IMPACT OF INFERRED COMORBIDITY NETWORKS

ON HEALTH OUTCOMES

By

PANKUSH KALGOTRA

Bachelor of Technology in Information Technology
National Institute of Technology
Raipur, CG, India
2011

Master of Science in Management Information Systems
Oklahoma State University
Stillwater, Oklahoma
2013

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2017

IMPACT OF INFERRED COMORBIDITY NETWORKS

ON HEALTH OUTCOMES

Dissertation Approved:

Dr. Ramesh Sharda

Dissertation Adviser

Dr. Dursun Delen

Dr. Andy Luse

Dr. Bruce Benjamin

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Ramesh Sharda for his guidance, encouragement and support as my mentor, my advisor and being the iron pillar needed for research. I have been fortunate to have Dr. Sharda for his timely care and infinite considerations to shape my research career as a Ph.D. Thanks to Dr. Luse, my committee member and a friend, for his help in data analysis in second chapter and inputs in other chapters. Many thanks to Dr. Dursun Delen and Dr. Bruce Benjamin for their comments that helped improve this dissertation significantly.

I would also like to thank Dr. William Paiva and his team at Center for Health Systems Innovation (CHSI) for extending the access to the Cerner Corporation's Health Facts database, a critical element used for my dissertation work.

I would also like to recognize Vijay Singh, Nandan Moza, Shubham Singh, Rupesh Agarwal, Rajeev Gangwar, Jayant Sharma, Ankita Khurana and Peeyush Patel for their critical inputs on various sections of the dissertation.

I would also like to thank my parents, Santosh Kumari and Kuldeep Raj Kalgotra, younger sister and brother, and extended family members. They helped me with moral support and encouraged me with their best wishes when I needed the most. Finally, thanks to Nitika Sharma for her love and support.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: PANKUSH KALGOTRA

Date of Degree: JULY, 2017

Title of Study: IMPACT OF INFERRED COMORBIDITY NETWORKS ON HEALTH OUTCOMES

Major Field: MANAGEMENT SCIENCE AND INFORMATION SYSTEMS

Abstract: High dimensionality in Big Data can be modeled using network approach. The traditional networks (e.g. online social network) are explicit and easily observed. However, there are certain networks that are implicit and exist by virtue of some underlying collective behavior. Our focus is on these implicit networks, which can be inferred from the secondary data using statistical modeling. An example of such a network is a comorbidity network. In a comorbidity network, diseases form connections based on their co-occurrences in patients. We use data on the health history of 24.7 million patients recorded in US hospitals (2000-2016) to infer comorbidity networks. Since most statistical models depend upon sample size, it is important to study how sample size affects the structure of an implicit network. We study the impact of sample size on comorbidity networks developed using Pearson's Correlation Coefficient (PCC) and Salton Cosine Index (SCI). We present a comparative analysis and show that a network developed using SCI is robust to sample size as compared to the PCC.

Our first study of comorbidity networks employs descriptive analytics. We investigate how comorbidity networks are different across population groups. We compare networks based on gender, race and insurance types. Our analysis at the comorbidity level presents health disparities across population groups.

These disparities across population groups are considered to study the impact of comorbidity network on patients' hospital length of stay in second study. We develop an explanatory and predictive model to estimate length of stay using features extracted from comorbidity networks and compare with the extant models. We show that our model outperforms the existing models.

Finally, we study how an implicit network can help theorize certain phenomenon related to it. With respect to the comorbidity network, we theorize clique property of a network as trap state. The trap state is hypothesized to be related to mortality risk of a patient. We identify eighteen such cliques in a comorbidity network.

This dissertation contributes to network science, analytics and healthcare literature but the theory, models, algorithms, and processes developed are generalizable to other inferred networks.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 1

1.1. PROBLEM DOMAIN 3
1.2. SCOPE OF THE DISSERTATION 5
1.3. DATA SOURCE 6
1.4. OUTLINE OF THE DISSERTATION 7

II. IMPACT OF SAMPLE SIZE ON AN IMPLICIT NETWORK 9

2.1. INTRODUCTION 10
2.2. METHODOLOGY 12
2.2.1. NETWORK 12
2.2.2. NETWORK TOPOLOGIES 14
2.2.3. NETWORK METRICS 15
2.2.3.1. NODES AND EDGES 15
2.2.3.2. DEGREE AND WEIGHTED DEGREE CENTRALITY 16
2.2.3.3. BETWEENNESS CENTRALITY 17
2.2.3.4. CLOSENESS CENTRAILITY 17
2.2.3.5. EIGENVECTOR CENTRALITY 18
2.2.3.6. CLUSTERING COEFFICIENT 18
2.2.3.7. NETWORK DENSITY 19
2.3. ILLUSTRATION 19
2.3.1. COMORBIDITY 19
2.3.2. COMORBIDITY NETWORK 20
2.4. ANALYSIS AND RESULTS 22
2.4.1. SAMPLING AND COMORBIDITY NETWORKS 22
2.4.2. EFFECT OF SAMPLE SIZE ON THE OVERALL STRUCTURE OF COMORBIDITY NETWORK 27
2.4.3. EFFECT OF SAMPLE SIZE ON NETWORK METRICS 31
2.5. DISCUSSION AND CONCLUSION 34

III. EXAMINING HEALTH DISPARITIES BY GENDER: A MULTIMORBIDITY NETWORK ANALYSIS OF ELECTRONIC MEDICAL RECORD 37

3.1. INTRODUCTION 38

3.2. METHOD AND ANALYSIS 39
3.2.1. DATA DESCRIPTION 40
3.2.2. MEASURING MULTIMORBIDITY 41
3.2.3. MULTIMORBIDITY NETWORK 41
3.2.4. NETWORK METRICS 43
3.3. RESULTS AND DISCUSSION....43
3.3.1. COMPARISON OF MALE AND FEMALE MULTIMORBIDITY NETWORKS....43
3.3.2. NETWORK PROPERTIES 44
3.3.3. ORGAN LEVEL NETWORK COMPARISON....46
3.2. CONCLUSIONS....51

IV. HEALTH ANALYTICS LEAD TO MORE QUESTIONS: A COMORBIDITY LENS APPROACH....54

4.1. INTRODUCTION....55
4.2. COMORBIDITY NETWORK ANALYSES OF RACES 56
4.3. COMORBIDITY NETWORK ANALYSES OF MEDICAID AND NON-MEDICAID PATIENTS 60
4.4. COMORBIDITY NETWORK ANALYSES OF MEDICARE AND NON-MEDICARE PATIENTS 61
4.5. DISCUSSION 63

V. WHEN WILL I GET OUT OF THE HOSPITAL? MODELING LENGTH OF STAY USING COMORBIDITY NETWORK....65

5.1 INTRODUCTION....66
5.2. LITERATURE REVIEW....71
5.2.1. LENGTH OF STAY....71
5.2.2. LENGTH OF STAY AND COMORBIDITY 73
5.3. MODEL DEVELOPMENT 76
5.3.1. BASELINE MODELS....76
5.3.2. MODELING USING COMORBIDITY NETWORK....78
5.3.2.1. COMORBIDITY NETWORK....78
5.3.2.2. NETWORK METRICS....81
5.3.2.3. EXPLANATORY AND PREDICTIVE MODELING USING COMORBIDITY NETWORK....84
5.4. ANALYSIS AND RESULTS 87
5.4.1. EXPLANATORY MODELING RESULTS AT HOSPITAL VISIT LEVEL....92
5.4.2. PREDICTIVE MODELING RESULTS AT HOSPITAL VISIT LEVEL 96
5.5. DISCUSSION AND CONCLUDING REMARKS 97

VI. DIAGNOSES FORM TRAPS: IDENTIFYING MORTALITY RELATED CLIQUES IN COMORBIDITY NETWORK....101

6.1. INTRODUCTION....102
6.2. BACKGROUND....105
6.3. METHOD....107

6.3.1. DETECTING CLIQUES....109
6.4. DATA DESCRIPTION AND PREPARATION....112
6.5. RESULTS....114
6.6. CONCLUSIONS....120

VII. CONCLUSIONS....122

7.1. CONTRIBUTIONS AND GENERALIZABILITY....122
7.2. FUTURE WORK....124

REFERENCES....126

APPENDICES....139

APPENDIX A. PERFORMANCE OF THE HOSPITAL LENGTH OF STAY MODELS FOR PATIENTS WITH A SPECIFIC PRIMARY DIAGNOSIS....139

LIST OF TABLES

Table Page

Table 2.1. Definitions of network measures 19
Table 2.2. Comorbidity networks nodes and edges 26
Table 2.3. Features of networks related to their topologies 29
Table 2.4. Network measures and their interpretation in our context 32
Table 2.5. Comorbidity networks properties 33
Table 3.1. Gender multimorbidity networks properties 46
Table 3.2. ICD-9 code classification 47
Table 3.3. Class associations in Female and Male Networks 48
Table 4.1. Race comorbidity networks properties 57
Table 4.2. Comorbidities across races 59
Table 4.3. Medicaid and non-Medicaid comorbidity networks properties 61
Table 4.4. Medicare and non-Medicare comorbidity networks properties 62
Table 5.1. Different categories of networks based on the purpose of formation. 70
Table 5.2. A review of selected papers on length of stay and comorbidity 72
Table 5.3. Network measures and their interpretation in our context 83
Table 5.4. An algorithm to add predicted comorbidities of the known diseases at the point of admission 86
Table 5.5. Network properties. 92
Table 5.6. Variable description. 93
Table 5.7. Linear models for length of stay at hospital visit level. 94
Table 6.1. An algorithm to find diagnoses forming cliques with high mortality rate 111
Table 6.2. Mortality rate with and without cliques. 116
Table 6.3. Robustness check on validation dataset. 120

LIST OF FIGURES

Figure Page

Figure 1.1 Venn diagram explaining scope of the dissertation....6
Figure 2.1. Network 1....15
Figure 2.2. Network 2....16
Figure 2.3. Network 3....17
Figure 2.4. Network 4....17
Figure 2.5. Network 5....18
Figure 2.6. Flowchart of data preparation and analysis....23
Figure 2.7. Pearson's Correlation Coefficient vs. Salton Cosine Index....24
Figure 2.8. Comorbidity Network. 25
Figure 2.9. Density of networks from different sample sizes using Correlation and Salton Cosine....27
Figure 2.10a. Number of nodes in largest connected component....31
Figure 2.10b. Average path length in PCC network....31
Figure 2.10c. Average path length in SCI network....31
Figure 2.10d. Average clustering coefficient in PCC network....31
Figure 2.10e. Average clustering coefficient in SCI network....31
Figure 2.11a. Number of edges....33
Figure 2.11b. Average Degree....33
Figure 2.11c. Average Weighted Degree....33
Figure 2.11d. Average Betweenness....33
Figure 2.11e. Average Closeness....34
Figure 2.11f. Network Density....34
Figure 2.11g. Average Clustering Coefficient....34
Figure 2.11h. Average Eigenvector Centrality....34
Figure 3.1a. Female Multimorbidity Network....45
Figure 3.1b. Female Multimorbidity Network....45
Figure 3.2a. Female Organ Comorbidity Network....48
Figure 3.2b. Male Organ Comorbidity....48
Figure 4.1a. Caucasian Network....58
Figure 4.1b. African- American Network....58
Figure 4.1c. Asian Network....58
Figure 4.1d. Hispanic Network....58
Figure 4.1e. Native Network....58
Figure 4.1f. Pacific Network....58
Figure 4.2a. Medicaid Network....61

Figure 4.2b. Non-Medicaid Network 61
Figure 4.3a. Medicare Network 62
Figure 4.3b. Non-Medicare Network 62
Figure 5.1. Data Processing and Modeling 89
Figure 5.2a. Female Comorbidity Network 91
Figure 5.2b. Male Comorbidity Network 91
Figure 5.3. Average improvement in variance explained in LOS in clusters of patients based on type of their primary diagnosis 96
Figure 5.4. Improvement in predictive power in different clusters of patients based on primary disease category due to comorbidity matrix 97
Figure 6.1. A clique/triangle of three diseases with their joint impact on mortality 104
Figure 6.2. Data processing 113
Figure 6.3. A clique/triangle of three diseases with their joint impact on mortality 115
Figure 6.4a. Clique 1 116
Figure 6.4b. Clique 2 116
Figure 6.4c. Clique 3 117
Figure 6.4d. Clique 4 117
Figure 6.4e. Clique 5 117
Figure 6.4f. Clique 6 117
Figure 6.4g. Clique 7 117
Figure 6.4h. Clique 8 117
Figure 6.4i. Clique 9 117
Figure 6.4j. Clique 10 117
Figure 6.4k. Clique 11 117
Figure 6.4l. Clique 12 118
Figure 6.4m. Clique 13 118
Figure 6.4n. Clique 14 118
Figure 6.4o. Clique 15 118
Figure 6.4p. Clique 16 118
Figure 6.4q. Clique 17 118
Figure 6.4r. Clique 18 118
Figure 6.5. Mortality rate with and without clique 119

CHAPTER I

INTRODUCTION

Across the disciplines, there are several phenomena that occur at a level that is not visible explicitly. It is a challenging problem to detect such hidden phenomena or structures, which are characterized by some implicit underlying behavior. The data analytics approach analogous to the grounded theory methodology can be applied to discern such hidden structures "from the data". In grounded theory methodology, the broader idea is to discover features from the data (Glaser & Strauss, 2009) where hypotheses are not pre-formulated but emerge from the data.

With the advancements in Information Systems, the collection, storage and analysis of large datasets are possible, which provide opportunities to discover hidden structures related to a particular phenomenon. Research using large datasets equivalent to the population has several advantages. First, the availability of large datasets mitigates issues related to the small sample size in research. And second, the conclusions from data can be validated across multiple samples and thus, their generalizability can be verified.

Extremely large datasets are characterized as Big Data. These are known to include high volume, high velocity of data collection and often have high dimensionality with a large variety. Due to high dimensionality in Big Data, model building is challenging and requires much computational power. Big Data analytics is thus focused on taming this beast requiring much data storage and analytics capacity to handle the large variety.

One of the models that can present a high dimensional space in a summarized manner is the network model. A network comprises of nodes connected to each other based on a well-defined relationship. In traditional networks such as an online social network, nodes (in case of an online social network, nodes are users), form a network based on their decision to connect to each other. The connections between friends or users on online social networks such as Twitter or Facebook are explicit and visible through features such as "followers", "friends", "likes", etc. However, there are several networks in which the interactions are implicit and it is not easy to draw links between nodes. We call these networks implicit because the relationships between nodes exist by virtue of some underlying exchanges (Roth et al. 2010). One common example of an implicit network is the network formed by collaborative filtering in recommender systems (Konstas, Stathopoulos, & Jose, 2009). Recommender systems create virtual connections between users or products based on their common characteristics. Another example of an implicit network is the ingredient network in which ingredients are connected based on their co-occurrences in different recipes (Teng, Lin, & Adamic, 2012). One more interesting implicit network is the language network emerged from the co-occurrences of words in a sentence, semantics, and syntactic (Solé et al. 2010; Liu, 2009). Implicit networks are also common in medical science. The biological networks such as a protein network (Weston et al., 2004), brain network (Van Den Heuvel & Pol, 2010), comorbidity network (Hidalgo et al. 2009), phenome-genome network (Butte & Kohane, 2006) and many others are all created through some underlying relationships and thus, are implicit networks.

The focus of this dissertation is the implicit networks which are inferred from historical patterns. These networks are data-driven and are inferred theoretically using mathematical formulas. To create relationships between nodes in a network, joint probabilities, co-occurrences, or similarities between the nodes are used (Hidalgo et al., 2009; Roth et al., 2010; Teng et al., 2012).

Therefore, it entirely depends on how a researcher defines a relationship between nodes. Most often, a similarity index is used to define a relationship between two nodes mathematically.

Since relationships emerge from data, the size of sample is an issue in the inferred networks. If the definition of a relationship between two nodes depends on sample size, it can result into an invalid and unreliable network. Therefore, to define a valid relationship between two nodes in a network, it is important to study the impact of sample size on inferred networks developed using different similarity indexes. This gives rise to the first research question of this dissertation:

> *Research Question 1: What is the impact of sample size on the structure of an inferred network created using a similarity index?*

The traditional networks (e.g. social network) have been shown in the past to be related to performance outcomes of the network source (Coleman, 1988; Provan and Sebastian 1998). For example, Sparrowe et al. (2001) found individual job performance was positively related to position of an employee in the advice network. Similarly, in this study, we study how a structure or network emerged implicitly from the unintended actions of source impact the performance outcomes of source? This broader question is studied in the context of US health explained in next section

1.1. PROBLEM DOMAIN

The broader problem domain of this dissertation is the health of US population recorded in hospitals electronically. The health history recorded in an Electronic Medical Record (EMR) includes different types of clinical information such as lab procedures, medications, diseases diagnosed and other hospital related variables. In this dissertation, we are specifically interested in the collective behavior of diseases in patients. We use network approach to study this underlying behavior. The network studied here is a network of diseases where diagnoses are

related to each other based on their co-occurrences in millions of patients admitted in US hospitals. To define a connection between two diseases in the network, we use an important medical concept known as comorbidity. Comorbidity is a medical condition in a patient when he or she develops multiple diseases simultaneously. For instance, the presence of diabetes and depression simultaneously in a patient is a comorbidity.

The comorbidity network contains diseases linked to each other whereby representing a summarized underlying joint behavior of the diseases. This joint behavior can be different across different groups of patients, thus leading to different health consequences. The health disparities across population groups can be caused by genetic, hormonal, physiological, behavioral, and sociocultural factors. Therefore, it is important to understand how diseases form relationships across different population groups. This provokes the second research question of this dissertation:

> *Research Question 2: How do diseases co-occur differently and form different network structures across population groups?*

The second research question discussed above is entirely exploratory and set the stage to find how the underlying interactions of diseases can affect some health outcomes of patients. Because the foundation of our network is co-occurrence of diseases, it can help predict other likely diseases in a patient in future based on his current condition. Thus, we use this idea to understand how network can be used to ex-ante predict the health outcomes, specifically the hospital length of stay. This gives rise to the third research question of this dissertation:

> *Research Question 3: How does the implicit relationships among diseases help ex-ante predict the health outcomes, specifically the hospital length of stay?*

The comorbidity network embeds risk in its structure, analogous to the social capital in social network (Coleman, 1988). This structural risk is not identified and analyzed yet in medical

literature, particularly with respect to comorbidity. Moreover, this structural risk can be used to theorize medical concepts. We use the structural risk embedded in comorbidity network to understand mortality, which is an important health outcome. We use an important structural property of a network known as a clique (a sub-network where all nodes are adjacent to each other) to explain mortality. Because a clique has maximum possible interactions among nodes, its presence in a patient can be critical due to high risk. It indicates a trap state in a patient from where the exit is difficult. Identifying such clique can help physicians take preemptive actions related to the health of a patient. Therefore, the final research question is related to the impact of cliques on mortality.

Research Question 4: Can we identify clique as trap state where its presence in a patient increases mortality risk?

1.2. SCOPE OF THE DISSERTATION

This interdisciplinary dissertation draws upon healthcare, network science and analytics/Information Systems literature. The healthcare problems are studied by applying network theories and using Information Systems tools and techniques. Figure 1.1 presents a Venn diagram describing the scope of this dissertation at the intersection of three areas: healthcare, network science and analytics/IS.

In the healthcare area, we enhance the understanding of comorbidity and its impact on the health outcomes such as patients' hospital length of stay and mortality. Our study applies network science concepts to study the comorbidities. Studying disease associations or comorbidities generates more insights than studying diseases independently.

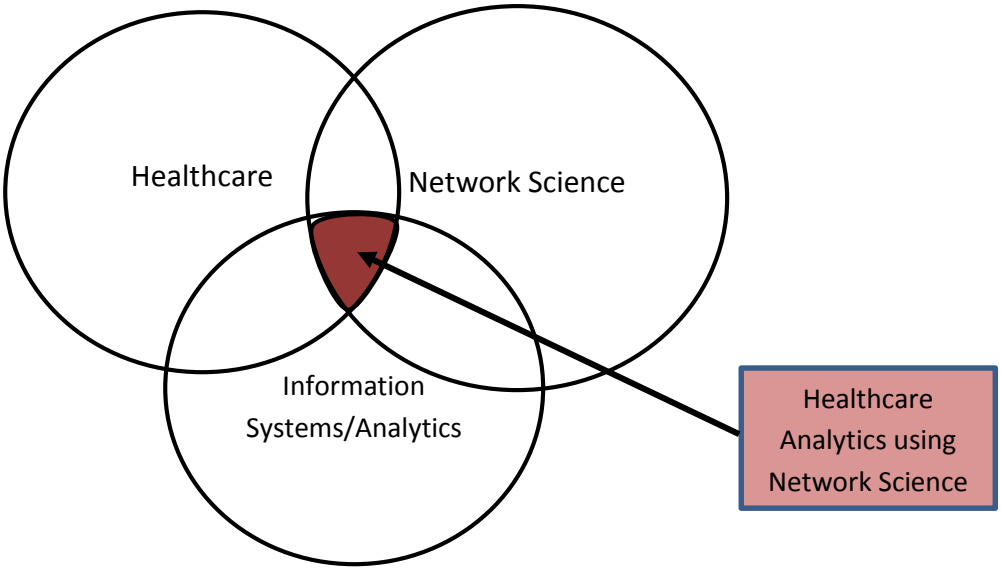


Figure 1.1 Venn diagram explaining scope of the dissertation

We apply analytical tools and techniques of the IS field to develop models, measures and algorithms, which are the explicit IT artifacts (March & Smith, 1995; Von Alan et al. 2004). The measures and models are expected to augment the performance of current Information Systems used to predict length of stay and mortality rate. The analytics has a transformational impact on the healthcare discipline (Agarwal & Lucas Jr, 2005; Markus & Mao, 2004).

1.3. DATA SOURCE

We obtained data from the Center for Health Systems Innovation (CHSI), a center at Oklahoma State University which houses data provided by Cerner Corporation, a major Electronic Medical Record (EMR) provider. The data warehouse contains records of visits of 58 million unique patients across 662 US hospitals (2000-2016). It includes more than 84 million admissions, emergency and ambulatory visits. It is the largest and industry's only relational database that includes comprehensive records with pharmacy, laboratory, clinical events, admission and billing data. Diagnoses are classified according to the International Classification of Diseases, 9th

Revision, Clinical Modification (ICD-9-CM). This data warehouse has recently been made available to OSU's Center for Health Systems Innovation through a gift from Cerner Corporation, a major EMR provider. The database also includes more than 2.4 billion laboratory results and more than 295 million orders for nearly 4,500 drugs categorized by name and brand. It is one of the largest compilation of de-identified, real-world, HIPAA-compliant data of its type that can permit such a large scale network analysis. The use of this massive dataset is one of the strengths of our study.

1.4. OUTLINE OF THE CHAPTERS

The rest of the dissertation is organized as follows. In Chapter 2, we answer the first research question using dataset described in the previous section. It describes the method to create network using Pearson's correlation coefficient and Salton Cosine Index in addition to a process to find statistical significance of the relationship using Salton Cosine Index. This chapter provides an appropriate index to define a relationship between two diseases, which is used in all other chapters later.

Following the method described in Chapter 2, in Chapter 3, we create two different comorbidity networks for men and women to find health disparity by gender. This comparison shows which comorbidities are more prevalent in one population group and not in others. This study is under review at a journal.

Following the same approach as Chapter 3, comorbidity network differences are discovered in races and different insurance holders in Chapter 4. The comparison between different population groups results into several research questions for medical, economics, social, public health, policy and analytics researchers. The third and fourth chapters are the responses to our second research question.

In Chapter 5, we answer the third research question. We extend the applicability of the network to create explanatory and predictive models to estimate the patient's length of stay. To ex-ante predict the length of stay, we only use information available at the point of admission. In addition, we also compare our models with the extant models and show our models perform better.

The fourth research question is addressed in Chapter 6. We use the clique concept to understand mortality risk embedded in the structure of a comorbidity network. A clique forms a trap state and thus, its presence in a patient is likely to increase mortality risk.

Finally, in Chapter 7, we conclude by discussing the contributions, generalizability and future work of the models, processes and algorithms developed in this dissertation.

CHAPTER II

IMPACT OF SAMPLE SIZE ON AN IMPLICIT NETWORK

ABSTRACT

Networks can be observed in different problem domains. Some networks are explicit where members make direct connections (e.g. Facebook network), whereas other networks are formed through some underlying implicit relationships, which are not directly visible (e.g. collaborative filtering network). Since implicit networks are present in almost every field of science and developed from a sample of some population, it is necessary to understand how sample size influences their structures given that the conclusions from network analysis can be biased if a network does not represent true relationships. The purpose of this paper is to understand how sample size impacts the structure of an implicit network. We compare the networks created using two indexes: Pearson's Correlation Coefficient (PCC) and Salton Cosine Index (SCI). For demonstration, we present an implicit network called a comorbidity network. The networks created using PCC and SCI from a large dataset containing health records of 22.1 million patients are compared based on their overall topologies and node centralities. The results show that the network formed using SCI is less affected by the sample size as compared to the network created using PCC. With respect to the overall structure of a network, the comorbidity network using SCI follows a small-world topology irrespective of the sample size; however, the structure of network using PCC is inconsistent in its structure. Regarding node centralities, the betweenness centrality

of the network is most affected by sample size. Our analysis is valuable as it establishes a need for choosing a right measure to create an implicit network for making valid conclusions.

2.1.INTRODUCTION

A network emerges from the interactions between elements or nodes (Euler, 1953). For example, online social networks (OSN) are one of the main research topics in Information Systems field. In an online social network, friends or users form a network based on their direct connections. The connections between friends or users on online social networks such as Twitter or Facebook are explicit and visible through features such as "followers", "friends", "likes", etc. These networks are easy to construct because one can easily define a relationship between two elements.

However, there are several networks in which the interactions are implicit and it is not easy to draw links between nodes. We call these networks implicit networks because the relationships between nodes exist by virtue of some underlying exchanges (Roth et al. 2010). One common example of an implicit network is the network formed by collaborative filtering in recommender systems (Konstas, Stathopoulos, & Jose, 2009). Recommender systems try to create virtual connections between users or products based on their common characteristics. Another example of an implicit network is an ingredient network in which ingredients are connected based on their co-occurrences in different recipes (Teng, Lin, & Adamic, 2012). Another interesting implicit network is the language network developed based on co-occurrences of words in a sentence, semantics, and syntactic (Solé et al. 2010; Liu, 2009). Implicit networks are also widely studied in medical science. The biological networks such as a protein network (Weston et al., 2004), brain network (Van Den Heuvel & Pol, 2010), comorbidity network (Hidalgo et al. 2009), phenome-genome network (Butte & Kohane, 2006) and many others are all created through some underlying relationships and thus, are implicit networks.

