# Mining High Impact Combinations of Conditions from the Medical Expenditure Panel Survey

Arjun Mohan

# Mining High Impact Combinations of Conditions from the Medical Expenditure Panel Survey

A Thesis Presented

By

ARJUN MOHAN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN INDUSTRIAL ENGINEERING AND OPERATIONS RESEARCH

September 2023

Mechanical and Industrial Engineering

Mining High Impact Combinations of Conditions from the Medical Expenditure Panel Survey

A Thesis Presented

By

ARJUN MOHAN

Approved as to style and content by:

DocuSigned by:
Hari Balasubramanian
D88DA3E999C4463...
Hari Balasubramanian

DocuSigned by:
Chaitra Gopalappa
2F548C0B4A0340E...
Chaitra Gopalappa, Member

DocuSigned by:
Ana Muriel
3A464F08FDE34A0...
Ana Muriel, Member

DocuSigned by:
Sundar Krishnamurty
3F8496455D304DC...
Sundar Krishnamurty, Department Head
Mechanical and Industrial Engineering

## ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Hari Balasubramanian, for the past years of his support, knowledge and mentorship. His trust in me as someone who could engage in research has been a major source of motivation and self-belief. Special appreciation is also due to Tan Rodrigues, who has been a constant source of support throughout the years I have worked on my degree. This thesis has had rocky months and their motivation and encouragement has kept me going. I am also very grateful to the members of my thesis committee, Dr. Chaitra Gopalappa and Dr. Ana Muriel, for their suggestions, comments and perspectives, which have helped me improve the quality of the thesis greatly. Dr. Balasubramanian and Dr. Muriel have been highly influential for my research interests and I will forever be grateful for the opportunities they have afforded me.

I would also like to thank Kalpak Kulkarni, Apoorva Hungund and Anirudh Natesan, whose friendship and assistance have helped me get through difficult days and have helped me through improving the versions of this manuscript.

## ABSTRACT

MINING HIGH IMPACT COMBINATIONS OF CONDITIONS FROM THE MEDICAL EXPENDITURE PANEL SURVEY

SEPTEMBER 2023

ARJUN MOHAN,

BE IN CHEMICAL ENGINEERING,

D.J. SANGHVI COLLEGE OF ENGINEERING, MUMBAI UNIVERSITY

MSIEOR, UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Hari Balasubramanian

The condition of multimorbidity — the presence of two or more medical conditions in an individual — is a growing phenomenon worldwide. In the United States, multimorbid patients represent more than a third of the population and the trend is steadily increasing in an already aging population. There is thus a pressing need to understand the patterns in which multimorbidity occurs, and to better understand the nature of the care that is required to be provided to such patients.

In this thesis, we use data from the Medical Expenditure Panel Survey (MEPS) from the years 2011 to 2015 to identify combinations of multiple chronic conditions (MCCs). We first quantify the significant heterogeneity observed in these combinations and how often they are observed across the five years. Next, using two criteria associated with each combination -- (a) the annual prevalence and (b) the annual median expenditure -- along with the concept of non-dominated

Pareto fronts, we determine the degree of impact each combination has on the healthcare system. Our analysis reveals that combinations of four or more conditions are often mixtures of diseases that belong to different clinically meaningful groupings such as the metabolic disorders (diabetes, hypertension, hyperlipidemia); musculoskeletal conditions (osteoarthritis, spondylosis, back problems etc.); respiratory disorders (asthma, COPD etc.); heart conditions (atherosclerosis, myocardial infarction); and mental health conditions (anxiety disorders, depression etc.).

Next, we use unsupervised learning techniques such as association rule mining and hierarchical clustering to visually explore the strength of the relationships/associations between different conditions and condition groupings. This interactive framework allows epidemiologists and clinicians (in particular primary care physicians) to have a systematic approach to understand the relationships between conditions and build a strategy with regards to screening, diagnosis and treatment over a longer term, especially for individuals at risk for more complications. The findings from this study aim to create a foundation for future work where a more holistic view of multimorbidity is possible.

KEY WORDS: multimorbidity, prevalence, expenditure, MEPS, unsupervised learning, association rule mining, hierarchical clustering.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| Acronym | Expansion |
|---|---|
| **MCC** | Multiple Chronic Conditions |
| **MEPS** | Medical Expenditure Panel Survey |
| **MEPS-HC** | Medical Expenditure Panel Survey- Household Component |
| **AHRQ** | Agency for Healthcare Research and Quality |
| **HHS** | United States Department of Health and Human Services |
| **NICE** | National Institute of Health Care and Excellence |
| **HCUP** | Healthcare Cost and Utilization Project |
| **CCC** | Collapsed Clinical Condition |
| **CCCR** | Collapsed Clinical Condition - Refined |
| **ICD-9** | International Classification of Diseases,9th Revision |
| **ICD-9-CM** | ICD-9, Clinical Modification |
| **CVD** | Cardiovascular Disease |
| **CNS** | Central Nervous System |
| **COPD** | Chronic Obstructive Pulmonary Disorder |
| **GI** | Gastrointestinal |
| **URI** | Upper Respiratory Infections |
| **HIV** | Human Immunodeficiency Virus |
| **NEC** | Not elsewhere classified |
| **OB** | Obstetrician |
| **WPGMA** | Weighted Pair Group Method with Arithmetic Mean Algorithm |

# CHAPTER 1

# INTRODUCTION

With an annually growing population of individuals with multiple chronic conditions (MCC), there is a growing need to better study and understand the patterns and effects of these cases. Although the prevalence rates of multimorbidity have shown a relative degree of stability over the years, as the baby boomer generation enters Medicare eligibility, this is a steadily growing cohort (Buttorff et al., 2017). From 2008-2014, the prevalence of MCCs has been approximately 40% of the US population with a direct correlation with age, and senior citizens form a majority of this cohort. It is a well-documented phenomenon that the number of multiple conditions an individual is affected by is directly proportional to healthcare expenditure (Hwang et al., 2001; Buttorff et al., 2017). In fact, 71% of the Total U.S. Healthcare Spending in 2010 was for patients with multiple chronic conditions (Gerteis et al., 2014). Additionally, they are also associated with a greater set of limitations and difficulties pertaining to the individual's independence and social and cognitive well-being.

According to the US Census Bureau, "By 2034 (previously 2035), there will be 77.0 million (previously 78.0) people 65 years and older compared to 76.5 million (previously 76.7 million) under the age of 18" (US Census Bureau, 2018). With that in mind, it is important to not only analyze the impact this demographic of patients has on the system, but also the quality of care they are afforded, as they are a set of individuals who have a lower quality of health and well-being due

to their conditions, and therefore face a greater variety of challenges over the course of their healthcare.

In a cross-sectional study by Wolff and colleagues that studied Medicare fee-for-service beneficiaries, inpatient hospitalizations and admissions were seen to have a positive correlation with the number of chronic conditions faced by a patient (Wolff et al., 2002). In a study of the elderly (>50 years of age) patients at three primary care practices in Ireland, a similar strong association was found by Glynn et al. between "...an increasing number of chronic conditions and the frequency of primary care consultations, hospital admissions and hospital out-patient visits among primary care patients" (Glynn et al., 2011).

Much of the existing research around multimorbidity has focused on the prevalence of the phenomenon itself. In recent years, demographic studies around the trends and patterns around multimorbidity have been conducted with a greater degree of interest, especially around combinations of conditions that have a high degree of prevalence (Ahn et al., 2020; Glynn et al., 2011). Although the population-wide prevalence has stayed largely stable, the fact remains that the total number of individuals presenting with multimorbidity has shown a general trend of growth. However, the majority of the research has not focused on understanding broader patterns and relationships between various conditions and how multimorbidity as a whole affects the system. Sorace et al. in their 2011 study found that the 100 most prevalent combinations of diseases in the year of 2008 accounted for 33% of Medicare Part A & B beneficiaries, but only 15% of the expenditure. Meanwhile, a different population, 32% of the beneficiaries,

accounted for nearly 80% of the Medicare expenditure. This population, unlike the former, is far more complex, and comprises more than 2 million unique combinations of diseases, as classified using International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes. Low prevalence or not, there is very clearly a great need to analyze the phenomenon of multimorbidity through a wider lens than merely prevalence.

With this in mind, we decided to use a multi-criteria approach to mine high-impact combinations of conditions, without focusing purely on prevalence in the system. Through observing the complexity in care that exists for individuals with complex multimorbidity, there was also a motivation for us to study the patterns between the conditions and condition groups that are frequently comorbid with each other, to better understand the relationships between these conditions and how they influence the development of multimorbidity.

Based on this, the goal of this thesis is to study multimorbidity through a lens of understanding systemic impact, and develop a means to interpret associations and relationships between conditions affecting individuals with multimorbidity. We do this using data from the Medical Expenditure Panel Survey (MEPS).

The MEPS is a survey conducted by the Agency for Healthcare Research and Quality (AHRQ), an agency under the purview of the United States Department of Health and Human Services (HHS). The MEPS - Household Component (MEPS-HC) is a survey of health care utilization and expenditure of a selected panel of households that can be used to obtain nationally representative

estimates for the civilian non-institutionalized U.S. population (Aizcorbe et al., 2011). With a nationally representative sample, it is generally considered a good source of data in the healthcare sector, especially because it can directly link and itemize the expenses from the various healthcare services (Sing et al. 2006), and additionally is the only dataset which captures the healthcare expenditures of uninsured individuals (Cohen et al., 2009). The MEPS has remained a prime source of data in the context of American healthcare, and is used for analysis across multiple domains of healthcare research.

In terms of the analysis, we used the MEPS-HC Event files, which documents the list of events pertaining to healthcare services for the relevant households of the survey. Of the event types, 4 event types were primarily studied viz. Inpatient Stays, Emergency Room visits, Outpatient care & Office-Based Medical Provider care, which represent the primary forms in which individuals receive health care. The event files also contain data pertaining to dental visits, medications prescribed, home health, and other expenses, but they were not considered as they are outside the scope of this study. For the purposes of this study, the data from the years 2011-2015 have been used.

First, we mine data from the MEPS-HC, with the aim of identifying unique combinations of conditions, and determining the prevalence of the combination and the expenses associated with having that combination. Using this data, we are able to create a distribution of the prevalence and expenditure for each combination of conditions. These combinations are grouped into sets, ordered as per their presence on Pareto fronts of this distribution, which we define as impact

levels. We achieve this grouping using an algorithm that identifies non-dominated fronts of a bicriteria optimization problem where one criterion is prevalence and the other, the load exerted on the system, mainly in terms of the expenditures of the median individual having that combination of conditions.

These sets, or impact levels, consist of both those groups which have extremely high prevalence leading to a large load on the system, as well as those groups which are low prevalence-high impact, which are those groups of conditions which, while present in a small set of the population, have a markedly higher degree of expenditures associated with them. Much of the work and policy decisions around multimorbidity tends to be focused on the former set of groups which are high impact due to the large population of patients that comprise them. This latter set of groups, while low in prevalence, has significant heterogeneity in terms of the conditions, but high impact due to the level of expense. This latter group in aggregate has non-trivially high prevalence — in 2014, patients with 5+ comorbidities represented 12% of the prevalence, but 41% of the total expenditure associated with patients presenting with MCC (Buttorff et al., 2017). High impact through expenses implies a high fragmentation of care for these patients, as the complexity of their medical profile requires a multitude of medical professionals to be involved in their care, as well as multiple medical procedures and forms of medication.

To find potential relationships between a large sample of data points, we chose to explore unsupervised learning methods that use a bottom-up approach, specifically association rule mining and hierarchical agglomerative clustering.

Association rule mining gives us a reasonably granular perspective of the associations between conditions and groups of conditions from a statistical point of view. Hierarchical agglomerative clustering gives us a bird's eye view of the various conditions in the system, showing us broader associations between the conditions on a systemic level. This lays the foundation to create a structured framework to understand multimorbidity through discrete associations between groups of conditions, while observing these associations in a holistic, system-wide manner. This foundation can be further built upon to create a systematic approach to better diagnose, and to develop better prognoses for individuals presenting with, and at risk of complex multimorbidity.

# CHAPTER 2

# REVIEW OF RELATED LITERATURE

## 2.1. Background

Since the 1990s there has been an awareness and a growth in the study of multimorbidity as a phenomenon in the population. While improved methods of surveying and data collection have improved the quality of the research in this space, it was recognized early on that there is a linear relationship between the number of chronic conditions an individual had and their healthcare expenditures, and expenses associated with multimorbidity were projected to rise (Hoffman et al, 1996).

The growing prevalence of multimorbidity has been observed outside America too, and some of the work done in Europe, Canada and Australia has been groundbreaking in terms of determining the areas of study in this phenomenon that require deeper investigation.

Researchers Van den Akker et al. (1998) studied the prevalence of multimorbidity in a sample from the Netherlands and found that prevalence increased over time in all age groups, but was particularly severe in participants aged 80 years and over. Factors that were found to contribute to multimorbidity included growing age, lower education, and public health insurance. In 2003, a study conducted in Canada witnessed a high prevalence of individuals in Saguenay, Quebec having 2 or more medical conditions, with high prevalence percentages among the age groups of ‘45 to 64 years’ and ‘65-years and older’. Upon further analysis, the rise of the prevalence of multimorbidity with increase in

age was observed here as well (Paez et al., 2009). Fortin et al. noted that individuals with multimorbidity “represent the rule rather than the exception” (2005).

In a systematic review of Australian studies wherein chronic diseases were included as national health priorities, Caughey et al. (2008) saw that participants with arthritis, asthma, cardiovascular disease (CVD) frequently faced multimorbidity. More than half of the elderly participants of the sample that had arthritis also had a comorbidity of hypertension, with others having CVD (20%), diabetes (14%), and mental health problems (12%). More than 60% of individuals with asthma also faced arthritis as a comorbidity, with others having CVD (20%) and diabetes (16%). About 60% of individuals with CVD also had arthritis, 20% had diabetes and around 10% of participants also had asthma or mental health problems.

In 2012, of a cohort of 15 million individuals with chronic conditions in the UK, approximately 45% of individuals presented with more than one condition. The National Institute of Health Care and Excellence (NICE), in 2016 published guidelines pertaining to the care and management of individuals with multimorbidity, recommending approaches that focus on the interactions between the conditions that the individual presents with, and the varying benefits and risks of using guidelines pertaining to single health conditions. In a 2018 study assessing the patterns of multimorbidity in middle aged and older adults using association rule mining, Zemedikun et al. (2018) found 3 primary clusters of

diseases, with 30 patterns in the multimorbid presentation of the 36 conditions included in the study.

A study which collected data from a cohort in Sweden, the highest comorbidity was seen with visual impairments and heart failure. Researchers Marengoni et al. noted that heart failure was rarely seen without any comorbidity. The prevalence found for many circulatory diseases, dementia and depression exceeded the expected prevalence put forth by the researchers, who noted that for individuals with multimorbidity, “there exists co-occurrence of diseases beyond chance” (Marengoni et al., 2009). Across these studies, it is apparent that a few common patterns begin to emerge globally, with growing prevalence of multimorbidity, and with the elderly being affected more adversely, understandably so.

Additionally, what we do notice is that in a lot of the earlier research into multimorbidity, the focus tended to be around the growth in the prevalence of the phenomenon, and not necessarily a more investigative analysis into what forms the multimorbidity took. In a 2013 study of the multimorbidity in 15 prevalent conditions for a cohort of Medicare beneficiaries, Salive found that the degree of multimorbidity grew with age, and noted that there is a need for a standardized method for measuring multimorbidity. Craig et al. (2023), in a cross-sectional survey of young adults aged 18-35 from Kenya, South Africa and the UK observed a greater degree of multimorbidity in individuals who perceived themselves as overweight or obese, where the degree of multimorbidity was determined by

assigning a multimorbidity risk score, with individuals with 3 or more conditions having the highest score.

## 2.2. Motivations For The Study

More extensive work has been done in recent years surrounding the study of dyad, triad and other small combinations of the most prevalent conditions in the overall population – some of these studies are detailed below. Additionally, conditions that find common prevalence in individuals presenting with multimorbidity tend to form the focus of general policy decision making. For example, conditions such as hypertension, diabetes, coronary artery disease, renal disease, depression and rheumatoid arthritis frequently occur in conjunction with other conditions in individuals presenting with multimorbidity, and a large body of work has been done in studying the patterns associated with them. The aforementioned conditions are among the most prevalent in the system overall, and in the analysis we have performed, feature frequently among the diseases in high impact groups of conditions.

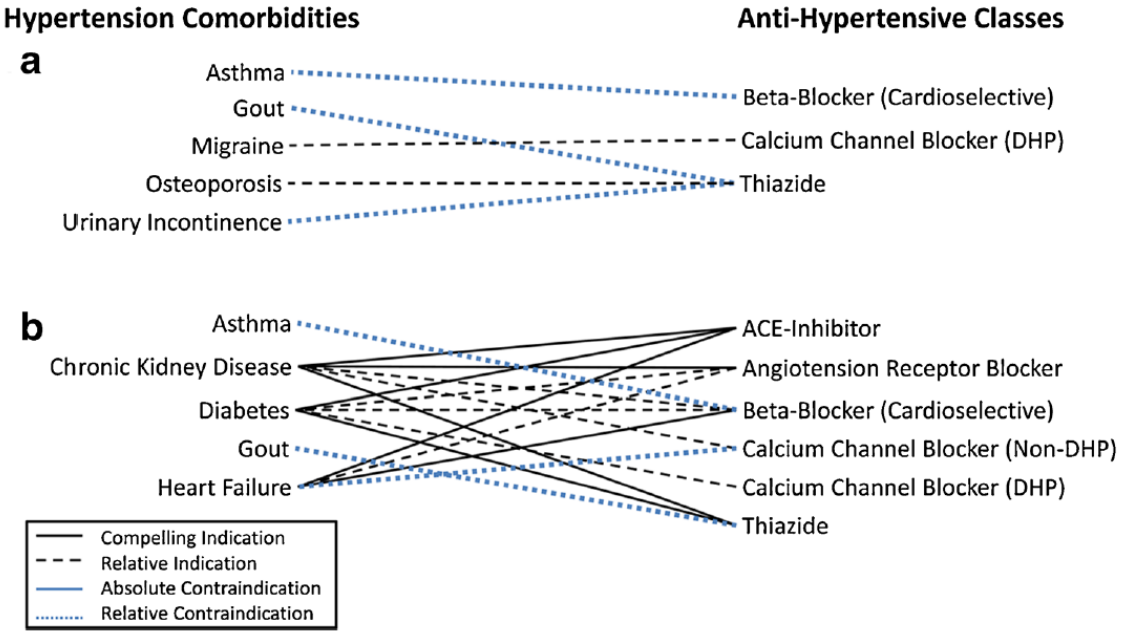
In 2013, Machlin & Soni and Lochner & Cox, found the most prevalent dyad of conditions in the US was hypertension and hyperlipidemia. The combinations of diabetes with hypertension or hyperlipidemia were also common among all sex and age groups (ranging from 21.0% for women aged 65 or older to 28.3% for men aged 65 or older). Additional common dyads for people aged 65 or older were coronary artery disease and hyperlipidemia (32.4%) and coronary artery disease and hypertension (31.4%) for men and hypertension and arthritis for women

(22.1%). Depression and hypertension or hyperlipidemia were also among the top 5 dyads for women aged 45 to 64 (19.1% and 15.6%, respectively) (Lochner & Cox, 2013; Machlin & Soni, 2013).

Zulman et al. in their analysis of multimorbidity and the nature of their healthcare life cycle noted many challenges which healthcare professionals have to consider when caring for individuals with multimorbidity. Over and above simply having their care fragmented, individuals also require a degree of collaboration or communication between their healthcare providers, as it is very frequently the case that the medications that are prescribed for the care of one condition can have an antagonistic effect on the individual, in combination with a medication prescribed for another one. It is also possible that the interaction between consuming multiple specific medications simultaneously could negate each other's effectiveness. These are all considerations to be made, which only become more and more complex as the order of multimorbidity increases. See Figure 1, reproduced from Zulman et al., for an example of such interactions.

**Figure 1**

*Variation in management complexity among two individuals with five comorbid conditions, but different levels of comorbidity interrelatedness. This figure illustrates variations in medication management complexity for two individuals with five comorbidities that occur with hypertension, but have vastly different degrees of complexity in terms of the interactions between the medications that are frequently prescribed for the care of those conditions. (Zulman et al., 2013).*




A cross-sectional study using 2009 and 2011 data from the MEPS focused on individuals with at least 2 of the following 4 of the most prevalent chronic conditions: arthritis, diabetes mellitus, heart disease, and hypertension. Adults with all 4 conditions had the highest average annual total expenditures ($20,016), whereas adults with diabetes/hypertension had the lowest annual total expenditures ($7,116) (Meraya et al., 2015). A qualitative element to this correlation was also observed, as heart disease and to a lesser extent diabetes mellitus (Baker et al., 2017) were linked more frequently to a higher annual expenditure, as they can often be linked with conditions outside the normally seen

comorbidities such as heart disease and hypertension. For instance, individuals having diabetes mellitus comorbid with musculoskeletal conditions can experience some of the most adverse effects to their quality of life and high levels of pain, while having higher expenses than most of the populace. It is pertinent to note that certain chronic conditions do have a higher degree of expenses associated with them (Vogeli et al., 2007), but the relationship between multimorbidity and impact on system and patient cannot be ignored. In 2014 it was observed that 20% of all individuals with multimorbidity had 5 or more conditions (King et al., 2018).

There has been a dearth of investigation into the expenses and impact of these low prevalence, high impact combinations of conditions, which, despite having a low weight in terms of the population of afflicted individuals, have disproportionately high expenses associated with them. As noted earlier, Sorace et al.'s analysis of disease combinations in the Medicare population determined that 32% of the beneficiaries had unique disease combinations that fell outside the 100 most common disease combinations. This population accounted for 79% of expenses and was composed of more than 2 million unique combinations of ICD-9 coded diseases. Building on these findings, a White Paper for Abt Associates, Inc. by Sorace et al. (2013) studied various combinations of conditions in terms of their prevalence, recognizing the 'long tail' of condition combinations that is populated by hundreds, if not thousands of unique combinations of conditions, which tend to be complex cases. Using multiple sources of data across various agencies, they observed that despite having above-average costs and healthcare needs, a substantial number of patients with less prevalent MCCs find themselves

excluded from clinical research studies due to a low prevalence of their specific combination of MCCs. Furthermore, with a much higher fragmentation of care, the complexity of the treatment of the conditions becomes all the more challenging. Therefore, understanding the patterns in which multimorbidity presents on both a qualitative as well as a quantitative level is one of the primary motivations of this study.

