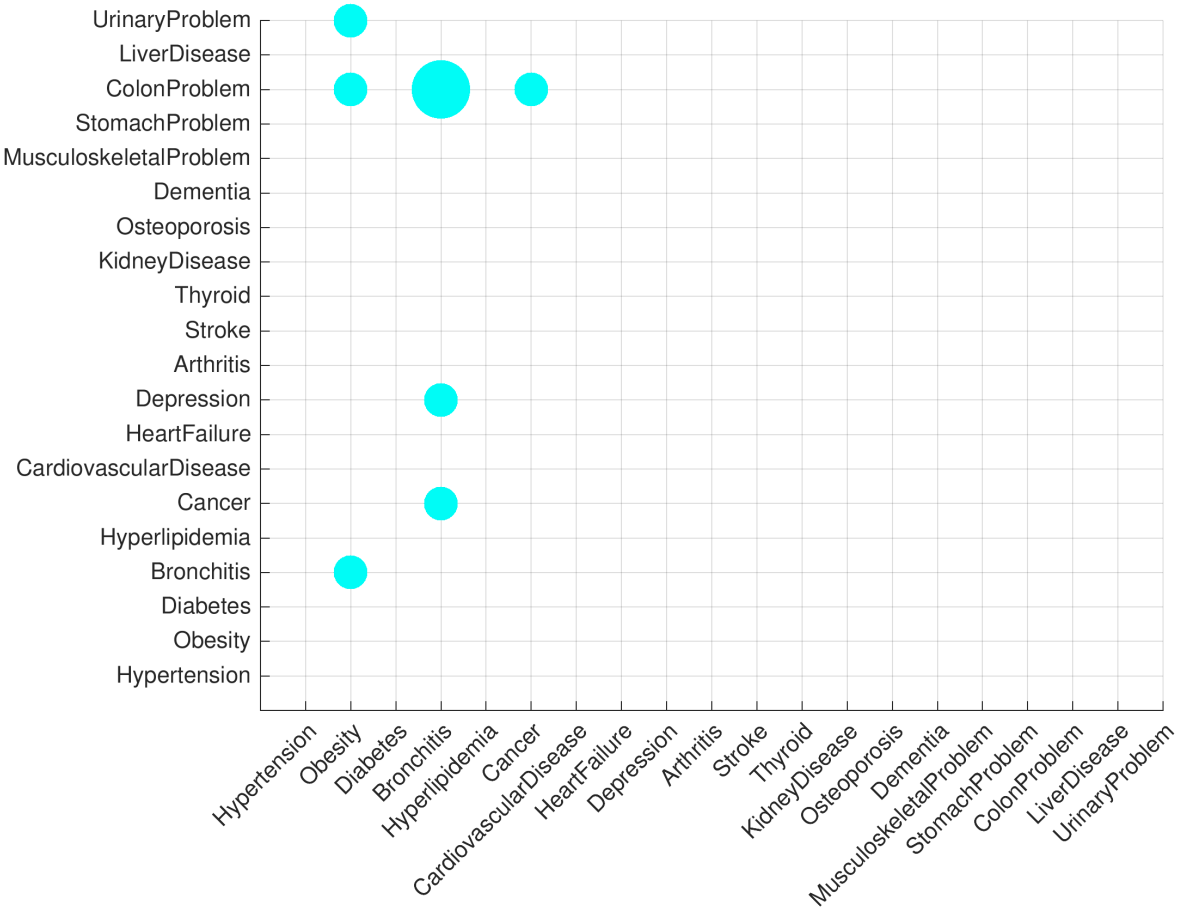


Figure A.5: Disease correlations of two coexisting diseases among children. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among children's population. There are nine patients in the database under *child* age group who are diagnosed with two diseases. Obesity, bronchitis and colon problem form the majority of disease combinations among children. The disease combination of bronchitis and colon problem is most commonly occurring disease pair affecting 3 children. The family income level of all patients in this age group belongs to high or highest income quintiles.

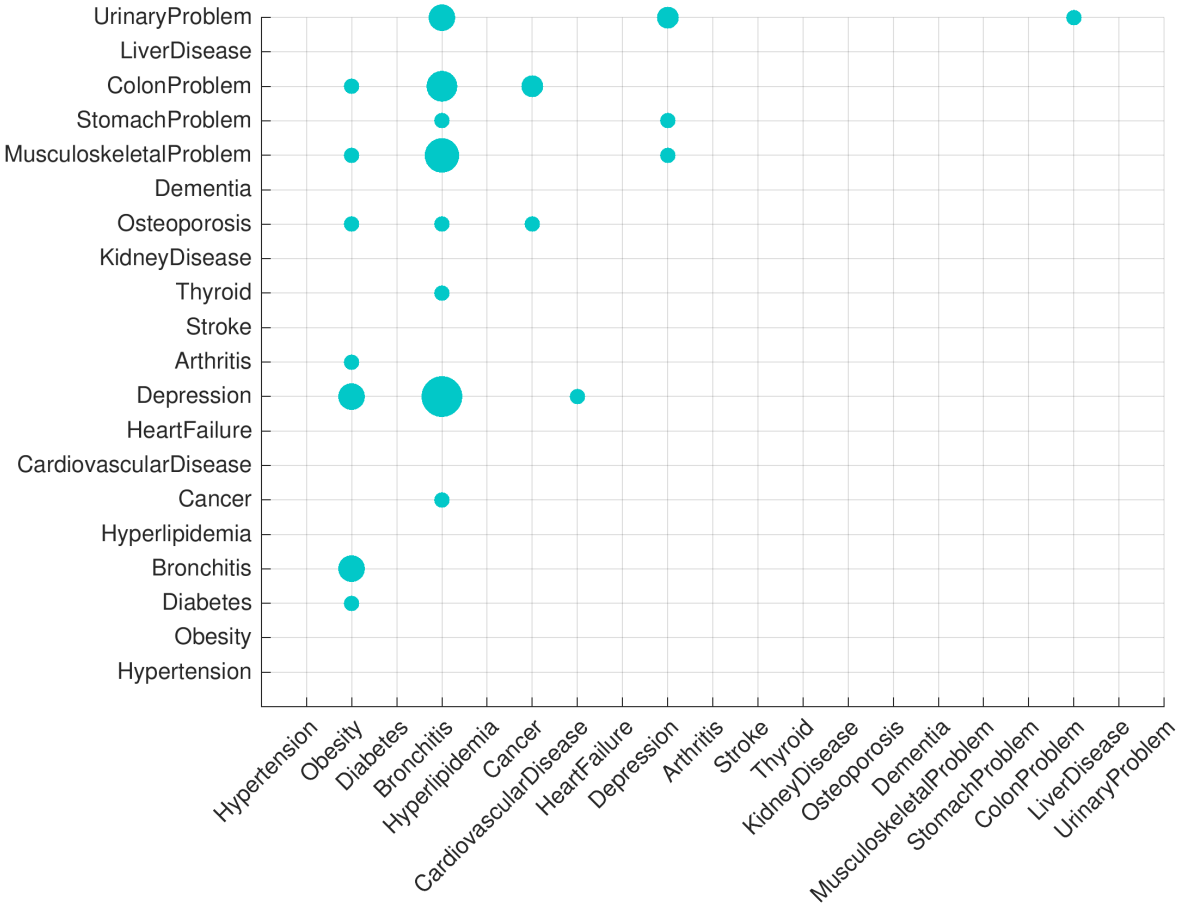


Figure A.6: Disease correlations of two coexisting diseases for adolescent. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among adolescent. Bronchitis is the dominant disease which coexists with many other chronic conditions among adolescent, and it correlates with a wider variety of diseases in adolescent compared to children. The family income level of all patients in this age group belongs to high or highest income quintiles.