Implicit networks are created using joint probabilities, co-occurrences, or similarities of nodes (Hidalgo et al., 2009; Roth et al., 2010; Teng et al., 2012). Therefore, it entirely depends on how a researcher defines a relationship between nodes. Mostly, researchers use a similarity index to define a relationship mathematically. One of the most common indexes is Pearson's correlation coefficient. It has often been used in author co-citation network analysis to find an intellectual structure in a given field (McCain, 1990). It has also been used in medical sciences for creating a network of diseases from electronic health records (Hidalgo et al., 2009; Divo et al., 2015). In contrast, some researchers have supported the use of other indexes such as cosine indexes over Pearson's correlation coefficient. Van Eck & Waltman (2008) argued that Pearson's correlation coefficient captures the linear relationship between two variables, which is not same as the commonality between two variables; therefore, it is not an appropriate measure to create a network. Instead, the authors suggested to use a cosine index to develop a network from the data. Similarly, Ahlgren, Jarneving, & Rousseau, (2003) also criticized the use of Pearson's correlation coefficient because it is sensitive to sparseness in the network. The authors argued that it results in low overlap between nodes. In addition, Pearson's correlation coefficient depends on the sample size and therefore it can influence the structure of a network (Ahlgren, Jarneving, & Rousseau, 2003).

Since implicit networks are present in almost every field of science and inferred from a sample of some population, it is necessary to see how sample size influences their structures. The conclusions from network analysis can be biased if the network is invalid. Therefore, it becomes important to study the behaviour of networks created using different indexes and different sample sizes. In this paper, the primary objective is to understand how sample size impacts the structure of an implicit network developed using different indexes. We want to study the impact of sample size on networks created using Pearson's correlation Index and a cosine index known as Salton

Cosine Index (Salton & McGill, 1986). Salton Cosine Index is unaffected by the sample size and only considers the co-occurrences and prevalence of nodes.

As mentioned earlier, we want to study how sample size impacts the structure of networks. The structure of a network can be measured using multiple network properties such as node centrality, clustering coefficient, density and others. In addition, the overall topology of a network (random, scale-free or small-world) can be assessed to understand the overall structure. Moreover, using a large real-world dataset, we offer useful and well-supported recommendations on desirable sample sizes for creating a valid implicit network.

To demonstrate our method and analysis, we demonstrate an implicit network known as comorbidity network. Comorbidity is a medical condition when two or more diseases are present simultaneously in a patient (Feinstein, 1970). Comorbidity networks have been mostly developed using inductive reasoning and are data driven. The relationships between diseases are inferred from the sample. We illustrate how sample size impacts the structure of a comorbidity network developed using Pearson's correlation coefficient and Salton Cosine Index.

The rest of the paper is organized as follows. In the next section, we elaborate on the mathematical formulations of a network, overall topologies, and network properties. Then, we explain comorbidity and a process to create comorbidity networks using Pearson's correlations coefficient and Salton's Cosine Index. We then explore and compare the disease associations in different sample sizes. Next, the results are discussed. Finally, we conclude by discussing implications.

2.2. METHODOLOGY

2.2.1. NETWORK

A network comprises nodes connected through defined edges. To create an undirected implicit network i.e. a network with no directions in the relationships, one has to define a transaction containing the related nodes. These transactions will be used to explain whether the connection between two nodes exists or not. A network *C* developed from *N* transactions is denoted by *C (D, E)* in which *D* is a set of *n* nodes and *E* is a set of edges.

An edge E_{ij} is created between two nodes d_i and d_j ($i, j = 1$ *to* n) where $i<j$ in the undirected network. Since we want to compare networks created using two indexes, we define an edge mathematically based on these two indexes i.e. Pearson's correlation coefficient (PCC) and Salton Cosine Index (SCI). We want to select the appropriate index to create a network.

In the network using PCC, the coefficient S_{ij} of an edge E_{ij} between nodes d_i and d_j is calculated as

$$S_{ij} = \frac{(c_{ij}*N)(c_i*c_j)}{\sqrt{(c_i*c_j)(N-c_i)(N-c_j)}} \quad \textbf{-(2.1)}$$

where c_{ij} is the count of transactions containing both *i* and *j* nodes, c_i is the count of transactions containing *i* and c_j is the count of transactions containing *j*. The maximum number of edges possible among *n* nodes is *(n(n-1)/2)*. However, we considered edges based on statistical significance of the PCC. We calculated T-statistic using S_{ij} of the edges as in equation 2.2. Following the most conservative approach, we used the c_{ij} (minimum of c_{ij}, c_i, and c_j) as the degrees of the freedom. Using the T-statistic, we developed networks at $\alpha=0.01$, $T>2.58$ *and* $cij>\sum c_{ij}/p$, where *p* is maximum number of pairs.

$$T=\frac{S_{ij}\sqrt{c_{ij}-2}}{\sqrt{1-S_{ij}{}^2}} \quad \textbf{-(2.2)}$$

In the network using Salton Cosine Index, $SCI_{ij,}$ of an edge between diseases d_i and d_j is calculated as in equation 2.3. It considers the individual prevalence of the two nodes (c_i and c_j) and their joint prevalence (c_{ij}).

$$SCI_{ij} = \frac{(c_{ij})}{\sqrt{(c_i * c_j)}} \quad \textbf{-(2.3)}$$

Unlike the correlation coefficient, this measure is unaffected by the sample size, *N*; however, it is difficult to find its statistical significance. Usually, a cut-off for SCI is defined. We use the relationship between PCC and SCI to find a cut-off for SCI as suggested by Egghe & Leydesdorff (2009). We present an approach that results into edges that are correlated significantly. The steps followed to find a cut-off are as follows:

Step 1. For each pair of nodes, calculate number of co-occurrences, Pearson's Correlation Coefficient and Salton Cosine Index in the population dataset (largest sample size)

Step 2. Find number of pairs (q) significantly correlated at $\alpha=0.01$ and $c_{ij} > \sum c_{ij}/p$, where p is maximum number of pairs

Step 3. Find Salton Cosine Index as the cutoff (S_c) where the number of pairs is equal to q and $c_{ij} > \sum c_{ij}/p$

Step 4. Use S_c as the cutoff to find edges in different sample sizes

We use the above process to create networks from samples of different sizes using PCC and SCI and then compare them. The effect of sample size is measured in terms of overall network topologies and node properties as explained in the next section.

2.2.2. NETWORK TOPOLOGIES

To understand the structure of a network as a whole, one can find the topology of a network. A topology is a global property of a network. Knowing the overall structure will specify the behavior of a network in a particular context. For example, epidemic spread depends on the network topology (Ganesh, Massoulié, & Towsley, 2005). In this research, we look at the overall structure of a network and understand its dependency on the sample size.

The most common topologies are random, scale-free and small-world. A network is called a random network if connections between a set of nodes are randomly connected using a defined probability (Erdos & Rényi, 1960; Erdős & Rényi, 1959). The degree of a random network follows binomial distribution. A network is called a scale-free network when a network contains hubs in it (Barabási & Albert, 1999). The degree distribution of the nodes in a scale-free network follow a power-law distribution. Finally, a network is called a small-world network when there are several clusters in a network, making the distance between nodes smaller (Watts & Strogatz, 1998). The degree of a small-world network can follow any distribution but the average clustering property is higher than the random network.

2.2.3. NETWORK METRICS

The structure of a network can be measured using several network metrics. A network has several inherent properties that can be observed at the node level. We describe multiple network properties briefly in the sub-sections below. The definitions are also listed in Table 1.1, which we will use to analyze the impact of sample size on a network in later sections.

2.2.3.1.NODES AND EDGES

Nodes are the elements among which relationships are studied. In Figure 2.1, the circles A, B, C and D are four different nodes that are related to each other.

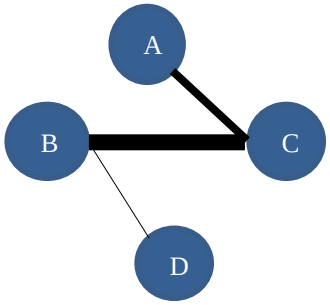


Figure 2.1. Network 1

It can be observed in Figure 2.1 that A is connected to C, C is connected to A and B, B is connected to C and D and finally D is connected to B.

These connections are represented by the lines or edges. These lines represent some relationships. So, before creating a network, there is a need to define the relationship between the nodes.

The edges in Figure 2.1 do not have directions and hence, it is an undirected network. In addition, the edges can also have weights. The weight represents the strength of a relationship. The edges in Figure 2.1 have weights and are represented by their thickness. The weight of the edge between B and C is the largest, followed by the edge between A and C, and then the edge between B and D have the smallest weight.

2.2.3.2.DEGREE AND WEIGHTED DEGREE CENTRALITY

An important property of a node in a network is its centrality. Centrality can be broadly defined as the importance of a node in the network. There are multiple ways to define a centrality. We report on four types of node centralities: degree, betweenness, closeness and eigenvector centrality.

Degree centrality is a simplest property of the nodes in a network. Degree of a node explains its number of direct connections (Freeman, 1979). Let us reconsider the Figure 2.1. Here degree of node A is 1, B is 2, C is 2 and D is 1. As discussed earlier, the edges are undirected in our networks. However, if an edge has direction, two types of degrees are there: in-degree (number of edges coming in) and out-degree (number of edges going out).

Moreover, if weights of the edges are considered to calculate degree, it is called the weighted degree of a node. Let us consider the network shown in Figure 2.2 where the weight of an edge between A and C is W_{ac}, B and C is W_{bc} and B and D is $W_{bd.}$ Then the weighted degree of a node is given by the sum of the weights of the direct connections.

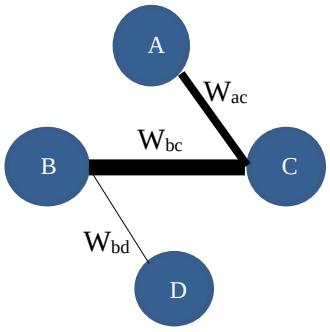


Figure 2.2. Network 2

Weighted Degree of A = W_{ac},

Weighted Degree of B = $W_{bc} + W_{bd}$,

Weighted Degree of C = $W_{bc} + W_{ac}$ and

Weighted Degree of D = W_{bd}

2.2.3.3. BETWEENNESS CENTRALITY

Another important network property of a node is its betweenness. The number of times a node is on a shortest path among all shortest paths (Freeman, 1979). In an undirected network, betweenness of a node *i* is

$$b_i = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}},$$

where σ_{st} is total number of shortest paths from node *s* to node *t* and $\sigma_{st}(i)$ is the number of those paths that pass through *i*. In Figure 2.3, betweenness of A is 15 because it is on every path from all other pair of nodes and there are total 15 paths. All other nodes have betweenness of 0.

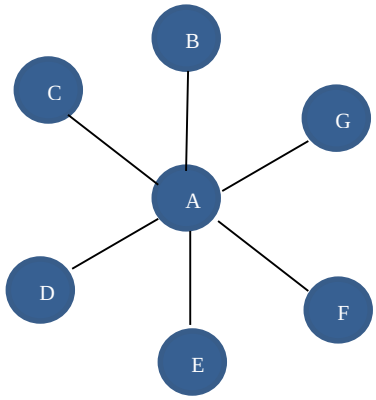


Figure 2.3. Network 3

2.2.3.4. CLOSENESS CENTRALITY

Closeness of a node *i* gives the average shortest distance of that node to all other nodes in the network. Closeness is a node *i* in the network of *n* nodes is given by

$$C_c(i) = \frac{\sum_{n-1} d(i,j)}{n-1},$$

where *d(i, j)* is the shortest distance between *i* and *j*.

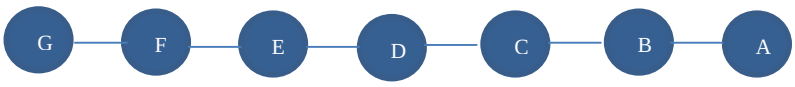


Figure 2.4. Network 4

In Figure 2.4, the closeness centrality of G is given by (1+2+3+4+5+6/6) = 3.5. Similarly, the closeness centrality of D is (1+2+3+1+2+3) = 2. It means that the average shortest distance of

node D is smaller than the node G. An inverse of the number is usually calculated to present that the higher the number, the higher the closeness.

2.2.3.5. EIGENVECTOR CENTRALITY

An eigenvector centrality of a node explains how well the direct connections of a node are also connected (Bonacich, 1987). It considers all the relationships in the network and assigns a relative score to every node. It can be understood as a degree centrality that spans the entire network.

2.2.3.6. CLUSTERING COEFFICENT

The clustering coefficient explains the small clusters formed by the nodes. The clustering coefficient of a node explains how well the neighbors of a node are connected (Watts & Strogatz, 1998). The clustering coefficient of a node, *i*, explains how well the direct connections of the node, *i*, are connected to each other. The clustering coefficient, C_i, of a node can be mathematically written as

$$C_i = \frac{2l_i}{k_i(k_i-1)},$$

where l_i is the number of links among the neighbors of the node *i* and k_i is the degree of a node *i*. In Figure 2.5, node *i* has three connections (A, B and C). Among three nodes, maximum three links are possible (A-B, B-C and A-C). However, only one link i.e. A-B is present. Hence, the clustering co-efficient of the node *i* is 1/3. Similarly, the clustering co-efficient of the other nodes can be calculated.

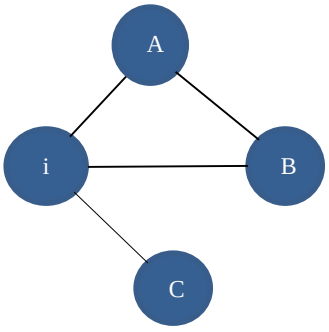


Figure 2.5. Network 5

2.2.3.7.NETWORK DENSITY

Density of a network is the proportion of edges present in the network. Density of a network with *n* nodes and E edges (undirected) is given by the ratio of number of edges present to the maximum number of edges possible.

Network Density = $\frac{E}{n(n-1)/2}$

Table 2.1. Definitions of network measures	
Network Measure	**Definition**
Nodes	Nodes or vertexes are the elements among which relationships are studied.
Edges	An edge represents the relationship between nodes.
Degree centrality	Degree of a node explains its number of direct connections (Freeman, 1979)
Weighted Degree	Degree calculated considering the strength of an edge.
Betweenness Centrality	Number of times a node is on a shortest path among all shortest paths (Freeman, 1979)
Closeness Centrality	Closeness of a node gives the average shortest distance of that node to all other nodes in the network (Freeman, 1979).
Eigenvector Centrality	Eigenvector centrality of a node explains how well the direct connections of a node are also connected (Bonacich, 1987).
Clustering Coefficient	The clustering coefficient of a node explains how well its neighbors are connected (Watts & Strogatz, 1998).
Network Density	Density of a network is the proportion of edges present in the network.

2.3. ILLUSTRATION

We illustrate the impact of sample size on the comorbidity network, an implicit network. We will first define comorbidity in our context and then explain the process to create a comorbidity network.

2.3.1. COMORBIDITY

Comorbidity is a medical condition when a patient is diagnosed with two or more diseases. Feinstein (1970) defined comorbidity as the presence of any other disease or complication in addition to the primary disease. The diseases present simultaneously can exist independently, or one disease causes another making them interdependent (Jakovljevic & Ostojic, 2013). These

conceptualizations do not consider the lifetime history of a patient but looks into the presence of diseases during a hospital visit. In other words, previous definitions focus on a much smaller timespan of a patient. Focusing on *current* patient information can help physicians to control comorbidities, but how the *history* of a patient is related to the current situation is not understood. If we look into the lifetime history of patients and find relationships between diseases, this can provide us additional understanding about comorbidities. In this paper, we delineate comorbidity considering the lifetime history of a patient rather than a single hospital visit. We define comorbidity as *the presence of multiple diseases in the lifetime history of a patient.* This definition has two advantages over previous definitions. First, the medical recording of a disease over multiple hospitals visits is only considered once. Considering the same disease as different across hospital visits can overestimate its presence and bias the analysis and conclusions. Second, our definition incorporates the impact of a disease on other diseases *across* multiple hospital visits, thereby incorporating the wider span of disease development. We use multiple similarity indexes to define a comorbidity that helps us to define it validly.

2.3.2. COMORBIDITY NETWORK

To create comorbidity networks, we used a real-world massive dataset from Electronic Medical Record (EMR). We obtained data from the Center for Health Systems Innovation (CHSI), a center at Oklahoma State University that houses data provided by Cerner Corporation, a major EMR provider. The data warehouse contains an EMR on the visits of more than 58 million unique patients across US hospitals (2000-2016). Among 58 million patients, nearly 24.7 million patients were diagnosed with at least one disease or a symptom. Moreover, there were 2.6 million patients who were coded only with symptoms. We did not use those patients in our analysis and extracted the remaining 22.1 million patient records for creating comorbidity networks.

We create a comorbidity network in which connections between diseases (nodes) are developed if diagnosed in the patients simultaneously. As noted in other implicit networks, a common way to define an association between two diseases is through their correlation in the database (Hidalgo et al., 2009). It can be useful to find the most correlated diagnoses but with a small sample, rare associations might not be captured because a correlation depends on the sample size (Egghe & Leydesdorff, 2009). Therefore, if the purpose is to find highly correlated diseases, Pearson's Correlation Coefficient can be used to find them if the sample size is sufficient. However, if the purpose is to find rare or less correlated disease associations, PCC should only be used with large sample sizes. In contrast, Salton Cosine Index (Salton & McGill, 1986) does not account for the sample size but only considers the co-occurrences and prevalences of the diseases forming an edge. The cosine index has been used in the past to find phenotype overlaps (Chen et al. 2015; Lage et al. 2007); however, we propose it for finding the strength of a comorbidity.

For this research, we require a transactional dataset to create a comorbidity network. In the past, a hospital visit in the EMR was considered as a transaction (Hidalgo et al., 2009), but as noted earlier, considering a hospital visit as a transaction to define a comorbidity has several shortcomings. Based on our definition of comorbidity, we consider the lifetime history of a patient as a transaction. A transaction contains multiple diseases diagnosed over time. The presence of multiple diseases in a patient throughout his lifetime are used to create associations between diseases.

In our comorbidity network, nodes represent diagnoses. In an EMR, diagnoses are classified using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). An ICD-9 code has three, four, or five digits (xxx.xx). The first three digits represent the broader category of a disease. The fourth and fifth digits represent the sub-divisions of a disease. For example, the ICD-9 code for viral hepatitis is 070. At the four-digit level (070.x), there are eight types of viral hepatitis and at the five-digit level (070.xx), two other viral hepatitis are

coded. We aggregated ICD-9-CM codes to the three-digit level. Thus, variations of the same disease were considered as one node in the network. For example, there were multiple types of viral hepatitis but there was only one node for this disorder in our network. There are both advantages and disadvantages of aggregation. An advantage is the reduction in measurement bias. In contrast, the disadvantage is the compromise of granularity as different classes of the same disease can have a dissimilar impact.

An edge was created between two diseases if they were comorbid. As there is no strong evidence regarding which disease leads to which other disease, we created an undirected network with no direction in the relationships. Using the process explained in the method section, we created twenty different comorbidity networks, ten each using Pearson's correlation coefficient and Salton Cosine Index from different samples described in the following sections.

2.4. ANALYSIS AND RESULTS

2.4.1. SAMPLING AND COMORBIDITY NETWORKS

The steps to find the influence of sample size on network structures are presented in a flowchart in Figure 2.6. First, information about patients, hospitals, types of visits, and diseases developed by the patients were joined for further analyses. The diseases were recorded according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). There were several hospital visits in which patients were not diagnosed with any type of disease at all. These patients and hospitals were not considered for further analyses.

In the second step, entries with invalid admission and discharge date/time and those with invalid entries for the disease were removed. We then aggregated the ICD-9 disease codes into three-digit codes. At this stage, we had approximately 22.1 million unique patients with sufficient information to perform analysis. After data cleaning and preparation, we created ten random samples of patients starting from 100% of the patients to as small as 1000 patients. The values of

the network measures can be evaluated by taking multiple random samples from the same pseudo population and analyzing the variation in the values, as suggested by Wolda (1981). We followed the same process suggested by the Wolda (1981) and drew ten samples. The ten random samples included: 1) 22.1 million patients (100%), 2) 11.1 million patients (50%), 3) 5.5 million patients (25%), 4) 2.75 million patients (12.5%), 5) 1.38 million patients (6.25%), 6) 500,000 patients, 7) 250,000 patients, 8) 100,000 patients, 9) 50,000 patients and 10) 1000 patients.

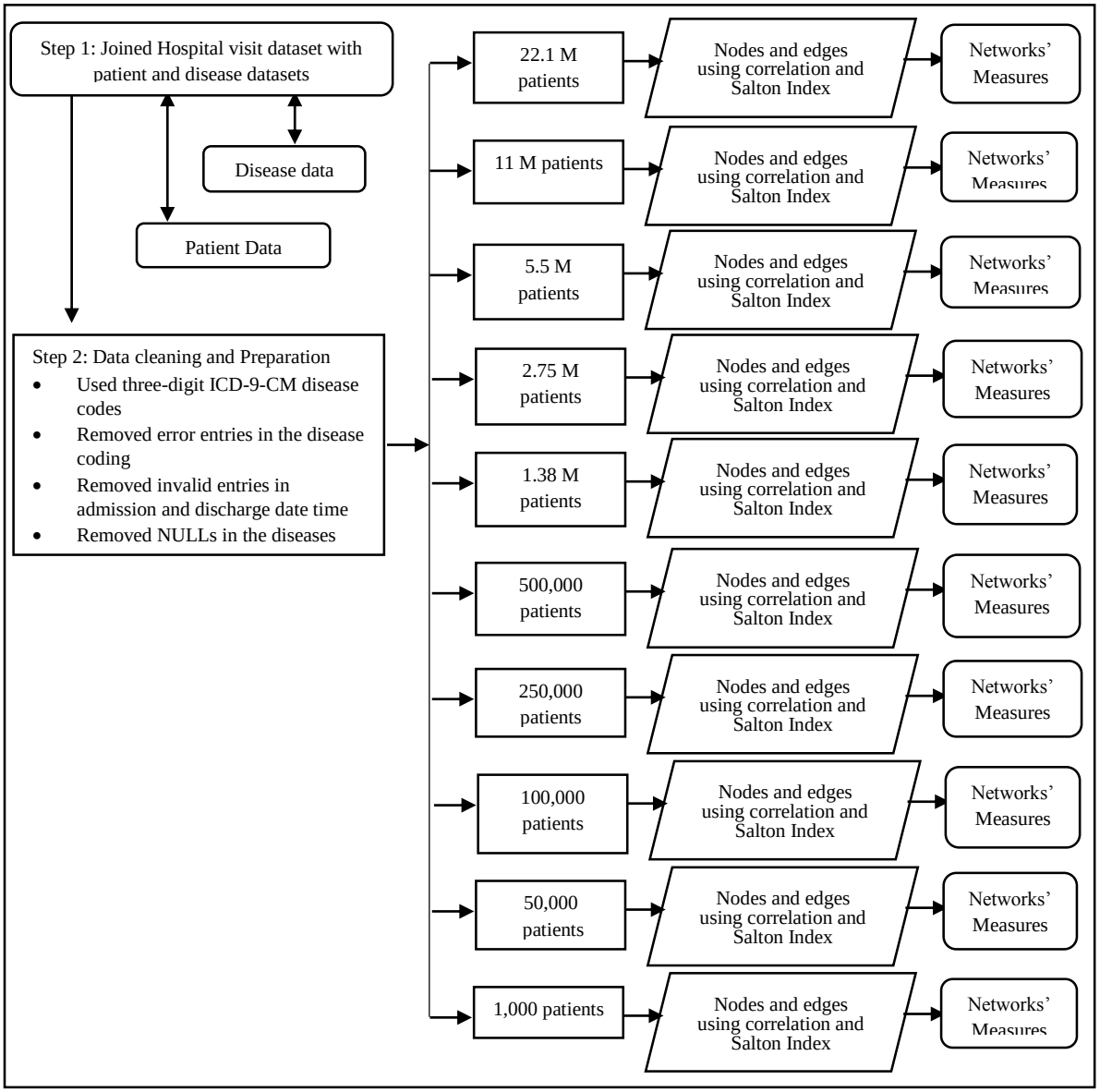


Figure 2.6. Flowchart of data preparation and analysis

From the ten random samples, comorbidity networks were created using PCC and SCI. PCCs of the edges and their statistical significance were calculated. From the comorbidity network of 22.1 million patients, we found 14,463 significant edges with Pearson's correlation coefficient significant at $\alpha=0.01$. In the network developed from 22.1 million patients with SCI, at 14,463 edges, the cutoff for SCI was 0.04. We compared the edges incorporated by the two indexes and found more than 95% of the edges to be common. The relationship between the two indexes is depicted in Figure 2.7. The majority of edges (more than 95%) at the SCI cutoff of 0.04 were also highly correlated ($p<0.01$). Therefore, from here onwards, we consistently use SCI cutoff of 0.04 for creating an edge in all comorbidity networks.

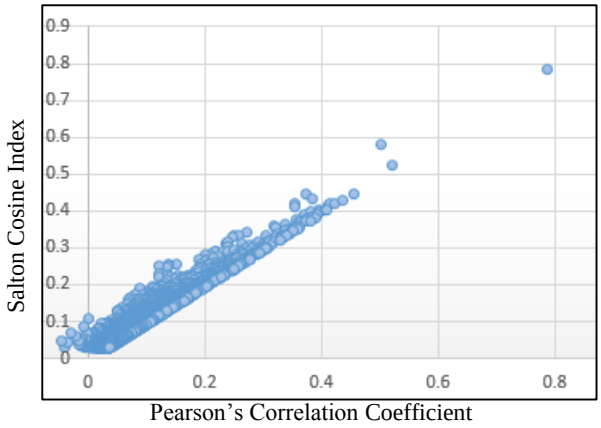


Figure 2.7. Pearson's Correlation Coefficient vs. Salton Cosine Index

We created twenty different networks, ten each using Pearson's correlation coefficient at $p<0.01$ and Salton Cosine Index with minimum value of 0.04. We present a visualization of one comorbidity network in Figure 2.8. In the visualization, the diseases are colored based on the 17 categories described by the ICD-9-CM. Size of a node represents its number of direct connections. It can be observed that some diseases are highly connected to other diseases whereas some are not connected at all. Some groups of diseases can also be observed indicating that the cluster of diseases are often diagnosed together in the patients.