For understanding the way to go about analyzing the data, a primary need was to find a way to consider more than just the prevalence as a dimension for observing the patterns in multimorbidity in the population. Studying the relationship between multimorbidity and out-of-pocket expenditure on medicines in 12 European countries, Palladino et al. (2023) found that nearly half the individuals presenting with multimorbidity presented with complex multimorbidity. With each additional chronic condition, there was found a 34% greater likelihood of incurring out-of-pocket expenditure on medication, with an average incremental expense of 26.4 euro for each chronic condition. This average expense was found to be 21.59% higher in the event that the additional condition affected an additional body system. Given this, it is not just important to find patterns in more complicated condition complexes in individuals presenting with multimorbidity, but also to observe the role of expenditure, while studying the types of conditions that are co-occurring in these condition complexes.

## 2.3. Exploration Of Unsupervised Learning Methods

To perform analysis on combinations of conditions that are present in the system, we decided to use association rule mining and clustering analysis. The specific methods we have chosen employ a “bottom-up” approach, which best suits this application — it is analogous with the gradual complication of the healthcare profiles that individuals with multimorbidity experience.

Association rules were initially conceptualized for discovering patterns in product sales in supermarkets (Agrawal et al., 1993). Here, a rule was formally defined as an implication of the form $X \Rightarrow Y$, where $X, Y \subseteq I$, a set of discrete items. In their definition of association rules, Agrawal et al. defined Y as a set of size 1, i.e., a single item. Building on this earlier work, Agrawal and Srikant (1994) created the apriori algorithm, which we have used for association rule mining in this study. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time, and groups of candidates are tested against the data.

To perform clustering analysis for the groups of conditions, we chose to follow an agglomerative hierarchical clustering approach, due to the fact that these methods employ a “bottom-up” approach. In such methods, all items to be clustered start as clusters composed of the individual items, forming clusters iteratively from these items and clusters gradually being clustered with each other until all items are aggregated. With this in mind, we had to determine the distance or dissimilarity measure to choose to differentiate between the conditions

themselves, and to select the clustering method or algorithm that best suited our use case.

Comparing the utilities of different dissimilarity/similarity measures or coefficients was imperative, as different measures in conjunction with different clustering methods can yield contradictory results (Bock, 2014; Rajagopalan & Robb, 2005; Meyer et al., 2004; Dalisefat et al., 2009). Given that our utility is to compare different conditions and determine the clusters in which they develop into more complicated disease complexes, set similarity measures were the more obvious choice. These measures are generally used to typify the associations between objects or items that can be defined in sets, and find applications in recommendation systems in social media analysis, species analysis in biology and so on (Rajagopalan & Robb, 2005). With a dearth of research that uses clustering to determine the relationships between conditions, particularly in longitudinal studies, we looked at applications of these measures in different contexts to find analogous measures that may be applicable for this study. Dalisefat et al. (2009) compared three similarity measures, namely the Sorensen-Dice coefficient, the Simple Matching coefficient and the Jaccard coefficient, to assess the variations caused in cluster analysis when using them. Using these coefficients to cluster seven silkworm species, their results demonstrated that the Jaccard and Sorensen-Dice produce extremely close results. They also demonstrated that these measures are of great use in datasets where there are many negative co-occurrences, as they do not consider these co-occurrences for similarity. Further, the Sorensen-Dice coefficient had greater utility in particularly sparse datasets, as

it gives additional weight to positive co-occurrences. Since our data is both sparse and has many negative co-occurrences in the set matching step, the Sorensen-Dice coefficient was selected.

With regards to selecting the linkage method or the clustering algorithm itself, we selected the WPGMA algorithm, which creates pairwise clustering, assigning equal weights to the groups/clusters formed (Sokal & Michener, 1958; McQuitty, 1966). The iterative step of the clustering algorithm was the main reason for our selection of the algorithm, where an equal weight is assigned to each cluster or item that is being compared (Belbin et al., 1992). This allows for an approach where the clusters, when compared for dissimilarity to other clusters or items, would be weighed equally regardless of the sizes of the clusters themselves. This is ideal for analyzing the associations between the conditions themselves.

# CHAPTER 3

## DATA

The data used for this study was sourced from the Medical Expenditure Panel Survey (MEPS), using the MEPS Household Component (MEPS-HC) Event Files, with the following 3 event types taken under consideration: Hospital Inpatient Stays, Emergency Room Visits and Office-Based Medical Provider Visits. The MEPS-HC collects data from a nationally representative sample of households through an overlapping panel design.

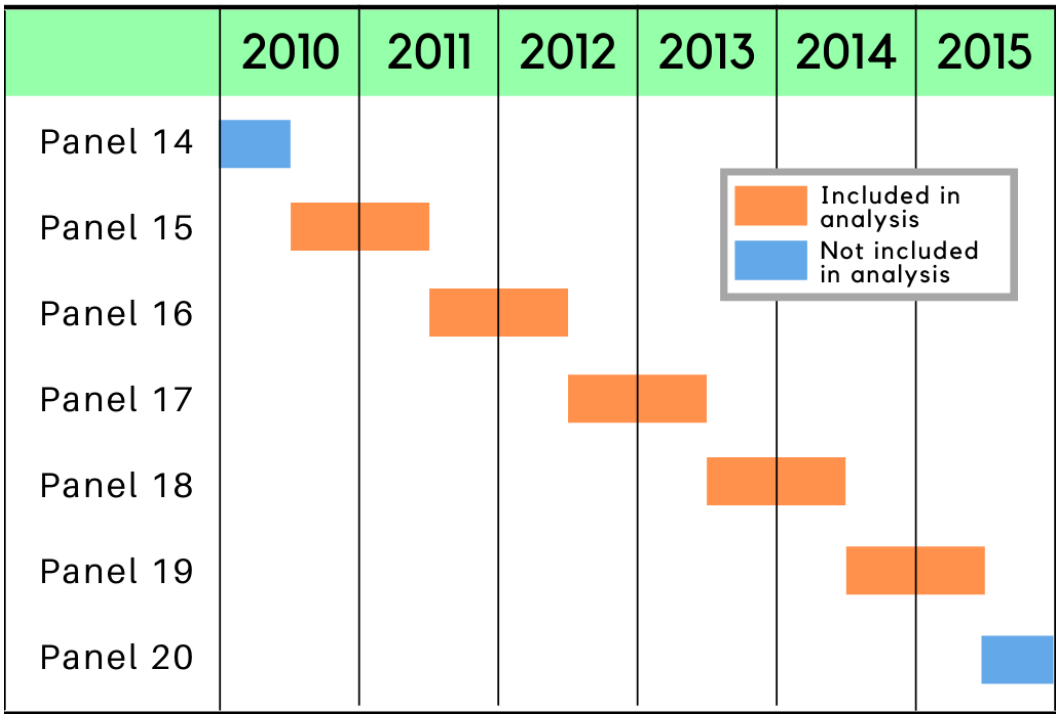
A new panel of sample households is selected each year, and data for each panel are collected for two calendar years (Agency for Healthcare Research and Quality, 2017). For the MEPS-HC Event files, this gives us a 2-year horizon over the healthcare history of an individual. For each event, the variable under consideration was the cost associated with the specific event, and for each individual, their weight in terms of the population was recorded. The annual expenditure for each individual was aggregated, and the individuals were then aggregated according to unique groups of conditions they were diagnosed with. This operation was performed separately on an annual basis for the years 2011-2015, as well as during the entire five-year duration under the scope of study, taking panels 15-19 into consideration, to get annual aggregates as well as a larger healthcare profile for the population over the course of the 5-year period.

For a more comprehensive analysis of the disease impact over multiple years, we picked the survey panels 15-19 as our source for the data, as each

year's survey includes individuals from 2 different panels, as this is a multi-panel survey conducted over 2 years. This is shown in the below figure:

**Figure 2**

*Panel Structure for the MEPS Panel Survey. Panels 15-19 are taken for 5-year analysis.*




Therefore, we drop the patients from panel 20 in the year 2015, and take the panel 15 data from the year 2010 to complete the healthcare history of the panel 15 patients from the year 2011.

The conditions are categorized using Collapsed Clinical Condition (CCC) codes, which are codes assigned to groups of conditions that have similar diagnoses. These codes are sourced from using the AHRQ's Clinical Classification Software on ICD-9 codes (Elixhauser et al., 2015), to get a set of diagnosis groups

that fall under similar buckets. ICD-9 codes are not directly used as they may create too much granularity by virtue of the specificity with which the diagnosis codes are designed. Individuals with insufficient data or no diagnoses were grouped under the bucket ‘ND’.

There are a total of 253 CCC Codes, and these codes were used for the grouping and to understand the degree of multimorbidity. At the same time, to allow for broader analysis, the Healthcare Cost and Utilization Project (HCUP) created a less granular set of codes, which group these conditions into 60 classes of conditions. These codes are called the CCC-Refined (CCCR) codes, and starting from 2016, along with moving from ICD-9 to ICD-10, these are the codes that are used for condition classification in the MEPS. Appendices 1 and 2 contain the two sets of condition codes, as well as the conversion map for CCC codes to CCCR codes.

These codes allow us to perform tasks like clustering with more interpretability, albeit at the cost of granularity with regards to the individual conditions, be it when we choose to use CCC codes instead of ICD codes, or CCCR codes instead of CCC codes.

# CHAPTER 4

# METHODOLOGY

The data is recorded on an event level during the survey, with each row corresponding to one event, and we aggregate the data by the individuals, taking the sum of the expenditure for each individual in the survey, and assigning condition codes to each individual. For each event, an individual may receive up to 4 diagnoses, which are assigned a code number. If no diagnosis is made, the code used is ‘-1’.

This patient grain data is then aggregated based on the unique groups of conditions each individual was diagnosed with. Individuals with no diagnoses, or insufficient information were assigned the code ‘ND’. The prevalence of each condition set was taken as the sum of the population weights of the individuals with that condition set. The expenses were aggregated by taking the total expenditures of the median individual with each group of conditions.

With this, we get the median expenses and total population weight associated with each unique group of conditions across the data.

**Figure 3**

*Data Flow with operations performed to transform the data.*

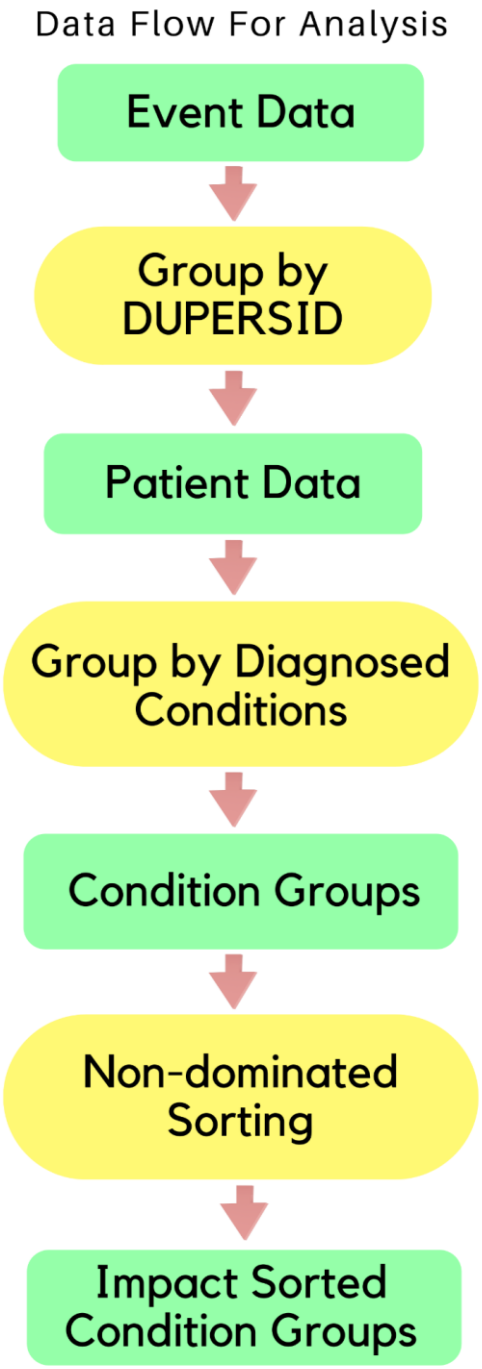
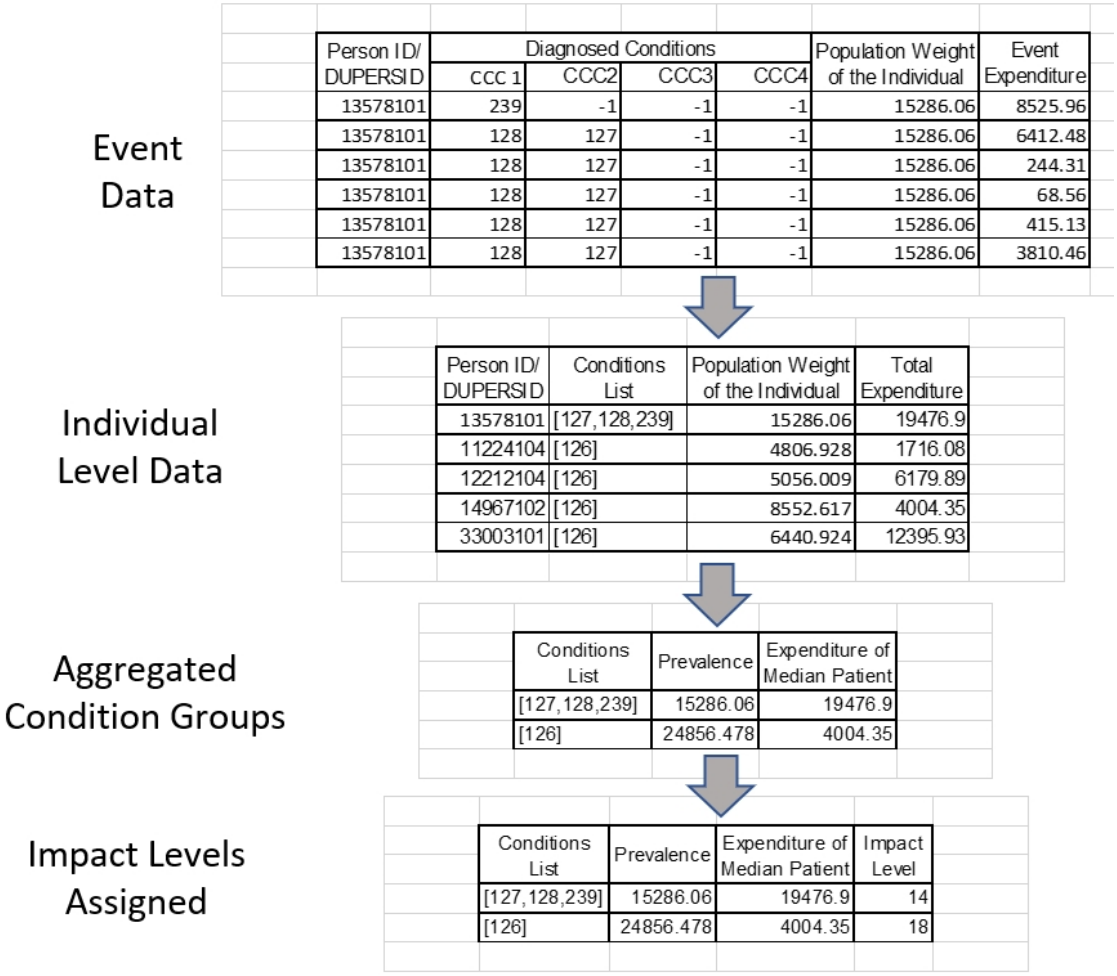


**Figure 4**

*An example of how event grain data is aggregated into condition group grain data.*




We now sort these condition sets into groups, ordered by their impact. Impact here is determined from the bicriteria optimization of the two variables, for each condition set. We take the prevalence of each set of conditions and the expenditures of a median patient presenting with them, and sort out such groups which are non-dominated in this distribution, thereby creating a Pareto front. In a bivariate distribution of points such as this, a point is considered to be non-dominated, or Pareto optimal, if none of the target variables — prevalence and

expenditure of the median patient in this case — can be increased or improved without decreasing or degrading the other.

Sets of condition groups which are non-dominated are hence determined to be those with the highest impact. The first such set of groups of conditions forms a Pareto front, and we can iteratively sort the remaining groups of conditions into subsequent impact levels based on this non-dominated sorting. The combinations of conditions are sorted into impact groups, numbered in order of the Pareto front they are a part of. Each impact group is hence a group of combination sets, which are non-dominated points on the expenditure vs prevalence distribution. This iterative sorting is performed using the ‘Skyline Operator’.

To determine the non-dominated fronts, we use an R implementation of the Skyline operation, as defined by Börzsönyi et al in their 2001 paper, from the package ‘rPref’. The operation uses a recursive sorting process to organize points into fronts that dominate the rest of the points in the dataset. In order to decide whether a point is a part of the dominating set or not, the algorithm chooses the tuple of points which has the maximum area covered by them in the direction of optimization. The package rPref has an implementation of the same algorithm, which was originally made for SQL, in R, and is used here to group the points into Pareto fronts, in order of their impact, defined as the result of this bicriteria optimization of prevalence and expenditure of the median.

The annual results will then be analyzed to study two things: first, identification of low prevalence high impact groups and investigation into the larger

systemic impact of these groups, and second, the trends in combinations of conditions over time, and whether the impact group sorting stays stable over the period.

With an aim to find patterns in the multimorbidity we use unsupervised learning techniques, viz. association rule mining and agglomerative hierarchical clustering.

For association rule mining, two approaches were taken, both using the apriori algorithm. The apriori algorithm is used to mine frequent itemsets and association rules, through the identification of items, or conditions in the case of our database, that occur frequently enough in the dataset as itemsets with one element. These sets of one can further be extended to larger and larger sets, provided they occur frequently enough in the dataset.

First, we define the items, or in our case the list of conditions as a set $C = \{c_1, c_2, \ldots, c_n\}$, where n is the number of conditions from the list of condition codes. From our initial analysis of the patient data, we have a list of itemsets, in our case the unique combinations of conditions $G = \{g_1, g_2, \ldots, g_m\}$. Here, each individual combination of conditions $g_i \subseteq C$. Using this, we can develop association rules, which are implications defined with an antecedent set X and a consequent set Y, as $X \Rightarrow Y$, where $X \subset C$, $Y \subset C$ and $X \cap Y = \emptyset$. These rules are interpreted through the measures – confidence, support and lift. A rule $X \Rightarrow Y$ holds with a confidence ‘c’, which is the ratio of combinations in G which contain $X \cup Y$ to the combinations in G which contain X. A rule’s support ‘s’, is defined by the frequency of $X \cup Y$ in

G. The lift of a rule ‘l’ is defined as the ratio of the observed support of the rule to the expected support if the two sets were completely independent.

Simply put:

$$Support\ (X)\ =\ P(X)\ = \frac{Itemsets\ containing\ X}{Total\ Itemsets}$$

$$Confidence\ (X \Rightarrow Y)\ =\ P(Y|X)\ = \frac{P(X \cup Y)}{P(X)} = \frac{Support(X \cup Y)}{Support(X)}$$
$$= \frac{Itemsets\ containing\ X \cup Y}{Itemsets\ containing\ X}$$

$$Lift\ (X\ \Rightarrow\ Y)\ = \frac{P(X \cup Y)}{P(X) \times P(Y)} = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)} = \frac{Confidence(X \Rightarrow Y)}{Support(Y)}$$

In the case of support and confidence, it can be straightforward to understand the utility of these measures. The support tells us the frequency at which a certain condition is fulfilled. The confidence tells us the probability of finding the consequent set in an itemset given that the itemset contains the antecedent set. The utility of the lift of a rule is in determining the dependence of the antecedent and consequent on each other. Unlike the confidence, which is independent of the dataset as a whole, the lift is a function of the number of itemsets. It can also be interpreted as the ratio of the confidence of the rule to the support of the consequent, i.e., the ratio of P(Y|X) and P(Y). This interpretation tells us that the lift helps understand whether a rule is misleading, even when the confidence of the rule is high. If the support of the consequent is higher than the confidence of the rule itself, then the rule may not indicate a strong association, as the high confidence stems from the high frequency of the consequent item in the

dataset. Therefore, the lift is useful as a measure as it adds more context to confidence measures. Looking at this from our understanding of conditional probability, what this means is that when the lift is equal to 1, the antecedent and consequent are independent of each other. Further, when the lift is greater than 1, it implies a positive correlation, with a higher value indicating a greater degree of dependence. Conversely, lift values that are lower than 1 indicate a negative correlation (Vu et al., 2019).

For the initial analysis, association rules were generated using the unique combinations of conditions scaled by their prevalence as itemsets, and determining the patterns in conditions that present together by identifying the frequent itemsets/condition groups in the dataset. This was done using the 'arules' package in R. This gives us a broad idea of the patterns of conditions; however, it is important to note that the algorithm is useful for understanding patterns in terms of prevalent sets of conditions of smaller size. While this information is certainly useful in the context of this study, we also would like to determine the patterns in which more complicated combinations of conditions with high impact present in relation to the more frequent combinations of conditions.

To this end, the apriori algorithm was used in conjunction with oversampling, using the impact level assigned to each unique combination of conditions to determine the degree of oversampling. Using the range of impact levels, a fuzzy set is created with a value between 0 and 1 assigned to each combination of conditions. This fuzzy value is then scaled to an integer value between 0 and 100 to determine how many times each combination of conditions

is sampled. For instance, if a combination is located in the first impact level or pareto frontier, it will be sampled 100 times for the apriori algorithm. If a combination is located in the lowest frontier, it will be sampled 1 time. This allows for higher sampling of combinations of conditions that are high in impact, even if low in prevalence. This allows such combinations to still be adequately represented in terms of the individual conditions that co-occur.

Therefore, to account for the impact of low prevalence high expense combinations of conditions, the unique condition sets were oversampled based on their impact levels. Condition sets present in higher pareto fronts were given a higher sampling rate to account for the larger impact they had in the system, due to the high expenses associated with presenting with those specific conditions.