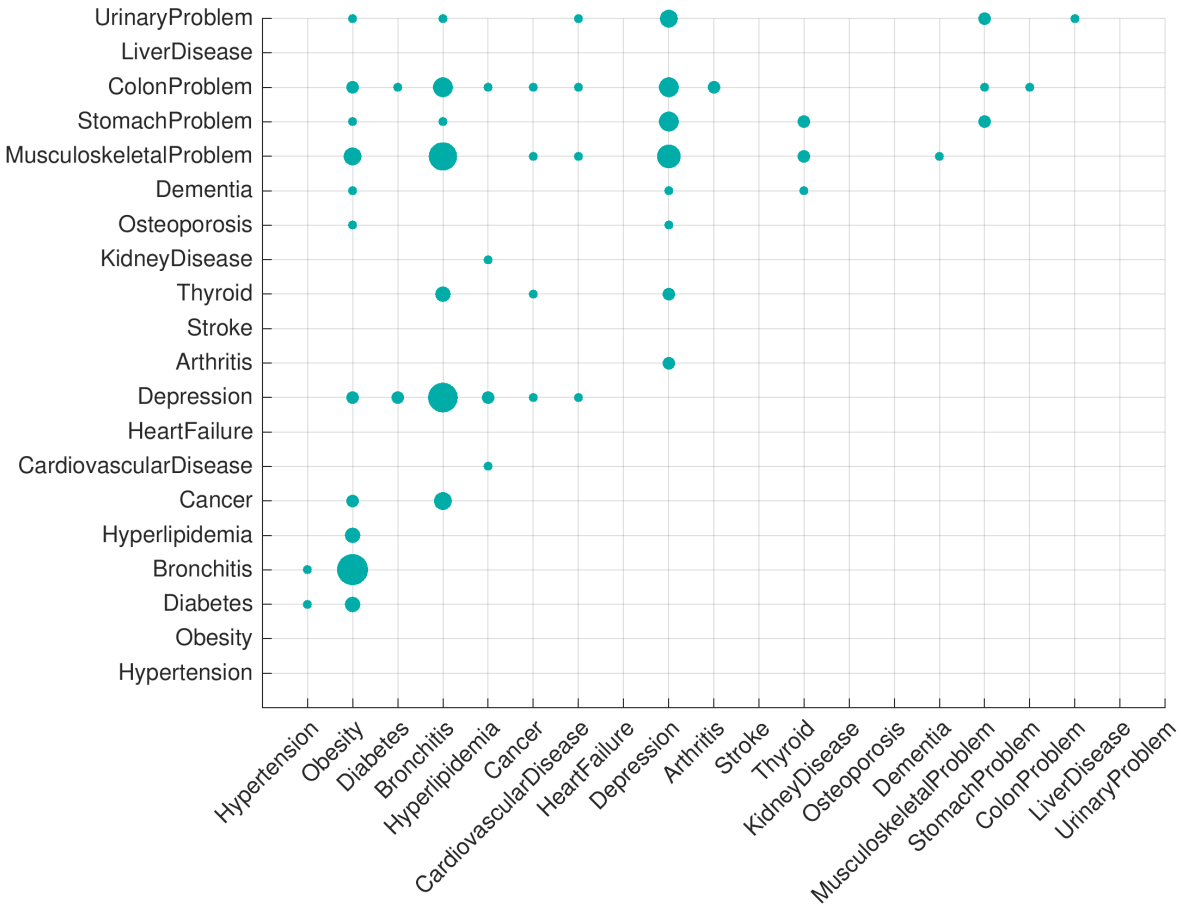


Figure A.7: Disease correlations of two coexisting diseases for young adults. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among young adults. Coexistence of hypertension with other chronic conditions starts to appear among patients from this age group. Bronchitis is still one of the dominant diseases mostly co-occurring with depression and musculoskeletal problem among young adults. Depression and obesity become the two other dominant diseases coexisting with a variety of chronic conditions. Correlation of thyroid problem with some chronic diseases starts to appear from this age group.

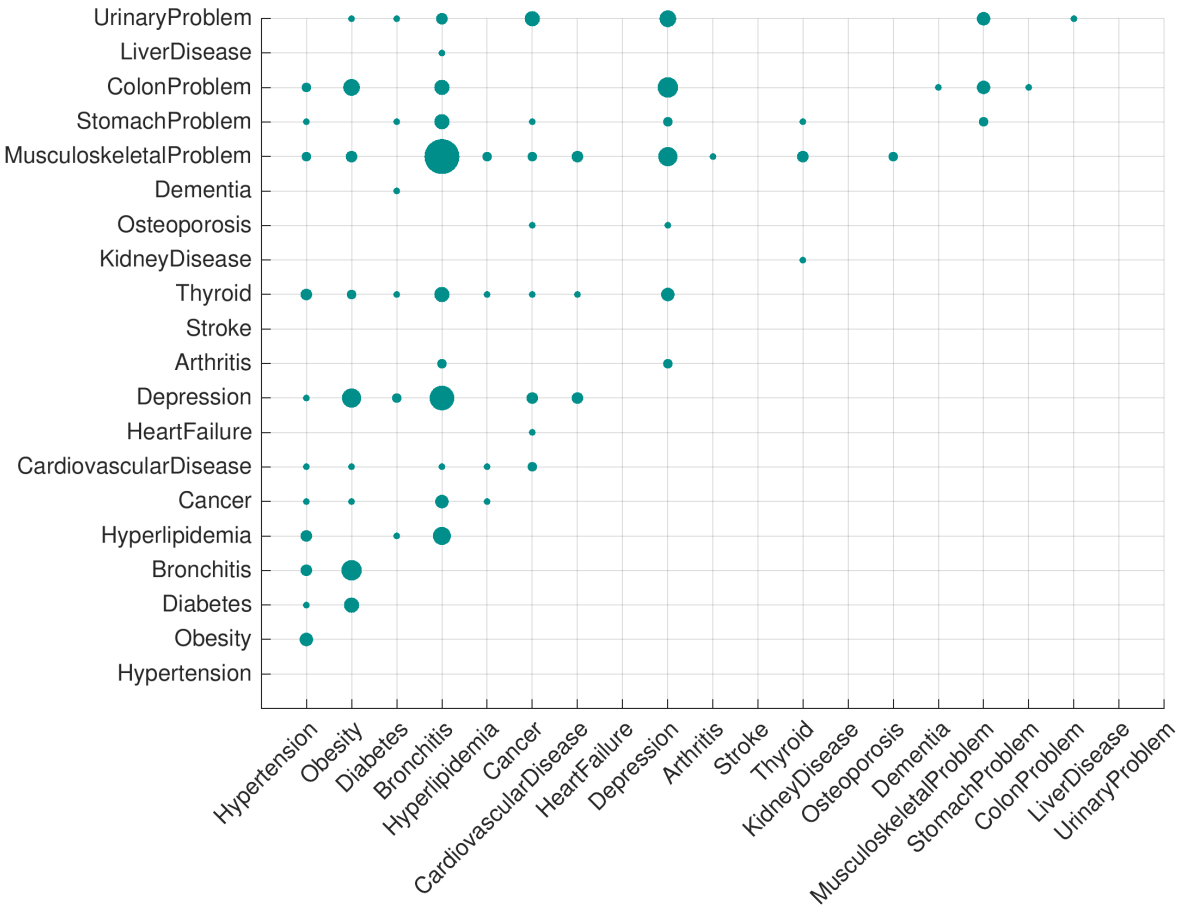


Figure A.8: Disease correlations of two coexisting diseases for adult patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among adults population. Bronchitis is still on the list of dominant diseases for this age group, coexisting with eleven other chronic conditions. Musculoskeletal problem becomes another dominant disease coexisting with ten chronic conditions. Hypertension correlates with many other chronic conditions in this age group and depression becomes less dominant compared to young adults age group.