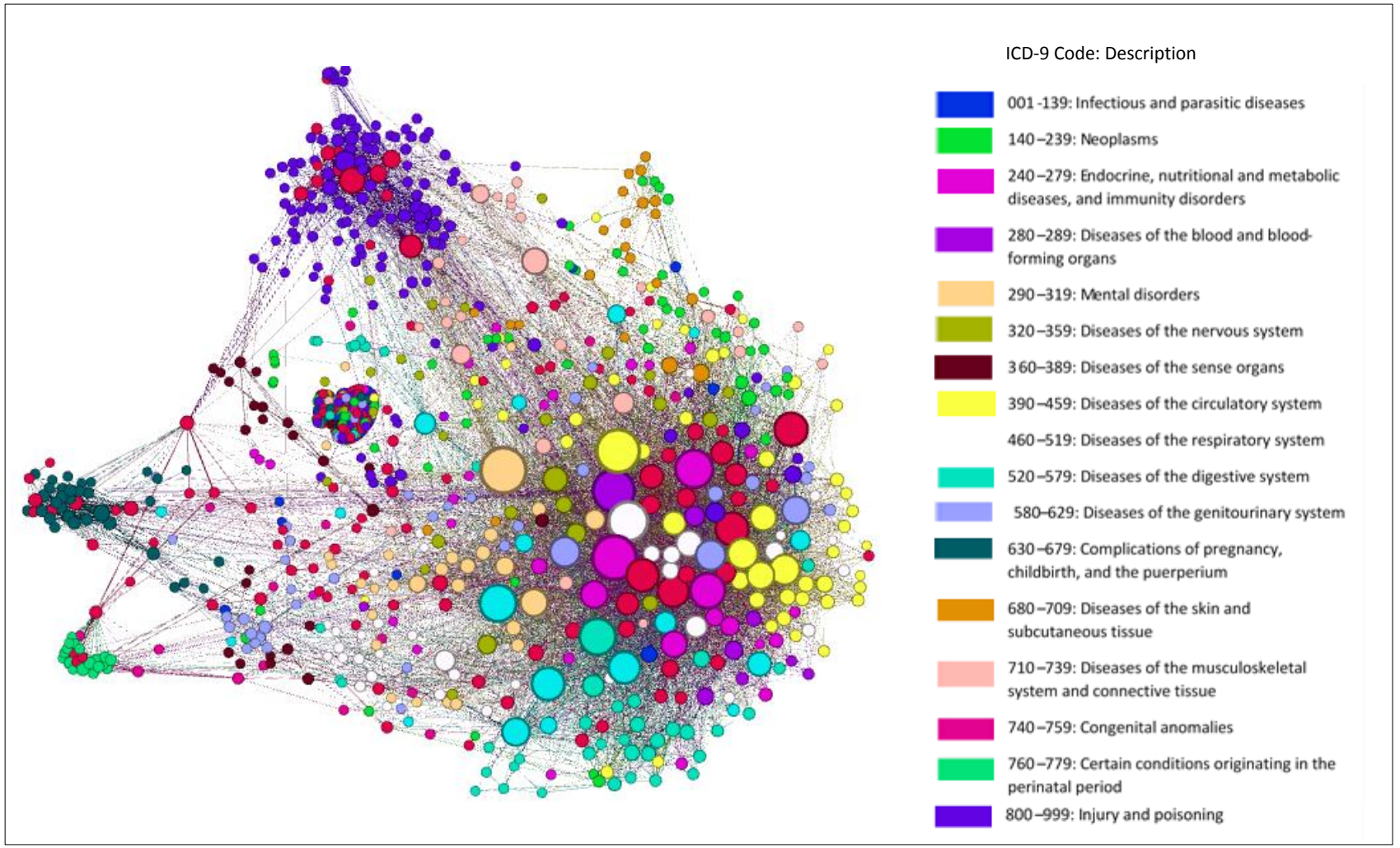


Figure 2.8. Comorbidity Network. A circle is a disease; an edge represents a comorbidity. Size of a node explains how well it is connected to other nodes.

The number of nodes and edges in all twenty networks are listed in Table 2.2. At the three-digit ICD-9 code level, there were 908 unique diseases or nodes in the network. The number of nodes remained the same until we decreased the random sample below 500,000 patients. The random sample of 250,000 patients contained almost 99% of nodes, the sample of 100,000 patients contained more than 97%, and the sample of 50,000 patients contained almost 95% of the nodes; however, the sample of 1000 contained substantially less diseases at only 63% of the total diseases.

Table 2.2. Comorbidity networks nodes and edges

No. of Patients	22.1 M	11 M	5.5M	2.75M	1.38 M	500,000	250,000	100,000	50,000	1000
Nodes	908	908	908	908	908	908	898	885	859	573
Edges										
PCC	14,463	11,088	8,354	6,120	4,284	2,506	1,632	866	508	5
SCI	14,463	14,311	14,283	14,238	14,195	14,178	14,073	13,769	13,502	5,935

While the number of nodes did not decrease substantially until the patient sample was below 50,000, the number of significantly correlated edges ($p<0.01$) in the network using PCC reduced with the decrease in sample size. The network created using 22.1 million patients comprised of 14,463 edges but the network using half of the patients contained 11,088 edges. Further, the network developed using 1000 random patients only had five significantly correlated edges (comorbidities or pairs of diseases).

On the other hand, the number of edges in the networks created using SCI did not change significantly with the number of edges remaining almost the same until the sample size decreased to 250,000 patients. In fact, looking at the density of the network, the density of networks using SCI remained the same throughout all the samples. On the other hand, the density drastically changed in the networks created using PCC. The change in density of the two types of networks with the indicated sample sizes are plotted in Figure 2.9.

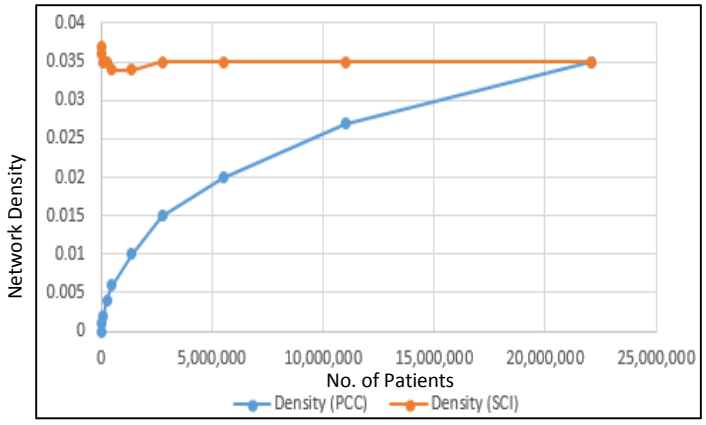


Figure 2.9. Density of networks from different sample sizes using Correlation and Salton Cosine Index

2.4.2. EFFECT OF SAMPLE SIZE ON THE OVERALL STRUCTURE OF COMORBIDITY NETWORK

Next, we demonstrate that the comorbidity network exhibits the small-world phenomenon (Watts & Strogatz, 1998). A network possesses a small-world property when multiple dense clusters are present in the network but the average path length (average distance between all pairs of nodes) is small, like a random network. The clustering coefficient, *C*, is a measure of a small-world network and explains the small clusters formed by the diseases. The clustering coefficient of a node explains how well the neighbors of a node are connected (Watts & Strogatz, 1998). With respect to the comorbidity network, the clustering coefficient of a disease, *d*, explains how well the direct connections of the disease, *d*, are connected to each other. The average clustering coefficient of the comorbidity network developed using 22.1 million patients was 0.487 (including all nodes). This means that on average 48% of the links were present among the neighbors of every node. The clustering coefficient of a node *d* can be mathematically written as

$$C_d = \frac{2l_d}{k_d(k_d-1)}, \qquad \textbf{-(2.4)}$$

where l_d is the number of links among the neighbors of the node d and k_d is the degree of a node d. For a network to possess small-world property, we require $n>>k>>ln(n)>>1$ (where n is the total number of nodes in the network) to make sure that the network is not disconnected into multiple sub-networks. In addition, two other conditions must be present: 1) the average path length of the network (P_{com}) needs to be approximately the same as the random network (P_{rand}) with the same parameters such as the number of nodes (n) and the average degree (k), and 2) the average clustering coefficient, C_{com}, of the network requires it to be greater than the average clustering coefficient of an equivalent random network (C_{rand}).

To calculate the small-world property of all our networks, we focused on the giant connected component of each network. A giant component contains the maximum number of connected nodes either directly or indirectly connected. For example, the largest connected component of the network with 22.1 million patients using Pearson's correlation coefficient contained 624 nodes or diseases (n) with the average degree of this largest connected component being 46.3 (k). The average path length and clustering coefficient of a random network with n=624 and k=46.3 can be calculated as $P_{rand} \sim ln(n)/ln(k)$ and $C_{rand} \sim k/n$ respectively.

$$P_{rand} \sim ln(n)/ln(k) \sim 1.68 \quad \textbf{-(2.5)}$$

$$C_{rand} \sim k/n \sim 0.074 \quad \textbf{-(2.6)}$$

The actual average path length of our network is 2.452 ($>\sim P_{rand}$) and the actual average clustering coefficient is 0.69 ($>>C_{rand}$). These numbers meet the requirements for the small-world property and hence, we prove that our comorbidity network followed the small-world topology. The largest connected component's number of nodes (n), average degree (k), average path length (P_{com}), average clustering coefficient (C_{com}), as well as the random network average path length (P_{rand}) and average clustering coefficient (C_{rand}) are listed in Table 2.3.

Table 2.3. Features of networks related to their topologies						
Sample Size	**n**	**k**	**$P_{rand} \sim \ln(n)/\ln(k)$**	**$C_{rand} \sim k/n$**	**P_{com}**	**C_{com}**
22.1 Million Patients						
PCC	624	46.3	1.68	0.074	2.452	0.69
SCI	631	45.8	1.69	0.072	2.460	0.69
11 Million Patients						
PCC	591	37.5	1.76	0.060	2.620	0.67
SCI	624	45.8	1.68	0.073	2.450	0.70
5.5 Million Patients						
PCC	540	30.9	1.83	0.057	2.760	0.66
SCI	619	46.0	1.68	0.074	2.440	0.70
2.75 Million Patients						
PCC	483	25.2	1.91	0.052	2.900	0.61
SCI	617	46.0	1.68	0.075	2.440	0.70
1.38 Million Patients						
PCC	411	20.7	1.99	0.050	3.080	0.60
SCI	612	46.0	1.68	0.075	2.430	0.70
500,000 Patients						
PCC	310	15.8	2.08	0.050	3.158	0.59
SCI	592	47.8	1.65	0.080	2.380	0.72
250,000 Patients						
PCC	240	13.2	2.12	0.055	3.145	0.60
SCI	584	48.0	1.65	0.080	2.362	0.73
100,000 Patients						
PCC	171	9.80	2.25	0.057	3.930	0.55
SCI	565	48.66	1.63	0.086	2.328	0.73
50,000 Patients						
PCC	104	8.60	2.16	0.082	2.827	0.59
SCI	550	49.0	1.62	0.089	2.273	0.73
1000 Patients						
PCC	-	-	-	-	-	-
SCI	399	29.7	1.77	0.074	2.204	0.72

n-number of node
k-average degree
P_{rand}-average path length of random network
P_{com}- average path length of our comorbidity network
C_{rand}-average clustering coefficient of random network
C_{com}-average clustering coefficient of comorbidity network

We compared the change in network structures of all the networks developed using PCC and SCI. Figures 2.10a to 2.10e compare the different features of the network topologies. As sample size decreases, the number of nodes in the largest connected component of network using Pearson's correlation coefficient (PCC) decreases drastically. In contrast, the number of connected nodes in the network using Salton Cosine Index did not change until we decreased the sample size to 1.38 million patients. Further reducing the sample size, the number of connected nodes started

decreasing. With respect to the change in the path length of the networks, we can observe in Figures 2.10b and 2.10c that the difference between P_{rand} and P_{com} remained almost constant in the two types of networks. In the network using PCC, P_{rand} and P_{com} slightly increased with the decrease in sample size but the difference between the two remained constant until the sample size decreased to 250,000 patients. In contrast, in the network using SCI, P_{rand} and P_{com} remain almost the same with the decrease in sample size. The difference between the two remained constant until the sample size of 50,000 patients.

A similar trend in the clustering coefficients of the two types of networks can be seen in Figures 2.10d and 2.10e. In the network using PCC, there is a slight random variation in the C_{com} with the change in sample size, but the difference between C_{rand} and C_{com} remained relatively large to keep the overall structure intact. We found the same dynamics in the network using SCI. The clustering coefficient C_{com} remained the same in all samples and the difference between C_{rand} and C_{com} remained constant keeping the overall structure as same.

For calculating the small-world property of each network, we require $n>>k>>ln(n)>>1$. As sample size decreased in the networks using PCC, the difference between k and $ln(n)$ became small. This means the comorbidity network became disconnected and formed multiple disconnected sub-networks as the sample size decreased. This violates the requirement for calculating the small-world property of the network. Therefore, the decrease in sample size affects the small-world property in the network using PCC. However, in the networks using SCI, the small-world property persists throughout. Overall, SCI preserved the overall structure of the network with small sample size but PCC could not.

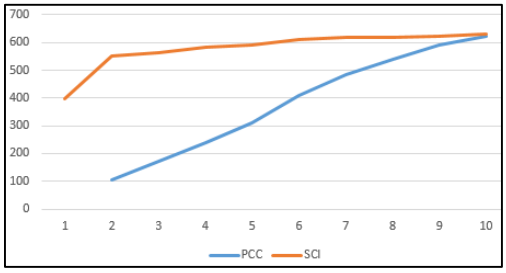


Figure 2.10a. Number of nodes in largest connected component

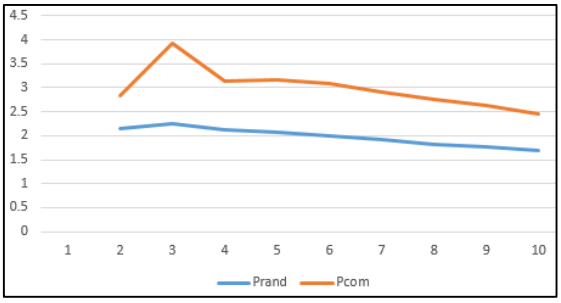


Figure 2.10b. Average path length in PCC network

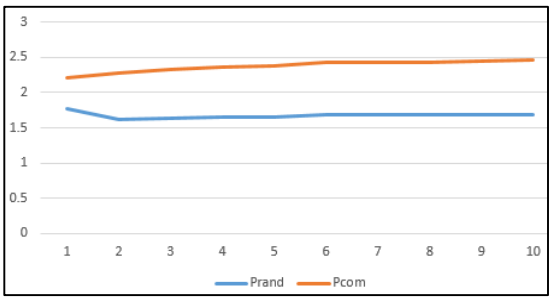


Figure 2.10c. Average path length in SCI network

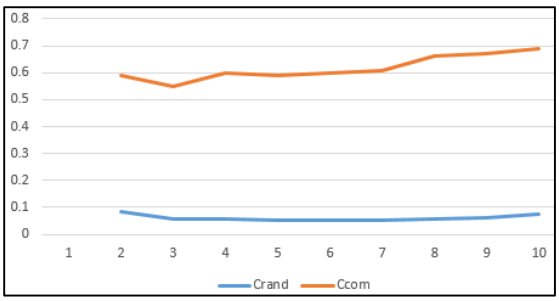


Figure 2.10d. Average clustering coefficient in PCC network

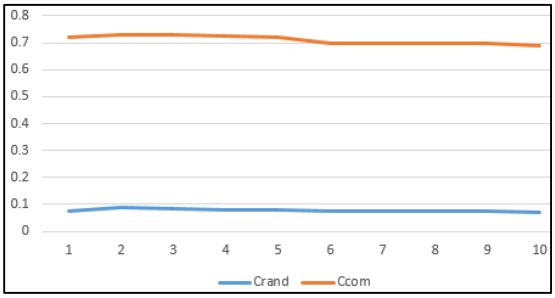


Figure 2.10e. Average clustering coefficient in SCI network

2.4.3. EFFECT OF SAMPLE SIZE ON NETWORK METRICS

When comparing the network metrics, we found interesting results (Table 2.4 lists the definitions of each metric along with the interpretation within the disease network context of our study). First, the influence of sample size on the measures is presented in Table 2.5 and Figures 2.11a to 2.11h. In the networks created using PCC, all measures decreased with the sample size as observed in the figures; however, in the networks using SCI, we found that the average degree, average weighted degree, average closeness, average clustering coefficient and average

Table 2.4. Network measures and their interpretation in our context		
Network Measure	**Definition**	**Interpretation in our context**
Degree centrality	Degree of a node explains its number of direct connections (Freeman, 1979)	Degree of a disease is the number of diseases directly connected.
Weighted Degree	Degree calculated considering the strength of an edge.	Degree considering comorbidity strength.
Betweenness Centrality	Number of times a node is on a shortest path among all shortest paths (Freeman, 1979)	Number of times a disease is a bridge between pairs of diseases.
Closeness Centrality	Closeness of a node gives the average shortest distance of that node to all other nodes in the network (Freeman, 1979).	Closeness centrality of a disease would represent how close a disease is to all other diseases in the network.
Eigenvector Centrality	Eigenvector centrality of a node explains how well the direct connections of a node are also connected (Bonacich, 1987).	How well the neighbors of a diseases are related to other diseases.
Clustering Coefficient	The clustering coefficient of a node explains how well its neighbors are connected (Watts & Strogatz, 1998).	Clustering coefficient of a disease explains how well the direct connections of the disease are connected to each other.
Network Density	Density of a network is the proportion of edges present in the network.	It explains how dense is the disease network.

eigenvector centrality of the networks remained constant until we decreased the sample size to 1000, where we see fluctuation in all the measures. Hence, the network using as small as 50,000 patients remains consistent with respect to the network measures. On the other hand, we observed that the average betweenness of the networks is the most inconsistent network measure among all. The average betweenness of the network did not change until we decreased our sample size to 1.38 million patients; however, it suddenly decreased with the smaller sample sizes. Hence, with the smaller sample size, average betweenness is not a valid measure to make conclusions.

Table 2.5. Comorbidity networks properties										
No. of Patients	22.1 M	11 M	5.5M	2.75M	1.38 M	500,000	250,000	100,000	50,000	1000
Average Degree Centrality										
PCC	31.857	24.42	18.4	13.48	9.436	5.52	3.635	1.957	1.183	0.017
SCI	31.859	31.52	31.46	31.36	31.267	31.229	31.343	31.116	31.437	20.72
Average Weighted Degree Centrality										
PCC	1.983	1.74	1.5	1.263	1.023	0.731	0.554	0.358	0.245	0.009
SCI	2.516	2.498	2.494	2.49	2.48	2.478	2.493	2.505	2.557	2.989
Average Betweenness Centrality										
PCC	310.78	311.1	282.94	243.69	193.49	114.1	68.72	48.29	11.74	--
SCI	319.9	311.2	303.96	301.88	293.41	265.39	258.23	239.1	223.82	166.9
Average Closeness Centrality										
PCC	0.297	0.267	0.246	0.218	0.188	0.15	0.135	0.084	0.089	--
SCI	0.298	0.296	0.295	0.294	0.294	0.289	0.289	0.29	0.297	0.323
Graph Density										
PCC	0.035	0.027	0.02	0.015	0.01	0.006	0.004	0.002	0.001	--
SCI	0.035	0.035	0.035	0.035	0.034	0.034	0.035	0.035	0.037	0.036
Average Clustering Coefficient										
PCC	0.487	0.445	0.4	0.333	0.282	0.218	0.172	0.116	0.092	--
SCI	0.487	0.486	0.484	0.484	0.484	0.476	0.478	0.471	0.473	0.5
Network Diameter										
PCC	6	6	7	8	9	9	8	11	8	1
SCI	6	6	6	6	6	6	6	5	5	5
Average Eigenvector Centrality										
PCC	0.146	0.126	0.11	0.092	0.076	0.058	0.046	0.033	0.027	--
SCI	0.143	0.142	0.142	0.141	0.141	0.141	0.144	0.145	0.148	0.139

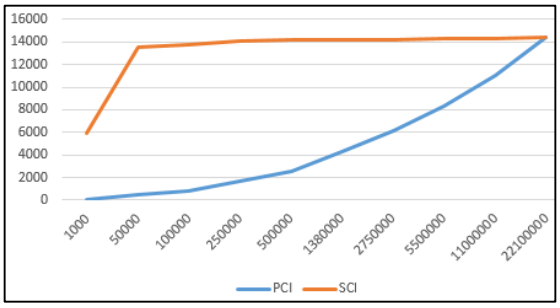


Figure 2.11a. Number of edges

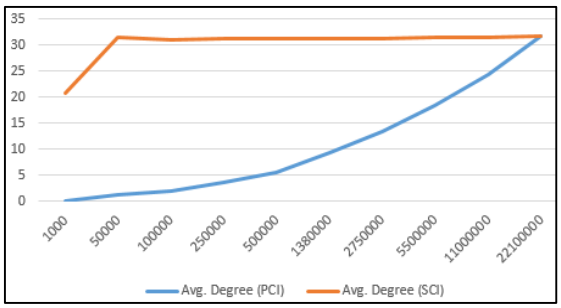


Figure 2.11b. Average Degree

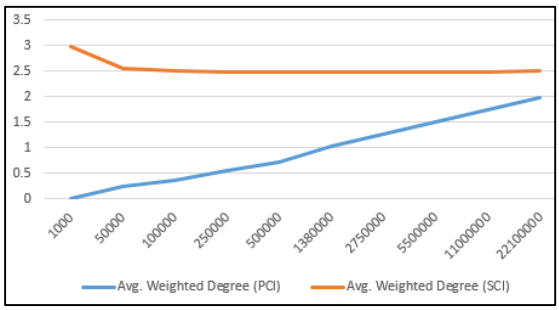


Figure 2.11c. Average Weighted Degree

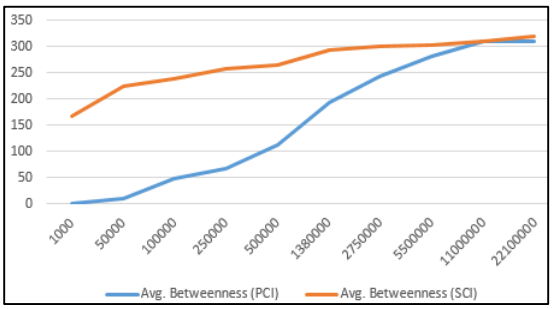


Figure 2.11d. Average Betweenness

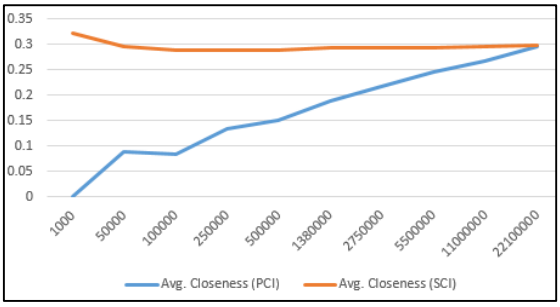


Figure 2.11e. Average Closeness

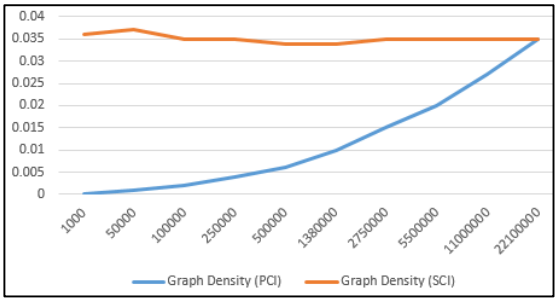


Figure 2.11f. Network Density

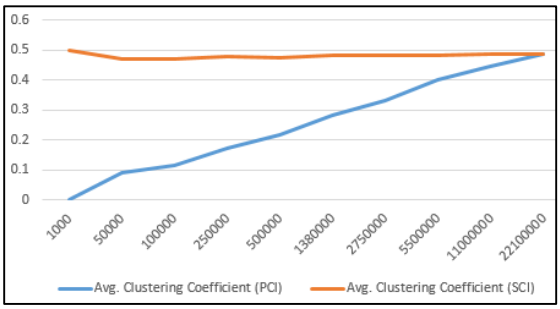


Figure 2.11g. Average Clustering Coefficient

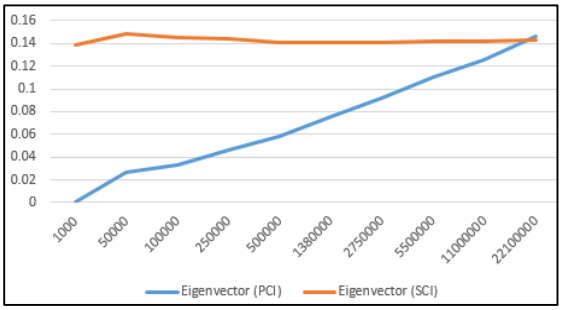


Figure 2.11h. Average Eigenvector

2.5. DISCUSSION AND CONCLUSION

In this paper, we set out to study the impact of sample size on an implicit network created using Pearson's correlation coefficient (PCC) and Salton Cosine Index (SCI). We found that PCC is not an appropriate index to draw relationships between nodes in an implicit network. The network properties and overall topology of the network using PCC get affected by sample size. On the other hand, we showed SCI to be an applicable measure for creating an implicit network because it does not depend on the sample size.

We observed that the decrease in sample size reduced the number of statistically significant correlated relationships between nodes. In other words, the number of edges in the network using PCC decreased with the decrease in sample size. The highly correlated nodes still existed in the small sample size but the rare connections did not. Therefore, if the purpose is to find highly correlated nodes, PCC can be used. However, if the objective is to make conclusions on rare

connections, PCC will not be able to catch those. In contrast, if SCI is used to make an implicit network, the same relationships can be observed in the network using small sample size that were seen in the network with large sample size. Therefore, it is recommended to use SCI if the sample size is small.

With respect to the overall structure of the network, the network using PCC became disconnected with the decrease in sample size. However, the network using SCI possessed the small-world topology in the networks from all sample sizes. Because our demonstration was on a small-world network, we note here that our conclusions are generalizable to other implicit networks that follow the small-world property. Small-world property is observed in the brain network (Bassett & Bullmore, 2006), language networks (Solé et al. 2010), social networks, actor-actor network, power-grid network, and many others (Watts & Strogatz, 1998).