Following this, we performed agglomerative hierarchical clustering to determine the similarity between individual conditions, so as to determine the conditions that show a greater degree of association with each other. To do so, the WPGMA (Weighted Pair Group Method with Arithmetic Mean) algorithm was implemented using the R package ‘hclust’, using the Sorensen–Dice similarity coefficient as the proximity measure for the clustering.

The Sorensen–Dice similarity coefficient is defined as:

$$SD = \frac{2 \cdot |X \cap Y|}{|X| + |Y|}$$

In our case, we have a list of unique combinations of conditions. To perform the above operation, we start by taking this list and binary encoding the condition

codes for each individual combination. Consequently, for each condition, we obtain a binary vector for each condition, populated by ones for each unique combination that contains the condition. This will give us *n* binary vectors for each condition $c_n$ of size *m*. For vector operations, the above equation can be expressed as:

$$s = \frac{2|x \cdot y|}{|x|^2 + |y|^2}$$

We perform this operation for each pair of conditions, which gives us a symmetric matrix of similarity values for each pair $S = \{s_{i,j}\}$. It is important to note here, this similarity matrix does not tell us that the conditions are necessarily similar to each other, but rather which conditions are more associated with each other, as what we are measuring here is the similarity coefficients of the binary vectors for the conditions as per the condition combinations that they are a part of. We then use this matrix to create a dissimilarity matrix $D_0 = \{d_{i,j}\}$, where each value $d_{i,j} = 1 - s_{i,j}$.

The WPGMA algorithm uses this dissimilarity matrix to construct an ultrametric dendrogram, where the distances (dissimilarity measures in our case) from each root node to the branch node are equal. This dissimilarity matrix is used to construct the dendrogram as:

i. First, we determine the two closest, or most similar conditions. This corresponds to the lowest dissimilarity value $d_{ij}$ from the matrix $D_0$. We connect these nodes to from a cluster $i \cup j$, corresponding to the conditions $i , j \in C$.

ii. The cluster thus formed, i ∪ j, is then used to recalculate the matrix $D_0$. The dissimilarity values for the next iteration, in the matrix $D_1$, are evaluated by using the pairwise mean of the average dissimilarity values. For instance, the dissimilarity value for a condition k with respect to i ∪ j would be calculated as $d_{(i\cup j),k} = \left(d_{i,k} + d_{j,k}\right)/2$.

iii. We perform the next iteration by repeating step 1 for $D_1$, joining new nodes or clusters accordingly, and creating subsequent dissimilarity matrices for each step. The dissimilarity values of each newly created matrix are used as the reference to select which nodes to join to form the cluster.

iv. The clustering thus produced is visualized using a rooted dendrogram.

# CHAPTER 5

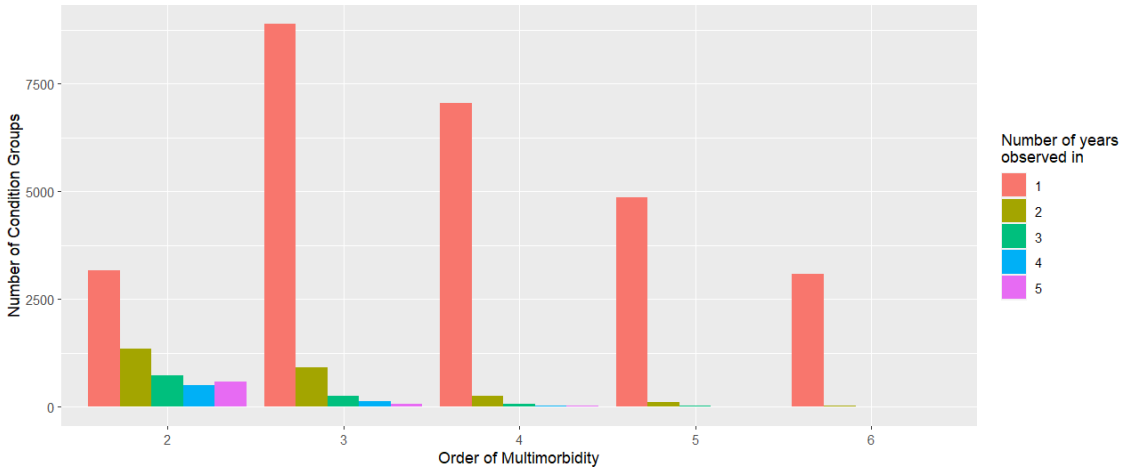
# RESULTS

Overall, across the 4 event types, 36,948 unique combinations of conditions were observed, using the Collapsed Clinical Condition (CCC) codes for the categorization of conditions. In comparison, using the Collapsed Clinical Condition - Refined (CCCR) codes, 23,183 unique combinations were found. This in and of itself is a relevant preliminary finding — the CCC codes represent around 260 individual codes that map to a specific condition, of which 210 to 230 were observed in each of the five years. The CCCR codes in comparison aggregate the CCC codes into 61 codes. Even after aggregating the clinical classification codes from over 250 conditions to just 61 codes, the number of unique combinations of conditions comes down by only about a third. This is a testament to the heterogeneity of the problem we are working on.

Of these unique combinations of conditions discovered, about 85% are combinations of up to 6 conditions, for both categorizations of the conditions. These conditions show a degree of repetitiveness over the period under study. Conditions with order of multimorbidity 5 or lower have appeared in all 5 years. However, the majority of the unique combinations of conditions have appeared in only 1 of the 5 years analyzed. The figures below show how many years the combinations of conditions with order of multimorbidity 6 and lower have repetitively appeared.

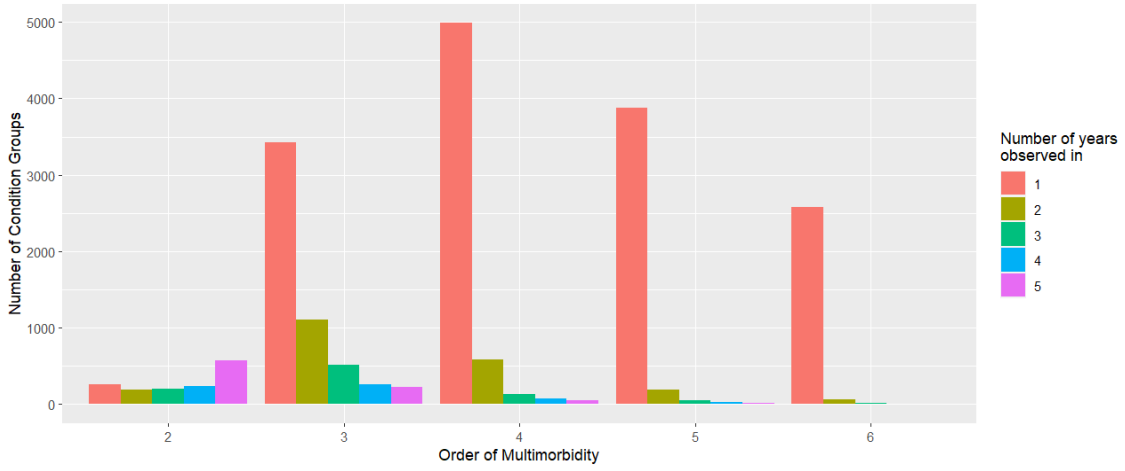
## Figure 5

*Continuity trends of unique combinations of conditions using CCC codes*




## Figure 6

*Continuity trends of unique combinations of conditions using CCCR codes*



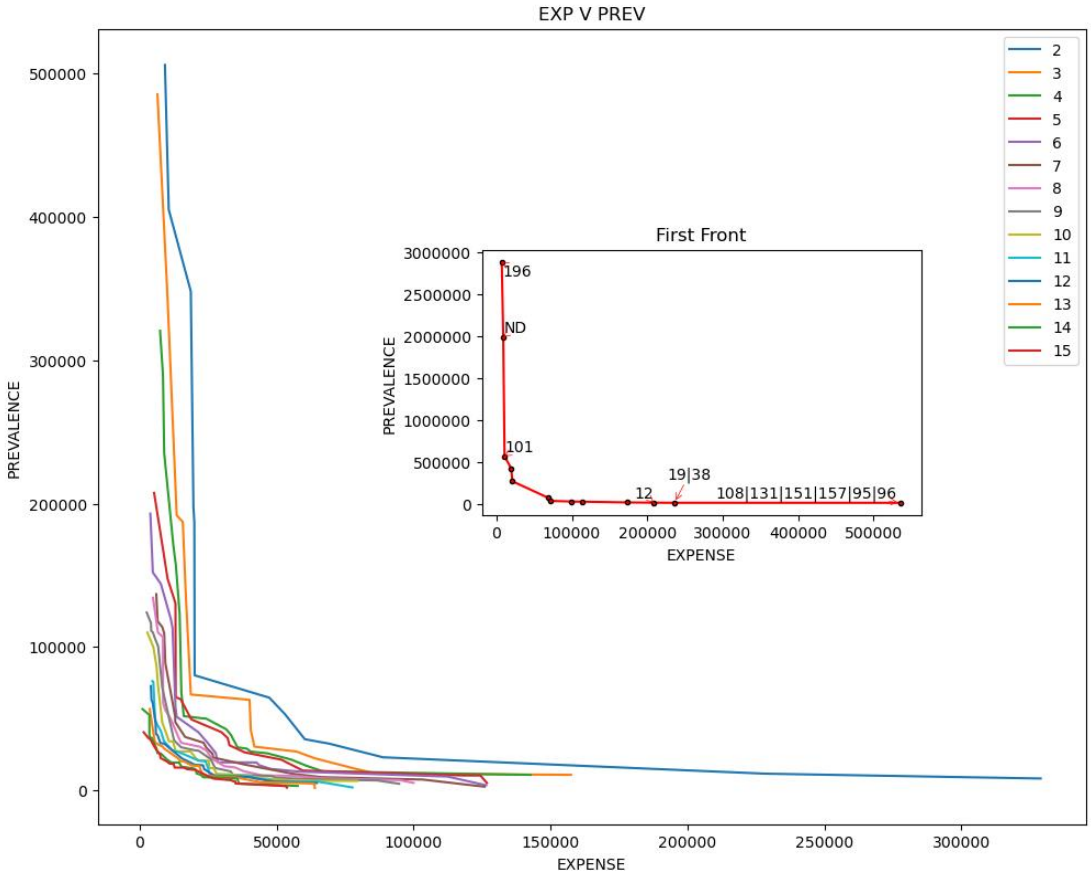
Using these unique combinations of conditions, we generated Pareto frontiers using the Skyline operation in R, using the package ‘rPref’. These frontiers allow us to classify the combinations on the basis of the impact they have on the system

As a result of how the individuals in the survey were assigned diagnoses, the analysis also includes people who received single diagnoses, and this is a sizable portion of the population as well. Looking at the distribution of the population weights versus the median expenses for each of the condition groups, we can see the fronts that get formed of various condition groups based on the level of impact they have on the system. For instance, these are the top 15 impact levels for condition groups of patients having inpatient visits in the year 2011.

The first front of disease groups that are non-dominated are shown separately, primarily due to the fact that individuals who have an undetermined diagnosis, and pregnancies completely skew the scale of the graph. However, we have chosen to show them regardless, as it is important to recognize that there is a bulk of such cases — ‘ND’ has a similar place in the bicriteria sorting of all of the datasets, while pregnancies are uniquely highly prevalent in the case of inpatient admissions.

**Figure 7**

*Expense vs Prevalence for Condition Groups, with impact levels/pareto fronts*




Similarly, codes pertaining to trauma-related codes, such as external injuries, fractures and wounds skew the scaling of the graph for emergency room events. Below are the conditions that are associated with the CCC codes shown in the first front of Figure 7:

**Table 1**

*Condition groups populating the first Pareto front or impact level, for inpatient stays surveyed in the year 2011.*

| Code List | Condition List | Prevalence | Expenses |
|---|---|---|---|
| 196 | Other pregnancy and delivery including normal | 2,883,667 | 7,113.48 |
| ND | Not Defined | 1,987,781 | 9,091.82 |
| 101 | Coronary atherosclerosis and other heart disease | 569641.9 | 10825.89 |
| 205 | Spondylosis; intervertebral disc disorders; other back problems | 424687.9 | 19553.13 |
| 230 | Fracture of lower limb | 270614.4 | 21253.37 |
| 203\|211 | Osteoarthritis \| Other connective tissue disease | 74694.67 | 68533.78 |
| 212 | Other bone disease and musculoskeletal deformities | 37481.19 | 71017.74 |
| 219\|224\|87 | Short gestation; low birth weight; and fetal growth retardation \| Other perinatal conditions \| Retinal detachments; defects; vascular occlusion; and retinopathy | 30256.65 | 99502.73 |
| 22 | Melanomas of skin | 28047.44 | 113340.8 |
| 106\|127\|133\|259\|59 | Cardiac dysrhythmias \| Chronic obstructive pulmonary disease and bronchiectasis \| Other lower respiratory disease \| Residual codes; unclassified \| Deficiency and other anemia | 19206.82 | 173493.8 |
| 12 | Cancer of esophagus | 16268.26 | 208517.4 |
| 19\|38 | Cancer of bronchus; lung \| Non-Hodgkin's lymphoma | 15032.31 | 235958.8 |
| 108\|131\|151\|157\|95\|96 | Congestive heart failure; nonhypertensive \| Respiratory failure; insufficiency; arrest (adult) \| Other liver diseases \| Acute and unspecified renal failure \| Other nervous system disorders \| Heart valve disorders | 14642.2 | 536194.7 |

Across the various types of events that were analyzed, there was a broad degree of multimorbidity observed. However, in terms of the categorization by impact, there were different patterns in the level of impact different combinations of conditions had in each event type.

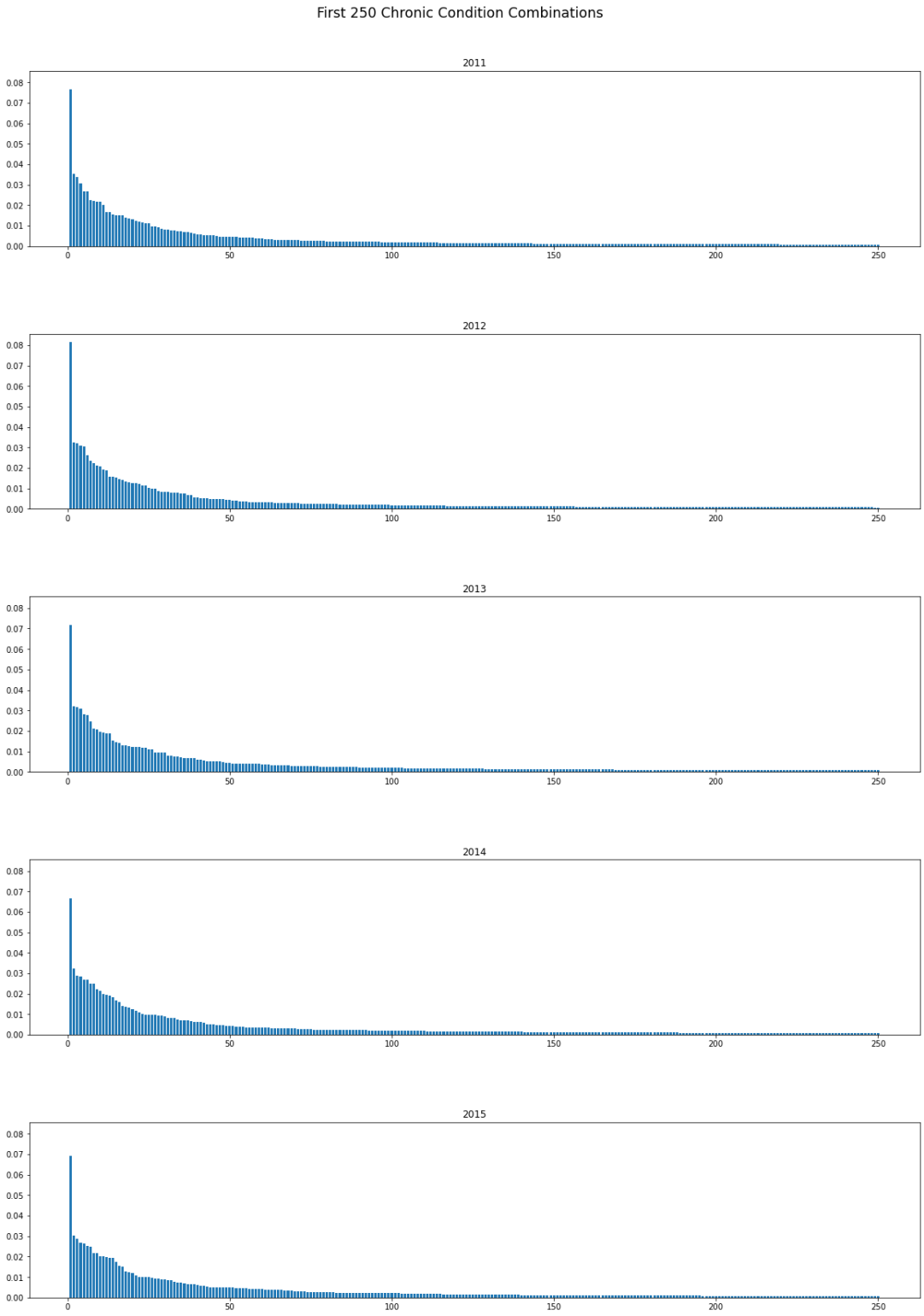
The impact levels are thus assigned as per the Pareto fronts of the distribution of the population weights and median expenses. Another observation that can quickly be made is that more complex forms of multimorbidity present near the right side, as the prevalence tends to be lower for these, while the expenditures are the primary driver of their impact.

It is important to note, that across the event types, the code ND remains one of the highest in prevalence, as the majority of events do lead to no diagnoses, or there may be a lack of sufficient information pertaining to whether the individual was diagnosed. As a result, it has a prevalence that is among the highest of all the groups. Similarly, the code 196 features among the highest in prevalence across all the event types, particularly office-based visits and inpatient visits, this is also as expected, since the code 196 pertains to pregnancies and normal births, which are routine across the board in the medical ecosystem.

In terms of the combined 5-year historical analysis of panels 15-19, there were varying patterns in the degree of impact across the four event types. Individuals with single conditions were highly prevalent in inpatient, outpatient and emergency type events, and high impact groups of conditions tended to usually be dyads and triads.

## Figure 8

*The Top 250 condition combinations in order of prevalence.*



A high order of multimorbidity was particularly associated with high impact despite relatively lower prevalence for office-based visits, with individuals having up to 12 comorbidities being present the highest Pareto front/impact level. While the higher impact of the complex cases in terms of the expense was seen more in office-based visits, there remained a larger aggregate impact from these complex cases with high order of multimorbidity. Individuals with 7 or more conditions make up around 10 million individuals, and approximately one third of these individuals lie in one of the upper impact fronts.

Inpatient visits and emergency events in particular were skewed by the first front, where events such as trauma-related injuries, fractures, pregnancies and so on, have extremely high prevalence due to their ubiquity. It may be pertinent to analyze the impact levels that come after it closer, especially in terms of what the expenses are for; some inpatient stays are routine while others can be disproportionately expensive, yet both cases represent the usage of the finite resource that is hospital beds.

When we look at the distribution of the prevalence of conditions over the years under analysis, shown here in Figure 8, we see a distribution much like that observed by Sorace et al., with smaller groups and individuals with single conditions having higher prevalence, but there being the presence of a long ‘tail’ of complex groups of multiple conditions. This ‘tail’ of low prevalence conditions is populated largely by dyads and triads, with the largest groups of combinations in the top 250 being around 6-8 conditions over the 5-year period. As seen in the condition groups shown previously, the groups with the highest complexity may

have lower prevalence relative to dyads or triads but have disproportionately high expenses associated with the kind of conditions they may have.

On the other hand, inpatient and outpatient event types tended to have combination sets of smaller sizes associated with higher impact. Only 2 condition groups of size > 3 were observed in the first front i.e., the highest impact level. In the case of office-based visits, there are 10 condition sets that have more than 3 conditions in the first impact level alone. This can be seen above in Figure 8, for the 250 most prevalent condition groups for office-based visits for each of the 5 years; approximately half of the condition sets shown here represent individuals in the population having multiple conditions.

Across the five-year horizon, there were, as expected, conditions and smaller groups which had high degrees of prevalence. Looking at the 5 dyads and triads with the highest impact (the 10 condition sets which presented in all 5 years under observation and had the lowest sum of impact level were selected here), we observe a relatively stable trend in terms of the prevalence, with a lot more fluctuation in terms of the expenses. These trends are visualized below, in Figures 9 and 10, with the condition groups in the legend being ordered by impact level.