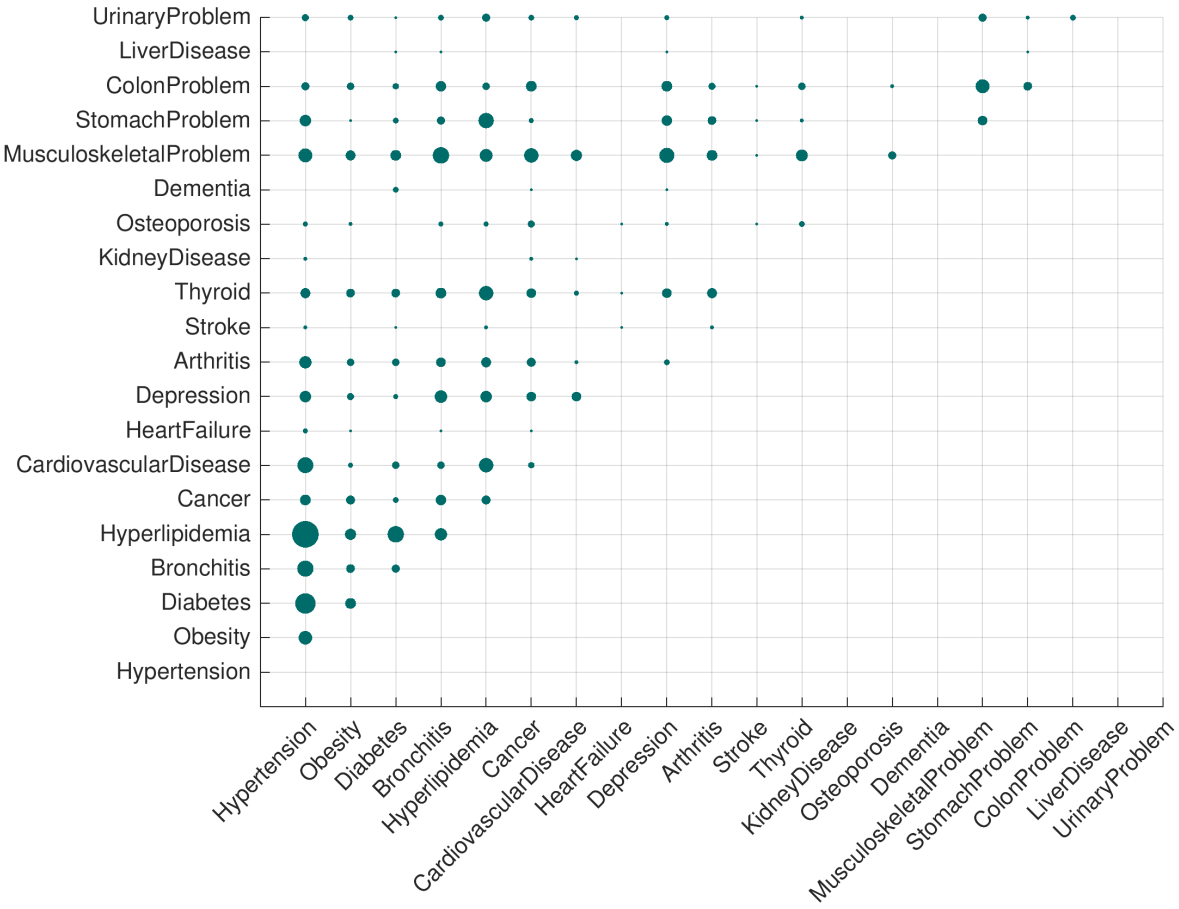


Figure A.9: Disease correlations of two coexisting diseases for middle age patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among middle aged patients. The variety of disease correlations become relatively larger compared to adults. Hypertension and musculoskeletal problem become two dominant diseases, coexisting with seventeen and eleven other chronic conditions respectively. Prevalence of coexistence of cancer with other conditions is greater compared to adults.

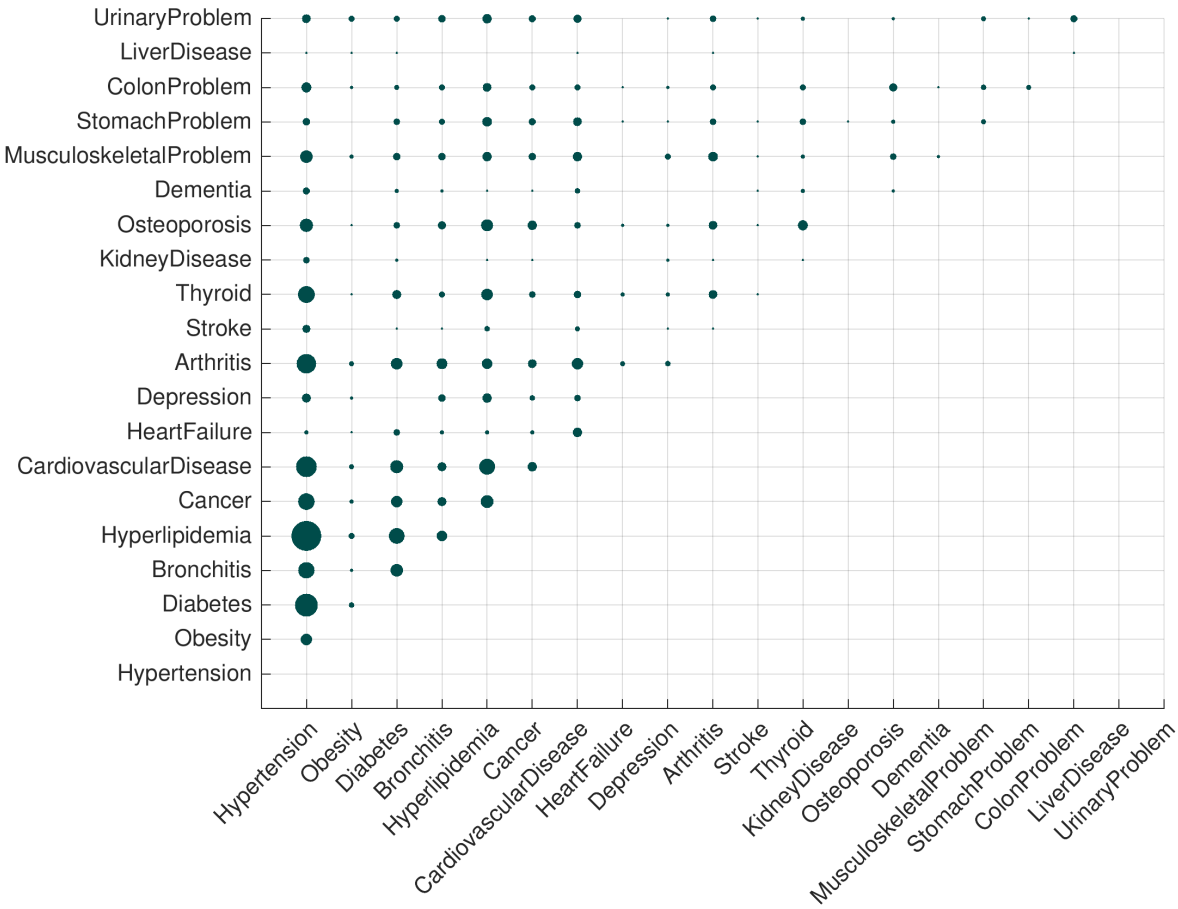


Figure A.10: Disease correlations of two coexisting diseases for elder patients. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pair among elder population. Hypertension becomes the most prevalent disease and co-occurs with all other chronic conditions.