Overall, we recommend researchers to consider SCI over PCC to create implicit networks. Our analysis is valuable to researchers studying networks as it establishes the need for choosing a right measure to create an implicit network for making valid conclusions.

The large dataset allowed us to study a much wider array of nodes. At the same time, the sample does not have to include all the millions of records to provide useful insights. In our analysis, we found that one can use a sample of 100,000 patients or 50,000 patients sample to study comorbidities or network properties (except betweenness) respectively. However, these numbers can vary with the type of network studied. Therefore, we encourage other researchers to perform the same analysis in other types of networks. Moreover, our network followed small-world topology but other networks may follow a different topology. One must find the specific topology of their specific network.

We add to the network theory by comparing the structure of networks developed using different sample sizes. Our recommendation to use SCI over PCC can help study the true relationships in

an implicit network. The use of SCI for creating a network preserves its structural properties even with smaller sample size.

CHAPTER III

EXAMINING HEALTH DISPARITIES BY GENDER: A MULTIMORBIDITY NETWORK ANALYSIS OF ELECTRONIC MEDICAL RECORD[1]

ABSTRACT

Multimorbidity health disparities have not been well examined by gender. Co-occurring diseases may be mutually deleterious, co-occurring independently, or co-occurring from a common antecedent. Diseases linked by a common antecedent may be caused by biological, behavioral, social, or environmental factors. This paper aims to address the co-occurrences of diseases using network analysis. We identify these multi-morbidities from a large Electronic Medical Record (EMR) containing diagnoses, symptoms and treatment data on more than 22.1 million patients. We create multimorbidity networks from males and females medical records and compare their structural properties. Our macro analysis at the organ-level indicates that females have a stronger multimorbidity network than males. For example, the female multimorbidity network includes six linkages to mental health, wherein the male multimorbidity network includes only two linkages to mental health. The strength of some disease associations between lipid metabolism and chronic heart disorders is stronger in males than females. Our multimorbidity network analysis by gender identifies specific differences in disease diagnosis by gender, and presents questions for biological, behavioral, clinical, and policy research.

[1] This paper is under review at a journal.

3.1 INTRODUCTION

Multiple ecological levels interact to influence disparities in health and health outcomes by gender. Health disparities observed between genders are caused by genetic, hormonal, physiological, behavioral, and sociocultural factors. Life expectancy at birth is notably longer for females at 81.4 years compared to males at 76.4 years (National Center for Health Statistics, 2016). During this longer lifetime, females are more likely to visit the hospital or health care provider, but less likely to die (Oksuzyan et al. 2008). Notably this male-female health-survival paradox is explained by chronic diseases which are most prevalent by gender: females are more likely to experience pain, reproductive cancers, and depression, while males are more likely to experience cardiovascular disease and diabetes (Case & Paxson, 2004). Additionally, when males and females are compared on the same chronic diseases, males may experience severe cases of chronic disease. Previous epidemiological studies of health disparities address individual diseases experienced by gender; however, most patients are diagnosed with multiple diseases. The goal of this paper is to explore disparities among males and females diagnosed with more than one disease, and present research and policy implications.

Two terms are often used to discuss the presence of more than one disease in a patient: comorbidity and multimorbidity. Comorbidity is a condition when an additional disease is diagnosed in presence of an index disease (Feinstein, 1970). Multimorbidity is defined as the coexistence of multiple chronic diseases and conditions in a patient (van den Akker, Buntinx, & Knottnerus, 1996; van den Akker et al., 1998). Throughout this manuscript we will use these terms interchangeably to denote co-occurrence of diseases, unless we need to specifically highlight the differences between comorbidity and multimorbidity. Previous studies on comorbidities have controlled for gender but rarely focused and reported differences in genders explicitly as pointed out by Short, Yang & Jenkins (2013). Further examination of comorbidities

by gender may be critically important for treatment of disease, and in identifying contraindications of common pharmaceuticals. The availability of large medical records affords the opportunity to study all possible disease relationships as observed in practice.

We adapt a network approach to model the multimorbidities (Euler, 1953). Networks are formed from the interactions between the elements or nodes. Network analysis has been used in health and medical literature to understand the interaction of genes (Goh et al, 2007; Ferrazzi et al. 2007), molecular involvement in disease (Barabási, Gulbahce, & Loscalzo, 2011), drug trials (Haslam and Perez-Breva, 2016), and historical epidemiological data on disease phenotypes (Hidalgo et al.,2009; Chen and Xu, 2014). Tai and Chiu (2009) applied association rule mining to create comorbidity network in ADHD patients using clinical database. Similarly, Chmiel, Klimek, and Thurner (2014) applied network approach to study the prevalence of different cluster of diseases over lifetime of genders. However, to the best of our knowledge, no one has applied this approach to study multimorbidity by gender in order to better understand health disparities.

In this paper, we develop and compare multimorbidity networks for males and females based on ICD-9 (International Classification of Diseases, Clinical Modification) codes of diagnoses. Our network comprises diseases connected based on the co-occurrences of diseases in 22.1 million patient records. The use of large dataset is another strength of our study. Knowing the relationships between diseases at the network level will enhance our understanding about disease associations at the patient population level.

3.2 METHOD AND ANALYSIS

In this section, we begin by describing the data and explaining how we measure the multimorbidity in our context. Next, we present a method to develop a multimorbidity network. Then, we briefly describe the properties of the network that can explain the position of a disease in a web of other diseases, and help us understand differences between males and females.

3.2.1 DATA DESCRIPTION

We obtained data from the Oklahoma State University Center for Health Systems Innovation (CHSI), which houses HIPAA compliant patient data provided by Cerner Corporation, a major Electronic Medical Record (EMR) provider. The data warehouse contains an EMR on the visits of 58 million unique patients across 662 US hospitals (2000-2016). We used information about the demographics of the patients, hospitals and disease diagnoses coded by ICD-9 system[2]. We removed several hospital visits in which patients were either not diagnosed with a disease or were marked only for symptoms. After data preprocessing, we had approximately 22.1 million unique patients with the sufficient information to perform analysis.

We extracted medical records for males and females in two different datasets from this pseudo-population dataset for comparing comorbidities by gender. The datasets were further cleaned based on the detected anomalies in particular category. For example, there were a few patients who were coded as a male during one visit and a female or null in another. Although males can also have breast diseases biologically, we removed the male patients diagnosed with such diseases with a suspicion that these are erroneously coded (ICD9: 610-612)[3]. We also removed males who were diagnosed with diseases such as inflammatory diseases of female pelvic organs (ICD9: 614-616)[4], and complications of pregnancy, childbirth, and the puerperium (ICD9: 630-679)[5]. Similarly, we removed female patients diagnosed with diseases of male genital organs (ICD9: 600-608)[6]. After cleaning the data, we had records of 12 million female patients and 9.9 million male patients. From the two samples, networks were created, one each for males and females.

[2] From the last quarter of 2016, the diagnoses in Cerner EMR are required to be coded in ICD-10 system. However, we did not consider the last quarter to maintain the consistency in our data analysis and considered only ICD-9 codes.

[3] There were 38,980 male patients with ICD-9 codes 610-612, which is 0.34% of the male database.
[4] 1,594 patients
[5] 20,009 patients
[6] 8,627 patients

3.2.2 MEASURING MULTIMORBIDITY

In the past, comorbidity and multimorbidity were largely defined at the cross-sectional level (Feinstein, 1970; Jakovljevic and Ostojic 2013). The chronic diseases, which we would not expect to go away in one hospital visit, could be overestimated from the medical records because they are recorded multiple times in an EMR. However, we delineate multimorbidity considering the lifetime history of a patient rather than a single hospital visit. We measure multimorbidity as the presence of multiple diseases in the lifetime history of a patient. This measurement has two advantages over previous definitions. First, the EMR recording of a disease over multiple hospitals visits is only considered once. Considering the same disease as different across hospital visits can overestimate its presence and bias the analysis and conclusions. Second, our definition considers the impact of a disease in one visit on subsequent visits. Therefore, it incorporates a wider span of disease developments. However, there is a concern of taking into account the association between diseases diagnosed across hospital visits occurring after long period of time. Given the relatively short time span of the database (17 years), short average length between first and last hospital visit in the database (527 days), average number of hospital visits of a patient being 5.1 (all types of visits including inpatient, outpatient, etc.) and statistical analysis on millions of patients, we mitigate the concern of false positives.

3.2.3 MULTIMORBIDITY NETWORK

A multimorbidity network developed from patients contains a set of nodes connected through edges. In our network, nodes represent diseases. In an EMR, an ICD-9 code of a disease has three, four or five digits (xxx.xx). The first three digits represent the broader category of a disease. The fourth and fifth digits represent the sub-divisions of the disease. For example, the ICD-9 code for personality disorder is 301. At four-digit level (301.x), there are ten types of personality disorders and at five-digit level (301.xx), two other specific personality disorders are coded. We aggregated ICD-9-CM codes to three-digit level. Thus, variations of the same disease

were considered as one node in the network. For example, multiple types of personality disorders mentioned above were aggregated into one node in our network.

An edge or connection between two diseases is created if these are comorbid. Since our focus is not to establish causality of a multimorbidity, we created a network with no direction in the relationships. For example, the comorbidity comprising congestive heart failure and rheumatic heart disease will be represented as an undirected edge between the two nodes representing the two diseases regardless of their causal relationship.

In the past, associations between diseases or comorbidities were modeled using a simple Pearson's correlation coefficient (Divo et al., 2015; Hidalgo et al., 2009). However, number of significant correlations is directly proportional to the number of observations used. Power to detect rare comorbidities is low because of the rareness of events. Therefore, to establish the right measure to model a comorbidity, we use a cosine index known as Salton Cosine index (Salton & McGill, 1986). SCI is immune to the total number of observations used (Ahlgren, Jarneving, & Rousseau, 2003) and measures the prevalence of a relationship between two diseases considering their individual prevalence. Salton Cosine Index, *SCI*, is calculated as in equation (1), where c_{ij} is the number of co-occurrences of diseases *i* and *j*, c_i is the prevalence of disease *i* and c_j is the prevalence of disease *j*. The cosine similarity has been used in the past to find phenotype overlaps (Chen et al. 2015; Lage et al. 2007). We propose this as an appropriate measure for finding the strength of a comorbidity.

$$SCI_{ij} = \frac{(c_{ij})}{\sqrt{(c_i * c_j)}} \qquad - \qquad (1)$$

Statistical significance of SCI was determined by assessing the relationship between correlation and SCI, because this approach has been suggested in the past to find the cut-off for SCI (Egghe & Leydesdorff, 2009). First, we determined the number of comorbidities significantly correlated

in a network created using Pearson's correlation coefficient. Then, we related the number of comorbidities in the network created using Salton Cosine Index and found a cut-off where number of significantly correlated comorbidities are equal in both networks. In the network from entire database using Pearson's Correlation Coefficient, at $p<0.01$, there were 14,463 significantly correlated comorbidities. Meanwhile, at the SCI cut-off of 0.04, the number of comorbidities were 14,463. Therefore, we used the cut-off of 0.04 for creating different networks for males and females. Then, the comparison between the networks was made using the network measures briefly described in the next section.

3.2.4 NETWORK METRICS

The structural properties of a network can be measured using several network metrics. These include degree, weighted degree, closeness and betweenness centrality (Freeman, 1979). In a multimorbidity network, the degree centrality of a disease (node) denotes the number of direct connections with other diseases. The weighted degree centrality of a disease considers the strength of the relationships with others and is calculated as a weighted sum of the strengths of the relationships. The closeness centrality of a disease determines an average number of steps it is away from other diseases in the network. A disease with higher closeness has a higher risk of being diagnosed with other diseases in less number of steps. Finally, the betweenness centrality of a disease describes its bridgeness. In other words, a disease with higher betweenness tends to be forming more bridges between other diseases.

3.3 RESULTS AND DISCUSSION

3.3.1 COMPARISON OF MALE AND FEMALE MULTIMORBIDITY NETWORKS

The visualizations of female and male multimorbidity networks are presented in Figures 3.1a and 3.1b respectively. In the visualization, the diseases are color coded based on the 17 categories/classes/organ systems described in the ICD-9 classification. These classes are also

listed in Figure 3.1. Size of a disease node represents an association with other disease(s), or its number of direct connections to other disease(s). The female multimorbidity network contains 300 diseases not connected to any other disease as compared to 265 diseases in the male multimorbidity network. In the female network, the diseases that are connected to at least one other disease in the network form three different sub-networks labelled as connected components. There is a primary cluster of diseases in the female network labelled as connected component-1 suggesting all diseases are associated to each other directly or indirectly. The two secondary clusters in the females were for burns (ICD9:941-945, 948, 949) and a pair of appendicitis codes (ICD9: 540-541).

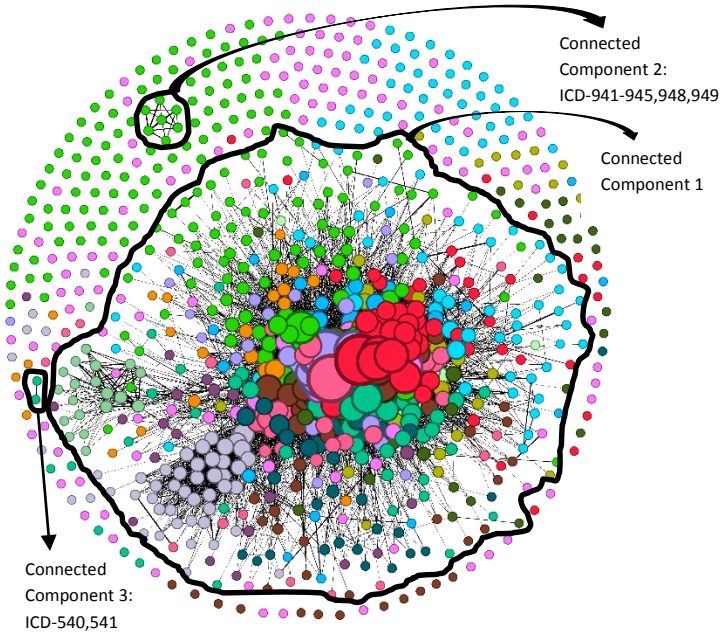
3.3.2 NETWORK PROPERTIES

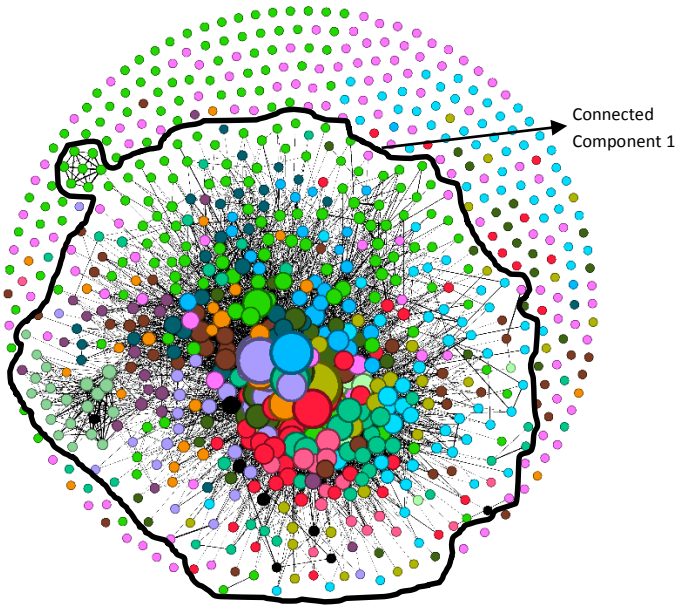
The properties of each network are listed in Table 3.1. The number of nodes or diseases in two networks are different as some diseases are unique to each gender. There were 839 diseases reported in males and 899 unique diseases in females at three-digit ICD-9 codes, that is, 7% more unique disease diagnoses in females. In the male network, there were 12,498 comorbidities as compared to 14,810 in females. Recall the edge strength denotes the magnitude of comorbidity. Out of all the edges detected above, 10,607 were common between both sexes. A pair-wise comparison of these 10,607 edge strengths indicates stronger comorbidities among females (t value=12.67, $p<0.0001$).

Although females have stronger and more comorbidities overall, we found some disease associations to be stronger in males than females.[7] These include disorders of lipid metabolism - chronic ischemic heart disease; disorders of fluid electrolyte and acid base balance - acute kidney

[7] It has to be noted that we focus on the top comorbidities based on their strength and not the frequency. In addition, comorbidities are discussed if they belong to distinct classes or organ systems listed in Figure 1 and Table.

failure; benign neoplasm of parts of digestive system and hemorrhoids - diverticula of intestine; diabetes - chronic ischemic heart disease; anemias - hypertensive chronic kidney disease; and





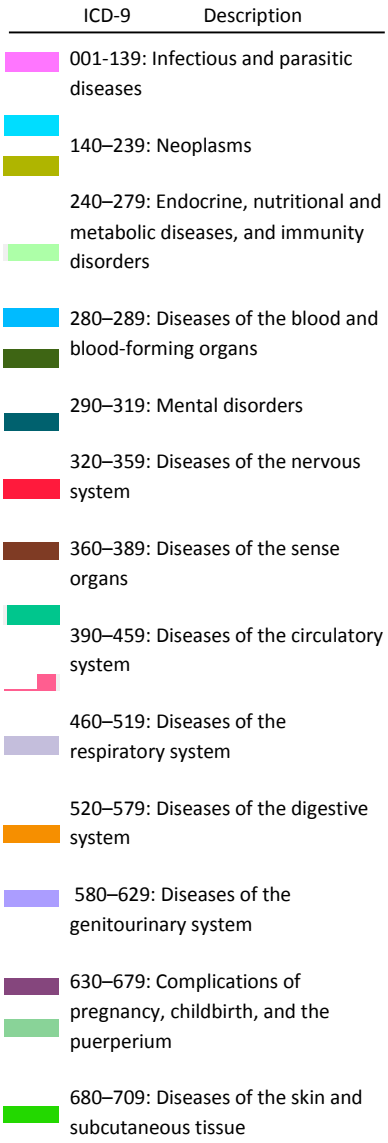


Figure 3.1a. Female Multimorbidity Network

Figure 3.1b. Male Multimorbidity Network

Table 3.1. Gender multimorbidity networks properties			
	Female	**Male**	**Pair-wise sample t test (Female-Male)**
No. of patients	12 M	9.9 M	N/A
Nodes (Diseases)	899	839	N/A
Edges (Comorbidities)	14,810	12,498	N/A
Avg. Degree (Degree of a disease is the number of diseases directly connected to it)	32.948	29.793	6.15, $p<0.0001$
Avg. Wt. Degree (Degree calculated as a weighted sum of the strength of the comorbidities)	2.592	2.365	4.50, $p<0.0001$
Avg. Betweenness (Number of times a disease is a bridge between pairs of diseases)	266.9	291.5	-1.78, $p=0.07$
Avg. Closeness (Closeness centrality of a disease would represent how close a disease is to all the other diseases in the network)	0.293	0.285	0.80, $p=0.42$
Graph Density	0.037	0.036	N/A

disorders of lipid metabolism – cardiac dysrhythmias. The average degree and weighted degree of the two networks were statistically different. Although the aggregated closeness and betweenness centralities of the two networks were not statistically different, we found several differences with respect to specific diseases in the two genders. For instance, acute upper respiratory infections and disorders of urethra & urinary tract form relatively more bridges between other diseases in females than males. On the other hand, the disorders of skin and subcutaneous tissue form a bridge between multiple other diseases more often in males than females.

3.3.3 ORGAN LEVEL NETWORK COMPARISON

We aggregated the relationships depicted in the networks in Figures 3.1a and 3.1b at the organ system level or class system categorized in the ICD-9 classification (See Table 3.2). We present two macro level networks at the class/organ system level in Figures 3.2a and 3.2b for females and males, respectively. The diseases of different classes were aggregated at the class level by adding up their weights (Salton Cosine Index). We highlight the connections between diagnoses of different classes if their aggregated weight is more than ten. This cut-off of ten is to study the

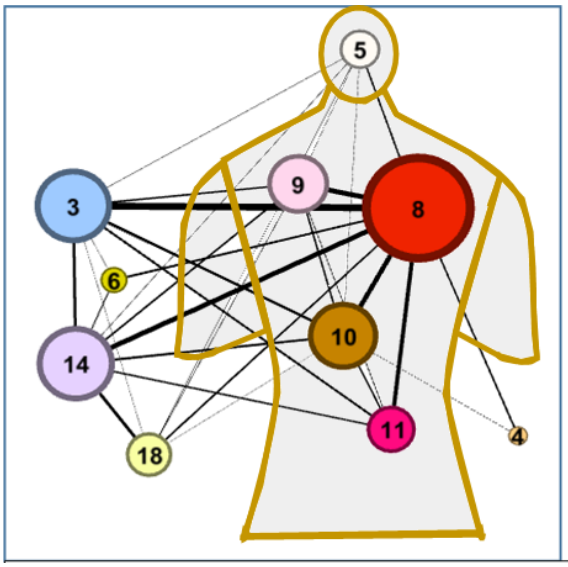
most prevalent relationships. However, one could select a lower cut-off to analyze the rare connections.

We present a unique way to visualize the relationships between disorders of different organ systems by creating an outline of a human body and mapping the categories of the diseases on it (See Figures 3.2a and 3.2b). In the ICD-9 classification, some categories can be directly related to the organ system present on a specific position in human body such as circulatory system (class-8), mental disorders (class-5), digestive system (class-10), respiratory system (class-9), and genitourinary system (class-11). However, other classes such as 1-4, 6-7, and 12-18 cannot be related to a specific organ system as listed in Table 3.2. The classes directly related to an organ system are mapped at the positions of the particular organ system in the human body. The classes that are not related to a specific organ system are presented outside the human sketch. The size of a node denotes the number of connections to other nodes. The width of an edge between two classes represents the aggregated weight (aggregated Salton Cosine Index) or the strength between them. The same connections can be observed in the Table 3.3 where a comparison is made between the two networks.

Table 3.2. ICD-9 code classification

Class No.	Description	ICD-9 codes range	Mapped on organ system
1	Infectious and parasitic diseases	001–139	No
2	Neoplasms	140–239	No
3	Endocrine, nutritional and metabolic diseases, and immunity disorders	240–279	No
4	Diseases of the blood and blood-forming organs	280–289	No
5	Mental disorders	290–319	Yes
6	Diseases of the nervous system	320–359	No
7	Diseases of the sense organs	360–389	No
8	Diseases of the circulatory system	390–459	Yes
9	Diseases of the respiratory system	460–519	Yes
10	Diseases of the digestive system	520–579	Yes
11	Diseases of the genitourinary system	580–629	Yes
12	Complications of pregnancy, childbirth, and the puerperium	630–679	No
13	Diseases of the skin and subcutaneous tissue	680–709	No
14	Diseases of the musculoskeletal system and connective tissue	710–739	No
15	Congenital anomalies	740–759	No
16	Certain conditions originating in the perinatal period	760–779	No
17[8]	Symptoms, signs, and ill-defined conditions	780–799	N/A
18	Injury and poisoning	800–999	No

[8] Not considered in the analysis



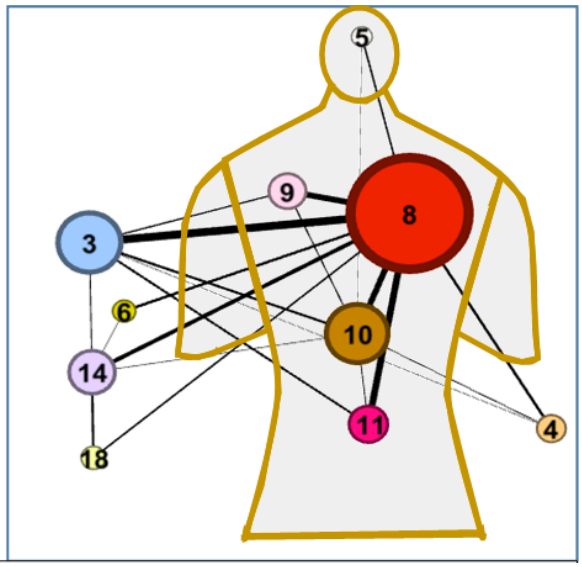


Figure 3.2a. Female Organ Comorbidity Network Figure 3.2b. Male Organ Comorbidity

Table 3.3. Class associations in Female and Male Networks

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18
Infectious and parasitic [1]																	
Neoplasms [2]																	
Endocrine, nutritional, metabolic, and immunity disorders [3]																	
Blood and blood-forming organs [4]			M														
Mental disorders [5]			F														
Nervous system [6]			F														
Sense organs [7]																	
Circulatory system [8]			FM	FM	FM	FM											
Respiratory system [9]			FM		F			FM									
Digestive system [10]			FM	FM	FM			FM	FM								
Genitourinary system [11]			FM					FM	F	FM							
Pregnancy, childbirth, and the puerperium [12]																	
Skin and subcutaneous tissue [13]																	
Musculoskeletal system and connective tissue [14]			FM		F	FM		FM	F	FM	F						
Congenital anomalies [15]																	
Perinatal period [16]																	
Injury and poisoning [18]			F		F			FM	F	F				FM			

** Class 17 is symptoms and thus not included in the analysis*

F Association only in Female Network

M Association only in Male Network

FM Association in both Networks

The Figures 3.2a-3.2b and Table 3.3 show which organ systems diseases are diagnosed simultaneously more often in different genders. The female network is clearly denser than the male network with more connections. Notably, there are several multimorbidities present in the female network not present in male network at the selected cut-off. These are highlighted in Table 3.3 and notated by an F in each area of comorbidity. There is only one males-specific comorbidity as compared to eleven comorbidities noted as significant only among females. For example, mental disorders in males are associated with the disorders of circulatory and respiratory systems. However, female patients with mental disorders are at risk of diagnoses belonging to multiple other organ systems such as circulatory respiratory, digestive and musculoskeletal systems in addition to the injury, poisoning, endocrine, nutritional, metabolic, and immunity related disorders. Similarly, the disorders of genitourinary system are strongly associated with the disorders of respiratory system, musculoskeletal system and connective tissue in females than males. The disorders of musculoskeletal systems are also more strongly connected to other disorders in women than men. The musculoskeletal disorders such as osteoarthritis are known to be more prevalent in females (Woolf & Pfleger, 2003) but other observed multimorbidity differences by gender need further research.

The above discussed relationships between diagnoses of different organ systems are more strongly connected in females than males. However, the comorbidity of endocrine, nutritional, metabolic diseases, and immunity disorders (3) with disorders of the blood and blood-forming organs (4) was only observed in males. Each connection needs further investigation so as to find the reasons for differences in genders. Recognition of these observed multimorbidities also may suggest greater precautions to be taken by patients themselves or the physicians to watch for related symptoms.