**Figure 9**

*Prevalence and median expense trends for 10 of the highest impact dyads over the 5 years.*

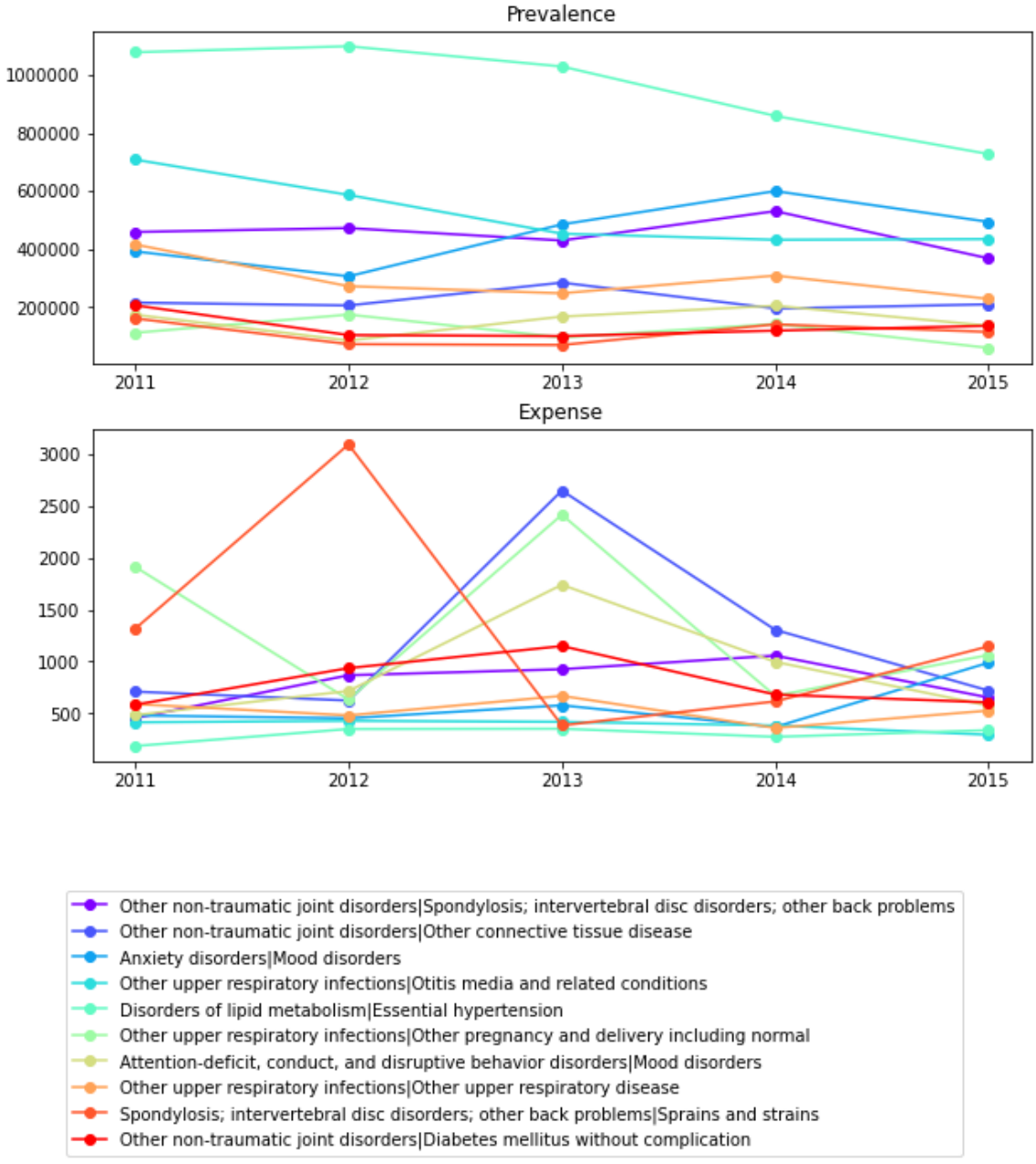
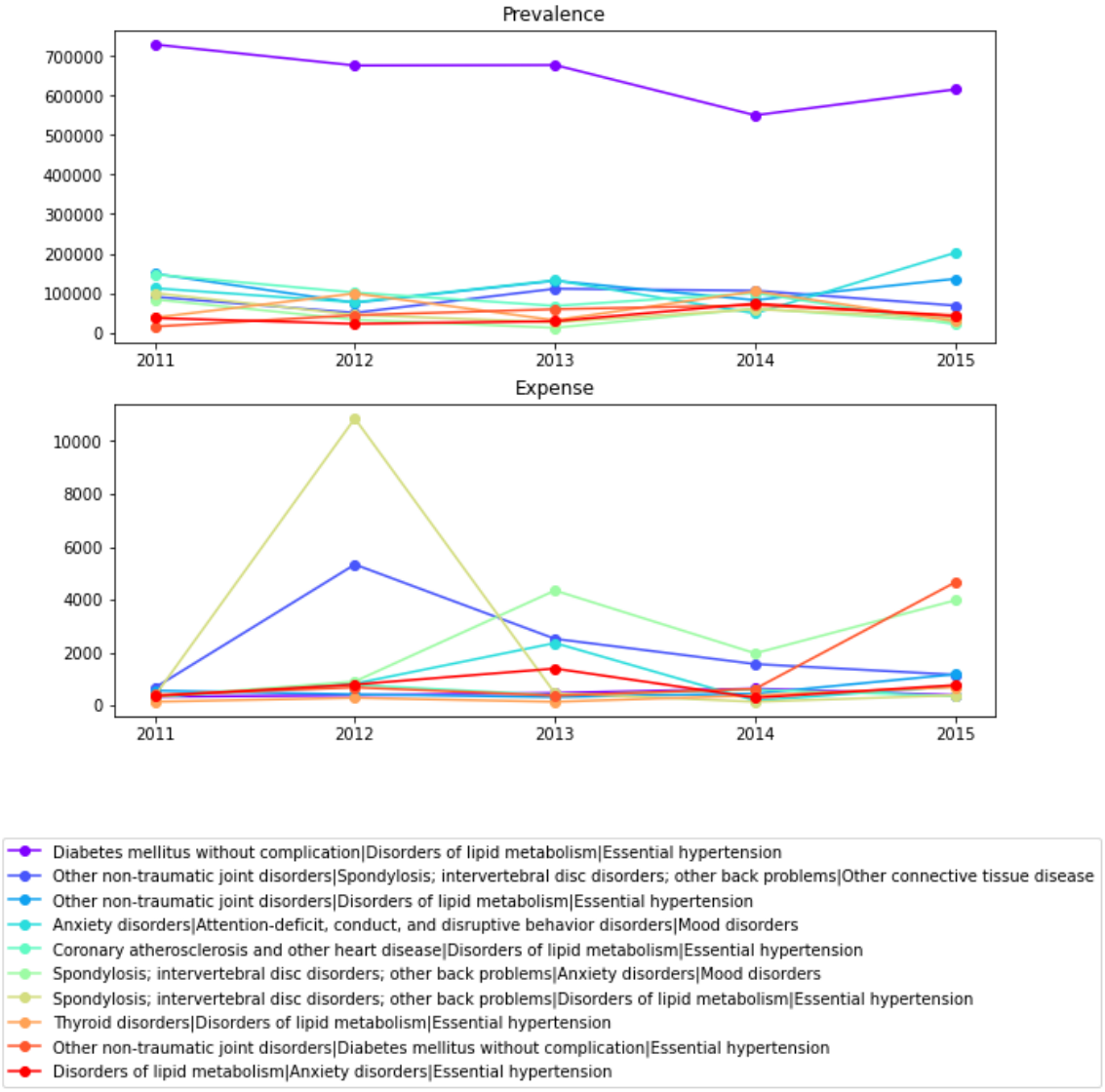


**Figure 10**

*Prevalence and median expense trends for 10 of the highest impact triads over the 5 years.*




Upon preliminary analysis of the dyads and triads, we can see very early on that as expected, there does tend to be a clustering around certain particularly high prevalence combinations such as lipidemia, hypertension, type 2 diabetes and heart disease, or joint disorders and other bone and back related problems, and so on.

These clusters tend to of course be extensively studied and are actively a part of public policy planning. Looking at some of the high impact groups of conditions (k>6), we see certain disease complexes that have particularly high expenses associated with them. This can be seen in the table below:

**Table 2**

*High impact condition groups (No. of conditions >6) observed over the period under study. These are groups of 6 or more conditions, which are present in the top Pareto front of the respective year that they were observed in.*

| Conditions List | Population Weights | Median Expense | Number of Conditions | Year |
|---|---|---|---|---|
| Skin and subcutaneous tissue infections \| Other inflammatory condition of skin \| Diabetes mellitus without complication \| Disorders of lipid metabolism \| Attention-deficit, conduct, and disruptive behavior disorders \| Mood disorders \| Schizophrenia and other psychotic disorders \| Essential hypertension | 28122.4 | 41448.59 | 8 | 2011 |
| Other upper respiratory disease \| Other non-traumatic joint disorders \| Spondylosis; intervertebral disc disorders; other back problems \| Other connective tissue disease \| Administrative/social admission \| Mood disorders \| Substance-related disorders \| Other nervous system disorders | 22458.31 | 77698.54 | 8 | 2011 |
| Other male genital disorders \| Skin and subcutaneous tissue infections \| Osteoarthritis \| Other non-traumatic joint disorders \| Cancer | 44650.28 | 13582.63 | 8 | 2011 |

| Conditions List | Population Weights | Median Expense | Number of Conditions | Year |
|---|---|---|---|---|
| of prostate \| Cancer of bladder \| Cancer of kidney and renal pelvis \| Essential hypertension | | | | |
| Coronary atherosclerosis and other heart disease \| Chronic obstructive pulmonary disease and bronchiectasis \| Asthma \| Esophageal disorders \| Other gastrointestinal disorders \| Heart valve disorders \| Essential hypertension | 43886.46 | 18829.3 | 7 | 2011 |
| Coronary atherosclerosis and other heart disease \| Other and ill-defined heart disease \| Acute cerebrovascular disease \| Other diseases of kidney and ureters \| Skin and subcutaneous tissue infections \| Other and unspecified benign neoplasm \| Diabetes mellitus without complication \| Other nutritional; endocrine; and metabolic disorders \| Essential hypertension | 22234.39 | 195295.2 | 9 | 2013 |
| Coronary atherosclerosis and other heart disease \| Phlebitis; thrombophlebitis and thromboembolism \| Esophageal disorders \| Other diseases of kidney and ureters \| Other non-traumatic joint disorders \| Fever of unknown origin \| Residual codes; unclassified \| Diabetes mellitus without complication \| Essential hypertension | 41217.71 | 18327.27 | 9 | 2013 |
| Other diseases of kidney and ureters \| Other skin disorders \| | 12257.66 | 128305.8 | 7 | 2014 |

| Conditions List | Population Weights | Median Expense | Number of Conditions | Year |
|---|---|---|---|---|
| Cancer of breast \| Thyroid disorders \| Disorders of lipid metabolism \| Cataract \| Essential hypertension | | | | |
| Coronary atherosclerosis and other heart disease \| Cardiac dysrhythmias \| Melanomas of skin \| Cancer; other and unspecified primary \| Malignant neoplasm without specification of site \| Thyroid disorders \| Diabetes mellitus without complication \| Disorders of lipid metabolism \| Cataract \| Essential hypertension | 15129.31 | 87503.46 | 10 | 2014 |
| Cardiac dysrhythmias \| Other lower respiratory disease \| Other upper respiratory disease \| Esophageal disorders \| Gastrointestinal hemorrhage \| Medical examination/evaluation \| Residual codes; unclassified \| Other endocrine disorders \| Deficiency and other anemia \| Essential hypertension | 4185.046 | 173102 | 10 | 2014 |
| Osteoarthritis \| Spondylosis; intervertebral disc disorders; other back problems \| Other connective tissue disease \| Other fractures \| Malignant neoplasm without specification of site \| Headache; including migraine \| Conditions associated with dizziness or vertigo \| Other ear and sense organ disorders \| Other nervous system disorders | 60677.08 | 14297.65 | 9 | 2014 |

| Conditions List | Population Weights | Median Expense | Number of Conditions | Year |
|---|---|---|---|---|
| Other gastrointestinal disorders \| Inflammatory diseases of female pelvic organs \| Other complications of birth; puerperium affecting management of mother \| Other pregnancy and delivery including normal \| Residual codes; unclassified \| Other endocrine disorders \| Developmental disorders \| Mood disorders \| Schizophrenia and other psychotic disorders \| Epilepsy; convulsions \| Blindness and vision defects \| Other eye disorders | 21672.89 | 51540.17 | 12 | 2015 |
| Other liver diseases \| Other diseases of kidney and ureters \| Skin and subcutaneous tissue infections \| Other connective tissue disease \| Residual codes; unclassified \| Anxiety disorders \| Mood disorders \| Retinal detachments; defects; vascular occlusion; and retinopathy | 26326.16 | 49203.61 | 8 | 2015 |
| Coronary atherosclerosis and other heart disease \| Cardiac dysrhythmias \| Bacterial infection; unspecified site \| Diabetes mellitus without complication \| Disorders of lipid metabolism \| Other infections; including parasitic \| Paralysis \| Essential hypertension | 2328.739 | 526743.7 | 8 | 2015 |

Over the 5 year period, the above table represents individuals who can, in total form a cohort of more than 300,000 individuals, all having a population

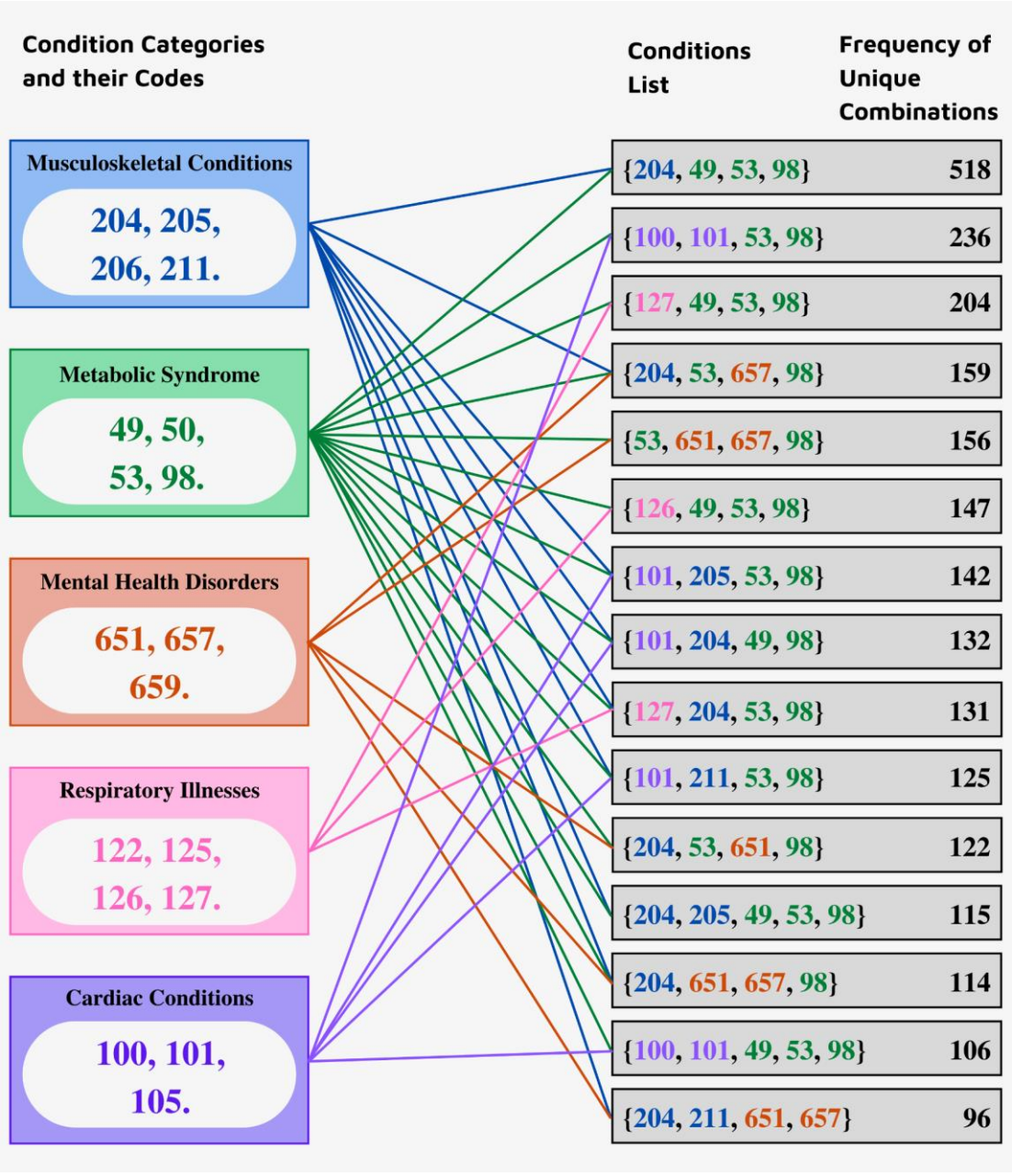
weighted annual expenditure of approximately $50,000. While these cases do not have prevalence values remotely close to that of the groups shown in figures 7 and 8, these are a set of individuals with high enough expenses to have such a high impact. The condition groups shown above are merely those condition complexes with the highest impact. In total, such combination groups with high order of multimorbidity number in the thousands. from 9,426,425 individuals in 2011 to 12,950,077 in 2015.

It is also worth noting that the dyads and triads of conditions that we have previously looked at frequently occur as subsets of the more complex cases. This means that individuals presenting with a high order of multimorbidity are often suffering from conditions from multiple commonly occurring dyads and triads. The most frequent such combination is that of hypertension (CCC code 98), diabetes mellitus (CCC code 49) and lipid disorders, or high cholesterol (CCC code 53). This triad, also commonly known as metabolic syndrome, occurs in conjunction with many others, and having an approach which details the way in which these conditions branch out to combine with other conditions allows us to qualitatively understand the relationships between these conditions.

For example, shown below in Figure 11 is a broad categorization of the condition codes in terms of the domain of illness they fall under. We can see these broad clusters of illnesses frequently presenting in conjunction with others in the condition groups as a whole, with a wide range of complexities. These individuals have a high degree of complexity of care, fragmentation of care, and a much-lowered quality of life as a result.

**Figure 11**

*Frequently occurring condition groups and the categories of healthcare they fall under. This is not an exhaustive list, but merely five of the most commonly observed categories of conditions. These conditions from various categories present in multiple combinations of each other, and the goal of the study is to quantitatively define and understand these relationships.*




This shows us that even with the high degree of heterogeneity, there are patterns in the way the conditions present in individuals with multimorbidity. The

CCC codes are themselves an aggregation of ICD-9 codes, however, the sparsity of the data can make it difficult to adequately develop associations between the conditions. Therefore, to understand the relationships and patterns in the conditions as they present in individuals with multimorbidity, it was more prudent to use the CCCR codes as the basis for the analysis. Furthermore, event types such as inpatient stays and emergency events tend to be skewed by specific condition codes, as previously mentioned, while office-based visits offer a large data sample combined with a higher variance in terms of the events themselves, giving us a richer understanding of the fragmentation of care such individuals experience.

Therefore, the analysis of the combinations of conditions using unsupervised learning methods was performed using the combinations that were created using the Collapsed Clinical Condition - Refined (CCCR) codes, with office-based visits as the focus, considering the combined data of all 5 years.

To determine potential patterns in the groups of conditions that are co-occurring in the population, we used two unsupervised learning methods, viz. Association Rule Mining using the Apriori algorithm and Hierarchical Agglomerative Clustering using the WPGMA algorithm.

For the generation of the association rules, threshold parameters in terms of support and confidence have to be selected. Any association rule that does not meet these criteria will not be considered. Given the sparsity of the data,

conservative thresholds were selected – the support threshold was 0.001, and the confidence threshold was 0.5.

Using the apriori algorithm on these unique combinations of conditions for association rule mining, there were a multitude of patterns that were seen in the frequency of the conditions that tended to co-occur. This corresponds to not just the high prevalence dyads and triads, but also the associations of these smaller groupings with each other. Based on these association rules that are generated, it is for us as the analysts to select conditions or selection criteria to determine rules that are interesting and worth looking into.

Given the lax thresholds that were set to generate the association rules, we get thousands of association rules that form complex networks of conditions. For mining these associations from the list of generated rules, it is necessary for us to set conditions and subsets to find information that is useful and worth studying and looking into. To this end, we set a baseline condition for the rules of interest to have lift > 1.5 and confidence > 0.6. This will return a subset of the generated rules where the associations being observed will have a strong positive correlation, with a strong likelihood associated with them. The visualization of the association rules is performed using the R package ‘arulesViz’. Even with the aforementioned conditions set to prune the overall list of association rules, the networks of associations can be extremely complicated, and involve a multitude of conditions from multiple different body systems. To better understand these networks, we used an inbuilt feature of the package to build an interactive network graph. Individual condition nodes and rule nodes can be selected to individually view their

associations and relevant conditions respectively. An example of a simple case of such a network graph is shown in the below, in Figures 12 and 13.

This is a simple network of associations, where the pruned association rules mentioned above are subset as per the presence of one of the codes pertaining to metabolic syndrome – hypertension (22), hyperlipidemia (7) and diabetes mellitus (5) – ordered by support.

The rules, colored in shades of red, follow a shading gradient according to the support of the rule i.e., the redder the node, the higher the support for the rule relative to the rules in the network. The blue rectangular nodes represent the conditions, and arrows in the network are directed to and from the rules depending on whether they are the antecedent or the consequent with regards to the rule.