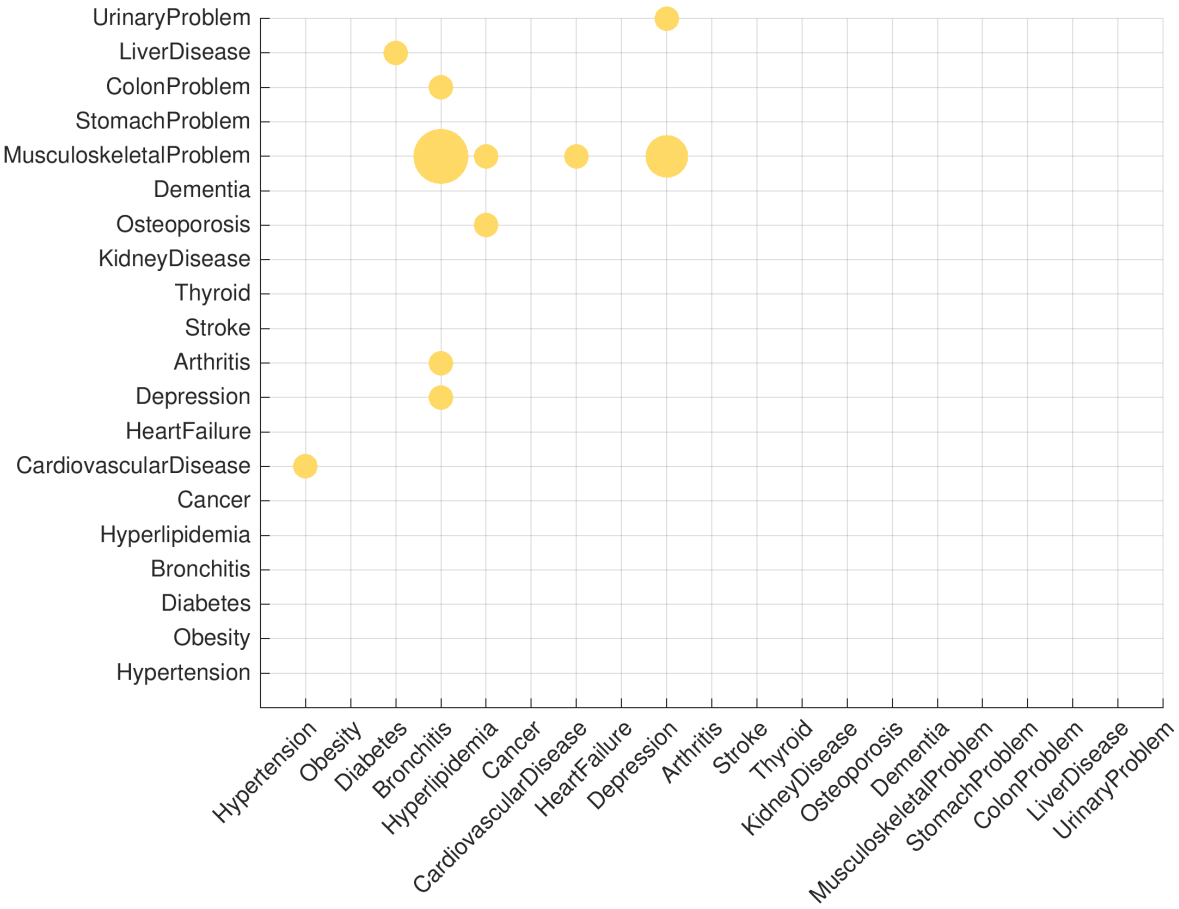


Figure A.11: Disease correlations of two coexisting diseases for patients belonging to moderate income quintile. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pairs. The most dominant disease correlation belongs to the pair of musculoskeletal problem and bronchitis. The pattern of disease correlations among patients with socioeconomic score 3, is different from those who belong to group of socioeconomic score 4 or 5, and the reason is small population of patients belonging to socioeconomic score 3.

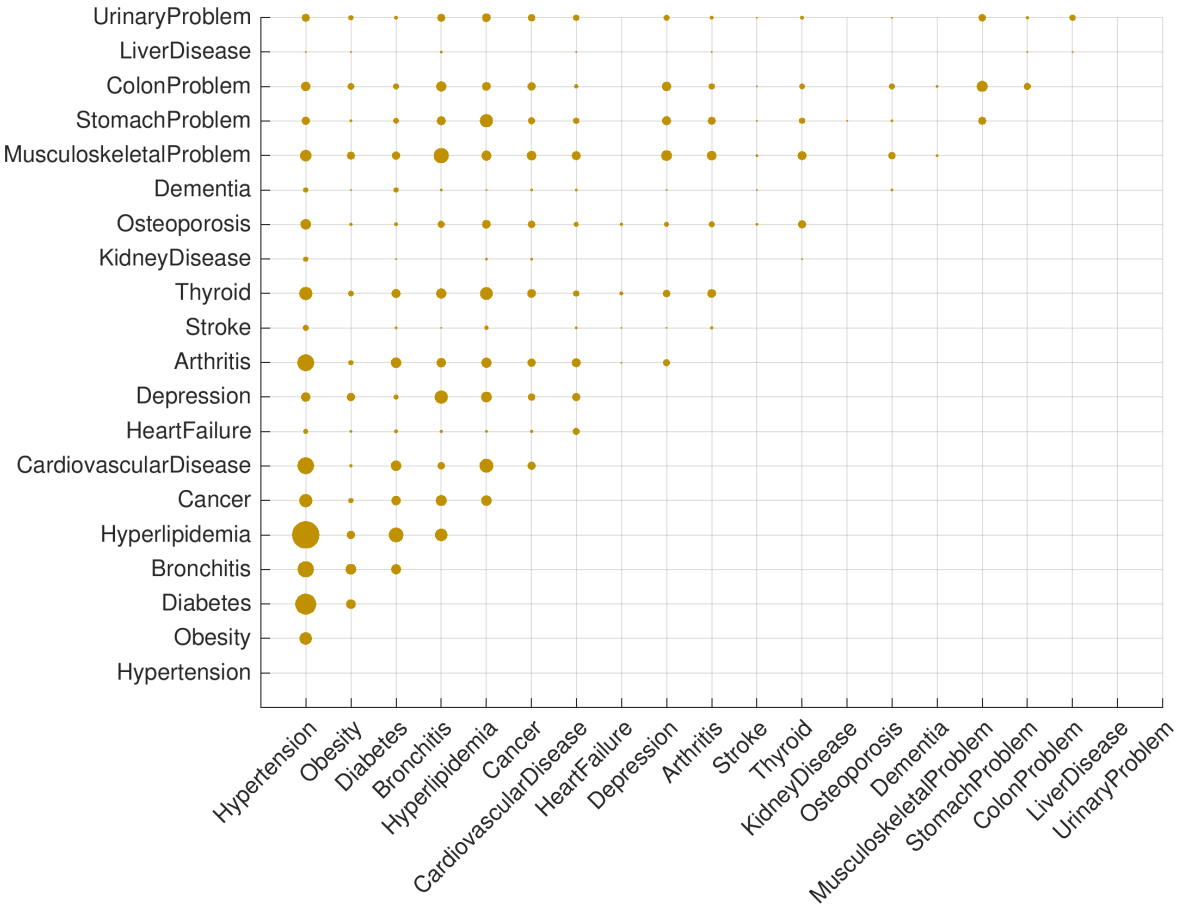


Figure A.12: Disease correlations of two coexisting diseases for patients belonging to high income quintile. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pairs.