Our observed networks of comorbidities from the EMR data confirm the prevalence of higher comorbidities in females than males as supported by the previous research (Blazer, et al. 2002;

Moller-Leimkuhler, 2007). Notably, this study conforms previous work which identified a greater proportion of diagnosis of reproductive cancers and mental health diagnoses among females. However, contrary to previous research, we also note that the strength of some comorbidities are stronger in males than females.

Multimorbidity networks were different may be due to gender differences in care seeking behaviors among females because a greater frequency of care seeking behavior in females increases the risk of multiple disease diagnoses (Corrigan, 2004). Moreover, the disparity between mental health multimorbidities in males and females is striking: perhaps physician implicit bias (Chapman, Kaatz & Carnes, 2013) and patient care seeking behaviors play a role in the diagnosis of mental health disorders by gender. Previous research suggests that social factors discourage men from seeking mental health care (Corrigan, 2004). Therefore, the absence of strong multimorbidities with mental health among men was expected.

Notably, there was only one male dominated comorbidity in the network, disorders of blood and blood-forming collectively and disorders related to the endocrine, nutritional, metabolic, and immunity collectively, which were found more strongly connected in males than females. There could be a few potential explanations for this relationship: HIV infection or obesity. HIV infection is still the highest among men, and would be comorbid with an immunity diagnosis. Factors associated with obesity were strongly represented in the initial male multimorbidity network, many of these linked lipids, heart disease, diabetes, and digestive neoplasms; therefore, it is most likely that these diagnoses are linked to obesity, which has multiple antecedents addressed by public health.

Analysis of the network centralities (particularly weighted degree) suggested acute kidney failure, chronic kidney disease and chronic ischemic heart disease to be more strongly connected to other diseases in males than females. Moreover, diabetes mellitus emerged to be one of top diseases in

males in terms of closeness (but the closeness number of diabetes in males was still smaller than females). Diabetes diagnosis is typically associated with overweight and obesity, and is often multimorbid with cardiovascular and other diagnoses related to overweight and obesity.

3.4 CONCLUSIONS

Better understanding of multimorbidity networks may allow for better screening and identification of diseases among patient populations, accounting for uniqueness for males and females in research measuring multimorbidity. These networks may improve health outcomes and reduce healthcare costs associated with hospital length of stay and readmission. The impact of comorbidity on the health outcomes has been studied in the past, but different network related properties have not been discussed in the public health literature. We shall establish the relationships between these concepts as a part of our future research.

Our study contributes both to the method and practice. With respect to the method contributions, we presented a novel approach to study multimorbidities at a population level. The network approach allowed us to study all the multimorbidities at once. Our paper is one of the first to apply a network approach to understand public health, particularly in the context of comorbidity/multimorbidity. The knowledge extracted from the large historical data can improve clinical decisions and outcomes as discussed by Tierney (2001).

The analysis presented in this study has several practical implications. We mainly developed insights for health researchers. However, our study has implications for policy makers. In 1993, National Institute of Health (NIH) Revitalization Act was passed to encourage researchers to include women and minorities in clinical trials. Our analysis validates the disparities in diagnoses by genders, and thus we reinforce the need for considering the gender multimorbidities in clinical trials. In addition, education at every level should reinforce teaching of multimorbidity differences across population groups. We provide evidence to the gender disparities in public

health through multimorbidity lens and support the global calls by Ovseiko, et al. (2016), Johnson, et al. (2014) and other thought leaders to recognize gender differences in health research.

This study has few limitations. First, our multimorbidities were based on the electronic health records and therefore, only the diagnoses recorded in specific hospitals was included. It is perhaps impossible to record lifetime history of a human in medical records. Hence, this limitation remains in all studies based on medical records. Second, we focused on the gender. However, we also recognize that the health disparities exist based on race and ethnicity (Fine, Ibrahim & Thomas, 2005). Studying such disparities is part of our current research. Third, we only discussed simple network metrics such as degree, closeness and betweenness centrality. However, other complex measures such as clustering coefficient, cliques, clubs, eigenvector centrality, etc. can provide more information about the multimorbidities. Next, the differences were reported if the diseases were of different organ systems. However, the comorbidities related to the same organ system can also help enhance our understanding about the multimorbidities. We will explore these in future research. We also note that there were thousands of comorbidity differences in different population groups and we could not report them in this paper. Instead, we have attached supplementary materials containing information on relationship of every disease with others. One can focus on one particular disease and find multimorbidities in our provided material.

Notwithstanding these limitations, our study shows that Big Data and advanced analytics of large information can help gain new insights previously hard to discern (Tierney, 2001). We showed that advanced analytics methods such as network analysis can provide additional dimensions to understand the public health. Our study analyzed a dataset of millions of patients where diseases form a network and suggest that the structure of a network can have several implications. Moreover, there are several differences in different population groups in terms of multimorbidity network that should be considered while dealing with the comorbidities. Our study opens up an

exciting and important area of research for policy makers, economists, social scientists and medical experts to treat different groups of population differently.

CHAPTER IV

HEALTH ANALYTICS LEAD TO MORE QUESTIONS: A COMORBIDITY LENS APPROACH

ABSTRACT

As we amass more data, we have an opportunity to analyze a pseudo population to better understand differences in health across population groups. The way patients belonging to different population groups develop comorbidities can have a major impact on their health outcomes. The differences in the diagnoses associations across populations groups can be examined by studying comorbidities found in the historical Electronic Medical Record (EMR). In this chapter, we apply the data analytics approach to extract knowledge about the comorbidities rooted in EMR. To model comorbidities, we draw on the network theory and develop multiple comorbidity networks based on co-occurrences of diseases in different population groups. We create and compare comorbidity networks for different races, Medicaid/non-Medicaid patients and Medicare/non-Medicare patients. This leads to developing multiple research questions that need to be explored in the future research. The interesting findings and theory implications are discussed.

4.1 INTRODUCTION

The sample size in many past studies looking at the impact of diagnoses on health has been an issue. New questions may emerge by knowing more about the diagnoses and their interactions among each other from a much larger sample that is more reflective of nearly the size of the population. Due to the lack of availability of sufficient data and advanced technologies, past clinical research has largely focused on studying the impact of diseases on fewer patients. The conclusions derived from studying fewer patients might not be rigorous, complete and generalizable.

Due to the acceptance of Electronic Medical Record by the hospitals, availability of health data for pseudo population is now possible. This gives the opportunity to apply Big Data technologies and techniques for analyzing such large datasets and ask new questions. Due to availability of massive datasets, it is now possible to study all possible diagnoses and their interactions at the same time. The interaction of a disease with other diseases may have different consequences. Studying diseases at the relational or interactional level can provide additional insights about their joint impact on the health or health related metrics such as the expected hospital length of stay, readmission rate, etc.

The objective of this chapter is to extend the analysis performed in Chapter 3 to other groups of population based on race and insurance type. For race, six different comorbidity networks are compared: Pacific Islander, Asian, Caucasian, African American, Hispanic and Native American. For insurance type, first a comparison between Medicaid and non-Medicaid patients is made. Then, networks for Medicare and non-Medicare patients are created and compared. These comparisons will help generate new research questions because such comparisons at the comorbidity level have not been possible in the past due to the lack of data.

4.2 COMORBIDITY NETWORK ANALYSES OF RACES

Is race a socially constructed term or can it be characterized genetically? There has been a long argument going on for years among public health researchers. There are two groups in public health; one of which supports the argument of social construction and another argue for genetic differences. Although two groups are ideologically apart, both agree that there are health disparities among different races. The research on race health disparity can be divided into three components as postulated by Fine, Ibrahim, and Thomas (2005): 1) identifying health disparities, 2) understanding the reasons for disparities, and 3) developing interventions to eliminate disparities. In this study, we contribute to the first line of research and identify differences in races using the administrative data of more than 22 million patients in US hospitals.

Previously, researchers have identified differences in diagnoses among races using administrative data. For example, Bresnahan et al. (2007) found babies born from African Americans mothers were more likely than whites to be diagnosed with schizophrenia. However, to the best of our knowledge, no one has comprehensively studied the differences in races through comorbidity lens.

For creating the race specific networks, we removed patients with ambiguous entries such as patients those were reported belonging to one race at one time and another at other time. In addition, the sample size difference in race was of higher moment of magnitude with 14 million Caucasians, 3.5 million Afro-Americans, 590,000 Hispanics, 400,000 Asians, 158,000 Native Americans and 25,500 Pacific Islanders. Since the lowest sample size was quite small among all the races (25,500 for Pacific Islanders), we randomly extracted almost equal number of patients as the second smaller sample i.e. 158,000, for all races. This led us compare the multimorbidity networks at the same level with balanced samples. Using each sample, a comorbidity network is created. Therefore, a total of six networks are created and compared.

Table 4.1 lists the properties of comorbidity networks of all races. All the networks have almost equal number of diagnoses except Hispanics. The number of unique diagnoses in Pacific Islanders is perhaps driven by the sample size but it forms the densest network of all with 15,064 pairs of diagnoses. Considering others, the number of diagnoses pairs in African-Americans were found highest followed by the Caucasian, Native Americans, Asians and Hispanics being the least. Now the question arises that can these results be attributed to differences in genetics in the races? Or are these difference due to the differences in the facilities available to diagnose or record such data to some races? Although researchers have both types of arguments to answer these questions, these need further explanation with respect to multimorbidity.

Table 4.1. Race comorbidity networks properties

African American	Caucasian	Hispanics	Asian	Native American	Pacific Islander
No. of patients					
176,093	173,282	167,086	159,975	157,880	25,414
Number of unique diagnoses					
892	884	872	889	887	829
Number of Connections					
15,871	13,389	6,390	8,623	11,560	15,064

In the literature, the impact of comorbidities on the health outcomes of different races has been studied. For instance, Olson and authors studied the impact of race and comorbidity on survival rate in endometrial cancer patients (Olson et al., 2012). However, how and why comorbidities differ across different races and reasons for the differences are not much explored. Is it because the clinical trials are dominated by white males? (Oh et al., 2015). This is a very strong argument and needs to be addressed in future research.

Like in Chapter 3, we present comorbidity network of different organ systems for each race in Figures 4.1a to 4.1f. To compare all differences at one place, we listed all connected in Table 4.2. We clearly see the differences in the networks. For example, the link between the disorders of

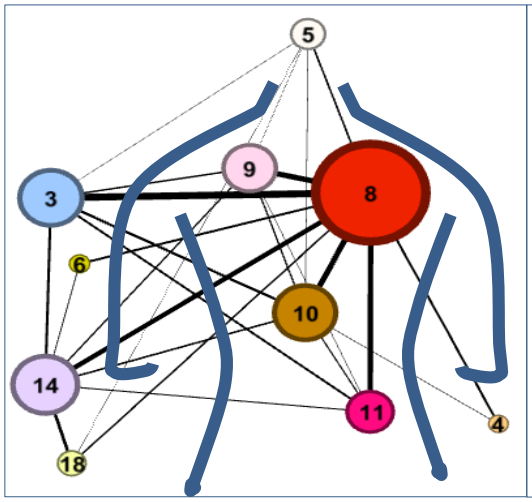


Figure 4.1a. Caucasian Network

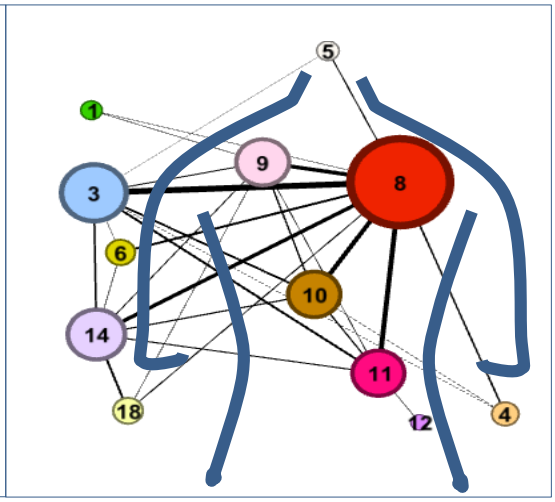


Figure 4.1b. African- American

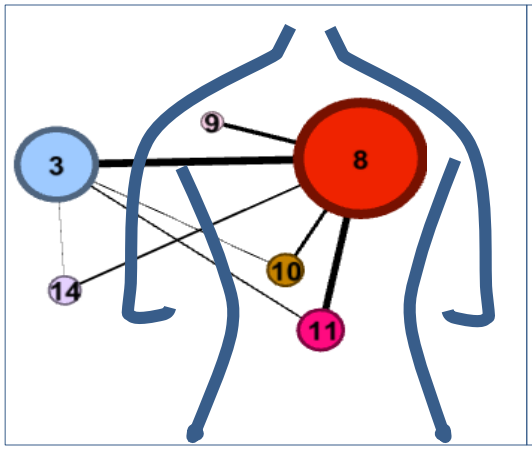


Figure 4.1c. Asian Network

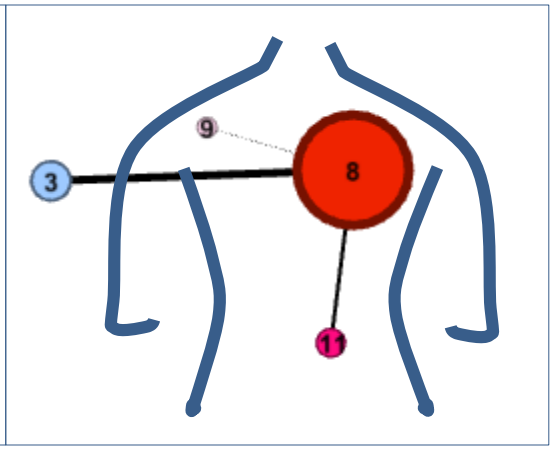


Figure 4.1d. Hispanic Network

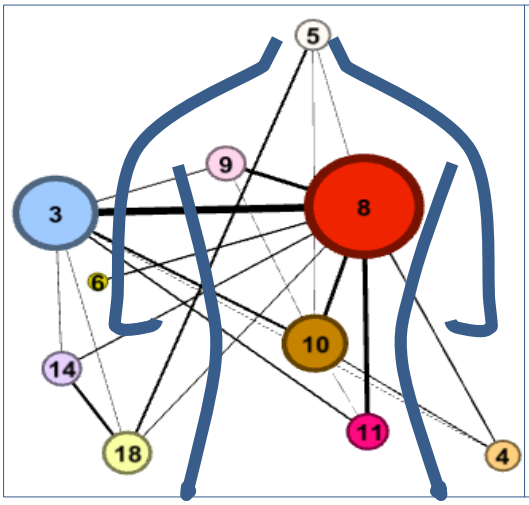


Figure 4.1e. Native Network

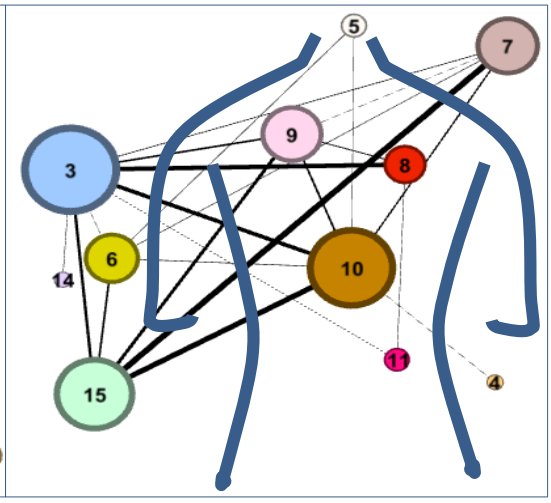


Figure 4.1f. Pacific Network

Table 4.2. Comorbidities across races												
F – African Americans	**A - Asians**		**C – Caucasians**		**H- Hispanics**		**N – Native Americans**			**P-Pacific Islanders**		
	4	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**14**	**15**	**18**
Infectious and parasitic [1]					F	F						
Neoplasms [2]												
Endocrine, nutritional & immunity disorders [3]	F N	F C	F P	P	F A C H N P	F C N P	F A C N P	F A C N P		F A C N P	P	N
Blood and blood-forming organs [4]					F C N		F C N P					
Mental disorders [5]			P		F C N	C	C N P					C N
Nervous system [6]				P	F C N		P			F C	P	
Sense organs [7]						P	P				P	
Circulatory system [8]						F A C H N P	F A C N P	F A C H N P		F A C N		F C N
Respiratory system [9]							F C N P	F C		F C	P	F
Digestive system [10]								F C N		F C	P	
Genitourinary system [11]									F	F C		
Pregnancy, childbirth, and the puerperium [12]												
Skin and subcutaneous tissue [13]												
Musculoskeletal system and connective tissue [14]												F C N
Congenital anomalies [15]												
Perinatal period [16]												
Injury and poisoning [18]												

F African-Americans C Caucasians A Asians H Hispanics P Pacific Islanders N Native Americans

the digestive system (10) and genitourinary system (11) is only present in African-American network and not in others. Moreover, the comorbidities involving disorders of sense organs are more prevalent in pacific islanders. Among all, the Hispanic is least dense than others. Does that mean Hispanics are healthier than others on an average or are there other reasons? These findings are supported by the Hispanic paradox, which argues that Hispanics enjoy mortality advantage (Markides & Eschbach, 2005). Can comorbidity lens explain more about the health of Hispanics? Sociologists and medical experts have to further research on these differences from the comorbidity length.

4.3 COMORBIDITY NETWORK ANALYSES OF MEDICAID AND NON-MEDICAID PATIENTS

We performed an interesting analysis to find comorbidity differences in poor and non-poor population. To do so, we extracted two samples; one with Medicaid patients and other without Medicaid. By Medicaid patients, we mean the patients who were enrolled in Medicaid programs throughout their lifetime and non-Medicaid patients mean that they were never enrolled in the Medicaid programs. There were about 15% patients in the dataset who were enrolled at least once in Medicare program. However, there were about 1.2 million patients (5.7%) patients who were always in Medicaid patients. We used only those patients assuming them as poor. To compare the poor with non-poor, we extracted an equivalent sized random sample of patients who were never enrolled in Medicaid programs at any point of time in their life.

The networks' statistics are listed in Table 4.3. The networks drawn from the samples show that the poor patients get diagnosed with 23% less number of comorbidities than the non-poor patients. For many years, it has been continuously reported that Medicaid patients get lower quality of care (Thompson et al. 2003; Yazdany et al., 2014). Due to the lower quality of care, do

poor patients not get diagnosed with all comorbidities? This clearly need further explanation from comorbidity perspective.

Table 4.3. Medicaid and non-Medicaid comorbidity networks properties	
Medicaid	**Non-Medicaid**
No. of patients	
1,253,513	1,277,998
Number of unique diagnoses	
906	907
Number of Connections	
7,792	10,147

To see the differences at the class/organ system level in two groups of patients, Figures 4.2a and 4.2b are presented. There are only two comorbidities in Medicaid network above our cutoff. This is a very critical issue and require immediate attention.

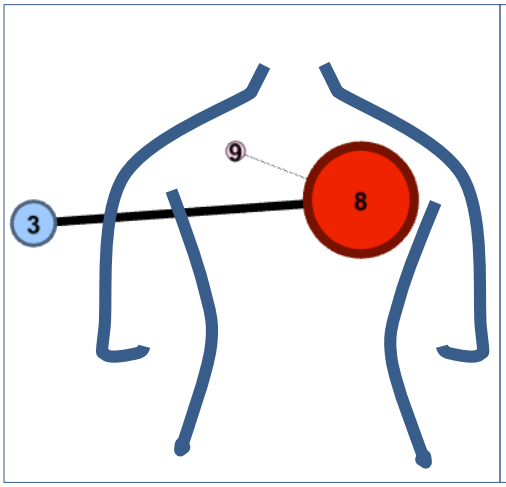


Figure 4.2a. Medicaid Network

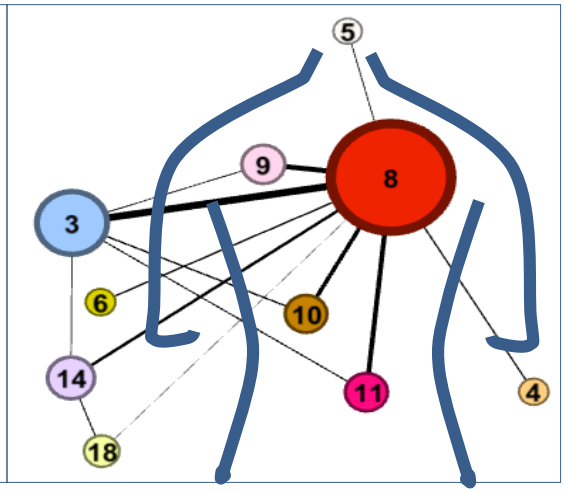


Figure 4.2b. Non-Medicaid Network

4.4 COMORBIDITY NETWORK ANALYSES OF MEDICARE AND NON-MEDICARE PATIENTS

Another comparison we performed was between Medicare and non-Medicare patients. This comparison mostly presented the comorbidity networks differences based on age. The Medicare

patients are at least 65 years of age. We had 3.4 million unique patients who were enrolled in any Medicare program. To compare the comorbidities in relatively older patients, we extracted an equal number of patients enrolled in other payer programs. It is well known that the number of diagnoses and comorbidities increases with age and we found similar results in comorbidity networks. Table 4.4 presents the network properties of the two networks. The number of comorbidities in Medicare patients were 72% more than the non-Medicare patients at our Salton Cosine Index cutoff of 0.04.

Table 4.4. Medicare and non-Medicare comorbidity networks properties

Medicare	Non-Medicare
No. of patients	
3,441,719	3,211,775
Number of unique diagnoses	
908	908
Number of Connections	
18,248	10,549

Similar to other comparisons, we compare the Medicare and non-Medicare comorbidities at the organ system/class level as shown in Figures 4.3a and 4.3b. As expected, there are huge differences between the two networks. A lot more edges are present at the organ level in Medicare patients as compared to the non-Medicare patients.

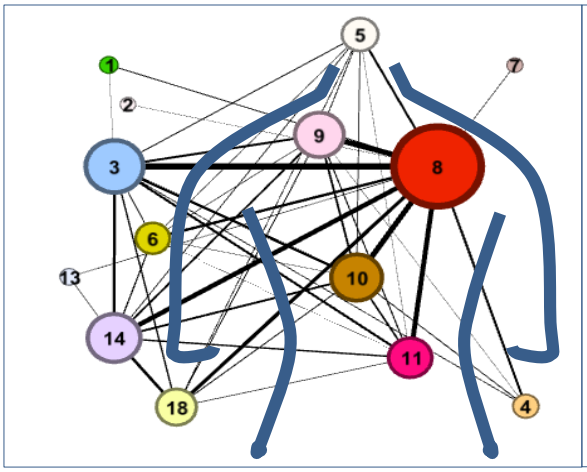


Figure 4.3a. Medicare Network

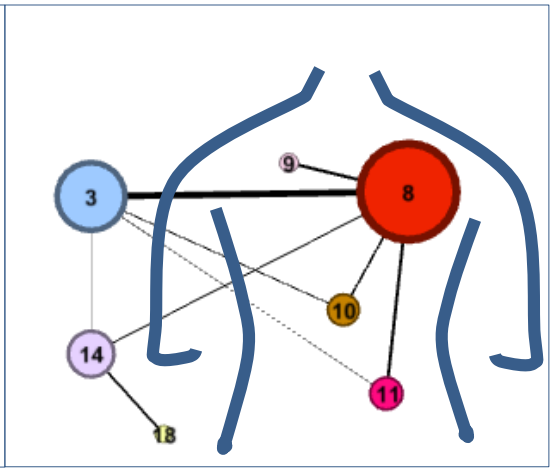


Figure 4.3b. Non-Medicare Network

4.5 DISCUSSION

We performed comprehensive descriptive analyses of the comorbidity networks across the US populations using the data analytics approach of extracting knowledge from the data. We have presented many comorbidity differences in the population groups. These differences are required to be analyzed by the researchers in the future. These differences raise many questions for the medical experts, social scientists, economists and policy makers to answer. The new interesting insights and answers to the questions can help medical experts improve their decisions.

We add to the data analytics literature by illustrating the power of descriptive analysis. Through data analytics, we dig into the data and based on the results, new questions emerged. This approach is similar to the overarching idea of Grounded Theory Methodology (GTM) (Glaser & Strauss, 2009; Glaser, Strauss, & Strutzel, 1968; Strauss & Corbin, 1967). In the past, GTM is discussed to be similar to the data analytics methodology. For instance, Müller et al. (2016) suggested to use the lens of GTM in data analytics studies. The authors summed up their argument as follows:

"IS researchers choosing to apply a BDA [Big Data Analytics] approach might want to consider some of the principles of grounded theory. Like grounded theorists, BDA researchers will spend an extraordinary amount of time on understanding the nature of the data, particularly if they have not collected them themselves." (Müller et al., 2016, p. 3)

Both data analytics and GTM involve discovering relationships between concepts from the data. There are methods available to perform data analytics such as CRISP-DM (Shearer, 2000) and SEMMA (Azevedo & Santos, 2008). However, these lack guidelines to connect the outcomes to the theoretical contribution. A methodology to perform rigorous and relevant data analytics study is required that can provide clear guidelines to add to the knowledge base. We believe GTM can provide rigorous guidelines to perform data analytics.

The analysis presented in this chapter has several practical implications. The effect of one disease on the other is explored in the past, however, we comprehensively studied all relationship between diagnoses in one picture. These types of analyses are only possible if such a large dataset equivalent to the population is available.

The descriptive analyses provided evidence that the relationships between diagnoses are different across patients. Based on the differences in relationships, we provided insights on the comorbidities in particular. Moreover, these relationships can have different impacts on the health outcomes of the patients such as mortality rate, length of stay and readmission arte. Following this, we establish the relationship of network properties with length of stay and mortality risk in Chapters 5 and 6 respectively.