**Figure 12**

*Top 10 association rules where at least one metabolic condition is present, ordered by support. The sampling for the generation of the association rules was done on the basis of the prevalences of each unique combination of conditions.*

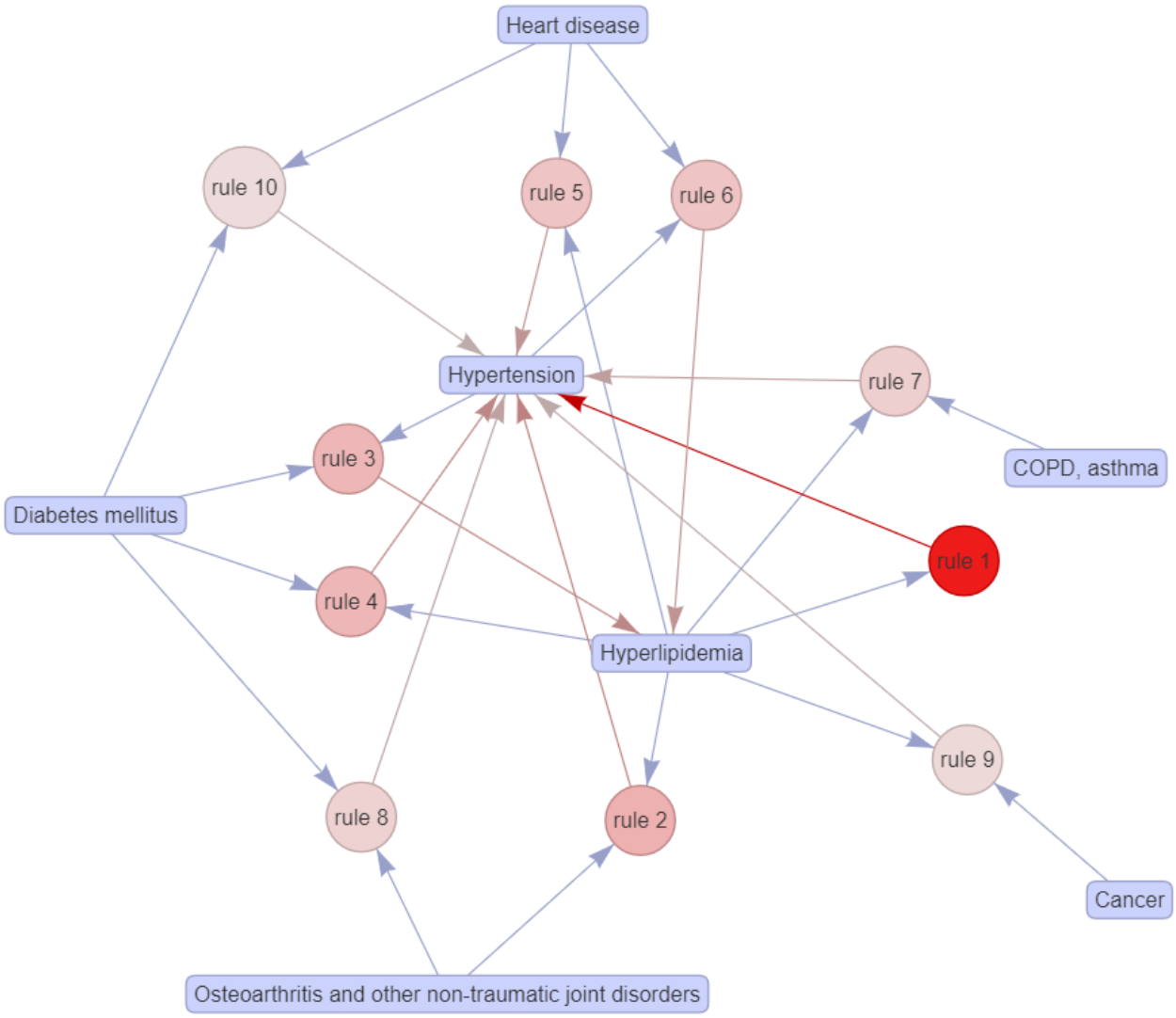
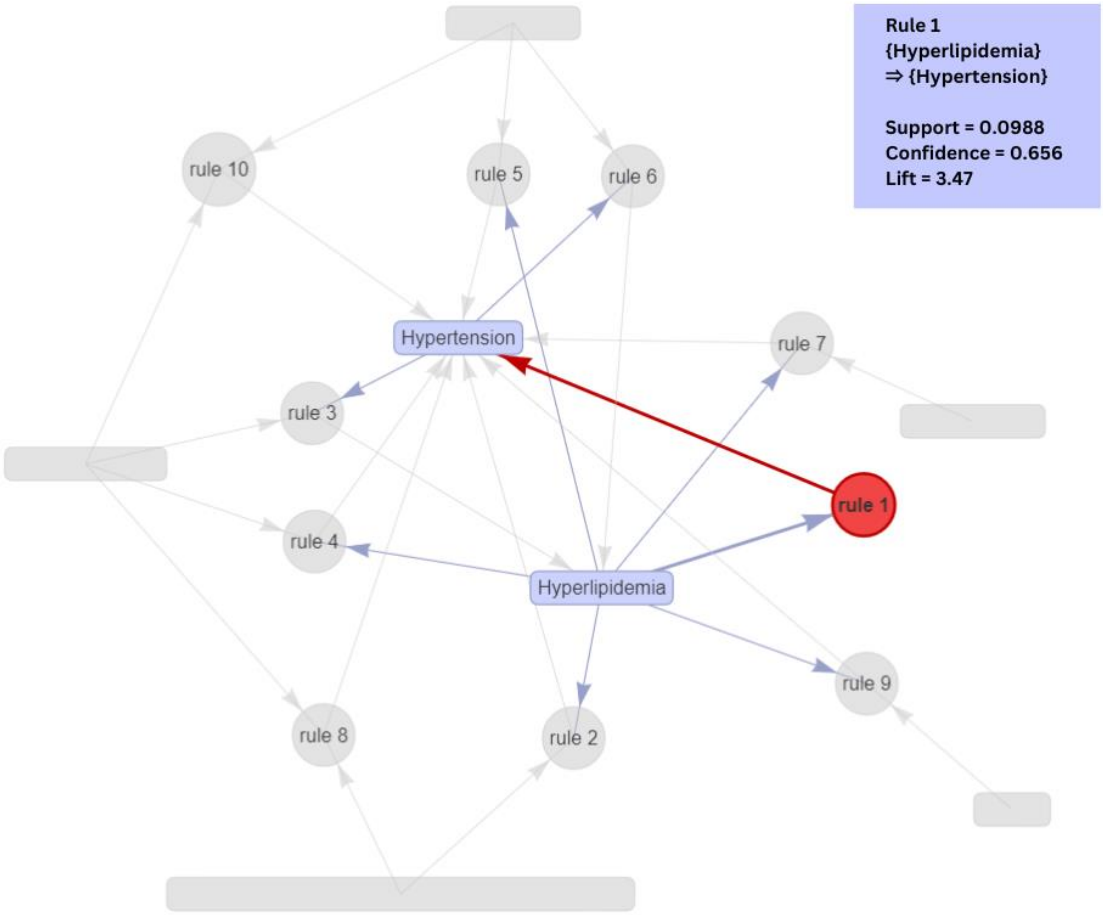


**Figure 13**

*Rule 1 from Figure 12. The association rule is highlighted, and the relevant measures are shown. The directions of the arrows indicate the antecedent and the consequent.*




Nodes can be highlighted to view more information about them as shown above. This allows us to study individually the rules that make up these networks. This is particularly useful when we wish to study more complex networks and understand the relationships between the conditions in more complex cases with higher orders of multimorbidity. For instance, in the above example, rule 1 shows the highest support, at 0.0988, meaning that 9.88% of the population has both these conditions. It also shows that there is a confidence of 0.656, meaning that

65% of individuals with hyperlipidemia also have hypertension. The lift of 3.47 shows a strong positive correlation between the two conditions.

Ordering the rules by support is a useful tool for validation, as it essentially selects rules preferentially on the basis of prevalence. However, merely selecting the top rules on the basis of support is unlikely to give us any interesting associations, or any new information. Ordering by lift, however, is an ideal and straightforward way to mine interesting rules, as the selection criterion is on the basis of degree of correlation, rather than prevalence.

Therefore, to mine interesting association rules and discover useful associations between conditions, we subset the pruned list of associations on the basis of various groupings of condition codes. These groupings are based on different affected body systems, such as metabolic conditions, musculoskeletal disorders, respiratory conditions and mental health disorders.

This analysis was performed using the unique combinations of conditions as itemsets, as previously mentioned, with 2 different types of sampling of the frequency of the itemsets. In the first stage, the itemsets were sampled in terms of the prevalence, while in the second stage the items were sampled in terms of the impact level the combination of conditions belongs to.

The next 4 pages contain interesting associations selected from various groupings of the association rules based on the presence of condition codes from the 4 previously mentioned body systems. These rules are generated using sampling of the itemsets on the basis of the prevalence.

Following this, the subsequent 4 pages contain similarly interesting associations, selected using the same criteria. However, here the sampling of the itemsets for the association rule generation is on the basis of the impact level of each individual combination of conditions. This leads to the generation of a different set of rules, showing us associations that would otherwise be less likely to be found when sampling according to the prevalence.

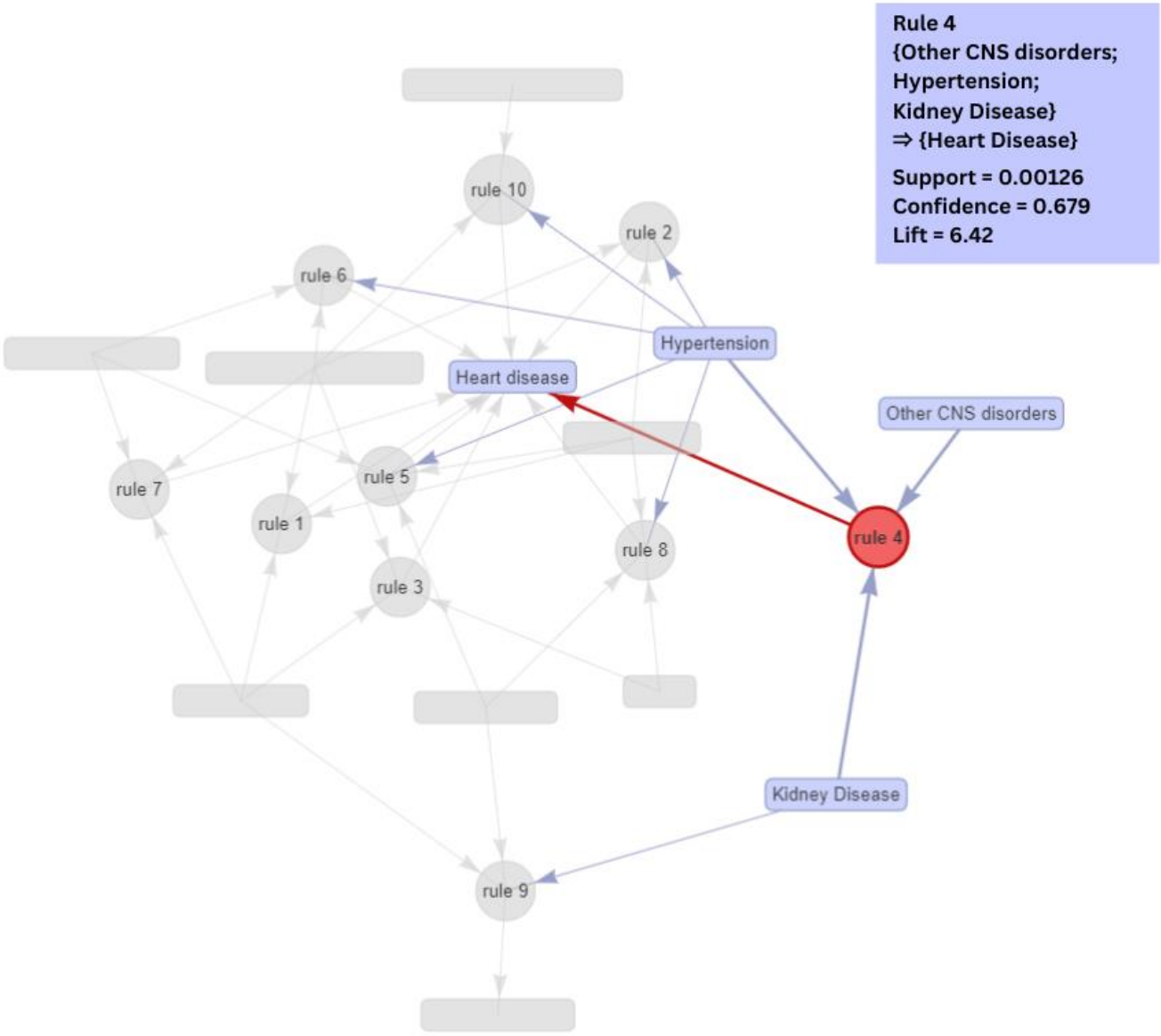



**Figure 14**

*A selected Association Rule. The rules in this network are the top 10 rules where at least one metabolic condition is present, ordered by lift. The sampling for the generation of the association rules was done on the basis of the prevalences of each unique combination of conditions.*

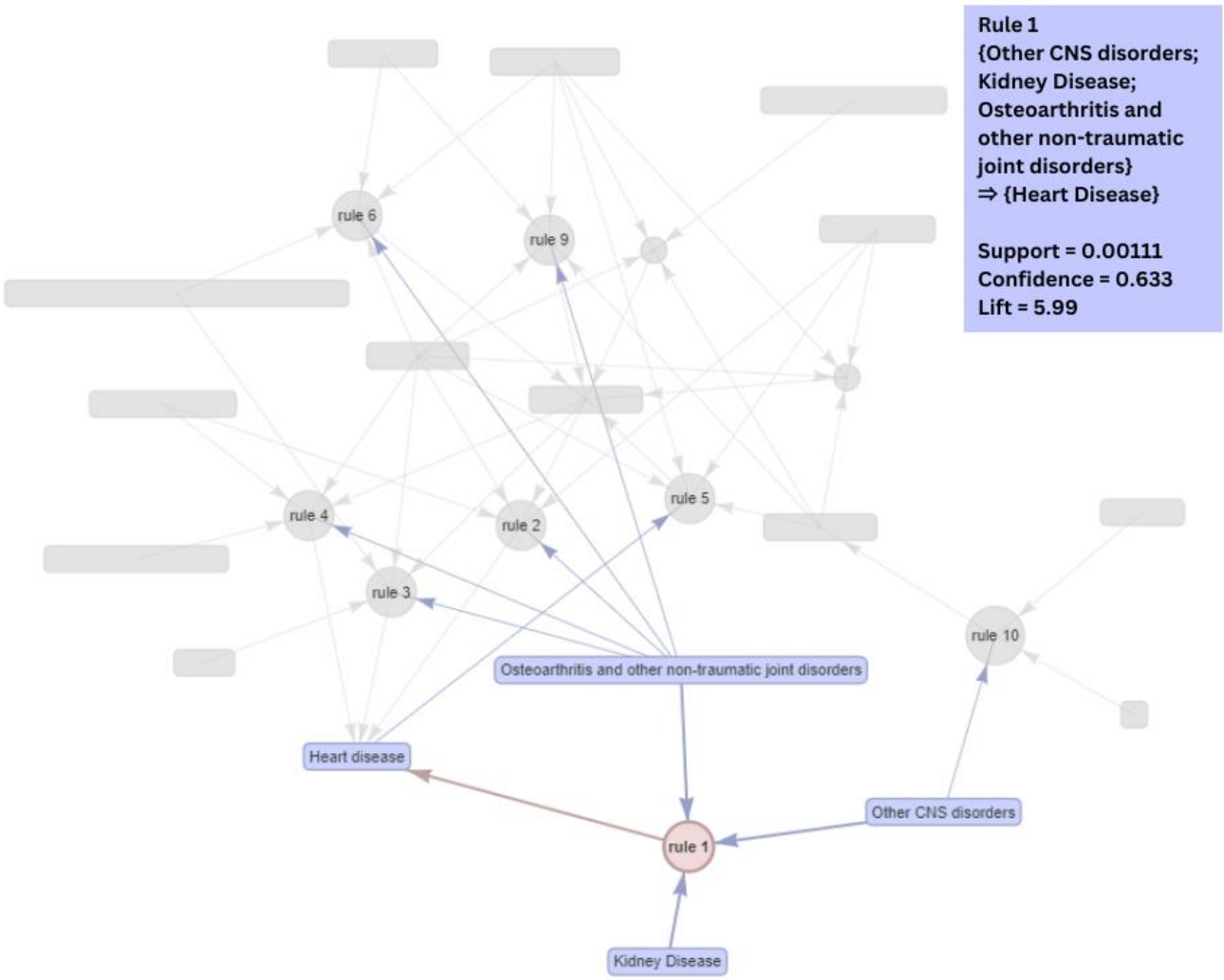



**Figure 15**

*A selected Association Rule. The rules in this network are the top 10 rules where at least one musculoskeletal condition is present, ordered by lift. The sampling for the generation of the association rules was done on the basis of the prevalences of each unique combination of conditions.*

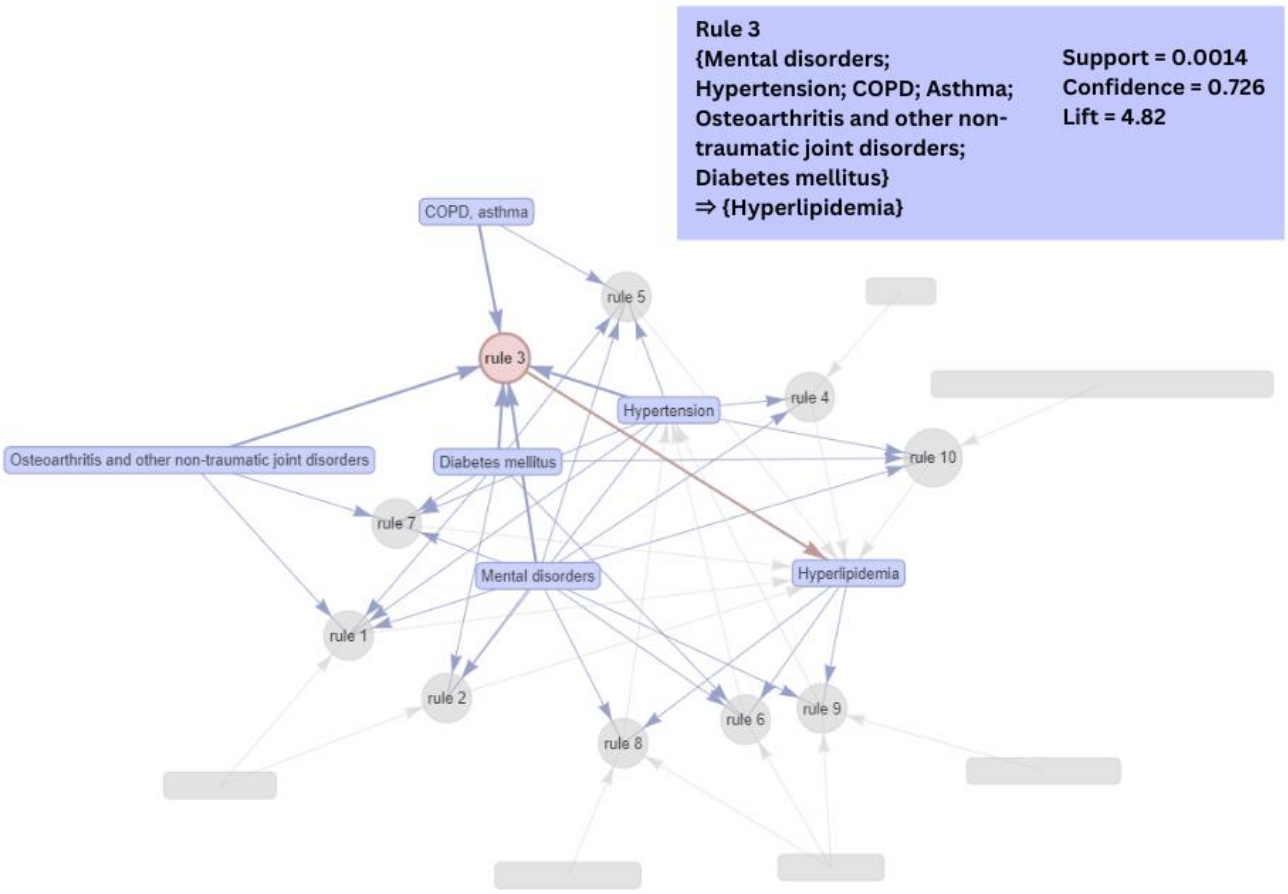



**Figure 16**

*A selected Association Rule. The rules in this network are the top 10 rules where mental disorders are present, ordered by lift. The sampling for the generation of the association rules was done on the basis of the prevalences of each unique combination of conditions.*

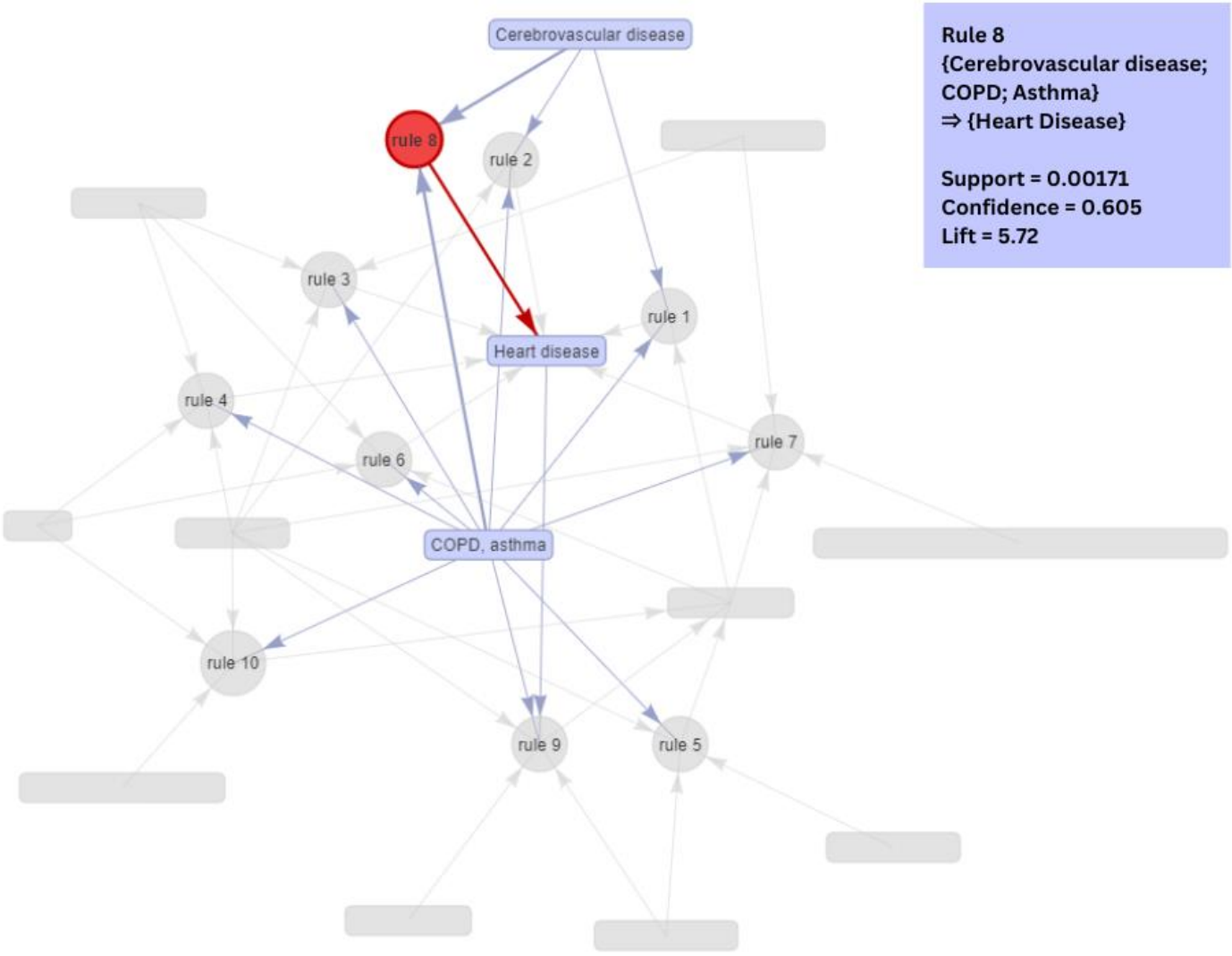



**Figure 17**

*A selected Association Rule. The rules in this network are the top 10 rules where at least one respiratory condition is present, ordered by lift. The sampling for the generation of the association rules was done on the basis of the prevalences of each unique combination of conditions.*