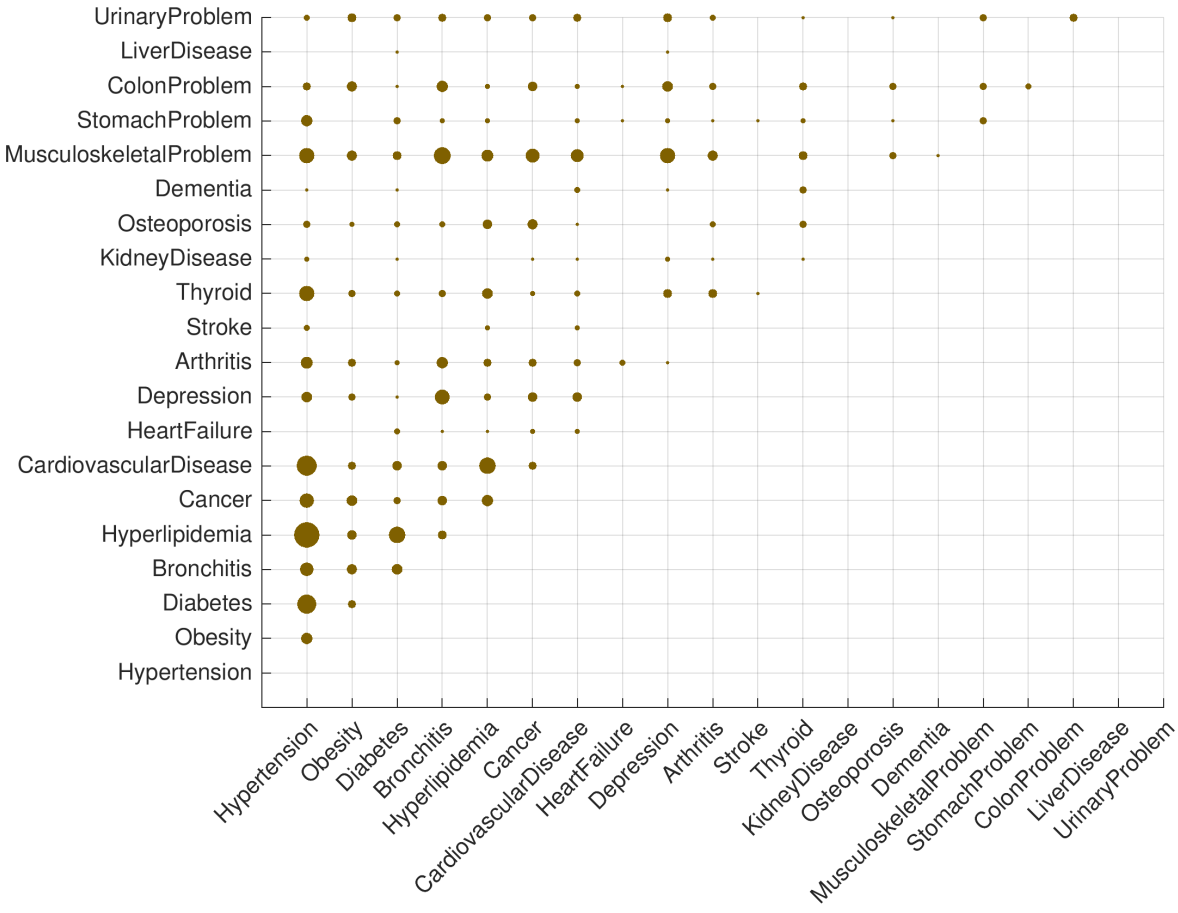


Figure A.13: Disease correlations of two coexisting diseases for patients belonging to highest income quintile. The diameter of each circle represents the relative percentage of occurrence of the corresponding disease pairs.

Appendix B

Modified K-modes Clustering Results

This appendix includes the clustering results of patient socio-demographic characteristics for patients who are diagnosed with the same two chronic conditions.

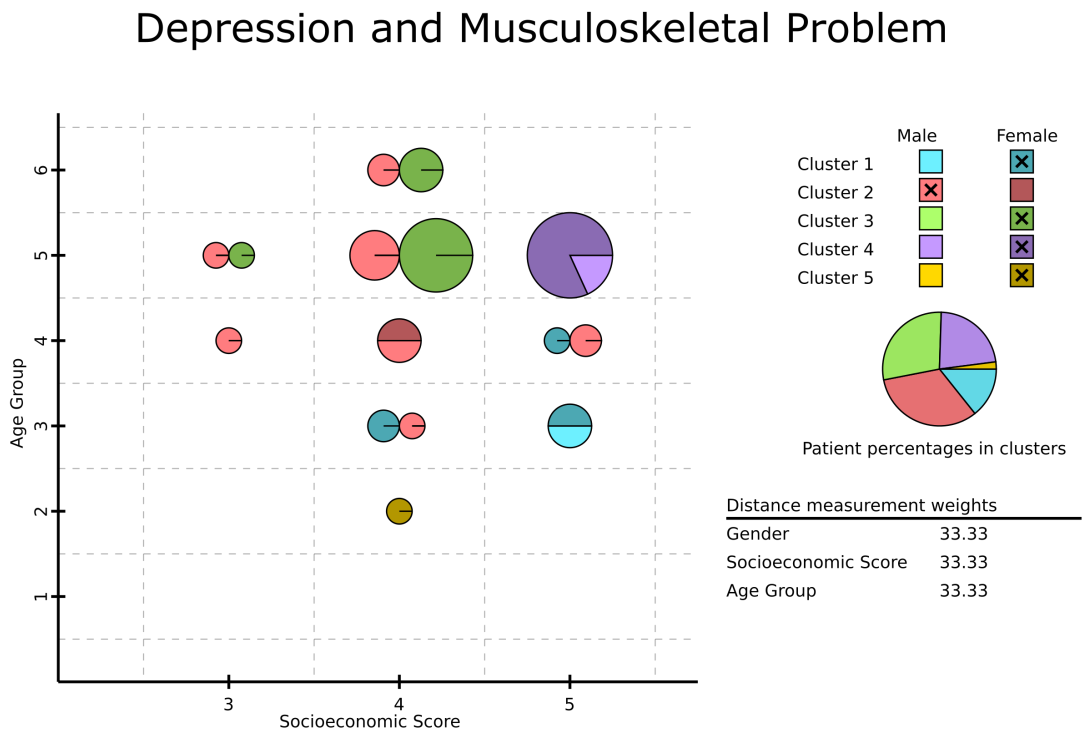


Figure B.1: Clusters of patients with depression and musculoskeletal problem. The clusters have appeared based on gender split. Every clusters represents the patients belonging to a few consecutive age groups and socioeconomic scores, which are mainly male or female. The second clusters holds the most patients who belong to all socioeconomic groups, are between age group 3 and 6, and are mainly male, while forth cluster specifies middle aged patients who belong to highest income quintile and are mainly female.