CHAPTER V

WHEN WILL I GET OUT OF THE HOSPITAL? MODELING LENGTH OF STAY USING COMORBIDITY NETWORK

ABSTRACT

A reliable and accurate estimate of the expected hospital length of stay (LOS) of a patient is important to patients, hospitals, and insurance companies. But predicting LOS is a complex, ill-structured, and dynamic decision-making problem. While recognizing that multiple factors interact with each other when predicting LOS, we specifically focus on the impact of co-occurrences of diseases in a patient (known in medical terms as comorbidities). Comorbidity has been used most often in the previous research to explain the length of stay. However, it has rarely been used to predict LOS, because the information about the entire hospital visit is required to know the actual comorbidities. To model and predict comorbidities from a large database containing medical history of patients, we create a comorbidity network in which co-occurring diseases form relationships. The network helps predict likely comorbidities at the point of admission based on the primary diagnosis of the patient. Because there is a gender disparity in comorbidity, we develop different networks for men and women using information on one million patient records in 662 US hospitals (2000-2016), The structural properties of the network are used to measure the comorbidities in a patient and create a model to explain and predict patients' LOS using another set of 2.2 million patient visits. The performance of our model is

compared with the extant models, which it outperforms. The theoretical and practical implications of our study are discussed.

5.1 INTRODUCTION

When will I get out of the hospital? This is the first question every patient asks when he or she is admitted to the hospital, because a longer stay increases costs in terms of health, time, and money. In addition to its importance to the patient, a reliable and accurate estimate of the expected length of stay (LOS) at the point of admission also helps healthcare providers and insurance companies. From a hospital's perspective, LOS is an important metric to measure the quality of care as discussed by Thomas, Guire and Horvat (1997). Prolonged stays also increase utilization of beds, care, staff, and equipment, and negatively affect the efficiency of patient flow systems. Given that hospital inpatient care constitutes nearly one-third (29%) of all healthcare expenses in the United States, it is important to correctly estimate LOS to manage workloads across departments and accurately plan for discharges to minimize readmissions. An early estimation of LOS is also crucial to insurance companies as it is directly related to the total payments made to providers. It can help with precertification (determining whether the selected medical services meet criteria for medical necessity under the member's benefits contract) and estimation of actuarial cost during admission. Given the importance of predicting LOS at the point of admission to different stakeholders, this paper attempts to create an explanatory and predictive model for estimating it. Because past attempts at predicting LOS resulted in low accuracy and limited applicability across multiple diseases, the problem remains open for further study.

Prediction of LOS, like many other healthcare problems, is a wicked, complex, ill-structured, and dynamic decision-making problem, as argued by Meyer et al. (2014). A problem may be defined as complex and ill-structured when its underlying state is unknown, it has multiple competing

outcomes, and it is affected by the interaction of multiple factors (Funke, 1991; Simon, 1973). Moreover, the environment under which it is tackled is dynamic and unpredictable, with time-delayed consequences for actions taken. Predicting LOS is also complex because the underlying reasons for a hospital stay may not be clearly identified. Moreover, several patient-level, disease-level, hospital-level, and unknown factors (e.g. medical injuries during hospitalization) interact with each other in this environment. It is important to study these multiple factors to predict LOS.

In this study, in addition to focusing on patient and hospital-level factors, we also examine the interactions of diseases with one another. To model complex interactions among diseases and symptoms in patients, we apply a network approach. This approach provides a theoretical understanding of the embeddedness of a disease or symptom within the web of diseases, and can present the complex structure of a system in a smooth and stable form. These terms were used by Dhar et al., (2014) when they applied the approach to convert a complex structure of products purchased together on Amazon.com into a network, relating structural properties with product demand.

A network is a representation of interconnected parts, or nodes, of a system linked through well-defined connections or edges. This approach cuts across all traditional disciplines of science including management, engineering and social sciences. We use it to create a network of diseases to understand the collective behavior of diseases and their effects on the health outcomes, particularly LOS. Our contribution is to employ network properties inferred from such analyses and then use those for predictive analytics through second level model building. In other words, properties derived from the network models are used further for predictive analysis of an outcome (i.e. LOS), which is exogenous to the network.

To know how diseases form relationships and interact with each other, we use information on comorbidities (a medical condition in which two or more diseases are diagnosed simultaneously

(Feinstein, 1970)). For example, the simultaneous presence of diabetes and pneumonia in a patient is a comorbid condition. Comorbidity is considered an important factor in health outcomes such as mortality, LOS, and readmission because two diseases jointly can have a different impact than the same two diseases individually. This joint impact is known as syndemicity, a term coined by Singer (1996).

In this paper, we develop a comorbidity network based on co-occurrences of diseases in a large number of patients. In our network, diseases form connections based on comorbidities. A connection in a comorbidity network has a different meaning than one in a traditional network, such as an online social network. In a social network, members make the decision to connect with others. In a comorbidity network, a connection represents an aggregation of comorbidities in patients, who are external to the network. An aggregation of a large number of patients provides a simple model of disease relationships. Developing such a model explicitly based on each disease's attributes would be incredibly complex and probably futile.

Comorbidity has been used most often in the previous research to explain the length of stay. However, it has been rarely used to predict LOS, because the information about the entire hospital visit is required to know the actual comorbidities for that specific patient. In contrast, our comorbidity network is used to predict the possibility of other diseases in the presence of an index disease, and therefore can be used to make predictions for LOS at the point of admission. To the best of our knowledge, this is the first such use of a comorbidity network for predicting LOS.

To develop our comorbidity network and make predictions, we use an Electronic Medical Record (EMR) containing information about more than 24.7 million patients across 662 US hospitals over 17 years (2000-2016). The use of this massive dataset is one of the strengths of our study. Due to the lack of availability of sufficient data and advanced technologies, past research largely

studied the impact of fewer diseases or fewer patients which may have resulted in conclusions that were not rigorous or complete.

In addition, we develop separate comorbidity networks for men and women. Previous studies controlled for gender but rarely focused on and reported differences in males and females explicitly, as noted by Short, Yang and Jenkins (2013). It is important to study gender differences because men and women might respond differently to a disease in the presence of another. Some thought leaders such as Ovseiko, et al. (2016) and Johnson, et al. (2014), have strongly argued for the recognition of gender differences in health research, and we respond to their call.

Our study contributes to the Information Systems literature by developing an information-based model for the healthcare industry that predicts LOS. This model improves decision-making outcomes and can be integrated to enhance existing Information Systems. Our models and algorithms generate business value from the data, which is one of the contributions of analytics research as deliberated by Agarwal and Dhar (2014). In addition, our approach is robust and easily implementable.

The network approach used in our study helped us create a new measure for quantifying and predicting the comorbidities in patients. As Shmueli and Koppius (2011) explained in their commentary, measure development is one of the roles of analytics in scientific research. Our measure enhances the understanding of the joint impact of diseases on LOS. In addition, we add to the existing knowledge by comparing our approach with competing models.

This study also contributes to the network science and analytics literature by presenting an application of how a network and its properties can be used for modeling. The network provides researchers with a new dimension that improves the contextual intelligence about comorbidities. The network illustrates the direct as well as indirect relationships among diseases, and the position of a disease in the network defines the risk associated with it, known as structural risk

(Coleman, 1988). The risk possessed by a disease can have several direct or indirect consequences on patients, hospitals, and providers. We measure structural risk through the network metrics such as node centralities. Our paper is one of the first to analyze this risk in the context of comorbidities and syndemics.

Our network is one of a class that emerges without the intentions of its source, often referred to as unintentional or implicit (Roth et al., 2010). Patients are the source, but the network is formed without their intentions. This class differs from traditional or explicit networks, which are developed from the intentional actions of its members. A list of networks is presented in Table 5.1, half of which are formed without the intentions of their source.

Table 5.1. Different categories of network based on the purpose of formation

Network Name	***Source of Network***
Intentionally Formed	
Online Social Network	Users
World Wide Web	Web pages and links
Road Network	Interconnected roads
Power Grid Network	Cables connecting power units
Airport Networks	Airlines connecting airports
Unintentionally Formed	
Co-citation Network	References in documents
Actor-Actor Network	Movies
Product-Product Network	Users purchases
Drug-Drug Network	Patients or hospital visits
Comorbidity Network	**Patients or hospital visits**

We take the application of implicit networks to next level, exploring how they impact the uncontrollable performance of the external source. In other words, can a collective behavior of the network formed unintentionally affect outcomes of the source? In our comorbidity network, positions of the diseases are used to predict LOS, which is an uncontrolled performance. This approach can be applied to solve problems in other domains. For instance, product co-purchase network can be used to predict future spending based on the current shopping cart.

In section 2, we review related works on LOS, comorbidities, and the research gaps this study attempts to fill. In section 3, we describe our baseline models and comorbidity network models, and include a demonstration of comorbidity network building. In section 4, we discuss our data processing and analysis, and address issues relating to Electronic Medical Record (EMR). We also present the explanatory and predictive power of the models proposed in this study. In Section 5 we discuss the implications of our results.

5.2 LITERATURE REVIEW

5.2.1 LENGTH OF STAY

The problem of predicting length of stay (LOS) has been studied for a long time. We present a list of competing studies and identify the current state of art in Table 5.2. The table indicates whether the predictive performance of the models developed by the studies is reported, and whether the models are applicable at the point of admission. The majority of these studies focused on the entire hospital visit, and therefore did not have predictive ability.

Previous models built to estimate LOS used common factors at the patient level and hospital level and some also considered external factors. Patient level factors included age, gender, race, and diseases diagnosed in the patients. Some studies also considered lab tests and procedures to explain LOS, such as Yang et al. (2010), Clague, et al. (2002), Liu et al. (2010) and Chertow, et al. (2005). There is some evidence that insurance type affects LOS, with Medicaid insurance holders having longer stays than others (Mainous, et al. 2011; Lopez-Gonzalez, et al. 2014). At the hospital level, the size of a medical unit, consultant, and clinical events occurring during the visit were found to be significant in some studies, e.g. Elixhauser et al. (1998). Huntley, et al. (1998) showed that LOS was affected by the number of previous admissions because it indicates that the patient's situation is critical. External variables affecting LOS included the day of the week when a patient was admitted. For instance, Carter and Potts (2014) found a longer stay for

Table 5.2. A review of selected papers on Length of Stay and Comorbidity							
Paper	**Setting**	**Sample Size**	**Comorbidity**	**Point of admission**	**Predictive Performance**	**Network Properties**	**Performance**
Chertow, et al. (2005)	Acute kidney injury	19,982	N	N	N	N	R-Square -33%
Lowell & Davis (1994)	Schizophrenia and Affective disorder	829	N	N	Y	N	Accuracy – 35%
Lopez-Gonzalez et al. (2014)	Medicaid, Uninsured and Private insurance	20.8 million	N	N	N	N	Not reported
Thombs, et al. (2007)	Acute burn injury in 70 burn centers	31,338	Y	Y	N	N	Not reported
Librero, et al. (1999)	Public healthcare system in 12 hospitals	106,673	Y	N	N	N	Not reported
Lyketsos, et al. (2002)	Psychiatric inpatients in a hospital	950	Y	N	N	N	Not reported
Furlanetto & da Silva (2003)	Inpatients in a general ward of a hospital	317	Y	N	N	N	Not reported
Mainous et al. 2011	Ambulatory care–sensitive (13 categories of diseases)	849,866	Y	N	N	N	Not reported
Carter & Potts (2014)	Knee operation	2,130	Y	Y	Y	N	Classification – 24.5%-76.9%
Hachesu et al. (2013)	Coronary artery disease	2,064	Y	N	Y	N	Classification - 96.4%
Liu et al. (2010)	All – one healthcare system (multiple hospitals)	155,474	Y	Y	N	N	R-square-14.6%
Clague, et al. (2002)	Hip fracture in a hospital	662	Y	N	N	N	R-square- 20.7%
Huntley, et al. (1998)	One Psychiatric center	769	Y	Y	Y	N	R-Square-17.6%
Rochon et al. (1996)	Spinal-cord injured	330	Y	N	N	N	R-square – 6.2%
Elixhauser et al. (1998)	Inpatients from 438 acute care hospitals	1.7 million	Y	N	N	N	R-square - 12%-39%
Yang et al. (2010)	Sepsis patients in one hospital	6,929	Y	N	N	N	R-square - 21%-34%
This Study	All types of patients in 662 hospitals across US	3 million	Y	Y	Y	Y	R-Square-38% (Max 72%) Accuracy: 65%

knee operation patients if admitted on Sunday, Tuesday, or Wednesday. Another factor considered by several studies was the discharge destination, which can affect hospital discharge decisions. Carter and Potts (2014) found shorter LOS for patients discharged to home as compared to those discharged to other facilities. The above brief review helped us determine the variables we used to create the baseline models in our study that are available at the point of admission, such as demographic, insurance type, admission type, and hospital size.

As discussed earlier, prediction of LOS is a wicked and ill-structured problem, as evidenced by the performance of models in previous studies. The explanatory power of these models is as low as 6% r-square for spinal-cord injury patients as found by Rochon et al. (1996) and as high as r-square of 39% for patients with low back pain by Elixhauser et al. (1998). With respect to the predictive power in terms of mean absolute error, again the accuracy is less than 35% on average. Some studies have also attempted to predict a range of LOS such as Carter & Potts (2014) and Hachesu et al. (2013). However, this approach is less practical because suggesting a range of LOS may not help physicians and insurance companies with decision-making.

5.2.2 LENGTH OF STAY AND COMORBIDITY

Comorbidity has been shown to be related to the LOS in the past, as the presence of two or more conditions impacts patients' stays because more care and resources are required to cure those conditions jointly. Many studies focused on one disease at a time, identifying the comorbidities related to a hospital stay. For instance, Hachesu et al. (2013) studied comorbidities of coronary artery disease, Yang et al. (2010) considered sepsis, Lowell and Davis (1994) included schizophrenia and affective disorder, Thombs, et al. (2007) examined burn injury, Lyketsos, et al. (2002), Furlanetto and da Silva (2003), and Huntley, et al. (1998) analyzed psychiatric-related comorbidities.

In the literature, there is no consensus on the definition of comorbidity. The concept of comorbidity is theorized using four distinctions by Valderas, et al. (2009). The first is based on the nature of the health conditions occurring concurrently. The second is based on the relative importance of the co-occurring conditions. In this case, one disease is given more importance than the others., and the presence of other conditions in addition to it is considered comorbidity. The third distinction is based on the chronology of development of the conditions. It is possible that multiple diseases develop concurrently or one disease leads to another, but it is not easy to draw causal relationships between them. Finally, the fourth distinction considers illness burden and patients' socioeconomic conditions that can play a role in the presence of multiple diseases. In these four distinctions, Valderas and colleagues considered patients' clinical and socioeconomic factors to conceptualize comorbidity. On the other hand, Jakovljevic and Ostojic (2013) defined comorbidity as a medical condition in three different ways based on only diseases diagnosed in patients. First, it is when two diseases are present simultaneously but independently. Second, it is when one disease causes another, making them interdependent. Third, it is the presence of multiple diagnoses regardless of their causal relationships.

These previous definitions of comorbidity do not consider the lifetime history of a patient, but rather look at the presence of diseases only during a hospital visit. Focusing on *current* patients' information can better help physicians to control comorbidities; however, how the *history* of a patient is related to the current situation is not understood. If we instead look into the lifetime history of patients and find relationships between diseases, this will provide additional understanding about comorbidities. This is the approach we adopt in this paper. We define comorbidity as *the presence of multiple diseases in the lifetime history of a patient.* This definition has two advantages over previous definitions. First, the medical recording of a disease over multiple hospitals visits is only considered once. Considering the same disease as different across hospital visits can overestimate its presence and bias the analysis and conclusions. Second,

our definition incorporates the impact of a disease on other diseases *across* multiple hospital visits, thereby considering a wider span of disease developments. Our definition is also useful for prediction purposes because a comorbidity developed in one patient during one hospital visit may be observed across two hospital visits in another patient. Therefore, considering a longer time span allows us to draw true relationships between diseases. There is some concern about considering the association between diseases diagnosed during hospital visits that occur with long intervals in between. Given the relatively short time span of our database (17 years), the short average length between first and last hospital visit in the database (527 days), the average number of hospital visits of a patient (5.1, including all types of visits), and the statistical analysis on millions of patients, we mitigate the concern of false positives.

To measure comorbidity, many comorbidity scales have been proposed. de Groot et al. (2003) identified twelve different indexes, concluding that the Charlson Index, the Cumulative Illness Rating Scale (CISR), the Index of co-existent Disease (ICED), and the Kaplan Index are reliable measures and can be used in clinical research. Since CISR, ICED, and Kaplan require clinical judgment and information, we consider only the Charlson Index when comparing the performance of our model because it is based on medical records.

The Charlson Index assigns weights to 19 different medical conditions. It was originally created for predicting mortality. Later, it was found to be related to LOS by several studies. For instance, Librero, Peiró, and Ordiñana (1999) conducted a simple bivariate analysis to see the increase in LOS with respect to the different levels of Charlson Index comorbidity scores, finding that LOS increases with the Charlson Index score. Rochon et al. (1996) also found a significant relationship between LOS and comorbidity indexes, however with low effect ($R^2 = 0.06$).

Past studies have rarely used comorbidity scales for prediction purposes because information about an entire visit is necessary to know all diseases in a patient. Some of those studies are listed

in Table 5.2. If information about comorbidities at the point of admission is not available, it is difficult to use it to make predictive models for LOS. Therefore, a mechanism is required to first predict the comorbidities and then use it for LOS predictions. Another limitation of the past research on comorbidity is that the existing scales are unable to consider all the diseases at the same time. Through our network approach, we can study all interactions at once.

From the literature review, we identify three research gaps. First, most past studies restricted their analyses to a few categories of diseases and patients, and their models are only applicable to the specific types of patients they studied. Our study aims to create a unified model for patients and to analyze the model performance for different categories of patients. Second, the Charlson Index considers only a subset of diseases. In this paper, we apply a network approach to measure all disease relationships from a database encompassing millions of patients in the United States. We also compare the model developed using our approach with the model using Charlson Index. Third, to the best of our knowledge no one has applied a network approach and used structural properties to predict LOS. Our study attempts to fill all of these research gaps.

5.3 MODEL DEVELOPMENT

In this section, we describe baseline models using the variables considered in past studies, and then describe how we develop a comorbidity network and use its properties to build models at hospital-visit level to explain and predict LOS

5.3.1 BASELINE MODELS

We created different baseline models to explain and predict LOS using the variables considered in the previous studies. From the preceding literature review, we identify the following information to build our baseline models: demographics (age, gender, race), visit (primary

diagnosis, patient type, admission type, number of previous admissions), hospital (size measured by number of beds), and insurance type. In addition, we also create a few new variables that can affect LOS. One measures the number of times a patient had already been admitted due to the same primary diagnosis, because it is possible that LOS decisions made by providers can be influenced by multiple admissions with the same diagnosis. Another variable is the number of different organ systems involved in the primary diagnoses and other diseases present during admission. This is important because LOS can be influenced by efforts and resources required to simultaneously cure diseases belonging to different organ systems. For instance, Braunstein et al. (2003) studied elderly patients with chronic heart failure and found non-cardiac comorbidities to be associated with health outcomes. We also calculate the Charlson Index from the known diseases at the time of admission to compare the performance of this extant index with our proposed measure. It was calculated as the weighted sum of scores of known diagnoses as suggested by D'hoore, Sicotte, and Tilquin (1993).

We build four hierarchical baseline models to study the impact of different factors. In the first linear baseline model, only patient demographics, visit characteristics (excluding primary diagnosis) and hospital information are used (See equation 5.1). The patient-level information includes age, gender and race; visit characteristics include patient type, admission type, insurance (payer), number of previous admissions (*visitNumber*) and number of previous admissions due to the same disease (*revisit*); and hospital data include size of the hospital in terms of number of beds. For each multi-level variable (such as race, hospital size, admission type and payer) a set of parameters is estimated. Each of the parameters $\beta_3, \beta_5, \beta_6$ and β_7 represents a set. For the variables having one level, (i.e. $\beta_1, \beta_2, \beta_4, \beta_8$ and β_9) a single value parameter is estimated.

Baseline 1: $LOS = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 race + \beta_4 patientType + \beta_5 admissionType + \beta_6 payer + \beta_7 hospitalSize + \beta_8 revisit + \beta_9 VisitNumber + e$ **-(5.1)**

The second baseline model is built by including the information about the primary disease of the patient and number of different organ systems or categories involved in the known diagnoses to the first baseline model (see equation 5.2). This model considers the impact of the primary reason for the visit on LOS. To compare the performance of our model with the existing model of comorbidity, a third baseline model is created to estimate LOS using a competing measure, Charlson Index (see equation 5.3). The final hierarchical baseline model is created using the all the variables defined in first three baseline models to see their joint relationship with LOS as presented in equation 5.4.

$$\textit{Baseline 2: LOS} = \textit{Baseline 1} + \gamma . \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix} + e \quad \textbf{-(5.2)}$$

$$\textit{Baseline 3: LOS} = \beta_0 + \alpha . chi + e \quad \textbf{-(5.3)}$$

$$\textit{Baseline 4: LOS} = \textit{Baseline 1} + \textit{Baseline 2} + \textit{Baseline 3} \quad \textbf{-(5.4)}$$

5.3.2 MODELING USING COMORBIDITY NETWORK

The traditional baseline models described above do not consider comorbidity; however, they represent important factors for explaining and predicting LOS. Since all comorbidities might not be known at the time of admission, we use the relationships between diseases through historical patterns to predict comorbidities related to the primary diagnosis. These relationships from the patterns can also help estimate LOS. In this sub-section, we explain the development of our comorbidity network and describe how network metrics such as centralities of a disease can impact LOS. Then, the models using network properties are described.

5.3.2.1 COMORBIDITY NETWORK

In a comorbidity network, diseases are connected to each other if they are likely to co-occur in a patient. As discussed earlier, comorbidity networks are implicit and a link between two diseases is

defined from the co-occurrences in a specified time interval. For our definition of a comorbidity, the lifetime history of a patient is considered as a transaction. This is analogous to creating a lifetime market basket for a buyer based upon multiple individual transactions. Of course, we recognize that the history may be incomplete. A patient may have gone to a hospital or hospitals that do not use same data collection system, and thus records may be missing. But this is still the best available compilation. A transaction may contain multiple diseases diagnosed over time, and in that case they will be used to discern associations among diseases. Let $T(d_1, d_2, d_3, \ldots d_t)$ denotes a transaction, where $d_1, d_2, d_3, \ldots d_t$ is a subset of all diseases $D(d_1, d_2, d_3, \ldots d_n)$ with $n \geq t$.

A comorbidity network developed from N patients is denoted by $C\ (D, E)$ where D is a set of n nodes and E is a set of edges. In our comorbidity network, nodes represent diagnoses. We use an Electronic Medical Record (EMR) in which conditions, including both diagnoses and symptoms, are classified as International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). An ICD-9 code has three, four or five digits (xxx.xx). The first three digits represent the broader category of a disease, and the fourth and fifth digits represent sub-divisions of a disease. For example, the ICD-9 code for personality disorder is 301. At the four-digit level (301.x), there are ten types of personality orders and at the five-digit level (301.xx), two other specific personality disorders are coded. We aggregated ICD-9-CM codes to the three-digit level. Thus, variations of the same disease are considered as one node in the network. There are advantages and disadvantages of aggregation. The advantage is reduction in the measurement bias. A disadvantage is the compromise to granularity as different classes of the same disease can have dissimilar impacts.

An edge E_{ij} is created between two diseases d_i and d_j ($\{d_i, d_j\} \in D$ and $i, j = 1$ *to* n) where $i<j$ as the network is undirected. Since the focus is on relationships based on co-occurrences and not the causality, we created edges between diseases with no direction. For example, an edge between

congestive heart failure and rheumatic heart disease represents an undirected connection between two nodes representing two diseases.

In the past, associations between diseases or comorbidities were modeled using a simple Pearson's correlation coefficient (Divo et al., 2015; Hidalgo et al., 2009). In the network using Pearson's correlation coefficient, the coefficient PCC_{ij} of an edge E_{ij} between diseases d_i and d_j is calculated as

$$PCC_{ij} = \frac{(c_{ij} * N)(i_c * j_c)}{\sqrt{(c_i * c_j)(N - c_i)(N - c_j)}} \quad \textbf{-(5.5)}$$

where c_{ij} is the count of patients containing both d_i and d_j diseases, c_i is the count of patients diagnosed with d_i and c_j is the count of patients diagnosed with d_j. However, the number of significant correlations is directly proportional to the number of observations used (N). The ability to detect rare comorbidities is lessened because of the rareness of events. Therefore, to establish the right measure to model a comorbidity, we use the Salton Cosine Index (SCI) (Salton & McGill, 1986). SCI is unaffected by the total number of observations used (Ahlgren, Jarneving, & Rousseau, 2003) and measures the prevalence of a relationship between two diseases considering their individual prevalence. Salton Cosine Index, w_{ij}, of two diseases d_i and d_j is calculated as an equation 5.6. The cosine similarity has been used in the past to find phenotype overlaps (Chen et al. 2015; Lage et al. 2007). We propose this as an appropriate measure for calculating the strength of a comorbidity.

$$w_{ij} = \frac{(c_{ij})}{\sqrt{(c_i * c_j)}} \quad \textbf{-(5.6)}$$

Unlike correlation coefficient, SCI measure is free from N. However, it is difficult to find statistical significance of SCI. To find a rigorous SCI cutoff, we present an approach that results in statistically significant edges between diseases. We use the relationship between correlation

and cosine index to find a cutoff for SCI as suggested by Egghe and Leydesdorff (2009), following the steps listed below (referred to as *Process 1* for reference during analysis).

Step 1. Calculate number of co-occurrences, correlations, and Salton Cosine Index for all pairs of diseases (p) in the pseudo-population dataset containing 24.7 million patients.

Step 2. Calculate T-statistic using PCC_{ij} of the edges as in equation 5.7. Following the most conservative approach, use the c_{ij} (minimum of c_{ij}, c_i, and c_j) as the degrees of the freedom. Using the T-statistic, develop a network at $\alpha=0.01$, $T>2.58$ and $c_{ij}>\sum c_{ij}/p$ to select correlations (pairs) occurring more than by chance. Find number of pairs (q) at $\alpha=0.01$, $T>2.58$ and $c_{ij}>\sum c_{ij}/p$.

$$T=\frac{PCC_{ij}\sqrt{c_{ij}-2}}{\sqrt{1-PCC_{ij}{}^2}} \quad \textbf{-(5.7)}$$

Step 3. Find Salton Cosine Index as the cutoff (w_c) where number of pairs is equal to q and $c_{ij}>\sum c_{ij}/p$.

Step 4. Use w_c as the cutoff to find statistically significant comorbidities and create networks.