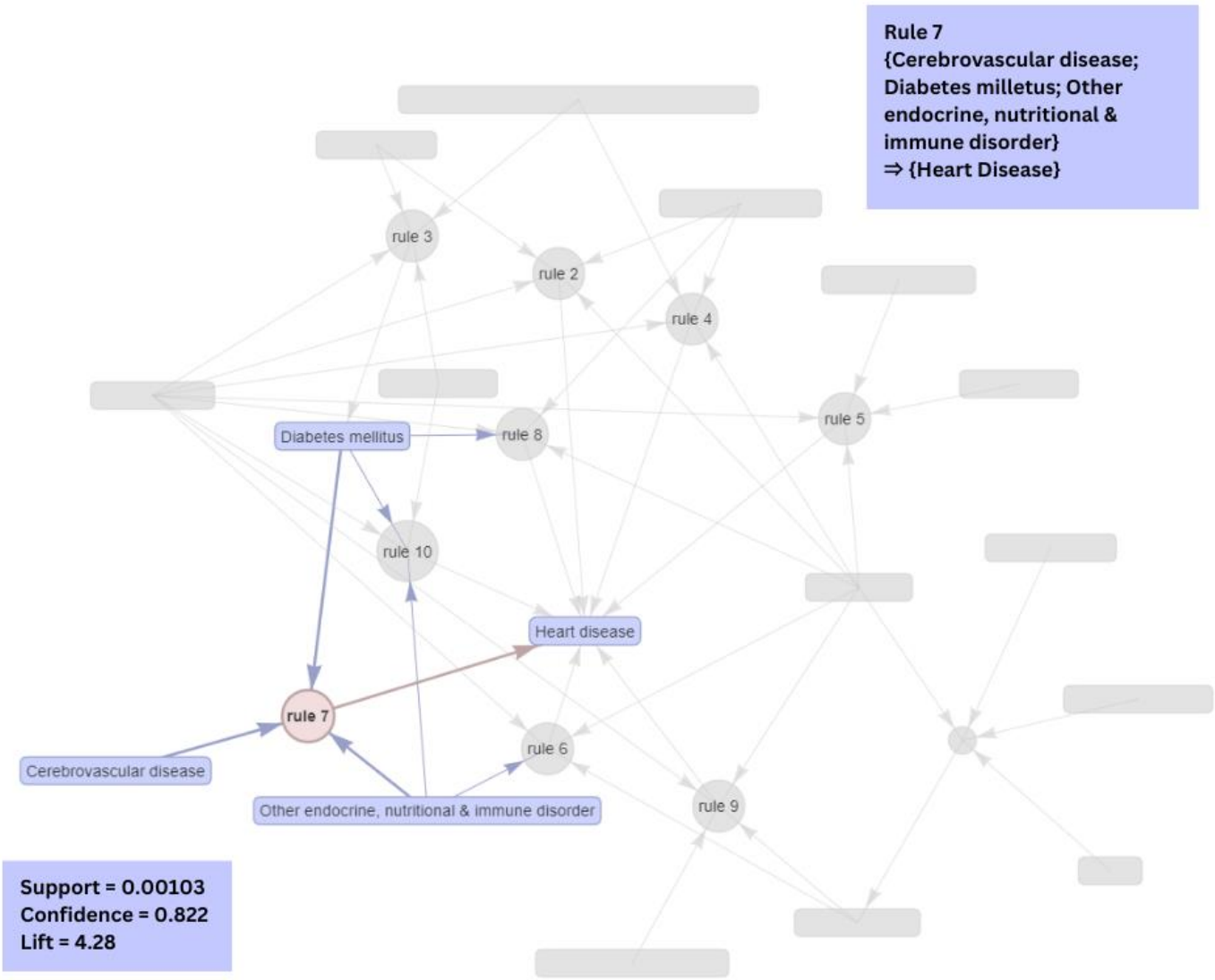



**Figure 18**

*A selected Association Rule. The rules in this network are the top 10 rules where at least one metabolic condition is present, ordered by lift. The sampling for the generation of the association rules was done on the basis of the impact level of each unique combination of conditions.*

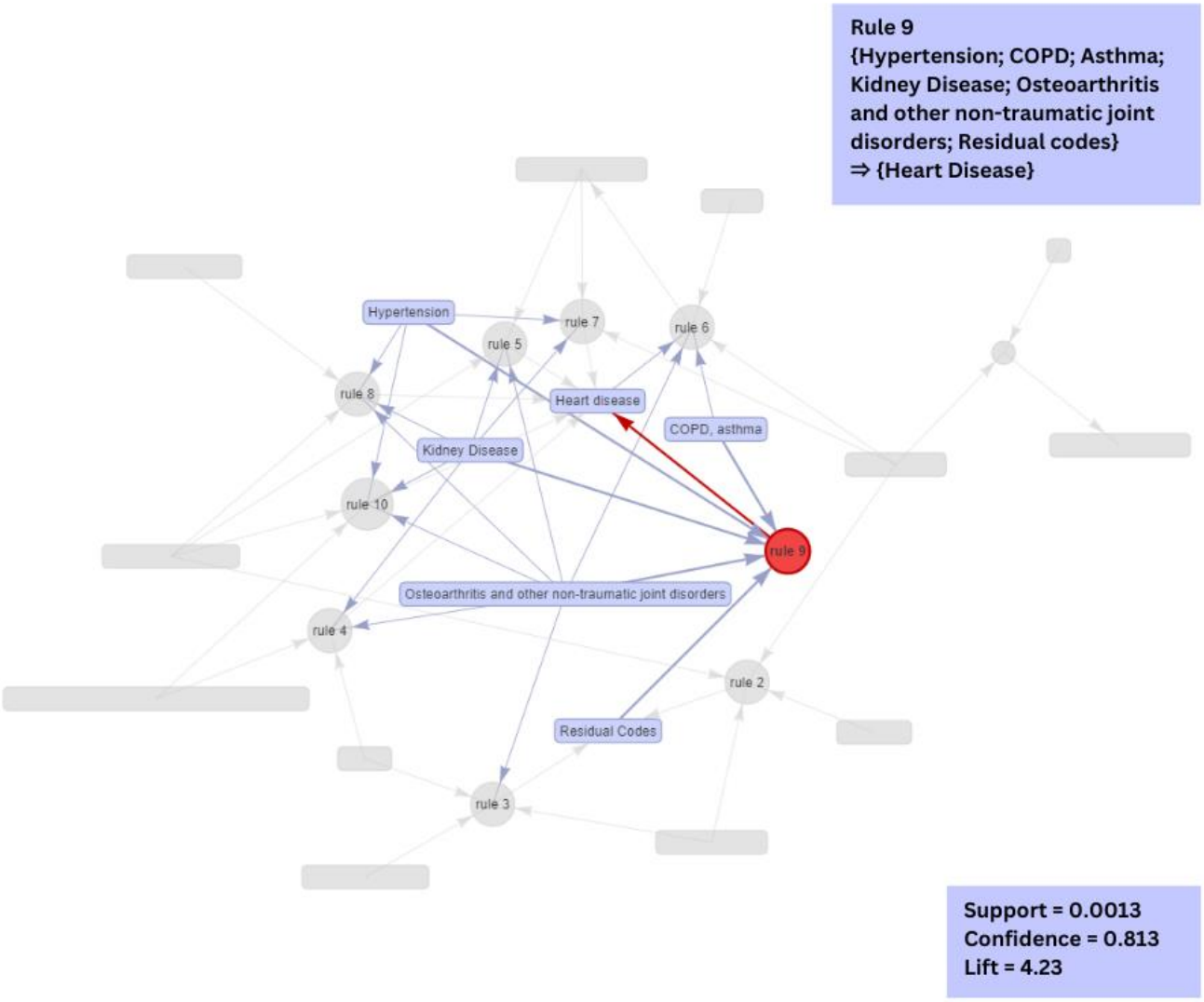



**Figure 19**

*A selected Association Rule. The rules in this network are the top 10 rules where at least one musculoskeletal condition is present, ordered by lift. The sampling for the generation of the association rules was done on the basis of the impact level of each unique combination of conditions.*

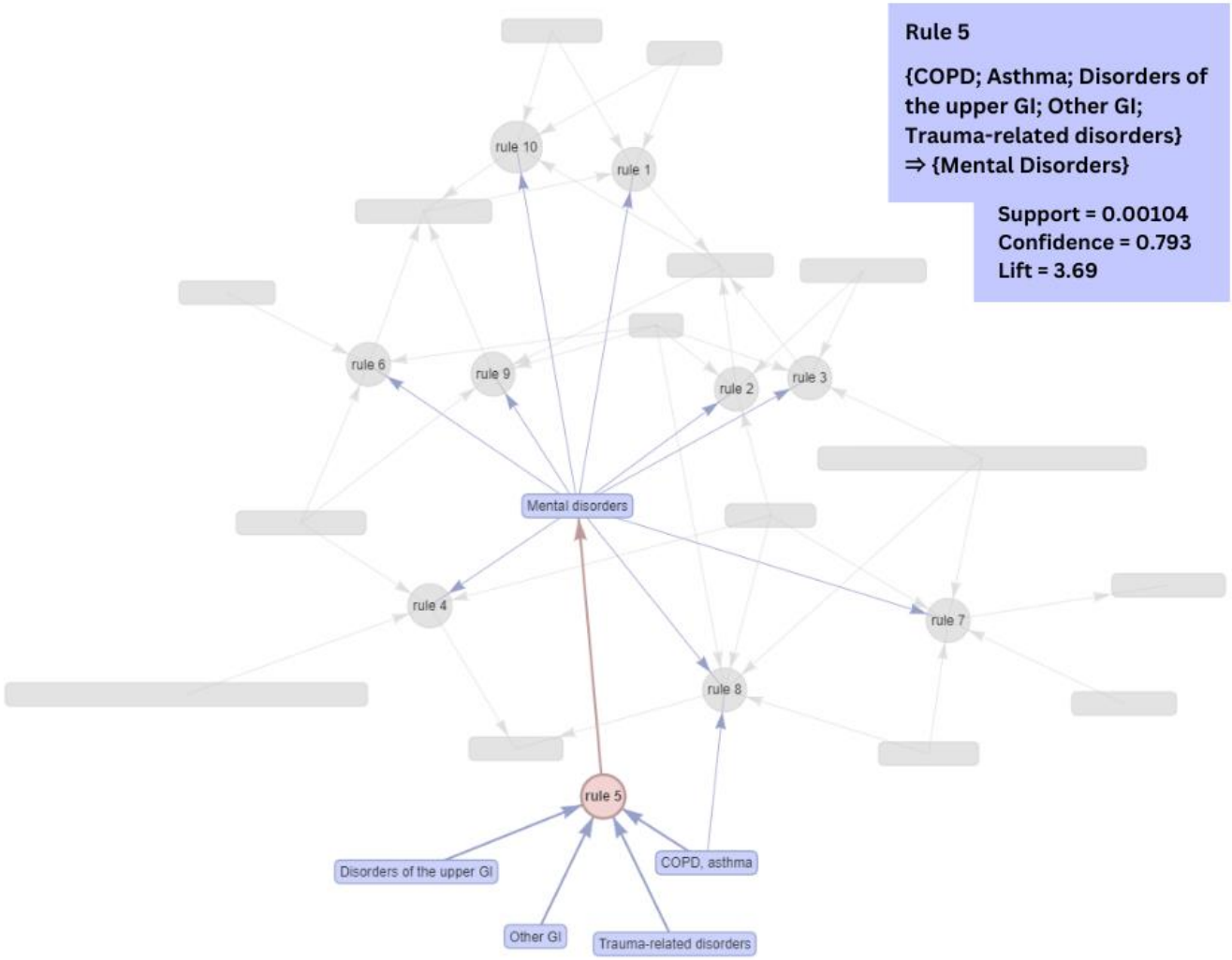



**Figure 20**

*A selected Association Rule. The rules in this network are the top 10 rules where mental disorders are present, ordered by lift. The sampling for the generation of the association rules was done on the basis of the impact level of each unique combination of conditions*

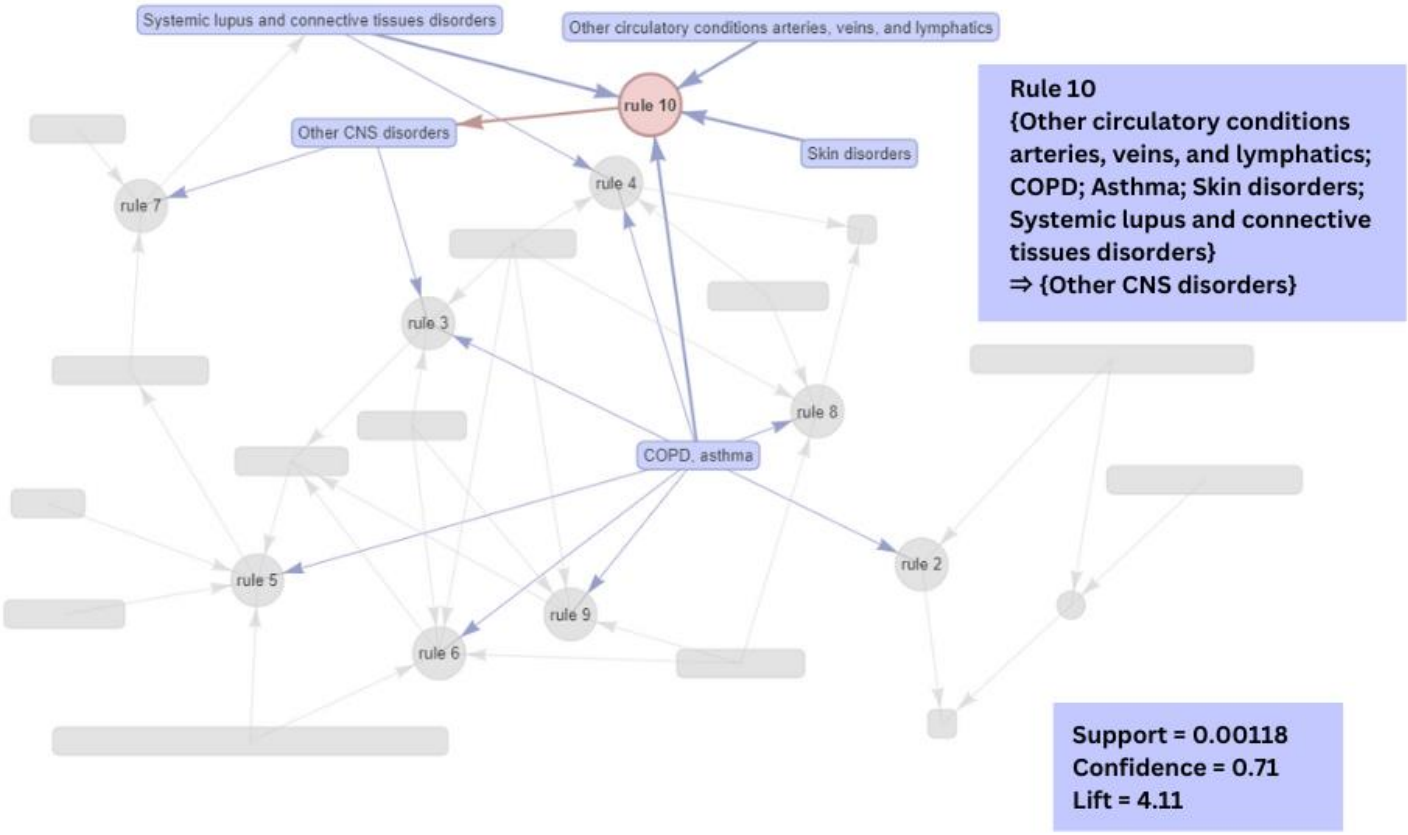


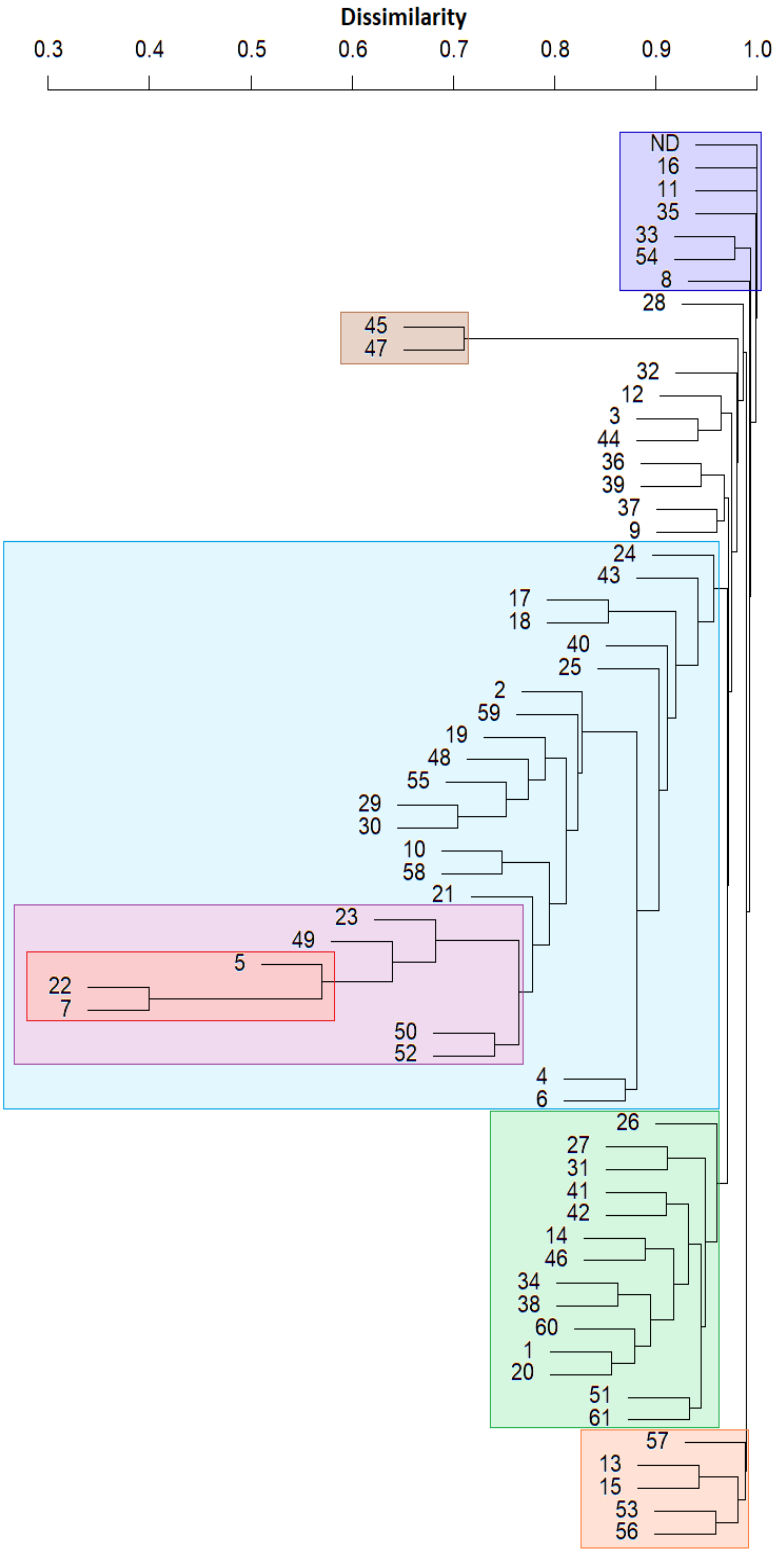

**Figure 21**

*A selected Association Rule. The rules in this network are the top 10 rules where at least one respiratory condition is present, ordered by lift. The sampling for the generation of the association rules was done on the basis of the impact level of each unique combination of conditions.*

Alongside the association rule mining, we also performed hierarchical agglomerative clustering using the Weighted Pair Group Method with Arithmetic Mean (WPGMA) algorithm, using the Sorensen-Dice coefficient as the dissimilarity measure. The distances or dissimilarities between each of the 61 condition codes was evaluated using the unique combinations of conditions observed over the 5-year period. The clustering has been visualized using a rooted dendrogram, and can be seen in Figure 22. This dendrogram shows the associations between the various conditions based on how often they tend to be comorbid with each other. A lower dissimilarity, or a higher similarity score implies that the conditions are more likely to be present together. Due to the sparsity of the data, most of the conditions are in the dissimilarity range of 0.8 to 1.0. However, the clusters in this range are merely reflective of the relative rarity of these conditions themselves, as compared to some of the other codes which are more prevalent.

From this dendrogram, what we observe is that most of the conditions fall into one of 4 clusters. These four clusters are marked in the dendrogram in brown, blue, green and orange. Apart from these, there is a group of conditions that are largely independent of the rest, and typically appear as singletons, or extremely rarely with some other condition.

**Figure 22**
*Rooted dendrogram showing the clustering from the WPGMA algorithm.*



The group of conditions marked in purple in Figure 22, primarily consists of conditions which have their dissimilarity score equal to or almost equal to 1. This essentially means that these conditions are largely independent of the rest of the conditions. In the case of the codes ND (No Diagnosis), 11 (CNS Infection) and 16 (Coma, brain damage), the dissimilarity score is 1.0, meaning that these codes have not been observed in a multimorbid complex in all of the data. The rest of the codes are largely independent, however, there are rare instances where they are comorbid with other conditions, as indicated by the dissimilarity value, which are almost equal to 1.0.

The brown cluster corresponds to codes related to pregnancy. These condition codes, 45 (Complications of pregnancy and birth) and 47 (Normal birth/live born), when presenting as a case of multimorbidity almost exclusively occur as a dyad of these codes. The dissimilarity values for this cluster are rather low, and this indicates a strong association between these two codes.

The orange cluster corresponds to codes that include epilepsy, paralysis and congenital anomalies. These condition codes seem unrelated to each other; however, they are in a completely separate cluster from all the other conditions. They are associated with a high dissimilarity score overall, meaning that these conditions are more likely than not to be diagnosed as a single condition, however, when they do show comorbidity, it is with the other conditions in this cluster. The heterogeneity of these codes does indicate that it might warrant further investigation.