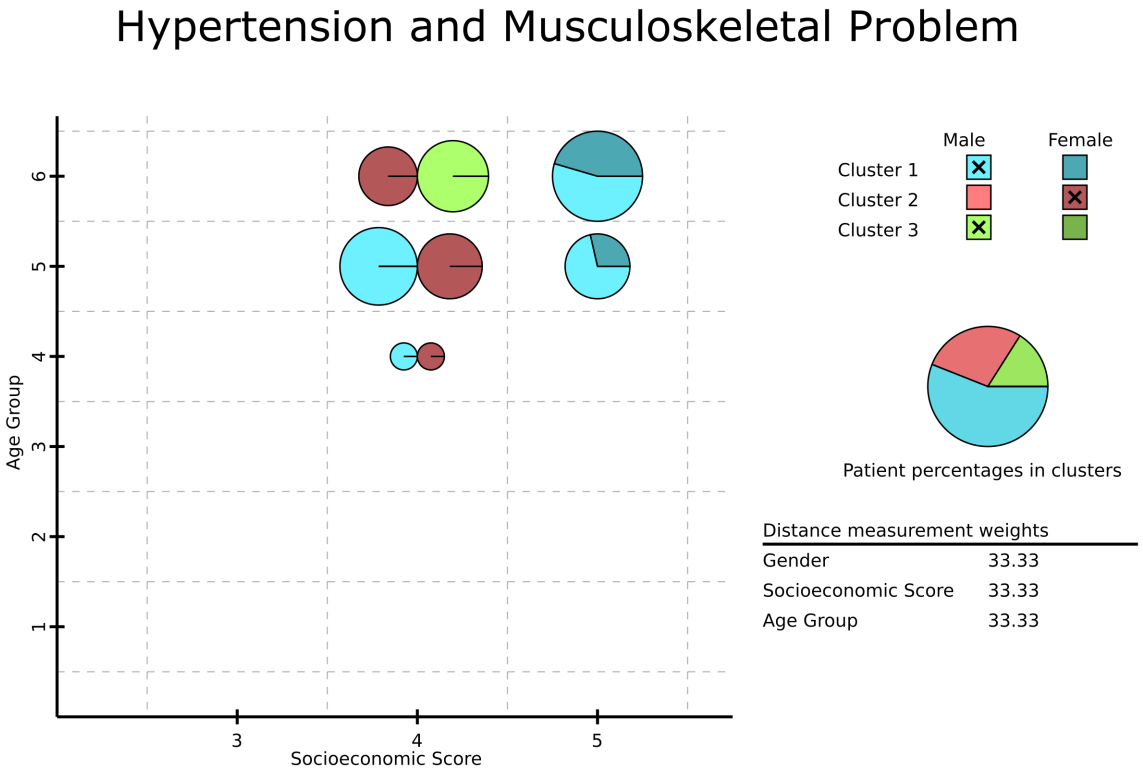


Figure B.2: Clusters of patients with hypertension and musculoskeletal problem. The data split occurs according to the three attributes age, gender and socioeconomic score. The second cluster refers to female patients who belong to socioeconomic score 4, while the third cluster refers to elder male patients of the same socioeconomic score.

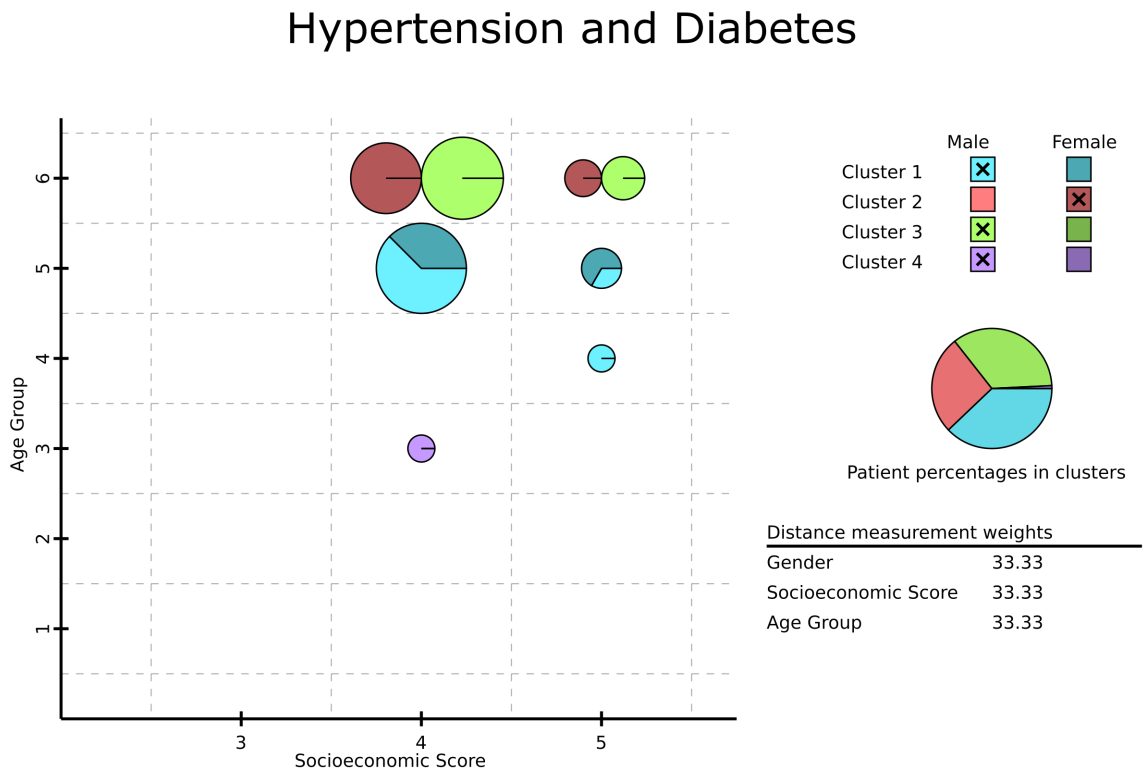


Figure B.3: Clusters of patients with hypertension and diabetes. The clusters split the data over age group and gender. Cluster 1 belongs to patients of socioeconomic score 4 and 5, who are either adult or middle aged, while clusters 2 and 3 belong to elder female and male patients respectively.

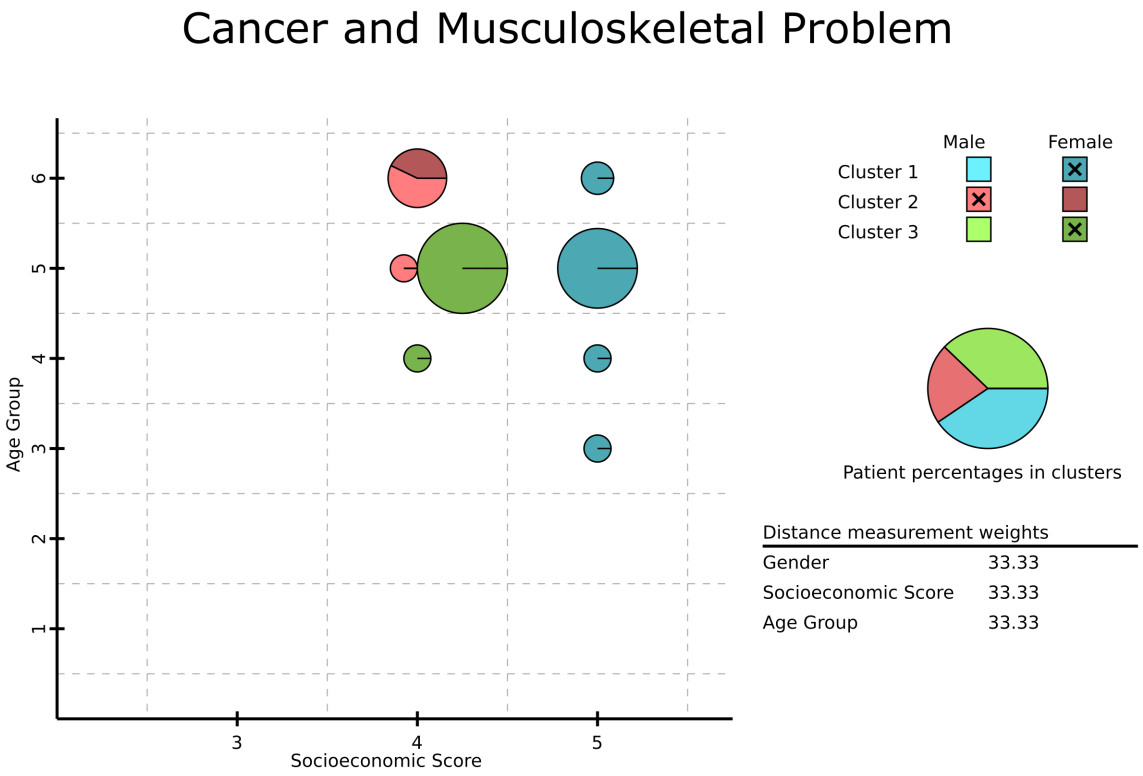


Figure B.4: Clusters of patients with cancer and musculoskeletal problem. Clusters mainly appear over socioeconomic score and gender. Cluster 1 belongs to female patients of 4 age groups who belong to highest income quintile, while clusters 2 and 3 refer to patients with high socioeconomic quintile and split males from females.