These steps are used to create networks. The network is represented by a matrix, D_{nXn} containing strength of connections between diseases, where the Salton Cosine Index for a pair (w_{ij}) of diseases indicates the strength between them.

$$D_{nXn}=\begin{bmatrix} 0 & w_{12} & \dots & w_{1n} \\ w_{21} & 0 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & \dots & & 0 \end{bmatrix}$$

5.3.2.2 NETWORK METRICS

The structural properties of a network can be measured using several metrics, including node centralities which define the importance of a node in the network. The most common centralities are degree, eigenvector, closeness, and betweenness as described in Table 5.3. In a comorbidity network, the degree centrality of a disease (node) denotes the number of direct connections with

other diseases (Freeman, 1979). If weights of the direct connections are considered, it is called weighted degree centrality, and calculated as a weighted sum of the strengths of direct relationships of a disease with others. The degree is a local property of a node, as only direct connections are considered. The direct connections of a disease, d_i, considered by the degree centrality, explain the other diseases likely to be diagnosed in the presence of d_i. Therefore, a higher degree of a disease increases the likelihood of it getting diagnosed with other diseases. A higher number of diagnoses predicts a longer stay because more resources are required for care.

The eigenvector centrality is an extension of degree centrality that incorporates the indirect connections of a node. It is a metric for the influence and power of a node in the network and measures how well its connected nodes are further linked (Bonacich, 1987). In a social network, a person with high eigenvector centrality is connected to the important people in that network. An eigenvector centrality of a disease, d_i, in a comorbidity network measures how well its neighbors (directly connected diseases) are connected further. A disease with higher eigenvector centrality connects to other diseases that are central in the network. Although conceptually different, the eigenvector and degree centrality are highly correlated measures (Valente et al. 2008). Therefore, we expect eigenvector to be related to LOS in the positive direction as degree centrality. It must be noted that we will use only eigenvector centrality for creating models because degree and eigenvector centrality are highly correlated measures.

The closeness centrality of a disease determines the average number of steps it is away from other diseases in the network (Freeman, 1979). Unlike degree centrality, closeness centrality is a global property of a disease and explains its centralization in the network. It is expressed in terms of distances (one connection is one distance) among the different nodes. The closeness centrality (c_i) of a disease, d_i, is calculated as the average shortest distance of d_i, to all other diseases in the network, if a path exists at all. A disease with higher closeness (indicated by a small closeness number) is relatively fewer steps away from other diseases in the network. Therefore, a patient is

likely to be diagnosed with closer diseases. This measure has been shown to be related to mortality by Hidalgo et al. (2009). However, we expect that the closer a disease is to other diseases, the longer the average LOS for patients with that disease.

Finally, the relationship betweenness centrality and LOS is studied. It is a global property of a disease, which describes its bridgeness in the network (Freeman, 1979). In other words, a disease with higher betweenness tends to form more bridges between other diseases. It is measured as the

Table 5.3. Network measures and their interpretation in our context

Network Measure	Definition	Interpretation in our context
Nodes	Nodes or vertexes are the elements among which relationships are studied.	Diseases are the nodes. Each disease or a node is a three-digit ICD-9 code.
Edges	An edge represents the relationship between nodes.	Relationships are comorbidities.
Degree centrality	Degree of a node explains its number of direct connections (Freeman, 1979)	Degree of a disease is the number of diseases directly connected to it.
Weighted Degree	Degree calculated as a weighted sum of the strength of the connections.	Degree calculated as a weighted sum of the strength of the comorbidities.
Closeness	Closeness of a node gives the average shortest distance of that node to all other nodes in the network (Freeman, 1979). Closeness of a node d_i is $c_i = \frac{\sum_{n-1} d(i,j)}{n-1}$, where $d(i, j)$ is the shortest distance between d_i and d_j.	Closeness centrality of a disease would represent how close a disease is to all the other diseases in the network.
Betweenness	Number of times a node is on a shortest path among all shortest paths (Freeman, 1979). Betweenness of a node d_i is $b_i = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$, where σ_{st} is total number of shortest paths from node s to node t and $\sigma_{st}(i)$ is the number of those paths that pass through d_i.	Number of times a disease is a bridge between pairs of diseases.
Eigenvector Centrality	A metric for influence of a node in the network measuring how well the direct connections of a node are further connected (Bonacich, 1987).	Measures how well a disease's connections are connected further.

number of times a disease is on the shortest path between other pairs of diseases. In a social network, the person with high betweenness is a point of connection between multiple communities and therefore has power to divide the network (Newman 2005). Similarly, a diagnosis with the higher betweenness in the comorbidity network acts as a bridge between other diagnoses. There are several symptoms that are connected to different categories of diseases; therefore, they are likely to act as bridges between other diseases and are consequently expected to have high betweenness. The patients with general symptoms are expected to have shorter stay as compared to the patients with actual diagnoses. Therefore, we expect betweenness of a disease to be negatively related to LOS.

5.3.2.3 EXPLANATORY AND PREDICTIVE MODELING USING COMORBIDITY NETWORK

Our aim is to add comorbidities to the baseline models using the network structure as defined in the previous section. As discussed in the Introduction, comorbidities vary by gender, therefore, we create different networks for males and females. A network of all diseases in males is created from a set of m number of diseases $D_m=\{d_1, d_2, \ldots d_m\}$, and a network from females is created containing f diseases $D_f=\{d_1, d_2, \ldots d_f\}$ where $m \neq f$ because both genders have unique diseases due to their biological differences. The network measures of each disease, such as betweenness (b_i), closeness (c_i), and eigenvector centrality[9] (v_i), are calculated separately for different genders.

We use the network measures of the diseases diagnosed in patients at the point of admission to know how the relationship between diseases and position of a disease in the web of diseases can explain and predict LOS. First, we create a model to see the relationship of network centralities and LOS, and then we use actual connections with other diseases for predictions. Let

[9] Degree and eigenvector centrality are highly correlated. In our dataset, correlation coefficient is 0.96.

KD_{1XKD}=$\{d_1, d_2, d_3, \ldots d_{KD}\}$ be the diagnoses known during the point of admission. For a manageable analysis, we restricted *KD<=4,* which includes more than 90% of visits in the data. A model for explaining and predicting LOS at the hospital visit level was created using the aggregated sum of their network metrics as presented in equation 5.8. Although the model is linear, the inputs are non-linear. This model describes how the structural positions of diagnoses known at the time of admission are related to LOS.

Comorbidity Model 1: $LOS=\beta_0 + \beta_1 \sum_{i=1}^{KD} b_i + \beta_2 \sum_{i=1}^{KD} c_i + \beta_3 \sum_{i=1}^{KD} v_i + e$ **-(5.8)**

The above model provides information about the network properties of diseases based on their structural positions but does not explain the actual association of the observed diagnoses with other likely diagnoses (which we call predicted comorbidities). For example, if a patient visits a hospital with d_1 and d_2 as known diagnoses, our aim is to search the comorbidity network and extract direct connections of d_1 and d_2 to use for modeling. The direct connections of a diagnosis can help predict other likely diagnoses and LOS. To add relationships to the model, we use the comorbidity matrix described earlier, D_{nXn}, which represents the strength of relationships between diseases. Since we create separate networks for each gender, we use two separate comorbidity matrices. The use of connections is an alternative to adding all observed diseases in the model. The addition of predicted comorbidities is more theoretically robust, and expected to have more explanatory and predictive power.

An algorithm created to extract the connections of the diseases in the model is presented in Table 5.4. Since a diagnosis has multiple associations, we restrict the number of comorbidities being considered. In our study, the top five weighted relationships of a disease are added in the model as explained in Code 5.1. As discussed, we restrict our analysis to patients with a maximum of four observed diseases, meaning five to twenty new values in terms of relationship strengths are added in the model as presented in Code 5.3 in Table 5.4. If two diseases have an association with

Table 5.4. An algorithm to add predicted comorbidities of the known diseases at the point of admission

Input: Array D_{nXn} denotes the matrix generated from the comorbidity network containing strength between diseases
Input: Array K_{1Xn} denotes an array containing all the possible diseases
Input: Array KD_{1Xm} contains the list of diseases known during admission
Output: Array COM_{1Xn} contains the final diseases and strengths added to the model
Let:
Scalar *MaxN* – denotes how many connections or comorbidities of a known disease are used in the model
Array RD_{1Xn} – An array used to create a ranking order of comorbidities for each disease

Initialize:
$D_{ij} = 0, \quad K_i = 0, \quad COM_i = 0, \quad KD_i = 0, \quad \forall i, j \in (1 \text{ to } n)$

Code 5.1: To rank the strength of comorbidities of each disease and consider only top *MaxN*. In our paper, *MaxN*=5.

```
For i=1 to n
  RD1Xn = Rank (Di, Descending)
    For j=MaxN+1 to n
      Dij = 0
    Loop
Loop
```

Code 5.2: Code for creating an array to label the known diseases as present

```
For i= 1 to m
  For j=1 to n
If KDi = Dj
      Kj = 1
    End
  Loop
Loop
```

Code 5.3: Code for creating a result arraying containing the strengths of comorbidities of known diseases

```
For i=1 to n
  For j=1 to n
    COMi = COMi + Dij * Kj
  Loop
Loop
```

the same disease, the strengths are added. The measure *COM* calculated in Code 5.3 for each hospital visit is added as a metric containing the predicted comorbidities in the models.

The model with the predicted comorbidities is presented in equation 5.9 where five likely comorbidities of the primary disease and five each for other known diseases are added from the matrix. A set of parameters are estimated where Ω (a set of n parameters) are the coefficients and COM_{1Xn} is an array of variables. This model contains the variables coming from the network built using secondary data. Finally, to see how *COM* adds value to the estimation and prediction of LOS, we control for the baseline model 4 in which all other factors are considered. The model was created by adding our variables with the baseline model as presented in equation 5.10.

Comorbidity Model 2: $LOS = \beta_0 + \Omega . COM_{1Xn} + e$ **-(5.9)**

Comorbidity Model 3: $LOS = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 race + \beta_4 patientType + \beta_5 admissionType + \beta_6 payer + \beta_7 hospitalSize + \beta_8 revisit + \beta_9 VisitNumber + \beta_{10} classes + \gamma . \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix} + \alpha . chi + \Omega . COM_{1Xn} + e$ **-(5.10)**

5.4 ANALYSIS AND RESULTS

We obtained data from the Center for Health Systems Innovation (CHSI), a center at Oklahoma State University which houses data provided by Cerner Corporation, a major Electronic Medical Record (EMR) provider. The data warehouse contains records of visits of 58 million unique patients across 662 US hospitals (2000-2016). We used information about the demographics of the patients, hospitals, types of visits, and diagnoses. The diagnoses are recorded according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). We removed hospital visits in which patients were not diagnosed with any disease or symptom, leaving approximately 24.7 million unique patients with sufficient information to perform our analysis.

The raw health records must be organized in a transactional form to create our comorbidity network and models. We follow the Transparent Reporting of a multivariable prediction model

for Individual Prognosis or Diagnosis (TRIPOD) guidelines created by Collins et al. (2015). The steps followed in processing the massive data containing medical records of 24.7 million unique patients are presented in a flowchart in Figure 5.1. The flowchart also presents the number of patients and hospital visits at each step in the process. First, information about the patients, admissions, diseases diagnosed, and hospital was merged. The patient level information included gender, age, and race. The information about the hospital visit included type of insurance held by the patient, type of visit (inpatient, outpatient, emergency, etc.), type of admission (elective, newborn, urgent, etc.), and hospital size.

At the disease level, the information available includes all diagnoses made during the visit, but not the exact date and time of disease development. Because it is difficult to know when a disease begins to develop in a patient and therefore, this is a fair limitation of the data. Nevertheless, priorities of diagnoses during a visit are still known, with the top priority diagnosis being the primary disease and the main reason for visiting the hospital. Secondary diagnoses are less important and annotated with lower priorities. The EMR also notes whether a disease is present at the point of admission. In our study, we used the information about all diagnoses to create the networks but used only primary diagnoses and diseases present during admission for creating explanatory and predictive models.

An integration of different datasets gave us medical records of 24.7 million patients diagnosed with at least one disease or symptom which were used to find a cut-off for Salton Cosine Index using *Process 1*. At SCI cutoff of 0.04, we found the number of edges in the comorbidity network developed using SCI is equal to the number of edges in the network using Pearson's Correlations Coefficient significant at $\alpha=0.01$ and c_{ij}>Average (c_{ij}). From this point, we consistently use the SCI cutoff of 0.04 for creating an edge in different comorbidity networks.

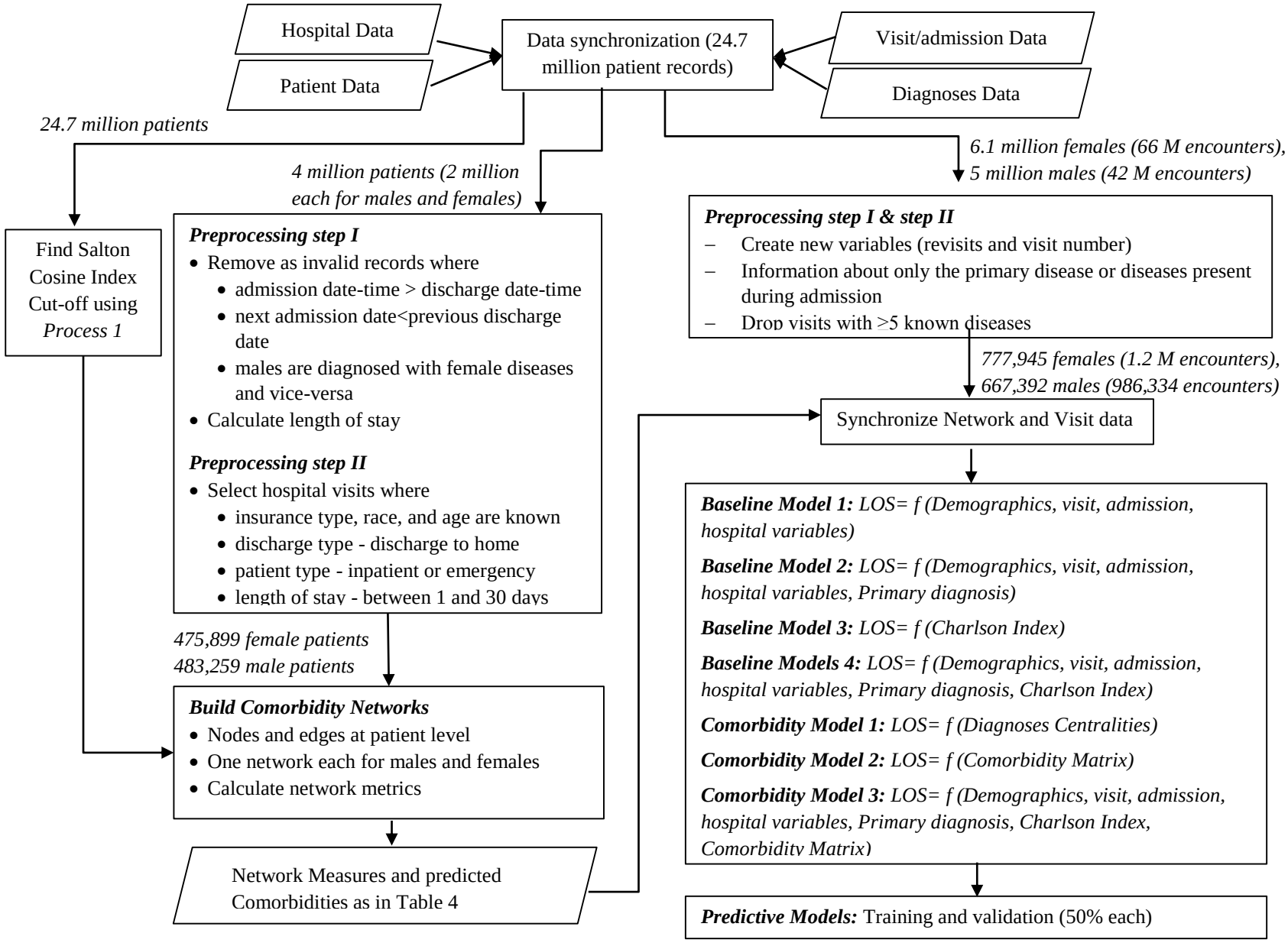


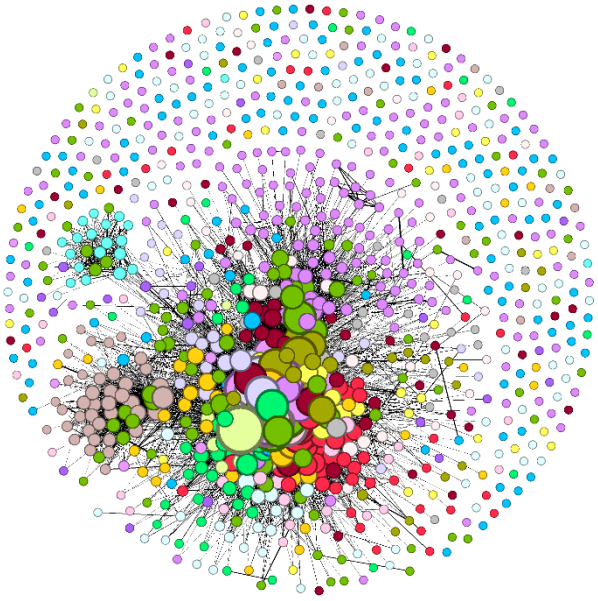
Figure 5.1. Data Processing and Modeling

Multiple random samples were extracted for creating networks and models. First, medical records of four million patients (sample 1) were extracted to build comorbidity networks and compute network properties. For modeling, an independent random sample of approximately 11 million patients (sample 2) was extracted. Because there were several data quality and integrity issues in the EMR, data cleaning was done on both samples, removing records with erroneous and suspicious coding. For instance, there were hospital visits that recorded the admission date and time later than the discharge date and time, and visits where the admission date and time were earlier than the previous discharge date and time. We found a few patients who were coded as male during one visit and female in another, and some males who were diagnosed with female diseases such as inflammatory disorders of female pelvic organs (ICD9–614-616), complications of pregnancy, childbirth, and the puerperium (ICD9–630-679) and diseases of breast (ICD9–610-612)[10]. Similarly, we noticed female patients diagnosed with diseases of male genital organs (ICD9–600-608). Considering these as suspicious entries, we removed them from further analysis. The challenges of using secondary data for research are well documented (Bellazzi and Zupan, 2008; Shmueli and Koppius, 2011), and we experienced the same. However, the size of our dataset helps mitigate these issues.

We restricted our analysis to inpatient and emergency visits that were discharged to home. Moreover, we considered hospital visits with LOS of at least one day and up to 30 days in order to remove outliers. Long-term hospital stays represent a small percentage of patients (Marazzi, et al. 1998) and predictive models for such patients may include other signatures. For instance, Spratt, et al. (2003) found the level of disability after stroke could predict the prolonged stays.

[10] Although males can also have breast disease biologically, we considering these records suspicious and removed them

After data cleaning, the sample (sample 1) was used to create networks. Recognizing the difference in diagnoses and comorbidities as argued by many health researchers such as Johnson, et al. (2014) and Ovseiko, et al. (2016), we created separate networks for females and males. The network visualizations of these networks are presented in Figures 5.2a and 5.2b respectively. The diseases are color coded according to the 19 categories/classes/organ systems as per the ICD-9



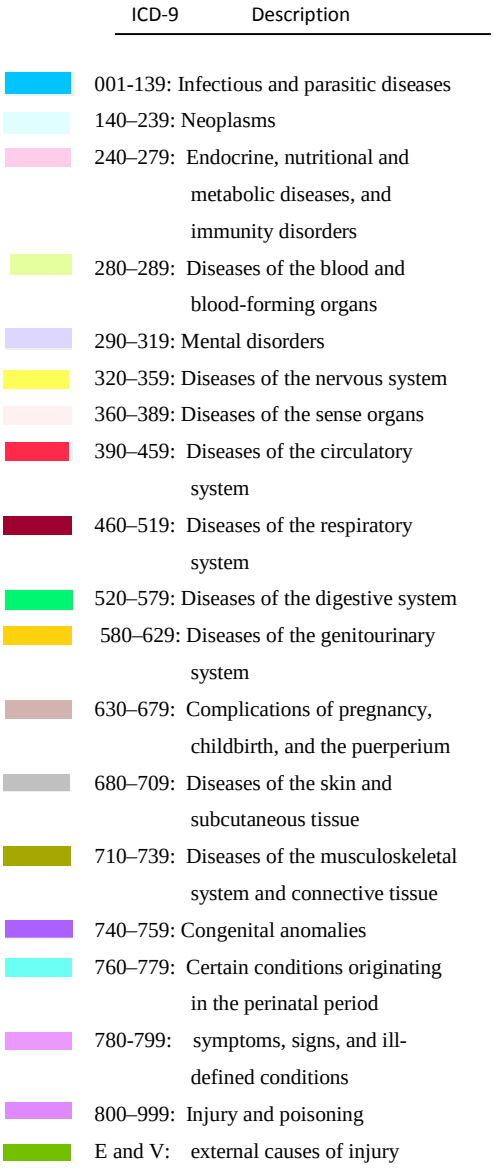


Figure 5.1b. Female Comorbidity Network

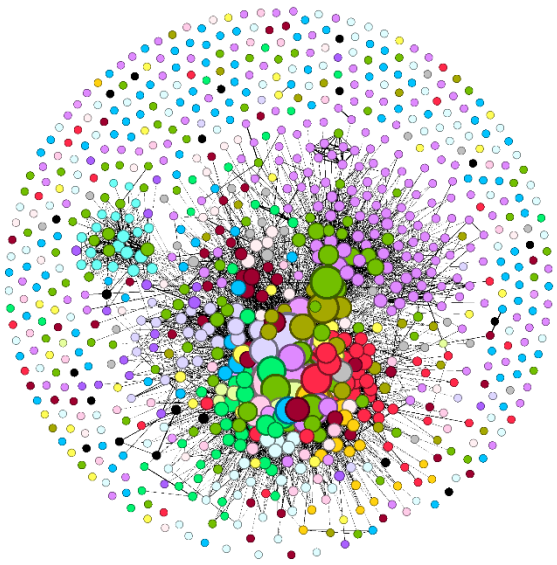


Figure 5.2b. Male Comorbidity Network

classification systems. Size of a disease node represents the number of direct connections to other diseases (degree centrality). The summary statistics of the network properties are listed in Table 5.5. The female comorbidity network was comprised of 1,013 diagnoses/nodes with 12,046 edges (comorbidities). The average degree of 23.8 indicates a disease is connected to approximately 24 other diseases in the network. The average closeness, betweenness, and eigenvector centralities are 0.25, 294 and 0.116 respectively. The male network had a slightly lower number of diseases and comorbidities (i.e. 956 and 11,065 respectively). However, the other metrics were not significantly different.

Table 5.5. Network properties

Name	Female	Male
Nodes	1,013	956
Edges	12,046	11,065
Average Degree	23.8	23.15
Average Weighted Degree	1.93	1.89
Average Betweenness	294.26	280.07
Average Closeness	0.25	0.25
Average Eigenvector	0.116	0.118

5.4.1 EXPLANATORY MODELING RESULTS AT HOSPITAL VISIT-LEVEL

At the visit level, we created multiple models to study the value added by the comorbidities to explain and predict LOS. Four baseline models from the traditional variables were created. The variable descriptions and descriptive results are listed in Table 5.6. The final dataset contained 2.2 million hospital visits of 1.45 million patients (55% female). The mean LOS was 1.68 days with 71% emergency visits and the remainder inpatients. Among all the visits, 22% were covered by Medicaid and 14.5% were self-paid. 6.1% of patients returned for one or more visits for the same primary reason.

The performance of all baseline models and comorbidity network models in terms of variance explained is listed in Table 5.7. The baseline model consisting of demographics, hospital, and admission type information explained 25% variance in LOS. The addition of primary disease of

Table 5.6. Variable description		
Variable Name	**Description**	**Descriptive Statistics**
Length of Stay(LOS)	Length of a hospital visit	Average=1.68 days
Age	Age of a patient recorded during hospital visit	Average: 30.4 years
Gender	Patient's gender	Visits- 55% Females Patients-53.8% Females
Charlson	Charlson index calculated as weighted sum of score of known diseases	Average: 0.09
Visit Number	A patient's number of visits as (inpatient or emergency)	Average of maximum: 1.62
Classes	Total number of categories of diseases diagnosed during admission	Average: 1.41
Race	Race of the patient. Six different binary variables were created: African-American, Caucasian, Hispanics, Asians, Native Americans and Pacific Islanders.	Afro-American: 24% Caucasian: 60% Hispanics: 5% Asians: 1.8% Native Americans: 1.6% Pacific Islanders: 0.19%
Hospital Size	Size of the hospital. Five different binary variables were created for hospital sizes: <5, 100-199, 200-299, 300-499 and 500+.	<5: 4.4% 6-99: 9.9% 100-199: 15.1% 200-299:27.9% 300-499: 25% 500+: 17.7%
Patient Type	Type of patient. A single binary variable is created for inpatient and emergency.	Inpatient: 28.2% Emergency: 71.8%
Admission Type	Type of admission. Five binary variables were created: emergency, urgent, elective, newborn and trauma center.	Emergency: 64.7% Urgent: 6.6% Elective: 7.9% New Born: 5.4% Trauma center: 0.24%
Payer	Insurance type during the visit. Two binary variables were created: Medicaid and self-pay visits.	Medicaid: 22% Self-pay: 14.5%
Revisit	Number of visits due to same primary diagnosis.	Second visit: 6.1% (88,615 patients came back at least once)

each visit increases the variance explained to 34%. This is the highest baseline model; as other models did not show any significant results. The third baseline model describing the relationship of the Charlson Index and LOS showed poor performance with R^2 of 1.8%. Because the Charlson Index considers only 19 categories of diagnoses, it does not explain the visits of all types of patients, and just 10% of the visits had a Charlson score of more than zero. A similar low effect

has been observed in the past by Rochon et al. (1996), in which the authors found that the Charlson Index could explain only 1.9% variance in LOS. As expected, the addition of the Charlson Index in the first two baseline models did not improve the performance significantly. However, the model built using network properties such as centralities (Comorbidity Model 1) did a better job (R^2=0.07), indicating that the network position of the observed diagnoses at the time of admission can explain the length of stay better than the Charlson Index. Moreover, the model using only the comorbidity matrix (R^2=0.25), Comorbidity Model 2, performed better than the Charlson Index and network metrics models. The best performing model used the comorbidity matrix to control for demographics, hospital information, and primary disease (R2=0.38), with a statistically significant improvement of about 4%, calculated using partial F-Test (F-value=163.3, p<0.0001). For a complex and ill-structured problem like LOS, this improvement is highly desirable.