The majority of the conditions, however, fall into one of two clusters, which are marked in blue and green in the dendrogram. The blue cluster primarily consists of chronic conditions, while the green cluster primarily consists of infections pertaining to the GI system and the respiratory system, and other associated conditions.

The chronic condition cluster consists of a variety of conditions that are frequently comorbid with each other, with the strongest association being between the codes pertaining to the triad of metabolic conditions – hypertension (22), hyperlipidemia (7) and diabetes mellitus (5) – and a moderately strong association with heart disease (23), and musculoskeletal conditions like osteoarthritis (49) and back problems like spondylosis (50). These associations can be considered subclusters of the larger chronic condition cluster due to the large difference in dissimilarity values compared to the rest of the cluster. This cluster also contains other noteworthy chronic conditions that are often seen in complex cases of multimorbidity, such as thyroid (4) and other endocrine disorders (6), bronchitis and upper respiratory infections (29), COPD and asthma (30) and kidney disease (40). It also is worth noting that the code pertaining to mental health disorders (10) is also in this cluster, and shows a strong association with the condition codes in the subcluster of heart disease, musculoskeletal conditions and metabolic disorders.

The infectious disease cluster contains multiple condition codes that pertain to infectious disease. These codes include aggregate codes such as infectious diseases (1), which itself is an aggregation of codes for diseases in the CCC

classification such as tuberculosis, hepatitis, viral infections and so on. The cluster also contains condition codes such as influenza (27), intestinal infection (31), urinary tract infections (41), GI disorders (34 and 38) and so on. Included in this cluster are also codes which are often given as diagnoses along the course of specifically diagnosing such infections, such as headache (14) and symptoms (60), which aggregates CCC codes for conditions like fever, nausea and vomiting, abdominal pain and so on, which are also associated with infections.

The remaining conditions have a loose association with the pregnancy, infectious disease and chronic condition clusters, but do not have the same degree of association as the members of the clusters do with each other. Their association is in fact with the supercluster that could be formed by merging these clusters together, as they stem from the same set of branching points of the rooted dendrogram. What this means is that these conditions are found comorbid with conditions from all three clusters and cannot necessarily be classified into one of the three clusters specifically. These conditions include codes for disorders of teeth and jaws (32), gallbladder, pancreatic and liver disease (39), other stomach and intestinal disorders (37) and so on.

# CHAPTER 6

# DISCUSSION OF RESULTS

Based on the aggregation of diagnoses of the individuals from the MEPS, we found 36,948 unique combinations using the CCC codes and 23,183 unique combinations using the CCCR codes over the years 2011-2015. These combinations had a varying degree of repetitiveness, but it is important to note that the data we are using for this study come from panel surveys that are conducted over two years, so it is difficult to gauge long term trends with the highest degree of accuracy on an individual level. Through our analysis we were also able to confirm the fact that the multimorbid cohort of the population is a growing demographic, over the 5-year period. We have also observed a large number of complex cases with a high order of multimorbidity – individuals with 7 or more conditions make up around 10 million individuals every year, and approximately one third of these individuals lie in one of the upper impact fronts.

We have also observed a great degree of heterogeneity over the course of this analysis, particularly when classifying the conditions in a more segmented manner, such as with CCC codes. This points to the challenges associated with such a study, which attempts to take a more holistic approach towards multimorbidity. The majority of combinations of conditions do not repeat year on year, and even in the case of the combinations of conditions which do show repetitiveness, there can often be significant variance in the expenses associated with their treatment. This variance could be caused by a variety of factors, ranging from changes in treatment norms, changes in prices for medication, development

of new medication and so on. What we have learned, however, is that to understand this heterogeneity, the experience and domain knowledge of physicians from multiple specialties needs to be combined with such a data mining approach. This will allow physicians to have a structured way to work with the many different high impact combinations of conditions that are made up of smaller complexes of conditions.

From the bicriteria analysis, we were able to categorize the various unique combinations of conditions into impact levels, giving us a multidimensional perspective with regards to the impact each of these combinations has on the system. This allowed us to not just have an idea of which complex cases of multimorbidity were worth exploring in more detail, but also gave us a means to approach the association rules mining in a manner that is not wholly dependent on the prevalence. With the approaches taken using association rule mining, we were able to develop a methodology for the analysis of such heterogeneous data, albeit rudimentary given the amount of data and its relative sparsity.

Future work on such a framework could use data sourced from healthcare facilities or administrations such as the Veterans Health Organization or Medicare, which service individuals who are more at risk of developing multiple conditions over their healthcare horizon, over longer periods of time. This allows such a framework to develop more concrete associations. The patterns observed through utilizing this framework of analysis across different domains such as the ones previously mentioned, or through comparing larger datasets that are

demographically segmented would also be a useful way to understand the different ways in which multimorbidity presents itself in different sections of the population.

There is also a great potential for such a framework to be used in conjunction with the knowledge that medical professionals have on their patients who present with complex multimorbidity. Using longitudinal data over longer periods to understand the progressions of the case histories of patients presenting with multiple conditions will allow physicians to use this framework to plan their work more easily and can inform multispecialty treatment approaches. This is especially pertinent for individuals who have a highly fragmented course of care and have to visit multiple specialists and undergo multiple procedures. This could help not just with diagnostics and decisions relating to the prognosis of individuals presenting with multimorbidity, but can also help with the planning of what medications to prescribe depending on the kinds of conditions such individuals might be at risk of contracting.

While the association rule mining gives us a somewhat granular perspective of the associations between groups of conditions, the hierarchical clustering using WPGMA gives us a bird's eye view of the various conditions in the system, and shows us the broader associations between the conditions on a system wide level. For instance, it is a known phenomenon that patients experience an increase in mental health disorders after developing musculoskeletal disorders (Adogwa et al., 2023; Duffield et al., 2017; Tazzeo et al., 2021), and this increase can be correlated with the increased degree and frequency of pain, and the reduced mobility that comes with it. This subset of conditions is itself frequently found to be comorbid

with metabolic disorders. The hierarchical clustering confirms this, and the association rules allow us to quantify the individual correlations between these groups of conditions. There is a potential for similar inferences to be drawn by combining the information derived from such data mining with clinical expertise.

While the work done in this study shows these patterns and associations, it is important to remember that there are limitations with regard to the nature of the data. If multimorbidity could be studied with data that considers more long-term information regarding an individual's case history, one could potentially take a Markovian approach, taking into account state transitions to estimate likelihoods of the next condition or disease an individual already presenting with multimorbidity could contract. Having a systematic approach to understanding the sequential evolution of multimorbidity would greatly enhance the decision-making process for physicians over longer periods of time. This is especially pertinent when we take into account the fact that both the prevalence and the expenses associated with healthcare are correlated with various demographic variables, geographical factors and access to healthcare.

The heterogeneity in the combinations of conditions makes the requirement for long term data imperative, and future research with the involvement of healthcare professionals in this space using such data will greatly aid the advancement in our understanding of multimorbidity. This is not just to develop better inferences, but also to develop better recommendations and tools for clinicians, especially primary care providers, to take preemptive measures rather than reactive measures.

# CHAPTER 7

# CONCLUSION

Through this study, we have been able to determine the large volume of condition groups that exist in the population using the MEPS-HC event data from the years 2011-2015. We were able to demonstrate that there is a high degree of heterogeneity in the phenomenon of multimorbidity in the population. To better understand the patterns and associations between conditions that individuals who present with multimorbidity suffer from, we used bicriteria analysis, association rule mining and hierarchical agglomerative clustering to develop both a system-wide overview of the associations between conditions as well as a more granular understanding of the associations between conditions. The findings from this study create a structured framework for deeper analysis of multimorbidity and lay the groundwork for future work to better define diagnostic and prognostic practices to enable better patient outcomes.

# APPENDICES

## Appendix 1

*CCCR Codes*

| MEPS collapsed condition category | CCCR Code |
|---|---|
| Infectious diseases | 1 |
| Cancer | 2 |
| Non-malignant neoplasm | 3 |
| Thyroid disease | 4 |
| Diabetes mellitus | 5 |
| Other endocrine, nutritional & immune disorder | 6 |
| Hyperlipidemia | 7 |
| Hemorrhagic, coagulation, and disorders of White Blood cells | 8 |
| Anemia and other deficiencies | 9 |
| Mental disorders | 10 |
| CNS infection | 11 |
| Hereditary, degenerative and other nervous system disorders | 12 |
| Paralysis | 13 |
| Headache | 14 |
| Epilepsy and convulsions | 15 |
| Coma, brain damage | 16 |
| Cataract | 17 |
| Glaucoma | 18 |
| Other eye disorders | 19 |
| Otitis media | 20 |
| Other CNS disorders | 21 |
| Hypertension | 22 |
| Heart disease | 23 |

| MEPS collapsed condition category | CCCR Code |
|---|---|
| Cerebrovascular disease | 24 |
| Other circulatory conditions arteries, veins, and lymphatics | 25 |
| Pneumonia | 26 |
| Influenza | 27 |
| Tonsillitis | 28 |
| Acute Bronchitis and URI | 29 |
| COPD, asthma | 30 |
| Intestinal infection | 31 |
| Disorders of teeth and jaws | 32 |
| Disorders of mouth and esophagus | 33 |
| Disorders of the upper GI | 34 |
| Appendicitis | 35 |
| Hernias | 36 |
| Other stomach and intestinal disorders | 37 |
| Other GI | 38 |
| Gallbladder, pancreatic, and liver disease | 39 |
| Kidney Disease | 40 |
| Urinary tract infections | 41 |
| Other urinary | 42 |
| Male genital disorders | 43 |
| Non-malignant breast disease | 44 |
| Complications of pregnancy and birth | 45 |
| Female genital disorders, and contraception | 46 |
| Normal birth/live born | 47 |
| Skin disorders | 48 |
| Osteoarthritis and other non-traumatic joint disorders | 49 |

| MEPS collapsed condition category | CCCR Code |
|---|---|
| Back problems | 50 |
| Other bone and musculoskeletal disease | 51 |
| Systemic lupus and connective tissues disorders | 52 |
| Congenital anomalies | 53 |
| Perinatal Conditions | 54 |
| Trauma-related disorders | 55 |
| Complications of surgery or device | 56 |
| Poisoning by medical and non-medical substances | 57 |
| Residual Codes | 58 |
| Other care and screening | 59 |
| Symptoms | 60 |
| Allergic reactions | 61 |

**Appendix 2**
*CCC Codes with CCCR Code Conversion*

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| 1 | Tuberculosis | Infectious diseases | 1 |
| 2 | Septicemia (except in labor) | Infectious diseases | 1 |
| 3 | Bacterial infection; unspecified site | Infectious diseases | 1 |
| 4 | Mycoses | Infectious diseases | 1 |
| 5 | HIV infection | Infectious diseases | 1 |
| 6 | Hepatitis | Infectious diseases | 1 |
| 7 | Viral infection | Infectious diseases | 1 |
| 8 | Other infections; including parasitic | Infectious diseases | 1 |
| 9 | Sexually transmitted infections (not HIV or hepatitis) | Infectious diseases | 1 |
| 11 | Cancer of head and neck | Cancer | 2 |
| 12 | Cancer of esophagus | Cancer | 2 |
| 13 | Cancer of stomach | Cancer | 2 |
| 14 | Cancer of colon | Cancer | 2 |
| 15 | Cancer of rectum and anus | Cancer | 2 |
| 16 | Cancer of liver and intrahepatic bile duct | Cancer | 2 |
| 17 | Cancer of pancreas | Cancer | 2 |
| 18 | Cancer of other GI organs; peritoneum | Cancer | 2 |
| 19 | Cancer of bronchus; lung | Cancer | 2 |
| 20 | Cancer; other respiratory and intrathoracic | Cancer | 2 |
| 21 | Cancer of bone and connective | Cancer | 2 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| | tissue | | |
| 22 | Melanomas of skin | Cancer | 2 |
| 23 | Other non-epithelial cancer of skin | Cancer | 2 |
| 24 | Cancer of breast | Cancer | 2 |
| 25 | Cancer of uterus | Cancer | 2 |
| 26 | Cancer of cervix | Cancer | 2 |
| 27 | Cancer of ovary | Cancer | 2 |
| 28 | Cancer of other female genital organs | Cancer | 2 |
| 29 | Cancer of prostate | Cancer | 2 |
| 30 | Cancer of testis | Cancer | 2 |
| 31 | Cancer of other male genital organs | Cancer | 2 |
| 32 | Cancer of bladder | Cancer | 2 |
| 33 | Cancer of kidney and renal pelvis | Cancer | 2 |
| 34 | Cancer of other urinary organs | Cancer | 2 |
| 35 | Cancer of brain and nervous system | Cancer | 2 |
| 36 | Cancer of thyroid | Cancer | 2 |
| 37 | Hodgkin's disease | Cancer | 2 |
| 38 | Non-Hodgkin's lymphoma | Cancer | 2 |
| 39 | Leukemias | Cancer | 2 |
| 40 | Multiple myeloma | Cancer | 2 |
| 41 | Cancer; other and unspecified primary | Cancer | 2 |
| 42 | Secondary malignancies | Cancer | 2 |
| 43 | Malignant neoplasm without specification of site | Cancer | 2 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| 44 | Neoplasms of unspecified nature or uncertain behavior | Cancer | 2 |
| 45 | Maintenance chemotherapy; radiotherapy | Cancer | 2 |
| 46 | Benign neoplasm of uterus | Non-malignant neoplasm | 3 |
| 47 | Other and unspecified benign neoplasm | Non-malignant neoplasm | 3 |
| 48 | Thyroid disorders | Thyroid disease | 4 |
| 49 | Diabetes mellitus without complication | Diabetes mellitus | 5 |
| 50 | Diabetes mellitus with complications | Diabetes mellitus | 5 |
| 51 | Other endocrine disorders | Other endocrine, nutritional & immune disorder | 6 |
| 52 | Nutritional deficiencies | Other endocrine, nutritional & immune disorder | 6 |
| 54 | Gout and other crystal arthropathies | Other endocrine, nutritional & immune disorder | 6 |
| 55 | Fluid and electrolyte disorders | Other endocrine, nutritional & immune disorder | 6 |
| 56 | Cystic fibrosis | Other endocrine, nutritional & immune disorder | 6 |
| 57 | Immunity disorders | Other endocrine, nutritional & immune disorder | 6 |
| 58 | Other nutritional; endocrine; and metabolic disorders | Other endocrine, nutritional & immune disorder | 6 |
| 53 | Disorders of lipid metabolism | Hyperlipidemia | 7 |
| 60 | Acute posthemorrhagic anemia | Hemorrhagic, coagulation, and disorders of White Blood cells | 8 |
| 61 | Sickle cell anemia | Hemorrhagic, coagulation, and disorders of White Blood cells | 8 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| 62 | Coagulation and hemorrhagic disorders | Hemorrhagic, coagulation, and disorders of White Blood cells | 8 |
| 63 | Diseases of white blood cells | Hemorrhagic, coagulation, and disorders of White Blood cells | 8 |
| 64 | Other hematologic conditions | Hemorrhagic, coagulation, and disorders of White Blood cells | 8 |
| 59 | Deficiency and other anemia | Anemia and other deficiencies | 9 |
| 65-75 | (older codes) | Mental disorders | 10 |
| 650 | Adjustment disorders | Mental disorders | 10 |
| 651 | Anxiety disorders | Mental disorders | 10 |
| 652 | Attention-deficit, conduct, and disruptive behavior disorders | Mental disorders | 10 |
| 653 | Delirium, dementia, and amnestic and other cognitive disorders | Mental disorders | 10 |
| 654 | Developmental disorders | Mental disorders | 10 |
| 655 | Disorders usually diagnosed in infancy, childhood, or adolescence | Mental disorders | 10 |
| 656 | Impulse control disorders, NEC | Mental disorders | 10 |
| 657 | Mood disorders | Mental disorders | 10 |
| 658 | Personality disorders | Mental disorders | 10 |
| 659 | Schizophrenia and other psychotic disorders | Mental disorders | 10 |
| 660 | Alcohol-related disorders | Mental disorders | 10 |
| 661 | Substance-related disorders | Mental disorders | 10 |
| 662 | Suicide and intentional self-inflicted injury | Mental disorders | 10 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| 663 | Screening and history of mental health and substance abuse codes | Mental disorders | 10 |
| 670 | Miscellaneous mental health disorders | Mental disorders | 10 |
| 76 | Meningitis (except that caused by tuberculosis or sexually transmitted disease) | CNS infection | 11 |
| 77 | Encephalitis (except that caused by tuberculosis or sexually transmitted disease) | CNS infection | 11 |
| 78 | Other CNS infection and poliomyelitis | CNS infection | 11 |
| 79 | Parkinson's disease | Hereditary, degenerative and other nervous system disorders | 12 |
| 80 | Multiple sclerosis | Hereditary, degenerative and other nervous system disorders | 12 |
| 81 | Other hereditary and degenerative nervous system conditions | Hereditary, degenerative and other nervous system disorders | 12 |
| 82 | Paralysis | Paralysis | 13 |
| 84 | Headache; including migraine | Headache | 14 |
| 83 | Epilepsy; convulsions | Epilepsy and convulsions | 15 |
| 85 | Coma; stupor; and brain damage | Coma, brain damage | 16 |
| 86 | Cataract | Cataract | 17 |
| 88 | Glaucoma | Glaucoma | 18 |
| 87 | Retinal detachments; defects; vascular occlusion; and retinopathy | Other eye disorders | 19 |
| 89 | Blindness and vision defects | Other eye disorders | 19 |
| 90 | Inflammation; infection of eye | Other eye disorders | 19 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| | (except that caused by tuberculosis or sexually transmitted disease) | | |
| 91 | Other eye disorders | Other eye disorders | 19 |
| 92 | Otitis media and related conditions | Otitis media | 20 |
| 93 | Conditions associated with dizziness or vertigo | Other CNS disorders | 21 |
| 94 | Other ear and sense organ disorders | Other CNS disorders | 21 |
| 95 | Other nervous system disorders | Other CNS disorders | 21 |
| 98 | Essential hypertension | Hypertension | 22 |
| 99 | Hypertension with complications and secondary hypertension | Hypertension | 22 |
| 96 | Heart valve disorders | Heart disease | 23 |
| 97 | Peri-; endo-; and myocarditis; cardiomyopathy (except that caused by tuberculosis or sexually transmitted disease) | Heart disease | 23 |
| 100 | Acute myocardial infarction | Heart disease | 23 |
| 101 | Coronary atherosclerosis and other heart disease | Heart disease | 23 |
| 102 | Nonspecific chest pain | Heart disease | 23 |
| 103 | Pulmonary heart disease | Heart disease | 23 |
| 104 | Other and ill-defined heart disease | Heart disease | 23 |
| 105 | Conduction disorders | Heart disease | 23 |
| 106 | Cardiac dysrhythmias | Heart disease | 23 |
| 107 | Cardiac arrest and ventricular fibrillation | Heart disease | 23 |
| 108 | Congestive heart failure; | Heart disease | 23 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| | nonhypertensive | | |
| 109 | Acute cerebrovascular disease | Cerebrovascular disease | 24 |
| 110 | Occlusion or stenosis of precerebral arteries | Cerebrovascular disease | 24 |
| 111 | Other and ill-defined cerebrovascular disease | Cerebrovascular disease | 24 |
| 112 | Transient cerebral ischemia | Cerebrovascular disease | 24 |
| 113 | Late effects of cerebrovascular disease | Cerebrovascular disease | 24 |
| 114 | Peripheral and visceral atherosclerosis | Other circulatory conditions arteries, veins, and lymphatics | 25 |
| 115 | Aortic; peripheral; and visceral artery aneurysms | Other circulatory conditions arteries, veins, and lymphatics | 25 |
| 116 | Aortic and peripheral arterial embolism or thrombosis | Other circulatory conditions arteries, veins, and lymphatics | 25 |
| 117 | Other circulatory disease | Other circulatory conditions arteries, veins, and lymphatics | 25 |
| 118 | Phlebitis; thrombophlebitis and thromboembolism | Other circulatory conditions arteries, veins, and lymphatics | 25 |
| 119 | Varicose veins of lower extremity | Other circulatory conditions arteries, veins, and lymphatics | 25 |
| 120 | Hemorrhoids | Other circulatory conditions arteries, veins, and lymphatics | 25 |
| 121 | Other diseases of veins and lymphatics | Other circulatory conditions arteries, veins, and lymphatics | 25 |
| 122 | Pneumonia (except that caused by tuberculosis or sexually | Pneumonia | 26 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| | transmitted disease) | | |
| 123 | Influenza | Influenza | 27 |
| 124 | Acute and chronic tonsillitis | Tonsillitis | 28 |
| 125 | Acute bronchitis | Acute Bronchitis and URI | 29 |
| 126 | Other upper respiratory infections | Acute Bronchitis and URI | 29 |
| 127 | Chronic obstructive pulmonary disease and bronchiectasis | COPD, asthma | 30 |
| 128 | Asthma | COPD, asthma | 30 |
| 129 | Aspiration pneumonitis; food/vomitus | COPD, asthma | 30 |
| 130 | Pleurisy; pneumothorax; pulmonary collapse | COPD, asthma | 30 |
| 131 | Respiratory failure; insufficiency; arrest (adult) | COPD, asthma | 30 |
| 132 | Lung disease due to external agents | COPD, asthma | 30 |
| 133 | Other lower respiratory disease | COPD, asthma | 30 |
| 134 | Other upper respiratory disease | COPD, asthma | 30 |
| 135 | Intestinal infection | Intestinal infection | 31 |
| 136 | Disorders of teeth and jaw | Disorders of teeth and jaws | 32 |
| 137 | Diseases of mouth; excluding dental | Disorders of mouth and esophagus | 33 |
| 138 | Esophageal disorders | Disorders of the upper GI | 34 |
| 139 | Gastroduodenal ulcer (except hemorrhage) | Disorders of the upper GI | 34 |
| 140 | Gastritis and duodenitis | Disorders of the upper GI | 34 |
| 141 | Other disorders of stomach and duodenum | Disorders of the upper GI | 34 |
| 142 | Appendicitis and other appendiceal conditions | Appendicitis | 35 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| 143 | Abdominal hernia | Hernias | 36 |
| 144 | Regional enteritis and ulcerative colitis | Other stomach and intestinal disorders | 37 |
| 145 | Intestinal obstruction without hernia | Other stomach and intestinal disorders | 37 |
| 146 | Diverticulosis and diverticulitis | Other stomach and intestinal disorders | 37 |
| 147 | Anal and rectal conditions | Other stomach and intestinal disorders | 37 |
| 148 | Peritonitis and intestinal abscess | Other stomach and intestinal disorders | 37 |
| 153 | Gastrointestinal hemorrhage | Other GI | 38 |
| 154 | Noninfectious gastroenteritis | Other GI | 38 |
| 155 | Other gastrointestinal disorders | Other GI | 38 |
| 149 | Biliary tract disease | Gallbladder, pancreatic, and liver disease | 39 |
| 150 | Liver disease; alcohol-related | Gallbladder, pancreatic, and liver disease | 39 |
| 151 | Other liver diseases | Gallbladder, pancreatic, and liver disease | 39 |
| 152 | Pancreatic disorders (not diabetes) | Gallbladder, pancreatic, and liver disease | 39 |
| 156 | Nephritis; nephrosis; renal sclerosis | Kidney Disease | 40 |
| 157 | Acute and unspecified renal failure | Kidney Disease | 40 |
| 158 | Chronic kidney disease | Kidney Disease | 40 |
| 160 | Calculus of urinary tract | Kidney Disease | 40 |
| 161 | Other diseases of kidney and ureters | Kidney Disease | 40 |
| 159 | Urinary tract infections | Urinary tract infections | 41 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| 162 | Other diseases of bladder and urethra | Other urinary | 42 |
| 163 | Genitourinary symptoms and ill-defined conditions | Other urinary | 42 |
| 164 | Hyperplasia of prostate | Male genital disorders | 43 |
| 165 | Inflammatory conditions of male genital organs | Male genital disorders | 43 |
| 166 | Other male genital disorders | Male genital disorders | 43 |
| 167 | Nonmalignant breast conditions | Non-malignant breast disease | 44 |
| 177 | Spontaneous abortion | Complications of pregnancy and birth | 45 |
| 178 | Induced abortion | Complications of pregnancy and birth | 45 |
| 179 | Postabortion complications | Complications of pregnancy and birth | 45 |
| 180 | Ectopic pregnancy | Complications of pregnancy and birth | 45 |
| 181 | Other complications of pregnancy | Complications of pregnancy and birth | 45 |
| 182 | Hemorrhage during pregnancy; abruptio placenta; placenta previa | Complications of pregnancy and birth | 45 |
| 183 | Hypertension complicating pregnancy; childbirth and the puerperium | Complications of pregnancy and birth | 45 |
| 184 | Early or threatened labor | Complications of pregnancy and birth | 45 |
| 185 | Prolonged pregnancy | Complications of pregnancy and birth | 45 |
| 186 | Diabetes or abnormal glucose tolerance complicating pregnancy; childbirth; or the puerperium | Complications of pregnancy and birth | 45 |
| 187 | Malposition; malpresentation | Complications of pregnancy | 45 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| | | and birth | |
| 188 | Fetopelvic disproportion; obstruction | Complications of pregnancy and birth | 45 |
| 189 | Previous C-section | Complications of pregnancy and birth | 45 |
| 190 | Fetal distress and abnormal forces of labor | Complications of pregnancy and birth | 45 |
| 191 | Polyhydramnios and other problems of amniotic cavity | Complications of pregnancy and birth | 45 |
| 192 | Umbilical cord complication | Complications of pregnancy and birth | 45 |
| 193 | OB-related trauma to perineum and vulva | Complications of pregnancy and birth | 45 |
| 194 | Forceps delivery | Complications of pregnancy and birth | 45 |
| 195 | Other complications of birth; puerperium affecting management of mother | Complications of pregnancy and birth | 45 |
| 168 | Inflammatory diseases of female pelvic organs | Female genital disorders, and contraception | 46 |
| 169 | Endometriosis | Female genital disorders, and contraception | 46 |
| 170 | Prolapse of female genital organs | Female genital disorders, and contraception | 46 |
| 171 | Menstrual disorders | Female genital disorders, and contraception | 46 |
| 172 | Ovarian cyst | Female genital disorders, and contraception | 46 |
| 173 | Menopausal disorders | Female genital disorders, and contraception | 46 |
| 174 | Female infertility | Female genital disorders, and contraception | 46 |
| 175 | Other female genital disorders | Female genital disorders, | 46 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| | | and contraception | |
| 176 | Contraceptive and procreative management | Female genital disorders, and contraception | 46 |
| 196 | Other pregnancy and delivery including normal | Normal birth/live born | 47 |
| 218 | Liveborn | Normal birth/live born | 47 |
| 197 | Skin and subcutaneous tissue infections | Skin disorders | 48 |
| 198 | Other inflammatory condition of skin | Skin disorders | 48 |
| 199 | Chronic ulcer of skin | Skin disorders | 48 |
| 200 | Other skin disorders | Skin disorders | 48 |
| 201 | Infective arthritis and osteomyelitis (except that caused by tuberculosis or sexually transmitted disease) | Osteoarthritis and other non-traumatic joint disorders | 49 |
| 202 | Rheumatoid arthritis and related disease | Osteoarthritis and other non-traumatic joint disorders | 49 |
| 203 | Osteoarthritis | Osteoarthritis and other non-traumatic joint disorders | 49 |
| 204 | Other non-traumatic joint disorders | Osteoarthritis and other non-traumatic joint disorders | 49 |
| 205 | Spondylosis; intervertebral disc disorders; other back problems | Back problems | 50 |
| 206 | Osteoporosis | Other bone and musculoskeletal disease | 51 |
| 207 | Pathological fracture | Other bone and musculoskeletal disease | 51 |
| 208 | Acquired foot deformities | Other bone and musculoskeletal disease | 51 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| 209 | Other acquired deformities | Other bone and musculoskeletal disease | 51 |
| 212 | Other bone disease and musculoskeletal deformities | Other bone and musculoskeletal disease | 51 |
| 210 | Systemic lupus erythematosus and connective tissue disorders | Systemic lupus and connective tissues disorders | 52 |
| 211 | Other connective tissue disease | Systemic lupus and connective tissues disorders | 52 |
| 213 | Cardiac and circulatory congenital anomalies | Congenital anomalies | 53 |
| 214 | Digestive congenital anomalies | Congenital anomalies | 53 |
| 215 | Genitourinary congenital anomalies | Congenital anomalies | 53 |
| 216 | Nervous system congenital anomalies | Congenital anomalies | 53 |
| 217 | Other congenital anomalies | Congenital anomalies | 53 |
| 219 | Short gestation; low birth weight; and fetal growth retardation | Perinatal Conditions | 54 |
| 220 | Intrauterine hypoxia and birth asphyxia | Perinatal Conditions | 54 |
| 221 | Respiratory distress syndrome | Perinatal Conditions | 54 |
| 222 | Hemolytic jaundice and perinatal jaundice | Perinatal Conditions | 54 |
| 223 | Birth trauma | Perinatal Conditions | 54 |
| 224 | Other perinatal conditions | Perinatal Conditions | 54 |
| 225 | Joint disorders and dislocations; trauma-related | Trauma-related disorders | 55 |
| 226 | Fracture of neck of femur (hip) | Trauma-related disorders | 55 |
| 227 | Spinal cord injury | Trauma-related disorders | 55 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| 228 | Skull and face fractures | Trauma-related disorders | 55 |
| 229 | Fracture of upper limb | Trauma-related disorders | 55 |
| 230 | Fracture of lower limb | Trauma-related disorders | 55 |
| 231 | Other fractures | Trauma-related disorders | 55 |
| 232 | Sprains and strains | Trauma-related disorders | 55 |
| 233 | Intracranial injury | Trauma-related disorders | 55 |
| 234 | Crushing injury or internal injury | Trauma-related disorders | 55 |
| 235 | Open wounds of head; neck; and trunk | Trauma-related disorders | 55 |
| 236 | Open wounds of extremities | Trauma-related disorders | 55 |
| 239 | Superficial injury; contusion | Trauma-related disorders | 55 |
| 240 | Burns | Trauma-related disorders | 55 |
| 244 | Other injuries and conditions due to external causes | Trauma-related disorders | 55 |
| 237 | Complication of device; implant or graft | Complications of surgery or device | 56 |
| 238 | Complications of surgical procedures or medical care | Complications of surgery or device | 56 |
| 241 | Poisoning by psychotropic agents | Poisoning by medical and non-medical substances | 57 |
| 242 | Poisoning by other medications and drugs | Poisoning by medical and non-medical substances | 57 |
| 243 | Poisoning by nonmedicinal substances | Poisoning by medical and non-medical substances | 57 |
| 259 | Residual codes; unclassified | Residual Codes | 58 |
| 10 | Immunizations and screening for infectious disease | Other care and screening | 59 |
| 254 | Rehabilitation care; fitting of prostheses; and adjustment of devices | Other care and screening | 59 |