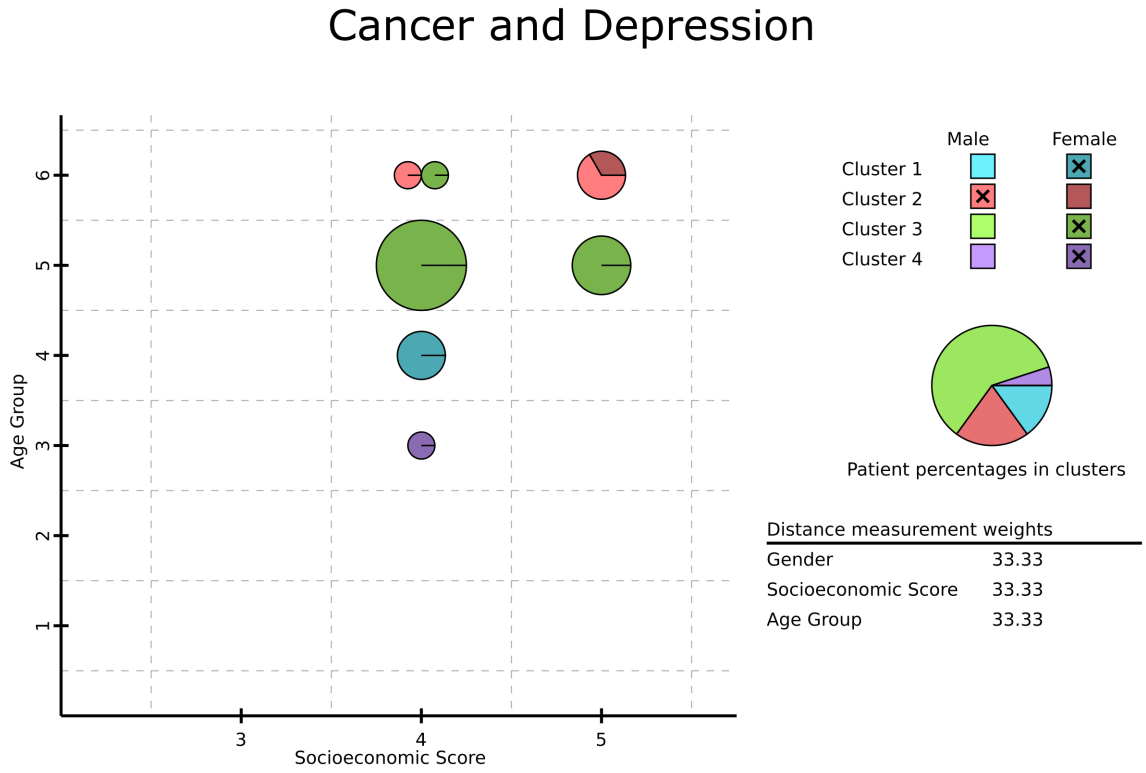


Figure B.5: Clusters of patients with cancer and depression. This disease pair mainly affects females according to out dataset. Clusters divide patients of different age groups and males of elder age group from elder females.

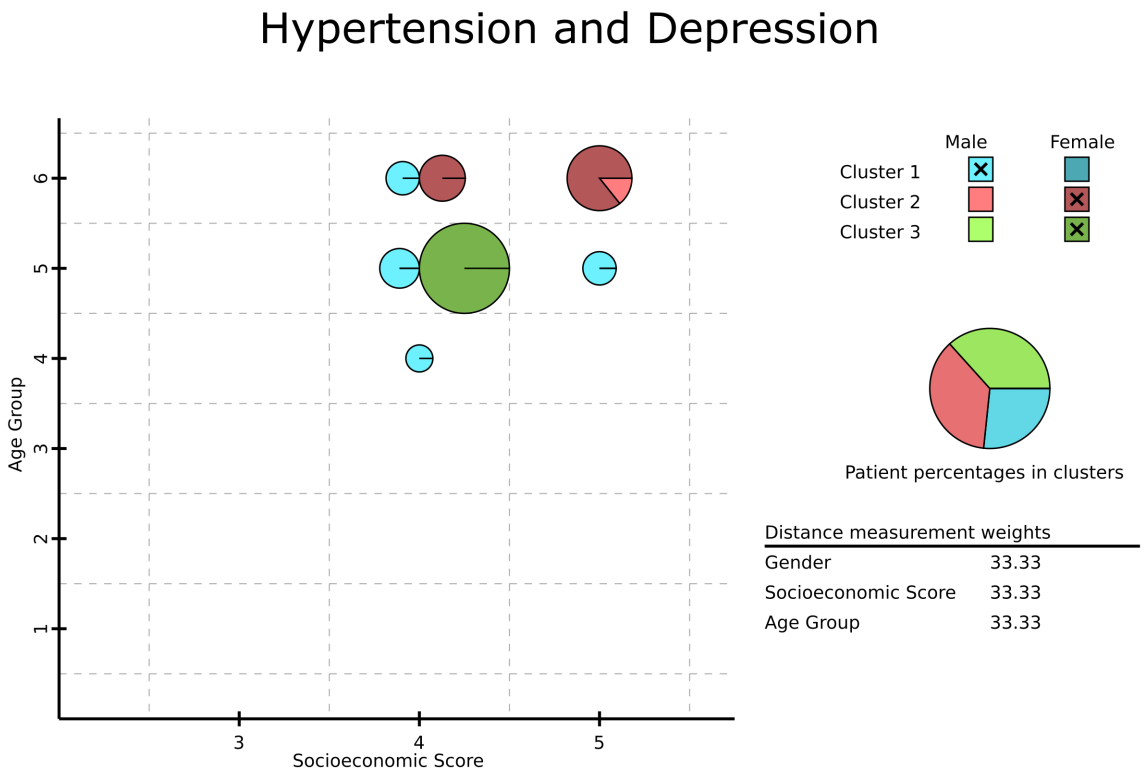


Figure B.6: Clusters of patients with hypertension and depression. Clusters divide the patients over gender and age group. Females of age group 5 and 6 are clustered in clusters 3 and 2 respectively, while most of the male patients who have this disease combination are clustered in cluster 1.

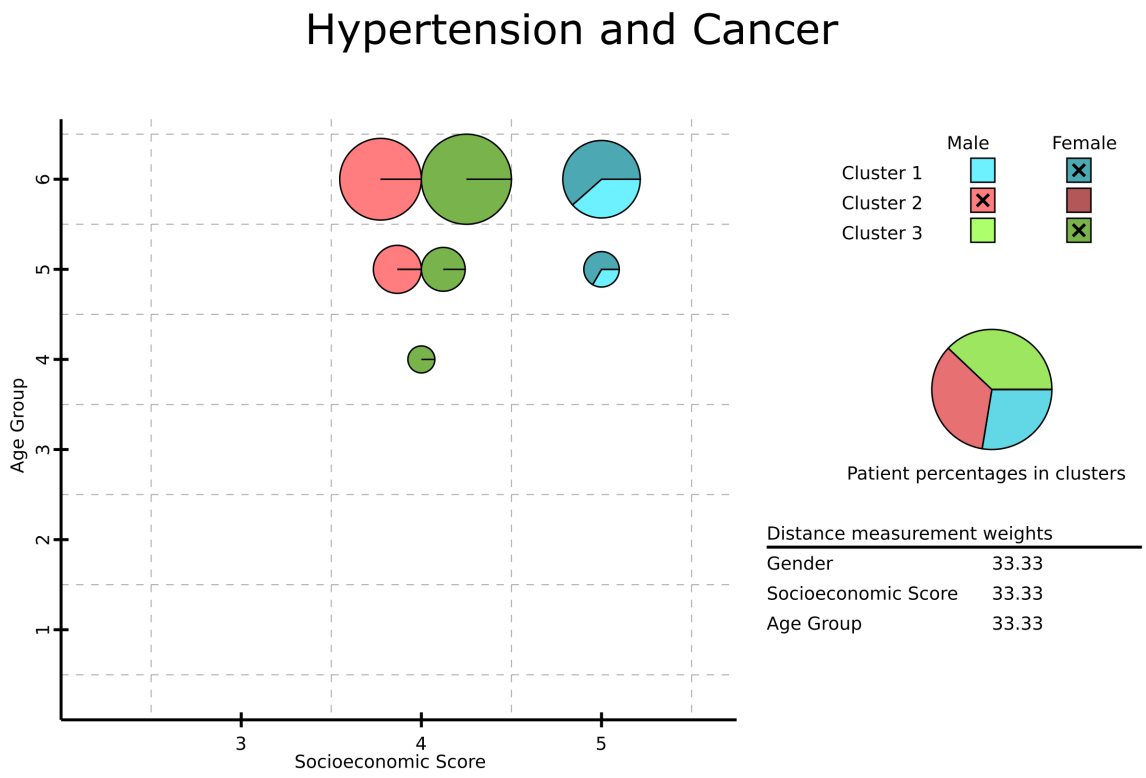


Figure B.7: Clusters of patients with hypertension and cancer. Clusters clearly appear over gender and socioeconomic score split. Patients from highest income quintiles are clustered separately from patients from high income quintile. Male and females who belong to socioeconomic score 4, have been divided into two clusters.