| CCC Code | CCC Description | MEPS collapsed condition category | CCCR Code |
|---|---|---|---|
| 255 | Administrative/social admission | Other care and screening | 59 |
| 256 | Medical examination/evaluation | Other care and screening | 59 |
| 257 | Other aftercare | Other care and screening | 59 |
| 258 | Other screening for suspected conditions (not mental disorders or infectious disease) | Other care and screening | 59 |
| 245 | Syncope | Symptoms | 60 |
| 246 | Fever of unknown origin | Symptoms | 60 |
| 247 | Lymphadenitis | Symptoms | 60 |
| 248 | Gangrene | Symptoms | 60 |
| 249 | Shock | Symptoms | 60 |
| 250 | Nausea and vomiting | Symptoms | 60 |
| 251 | Abdominal pain | Symptoms | 60 |
| 252 | Malaise and fatigue | Symptoms | 60 |
| 253 | Allergic reactions | Allergic reactions | 61 |

## BIBLIOGRAPHY

Adogwa, O., Reid, M. C., Chilakapati, S., & Makris, U. E. (2023). Clin-STAR corner: 2021 update in musculoskeletal pain in older adults with a focus on osteoarthritis-related pain. *Journal of the American Geriatrics Society*, *71*(8). https://doi.org/10.1111/jgs.18369

Agency for Healthcare Research and Quality. (2014). *Multiple Chronic Conditions Chartbook 2010 MEDICAL EXPENDITURE PANEL SURVEY DATA.* https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/prevention-chronic-care/decision/mcc/mccchartbook.pdf

Agency for Healthcare Research and Quality. (2017). *Medical Expenditure Panel Survey Content Summary of the Household Interview*. Ahrq.gov. https://meps.ahrq.gov/mepsweb/survey_comp/hc_data_collection.jsp

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, *22*(2). https://doi.org/10.1145/170035.170072

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases, Vol. 1215*, 487–499. https://doi.org/10.5555/645920.672836

Ahn, S., Hussein, M., Mahmood, A., & Smith, M. L. (2020). Emergency department and inpatient utilization among U.S. older adults with multiple chronic conditions: a post-reform update. *BMC Health Services Research*, *20*(1). https://doi.org/10.1186/s12913-020-4902-7

Aizcorbe, A., Bradley, R., Greenaway-Mcgrevy, R., Herauf, B., Kane, R., Liebman, E., Pack, S., & Rozental, L. (2011). *ALTERNATIVE PRICE INDEXES FOR MEDICAL CARE: EVIDENCE FROM THE MEPS SURVEY*. https://conference.nber.org/confer/2011/SI2011/PRCR/Aizcorbe_Bradley.pdf

Baker, T. A., Clay, O. J., Johnson-Lawrence, V., Minahan, J. A., Mingo, C. A., Thorpe, R. J., Ovalle, F., & Crowe, M. (2017). Association of multiple chronic conditions and pain among older black and white adults with diabetes mellitus. *BMC Geriatrics*, *17*(1). https://doi.org/10.1186/s12877-017-0652-8

Belbin, L., Faith, D. P., & Milligan, G. W. (1992). A Comparison of Two Approaches to Beta-Flexible Clustering. *Multivariate Behavioral Research*, *27*(3), 417–433. https://doi.org/10.1207/s15327906mbr2703_6

Bock, H.-H. (2014). Proximity Measures. *Wiley StatsRef: Statistics Reference Online*, *3*. https://doi.org/10.1002/9781118445112.stat06451

Borzsony, S., Kossmann, D., & Stocker, K. (2001). The Skyline operator. *Proceedings 17th International Conference on Data Engineering*. https://doi.org/10.1109/icde.2001.914855

Buttorff, C., Ruder, T., & Bauman, M. (2017). *Multiple Chronic Conditions in the United States*. RAND Corporation. https://www.rand.org/pubs/tools/TL221.html

Caughey, G. E., Vitry, A. I., Gilbert, A. L., & Roughead, E. E. (2008). Prevalence of comorbidity of chronic diseases in Australia. *BMC Public Health*, *8*(1). https://doi.org/10.1186/1471-2458-8-221

Cohen, R. A., Bernstein, A. B., Bilheimer, L. T., Makuc, D. M., & Powell-Griner, E. (2009, July). *Health insurance coverage trends, 1959-2007; estimates from the National Health Interview Survey*. Cdc.gov. https://stacks.cdc.gov/view/cdc/5454

Craig, A., Mtintsilana, A., Mapanga, W., Dlamini, S. N., & Norris, S. A. (2023). Socioeconomic position, perceived weight, lifestyle risk, and multimorbidity in young adults aged 18 to 35 years: a Multi-country Study. *BMC Public Health*, *23*(1). https://doi.org/10.1186/s12889-023-16234-1

Dalirsefat, S. B., da Silva Meyer, A., & Mirhoseini, S. Z. (2009). Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, Bombyx mori. *Journal of Insect Science*, *9*(71), 1–8. https://doi.org/10.1673/031.009.7101

Duffield, S. J., Ellis, B. M., Goodson, N., Walker-Bone, K., Conaghan, P. G., Margham, T., & Loftis, T. (2017). The contribution of musculoskeletal disorders in multimorbidity: Implications for practice and policy. *Best Practice & Research Clinical Rheumatology*, *31*(2), 129–144. https://doi.org/10.1016/j.berh.2017.09.004

Elixhauser, A., Steiner, C., & Palmer, L. (2015). *Clinical Classifications Software (CCS) for ICD-9-CM*. Www.hcup-Us.ahrq.gov. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

Fortin, M. (2005). Prevalence of Multimorbidity Among Adults Seen in Family Practice. *The Annals of Family Medicine*, *3*(3), 223–228. https://doi.org/10.1370/afm.272

Gerteis, J., Izrael, D., Deitz, D., LeRoy, L., Ricciardi, R., Miller, T., & Basu, J. (2014). *Multiple Chronic Conditions Chartbook*. Agency for Healthcare Research and Quality Publications.

Glynn, L. G., Valderas, J. M., Healy, P., Burke, E., Newell, J., Gillespie, P., & Murphy, A. W. (2011). The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. *Family Practice*, *28*(5), 516–523. https://doi.org/10.1093/fampra/cmr013

Hoffman, C. (1996). Persons With Chronic Conditions. *JAMA*, *276*(18), 1473. https://doi.org/10.1001/jama.1996.03540180029029

Hwang, W., Weller, W., Ireys, H., & Anderson, G. (2001). Out-Of-Pocket Medical Spending For Care Of Chronic Conditions. *Health Affairs*, *20*(6), 267–278. https://doi.org/10.1377/hlthaff.20.6.267

King, D. E., Xiang, J., & Pilkerton, C. S. (2018). Multimorbidity Trends in United States Adults, 1988–2014. *The Journal of the American Board of Family Medicine*, *31*(4), 503–513. https://doi.org/10.3122/jabfm.2018.04.180008

Lochner, K. A., & Cox, C. S. (2013). Prevalence of Multiple Chronic Conditions Among Medicare Beneficiaries, United States, 2010. *Preventing Chronic Disease*, *10*. https://doi.org/10.5888/pcd10.120137

Machlin, S. R., & Soni, A. (2013). Health Care Expenditures for Adults With Multiple Treated Chronic Conditions: Estimates From the Medical Expenditure Panel Survey, 2009. *Preventing Chronic Disease*, *10*. https://doi.org/10.5888/pcd10.120172

Marengoni, A., Rizzuto, D., Wang, H.-X., Winblad, B., & Fratiglioni, L. (2009). Patterns of Chronic Multimorbidity in the Elderly Population. *Journal of the American Geriatrics Society*, *57*(2), 225–230. https://doi.org/10.1111/j.1532-5415.2008.02109.x

McQuitty, L. L. (1966). Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. *Educational and Psychological Measurement*, *26*(4), 825–831. https://doi.org/10.1177/001316446602600402

Meraya, A. M., Raval, A. D., & Sambamoorthi, U. (2015). Chronic Condition Combinations and Health Care Expenditures and Out-of-Pocket Spending Burden Among Adults, Medical Expenditure Panel Survey, 2009 and 2011. *Preventing Chronic Disease*, *12*. https://doi.org/10.5888/pcd12.140388

Meyer, A. da S., Garcia, A. A. F., Souza, A. P. de, & Souza Jr., C. L. de. (2004). Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L). *Genetics and Molecular Biology*, *27*(1), 83–91. https://doi.org/10.1590/s1415-47572004000100014

National Institute for Health and Care Excellence. (2016, September 21). *Multimorbidity: Clinical Assessment and Management | Guidance | NICE*. Nice.org.uk; NICE. https://www.nice.org.uk/guidance/ng56

Paez, K. A., Zhao, L., & Hwang, W. (2009). Rising Out-Of-Pocket Spending For Chronic Conditions: A Ten-Year Trend. *Health Affairs*, *28*(1), 15–25. https://doi.org/10.1377/hlthaff.28.1.15

Palladino, R., Pan, T., Mercer, S. W., Atun, R., McPake, B., Rubba, F., Triassi, M., & Lee, J. T. (2023). Multimorbidity and out-of-pocket expenditure on medicine in Europe: Longitudinal analysis of 13 European countries between 2013 and 2015. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.1053515

Rajagopalan, S., & Robb, R. (2005). Assessment of similarity indices to quantify segmentation accuracy of scaffold images for tissue engineering: Medical Imaging 2005 - Image Processing. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, *5747*(III), 1636–1647. https://doi.org/10.1117/12.594654

Salive, M. E. (2013). Multimorbidity in Older Adults. *Epidemiologic Reviews*, *35*(1), 75–83. https://doi.org/10.1093/epirev/mxs009

Sing, M., Banthin, J. S., Selden, T. M., Cowan, C. A., & Keehan, S. P. (2006). Reconciling medical expenditure estimates from the MEPS and NHEA, 2002. *Health Care Financing Review*, *28*(1), 25–40. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4194973/

Sokal, R. R., & Michener, C. D. (1958). A Statistical Method for Evaluating Systematic Relationships. In *University of Kansas Science Bulletin* (Vol. 38). University of Kansas Science Bulletin. https://ia800703.us.archive.org/5/items/cbarchive_33927_astatisticalmethodforevaluatin1902/astatisticalmethodforevaluatin1902.pdf

Sorace, J., Millman, M., Rezaee, M. E., LeRoy, L., M. B. A. , White, A., Oppenheim, E., & Wasserman, M. (2013). Understanding the high prevalence of low-prevalence chronic disease combinations: Databases and methods of research. In *Office of Assistant Secretary for Planning and Evaluation (APSE)* . https://aspe.hhs.gov/reports/understanding-high-prevalence-low-prevalence-chronic-disease-combinations-databases-methods-research-0

Sorace, J., Wong, H., Worrall, C., Kelman, J., Saneinejad, S., & MaCurdy, T. (2011). The Complexity of Disease Combinations in the Medicare Population. *Population Health Management*, *14*(4), 161–166. https://doi.org/10.1089/pop.2010.0044

Tazzeo, C., Rizzuto, D., Calderón-Larrañaga, A., Roso-Llorach, A., Marengoni, A., Welmer, A.-K., Onder, G., Trevisan, C., & Vetrano, D. L. (2021). Multimorbidity patterns and risk of frailty in older community-dwelling adults: a population-based cohort study. *Age and Ageing*, *50*(6), 2183–2191. https://doi.org/10.1093/ageing/afab138

US Census Bureau. (2018, December 3). *Older People Projected to Outnumber Children*. The United States Census Bureau. https://www.census.gov/newsroom/press-releases/2018/cb18-41-population-projections.html

van den Akker, M., Buntinx, F., Metsemakers, J. F. M., Roos, S., & Knottnerus, J. André. (1998). Multimorbidity in General Practice: Prevalence, Incidence, and Determinants of Co-Occurring Chronic and Recurrent Diseases. *Journal of Clinical Epidemiology*, *51*(5), 367–375. https://doi.org/10.1016/s0895-4356(97)00306-5

Vogeli, C., Shields, A. E., Lee, T. A., Gibson, T. B., Marder, W. D., Weiss, K. B., & Blumenthal, D. (2007). Multiple Chronic Conditions: Prevalence, Health Consequences, and Implications for Quality, Care Management, and Costs. *Journal of General Internal Medicine*, *22*(S3), 391–395. https://doi.org/10.1007/s11606-007-0322-1

Vu, K., Clark, R. A., Bellinger, C., Erickson, G., Osornio-Vargas, A., Zaïane, O. R., & Yuan, Y. (2019). The index lift in data mining has a close relationship with the association measure relative risk in epidemiological studies. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-0838-4

Wolff, J. L., Starfield, B., & Anderson, G. (2002). Prevalence, Expenditures, and Complications of Multiple Chronic Conditions in the Elderly. *Archives of Internal Medicine*, *162*(20), 2269. https://doi.org/10.1001/archinte.162.20.2269

Zemedikun, D. T., Gray, L. J., Khunti, K., Davies, M. J., & Dhalwani, N. N. (2018). Patterns of Multimorbidity in Middle-Aged and Older Adults: An Analysis of the UK Biobank Data. *Mayo Clinic Proceedings*, *93*(7), 857–866. https://doi.org/10.1016/j.mayocp.2018.02.012

Zulman, D. M., Asch, S. M., Martins, S. B., Kerr, E. A., Hoffman, B. B., & Goldstein, M. K. (2013). Quality of Care for Patients with Multiple Chronic Conditions: The Role of Comorbidity Interrelatedness. *Journal of General Internal Medicine*, *29*(3), 529–537. https://doi.org/10.1007/s11606-013-2616-9