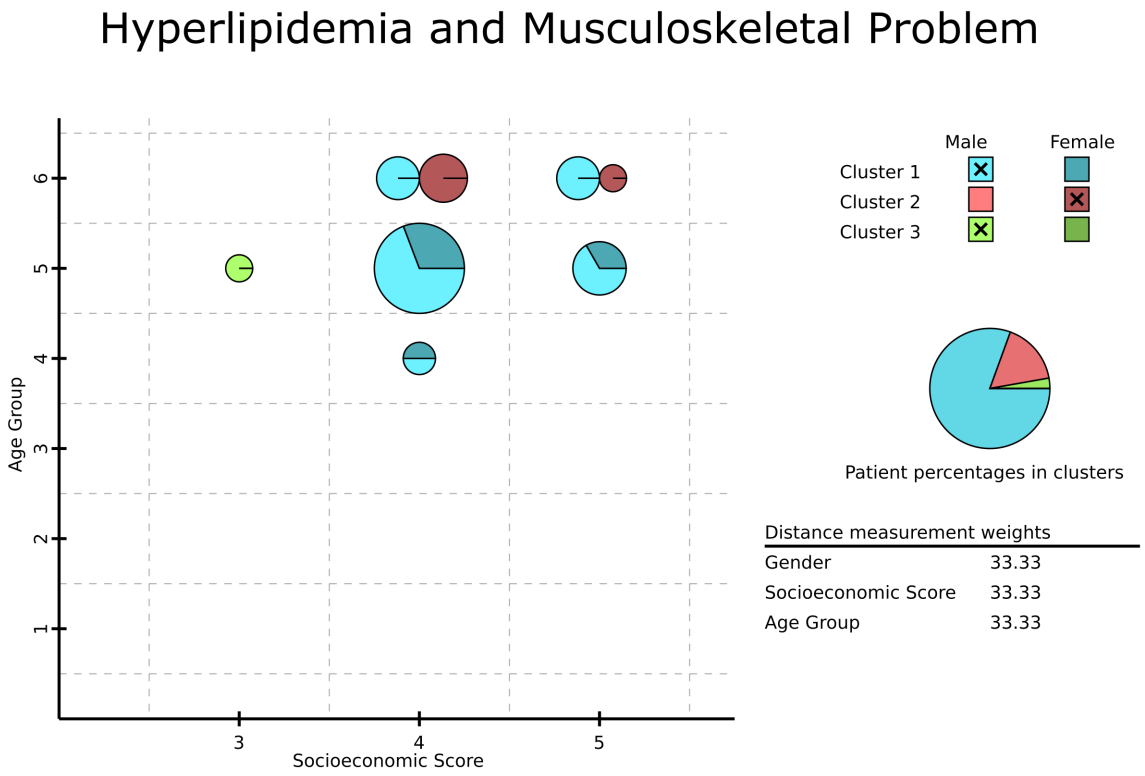


Figure B.8: Clusters of patients with hyperlipidemia and musculoskeletal problem. Clusters mainly divide the patients by gender and age group. However, this is not a perfect split of patients based on demographic characteristics.

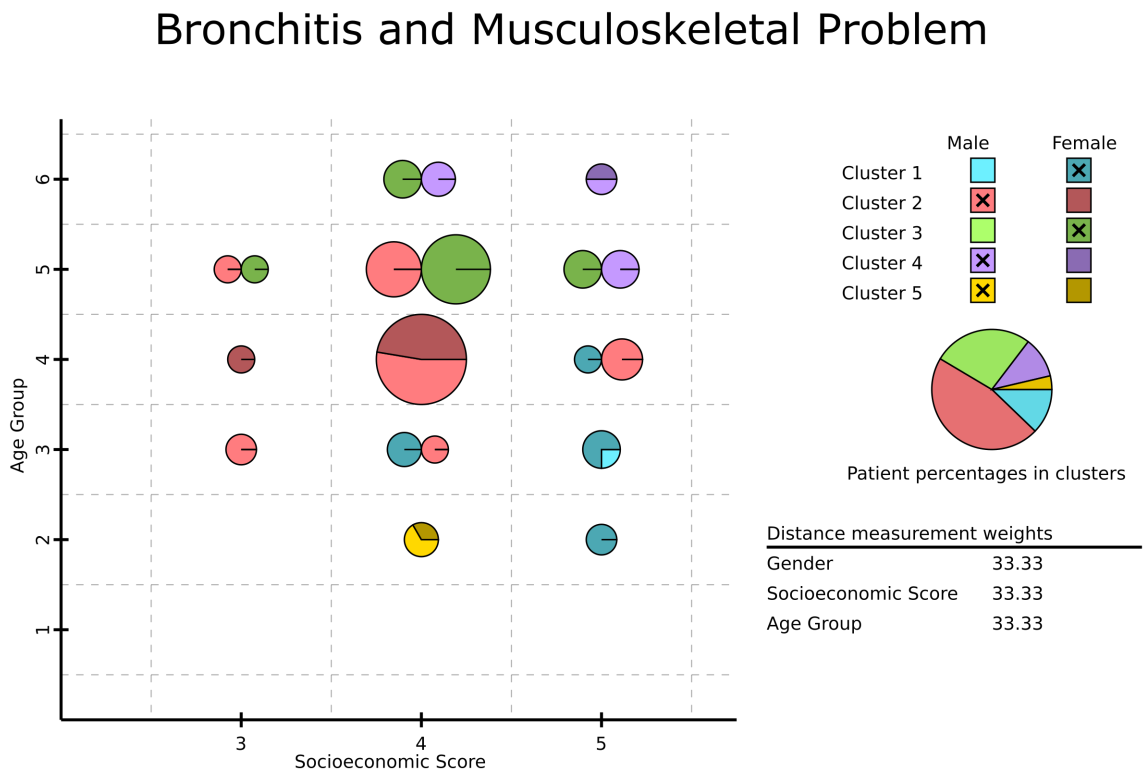


Figure B.9: Clusters of patients with bronchitis and musculoskeletal problem. Patients distribution over age groups and socioeconomic scores is similar to patients with depression and musculoskeletal problem. The division of patients into clusters is also similar in these two disease pairs. There are five clusters dividing the data: other than cluster 2 which mainly divides males from females, the other clusters belong to patients of one to three consecutive age groups.

Curriculum Vitae

Name:	Annette Megerdichian Azad
Post-Secondary Education and Degrees:	University of Western Ontario London, ON Computer Science 2015 - 2017 M.Sc. University of Tehran Tehran,IR Computer Science 2010-2014 B.Sc.
Honours and Awards:	WGRS 2015-2016
Related Work Experience:	Teaching Assistant The University of Western Ontario 2015 - 2017 Research Assistant The University of Western Ontario 2015-2017