Table 5.7. Linear models for length of stay at hospital visit level	
Model	**Model R-square**
Baseline 1: Demographics + Hospital variables + Visit variables	.25
Baseline 2: Demographics + diagnoses variables	.34
Baseline 3: Only Charlson Index (CHI)	.02
Baseline 4: Demographics + disease variables + CHI	.34
Comorbidity Model 1: Only Network Centralities	.07
Comorbidity Model 2: Only Comorbidity Matrix	.25
Comorbidity Model 3: Baseline 4 + Comorbidity Matrix	**.38**

Due to the availability of medical records from millions of patients, we were able to study all diseases together and created a common model for all types of patients. However, there are several diseases for which it is difficult to predict LOS. To identify them, we sliced the best model results (Comorbidity Model 3) into specific diagnoses to determine where comorbidities did a good job. Due to the size of our database, we were able to run different models for patients visiting hospitals due to different sets of diagnoses (d_1, d_2, … d_n). There were 319 types of patients based on the primary diagnoses. The hospital visits for which model performance was

low include those with a single live birth (ICD9-V30, R^2 = 0.07, n=103,027, mean 2.57 days), trauma to perineum and vulva during delivery (ICD9-664, R^2 = 0.07, n=10,292, mean 2.18), acute myocardial infarction (ICD9-410, R^2 = .09, n=7,737, mean 3.43), chronic ischemic heart disease (ICD9-414, R^2 = .096, n=6,974, mean 2.96) and others presented in Appendix A. The improvement due to the comorbidity matrix in each cluster of patients based on the primary diagnosis is also presented in Appendix A (*Comorbidity Model 3 - Baseline 4*).

To observe the contribution of comorbidity in different categories of patients, we aggregated the results of diagnoses models to nineteen clusters based on organ system or classes (see Figure 5.2), where $(d_1, \ldots d_n) \epsilon (O_1, \ldots O_{19})$. This aggregated analysis shows which organ system diagnoses are affected most by the comorbidity. The average improvement in the diseases of different organ systems is presented in Figure 5.3 (*Comorbidity Model 3 - Baseline 4*). The clusters of diseases for which comorbidities contribute significantly in explaining LOS are endocrine, nutritional, and metabolic diseases, and immunity disorders (approximately 25%). For example, the endocrine disorder diabetes has comorbidities such as cardiovascular disease, hypertension, and obesity that can affect health performance, primarily in older adults (Struijs et al. 2006; Kalyani et al. 2010). The other clusters affected by comorbidities include disorders of the blood and blood-forming organs (23%) and the nervous system (21%). Nervous system diseases such as Parkinson's disease, multiple sclerosis, epilepsy, and migraine are highly comorbid with other neuropsychiatric disorders, pain, and asthma (Ottman et al. 2011) and therefore can affect LOS.

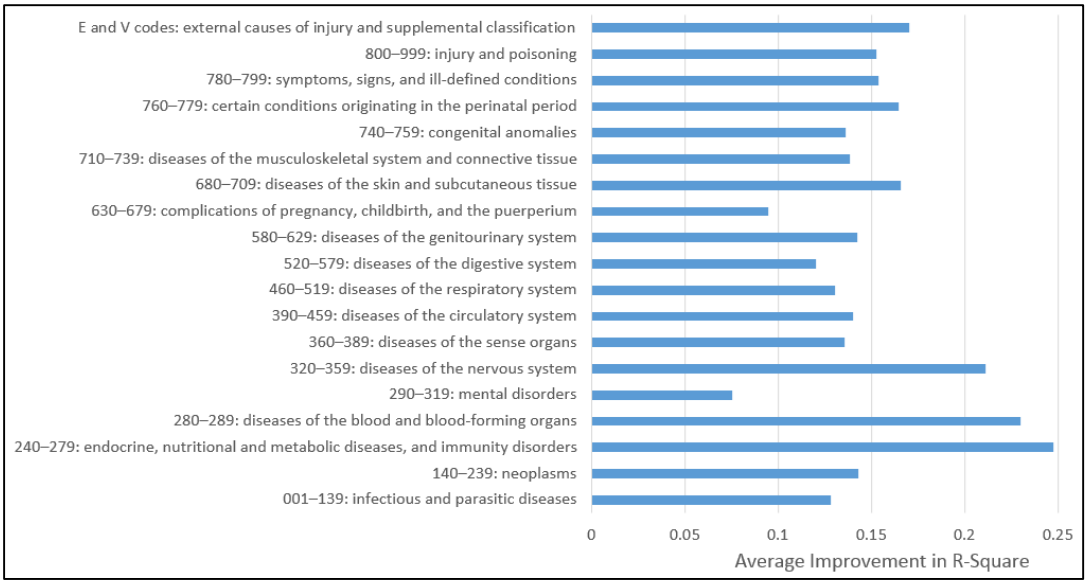


Figure 5.3. Average improvement in variance explained in LOS in clusters of patients based on type of their primary diagnosis

5.4.2 PREDICTIVE MODELING RESULTS AT HOSPITAL VISIT-LEVEL

The above models set the stage for predictive models. The dataset was randomly partitioned into training and validation sets, with half used to create a predictive model and half used for validation. This approach is common and has been followed in other studies such as Bardhan, et al. (2014). We ran general linear models, regression trees, and artificial neural network models. All of the models improved similarly when the comorbidity matrix was added to the baseline model. Regarding the general linear model, we observed an improvement in its predictive power due to the comorbidities in the overall model calculated in terms of average squared error and mean absolute percent error in the validation dataset. Overall, the average squared error improved from 2.34 to 2.26 and mean absolute percent error decreased from 37.5 to 36.6.

Although the overall improvement in the model seems low, we further analyze the results to see where comorbidity added predictive power. The absolute residuals and percent errors are separated for different primary diagnoses (d_1, d_2, … d_n). The absolute residuals and percent errors of *Baseline 4* model are subtracted from *Comorbidity Model 3*. Then, to understand the specific

cluster of diseases based on organ systems (O_1, O_2, … O_{19}), the improvement in residual and MAPE are presented in Figure 5.4 (*Comorbidity Model 3 - Baseline 4*). The predictive power due to the comorbidity improves greatly for patients with conditions originating in a perinatal period followed by patients with mental disorders. With the exception of patients admitted due to neoplasm, congenital anomalies, and diseases of skin and subcutaneous tissue, predictions for all other clusters of patients improved in terms of mean absolute percent error and residuals in predictions.

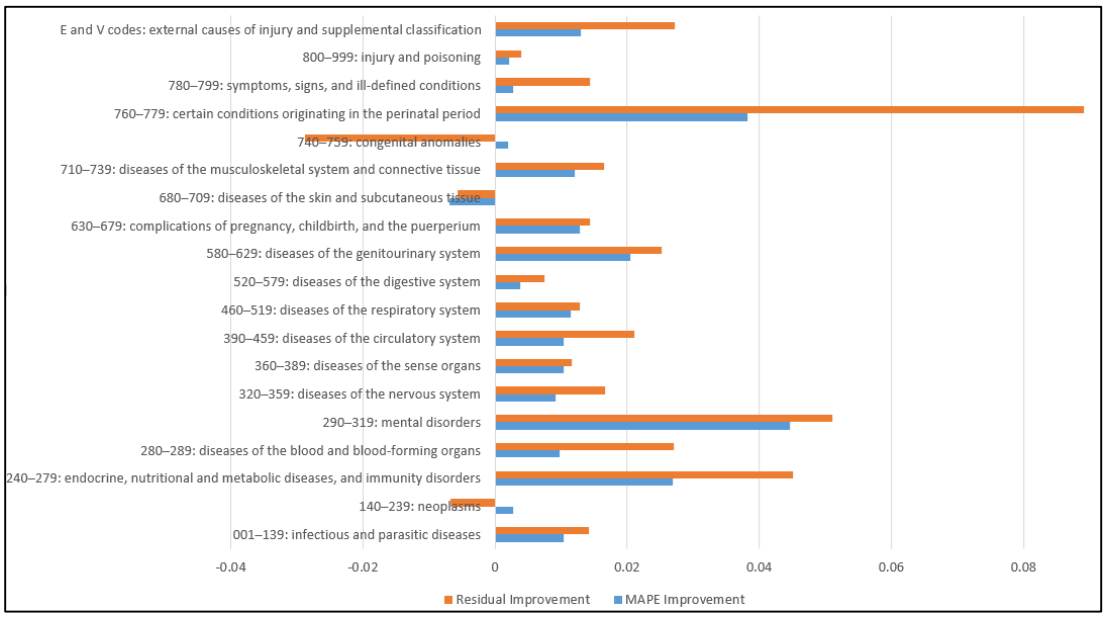


Figure 5.4. Improvement in predictive power in different clusters of patients based on primary disease category due to comorbidity matrix

5.5 DISCUSSION AND CONCLUDING REMARKS

Taken as a whole, our paper suggests that use of a massive dataset and analytics techniques such as network analysis, which help reduce bias (Shmueli, 2010), provide an opportunity to learn more deeply about the complex relationships between diseases. Our focus on both explanation and prediction results in a strong and applicable theoretical model as explained by Hong et al.,

(2013). In addition, building and assessing the proposed approach on mutually exclusive samples confirm its robustness.

Network properties were used to explain other diseases that can occur with a primary diagnosis, and help predict LOS. One of the strengths of our network method, and in particular the model using our comorbidity measure, is that it includes all possible diseases and can therefore capture all possible associations. It performed better than the existing Charlson Index, which is available for only 10% of hospital visits because it considers just 19 categories of diseases. The Charlson Index can only be calculated from observed comorbidities; however, our network method was able to forecast diseases that may occur in presence of another disease. Because we used information available at the point of admission, our approach has both explanatory and predictive ability.

Comparing our unified model to past studies, our model performed better overall, explaining about 38% variance in the data using the information present at the time of admission. Past studies using other comorbidity scales to explain LOS were able to account for less variance. For example, a model created by Rochon et al. (1996) for spinal-cord injury patients explained 6% variance, Chertow et al.'s (2005) model for acute kidney injury patients explained 33% variance, Liu et al.'s (2010) model from one healthcare unit explained 14% variance, and Huntley et al.'s (2014) model for patients in one psychiatric center explained only 17% variance.

With respect to predictive power, our overall model did better than the previous studies. The mean absolute percent error of our model was about 36%, meaning an overall accuracy of 64%. Consideration of comorbidities improved the accuracy of LOS prediction by about 1%. Although this improvement may appear small, its practical impact is significant. In the validation dataset, the comorbidity based prediction accuracy increased by about 14,500 days. The practical significance of Big Data Analytics can be derived in terms of dollars involved as suggested by Lin, Lucas Jr. and Shmueli (2013). Pfuntner, Wier, and Steiner (2013) estimated that the

aggregate cost for all hospital stays was $387.3 billion in 2011 with a mean of $10,000 per stay. Considering an average stay of 1.68 days in our data, 14,500 days are equivalent to about 8,630 visits, resulting in a better LOS forecast of $86 million. Therefore, the practical improvement due to the comorbidities is significant. Because the problem of LOS prediction is particularly notorious and ill-structured, any reasonable improvement in the accuracy is highly desirable. The use of our model to predict LOS can help hospitals better plan and allocate resources. In addition, the study of comorbidities can certainly help in clinical care and management (Valderas et al., 2009). Knowledge of the relationships among diseases can help predict and diagnose likely co-existing diseases, and disease network properties can help physicians prioritize diagnoses based on the position of a disease in the network. The relationship among diseases can help pharmaceutical companies consider multiple related diseases when developing new produce drugs.

This study has a few limitations. First, comorbidities were measured based on an EMR and therefore, only the diagnoses recorded in hospitals were included. It is perhaps impossible to capture someone's lifetime medical history in one record. Hence, this limitation exists in all studies that are based on medical records. Second, we considered gender differences while developing comorbidity networks; however, we also recognize that health disparities also exist based on race and ethnicity (Fine, Ibrahim & Thomas, 2005). Including such disparities is the next step of our research. Third, when inputting predicted comorbidities in the model, we restricted our analysis to the top five comorbidities of each known disease. Adding more diseases could further explain their relationships, and we want to explore this. Fourth, our model showed improvement with respect to the competing models, but there is still much room for improvement. We did not use patients' laboratory reports or consider other external factors such as hospital conditions and quality of physicians, which could make a significant difference. We

employed our proposed approach to predict LOS; in the future, we intend to use it to predict other medical outcomes such as readmission and mortality.

Notwithstanding these limitations, our study adds to the growing literature of analytics, comorbidity, and network science. The nascent measure and model we propose has tremendous potential to enhance existing information systems and improve decision making in healthcare and other related domains. Our approach is generalizable to similar problems in which unintended actions of individuals form a network pattern and impact their outcome.

CHAPTER VI

DIAGNOSES FORM TRAPS: IDENTIFYING MORTALITY RELATED CLIQUES IN COMORBIDITY NETWORK

ABSTRACT

Mortality rate is one of the important metrics of quality of care. Different biological, sociological and political factors might impact mortality risk in individuals. However, for the patients in hospitals, the primary reasons for a mortality are the diseases. Mostly, patient develop multiple diseases simultaneously but often only one diagnosis is considered as the primary reason for the death e.g. cancer, heart failure, etc. Because, multiple diseases jointly have a different impact than each of them independently, we focus on identifying the clusters of diseases related to mortality. We apply a network approach to create relationship between diagnoses based on their co-occurrences in the patients and then use the clique property to identify high risk cluster of diagnoses. To create a network of diseases, more than 8 million patient records stored in an Electronic Medical Record (EMR) are used. We identified eighteen mortality related cliques in the network and found that the mortality rate in the patients diagnosed with all the diseases in the cliques is significantly higher than the patients without all clique diagnoses. The results are validated on an independent set of 8 million patient records. The presence of the clique diagnoses in the patients can help physicians take preemptive decisions. The generalizability of our approach is discussed.

6.1 INTRODUCTION

For the hospitals, one of the most important metrics to measure quality of care is the mortality rate. Therefore, the prediction and explanation of mortality have been a topic of interest for the medical, bioinformatics and analytics researchers. The primary reasons for mortality of the patients are the specific diseases. However, in majority of the cases, patients get diagnosed with multiple diseases simultaneously. Therefore, it is important to understand how multiple diseases interact with each other that are eventually responsible for the casualty. The knowledge about the joint impact of multiple diseases on mortality will help physicians take preemptive decisions regarding the health outcomes of patients. In this paper, we identify clusters of diagnoses in the patients related to mortality using an information based approach.

When an additional diagnosis is present in a patient in addition to the primary disease, the medical condition is called comorbidity (Feinstein, 1970). For example, the simultaneous presence of diabetes and hypertension in a patient is a comorbid condition. Comorbidity has been considered as an important factor for mortality in the past because two diseases jointly can have a different impact than each of them independently. This joint impact of diseases on health is known as syndemicity, a term coined by Singer (1996). Most studies have considered one disease and its comorbidities at a time. However, how direct and indirect interactions of diseases are related to mortality is often not focused.

To model interactions of all diagnoses at one place, we have adapted network approach. This method can help explain the collective behavior of diseases and their impact on health. We create a comorbidity network where diagnoses form connections if these co-occur in patients. The network is inferred from the co-occurrences of diagnoses in large number of patients.

How to identify which combinations of diseases are critical so as to provide guidance to physicians? To address this, we have used an important property of network called clique to

identify such combinations. In terms of a social network, a clique is a part of the network where all individuals know each other and form a complete structure. All members of the clique in a social network may share common characteristics and information. The clique property of a network has several implications on the performance of the members such as trust, norms and obligations as discussed by Coleman (1988). In addition, Adler and Kwon (2002) argued that the tightly connected structure favors sharing of knowledge during uncertainty. Moreover, at the organizational level, Provan and Sebastian (1998) found that cliques are positively related to the network effectiveness.

In the comorbidity network, the clique is a complete sub-network which explains that the patients often get diagnosed with a collection of diseases together. A clique is a subset of the network in which each pair of diagnoses is adjacent. It represents a tightly connected hidden structure within a large network of diseases. Identification of a clique related to an outcome (mortality in our case) is a multi-step problem in such a network since the structure formation is not visible externally.

Pemantle and Skyrms (2004) explained a clique as a trap state. With respect to a game, a reinforced state is called a trap when a player is restricted to play a specific set of actions. Moreover, the traps decrease performance level of the player as explained by Roca, et al. (2010). Different researchers have identified traps in specific problem domains. For instance, Bonacich and Liggett (2003) identified traps in the network of gift exchanges where the members exchange gift within their own clusters. Similarly, in a comorbidity network, we hypothesize certain cliques of diagnoses represent trapping states, which are related to mortality. These cliques are expected to form stable equilibria for mortality. If the fully connected diseases are diagnosed in a patient, it is expected to indicate a critical condition. These groups of diagnoses are the topological traps analogous to the traps in games from where an exit is difficult. To the best of our knowledge, we are the first to use clique property to enhance our understanding about the relationship between comorbidity and mortality.

We first identify the diagnoses with high mortality risk from the data and then considering these as base diagnoses, we identify cliques around them, if any. To explain our hypothesis, we present one exemplar clique containing three diagnoses (d_1, d_2, d_3) in Figure 6.1 where d_1 is the base disease with high mortality rate. The degree or number of direct connections of d_1 is two. Here, we assume that d_1 has degree as two, but d_2 and d_3 may or may not have degree as two. Let N_1, N_2 and N_3 represent the number of patients having a particular disease d_1, d_1, d_3 respectively irrespective of the joint presence. The number of patients with all three diseases is $N_{123} = N_1 \cap N_2 \cap N_3$, where $N_{123} \leq N_1, N_2, N_3$). Because the underlying interactions of diseases in the patients with all three diseases are more than others, we hypothesize that mortality risk in N_{123} patients to be significantly greater than the mortality risk in N_1- N_{123} patients i.e. *MortalityRisk* (N_{123}) $\geq$ *MortalityRisk*(N_1- N_{123}). In other words, a patients diagnosed with all three diseases in a clique has high mortality risk than a patient not having all three diseases (i.e. <3 diseases). Since, our focus is on the base diagnosis and a clique around it, we hypothesize only about d_1. We recognize that the combinations other than a clique can also be related to mortality, however, we focus only on cliques of the base diagnoses in this study.

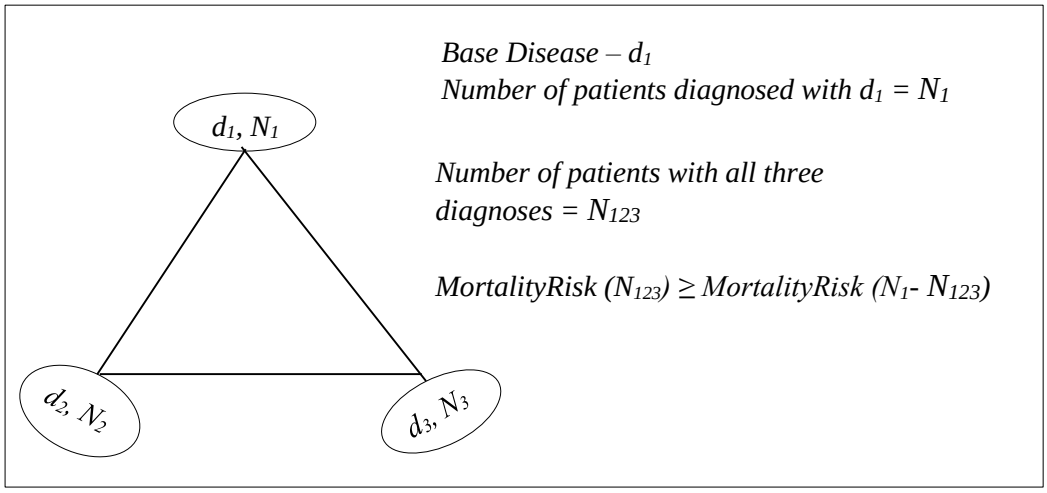


Figure 6.1. A clique/triangle of three diseases with their joint impact on mortality

We use an Electronic Medical Record (EMR) containing information of more than 24.7 million patients across 662 US hospitals in 17 years (2000-2016). The use of a massive dataset is one of the strengths of our study and overcomes the shortcomings of other studies where the sample size has been an issue for examining the interaction among diagnoses. Due to the lack of availability of sufficient data and advanced technologies, past research has largely focused on studying the impact of fewer diagnoses or fewer patients, which might make the conclusions derived to be less rigorous. Studying all diagnoses at one place can improve our understanding about the comorbidities.

6.2 BACKGROUND

Age is of course a primary reason for mortality with life expectancy of 78.8 years at birth (2015)[11]. According to the National Center for Health Statistics report, more than 75% deaths in United States take place after 75 years of age[12]. The leading causes of mortality in US are heart disease and cancer[13]. Although the primary cause of a death is a particular disease, patients suffering from such diseases also get diagnosed with multiple other comorbidities. For example, Ahluwalia et al. (2012) identified myocardial infarction, diabetes, chronic obstructive pulmonary disease (COPD), chronic kidney disease, dementia, depression, hip fracture, stroke, colorectal cancer and lung cancer to be significantly associated with increased hazard of dying in patients with heart failure. Similarly, Braunstein, et al. (2003) studied elderly patients with chronic heart failure and found non-cardiac comorbidities to be associated with mortality. The authors concluded that recognizing non-cardiac conditions in heart patients can improve health outcomes. Some authors have also studied how presence of unrelated diseases impact critical health outcomes. For example, Redelmeier, Tan and Booth (1998) argued that the presence of a critical

[11] https://www.cdc.gov/nchs/data/databriefs/db267.pdf
[12] https://www.cdc.gov/nchs/data/dvs/mortfinal2007_worktable23r.pdf
[13] https://www.cdc.gov/nchs/data/databriefs/db267.pdf

disease in a patient consumes most of the attention, as a result of which, the other unrelated existing diseases might get neglected. Obviously, neglecting a particular disease can worsen the condition of a patient and therefore, it is important to consider comorbidities while managing the primary disease.

The extant research focus on the impact of comorbidity on mortality in specific type of patients at a time. For example, Holguin et al. (2005) identified comorbidities related to mortality in COPD patients. The authors identified pneumonia, congestive heart failure, ischemic heart disease, thoracic malignancies, and respiratory failure to be associated with mortality. Similarly, Marrie, et al. (2015) found a significant effect of comorbidity on mortality in population with multiple sclerosis. Leontiadis et al. (2013) reviewed studies on the patients with peptic ulcer bleeding and found similar results. Zolbanin, Delen and Zadeh (2015) also showed that comorbidities improve the prediction performance of the models developed to forecast the survivability rate of the cancer patients.

In general, the interplay of diseases is studied by the medical researchers through gene interactions. Goh et al. (2008) and Bauer-Mehren et al. (2011) used common gene expressions to create connection between disorder. Lee et al. (2008) also followed the same approach to create a network of diseases and concluded that the connectedness of a disease with other diseases is related to higher risk of mortality.

Some studies have also used historical database of patients to create a network of diseases. For instance, Zhou et al. (2014) created a network of diseases based on the similar symptoms. Divo et al. (2015) created a comorbidity network from co-occurrence of diseases in COPD patients. Hidalgo et al. (2009) also applied a similar approach to create a comorbidity network and concluded that the patient developing diseases close to each other in the comorbidity network are

at high risk of dying sooner than others. However, to the best of our knowledge, no one in the past has identified cliques related to mortality in the comorbidity network.

6.3 METHOD

A network of diagnoses is inferred from the patient records. A connection between two diagnoses represents an aggregation of co-occurrences in the patients. An aggregation of a large number of patients reduces the complex interactions among diseases in a summarized network.

In our comorbidity network, a connection between two diagnoses is defined if these co-occur in a patient within a specified time interval. Analogous to the transactional database where a transaction contains a set of items, the records in an EMR are converted into transactions based on the defined time interval. For this, we define comorbidity as *the presence of multiple diseases in the lifetime history of a patient.* The measurements of comorbidity in the past do not explain how the *history* of a patient is related to the current situation. Therefore, we consider the lifetime history of a patient as a transaction containing unique diagnoses as the set of items. Of course, we recognize that the transaction may not contain all the diagnoses of the patient because he may have gone to a different hospital which does not use same data collection system. But this is the best available compilation. A transaction may contain multiple diagnoses. The presence of multiple distinct diagnoses in a patient are used to discern associations among diseases. Let $T(d_1, d_2, d_3, \ldots d_t)$ denotes a transaction, where $d_1, d_2, d_3, \ldots d_t$ is a subset of all diagnoses $D(d_1, d_2, d_3, \ldots d_n)$ with $n \geq t$.

A comorbidity network developed from N patients is denoted by $C\ (D, E)$ where D is a set of n nodes and E is a set of edges. In our comorbidity network, nodes represent diagnoses/symptoms. In the EMR, the diagnoses and symptoms are classified as International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). An ICD-9 code has three, four or five digits (xxx.xx). The first three digits represent the broader category of a disease, whereas the

fourth and fifth digits represent the sub-divisions of a disease. We aggregate these codes to their 3-digit level. For example, there are ten types of viral hepatitis (e.g. ICD-9: 070.0, 070.1, etc.) according to ICD9 classification, but these are aggregated to one 3-digit code i.e. one node. An advantage of aggregation is reduction in the measurement bias. In contrast, a disadvantage is the compromise to granularity.

An edge E_{ij} is created between two diseases d_i and d_j ($\{d_i, d_j\} \in D$ and $i, j = 1$ *to* n) where $i<j$ as the network is undirected. Since the focus is on relationships based on co-occurrences and not the causality, we create edges between diagnoses with no direction. To model the comorbidities statistically, we adapt a cosine index called Salton Cosine Index (SCI) (Salton & McGill, 1986). Salton Cosine Index, w_{ij}, of two diagnoses d_i and d_j is calculated as in equation 6.1, where c_{ij} is the number of co-occurrences of diagnoses d_i and d_j; c_i is the prevalence of diagnosis d_i; and c_j is the prevalence of disease d_j. The cosine similarity has been used in the past to find phenotype overlaps (Chen et al. 2015; Lage et al. 2007). We propose this as an appropriate measure for calculating the strength of a comorbidity.

$$w_{ij} = \frac{(c_{ij})}{\sqrt{(c_i * c_j)}} \qquad \textbf{-(6.1)}$$

Since there is no test to compute the statistical significance of the Salton Cosine Index (SCI), we use the relationship between Pearson's correlation coefficient and SCI to find a cut-off as suggested by Egghe and Leydesdorff (2009). We follow the following steps (we name these steps as *Process 1* for reference during analysis). First, we calculate the number of co-occurrences, correlations and Salton Cosine Index for each pair of diagnoses in the pseudo-population dataset containing 24.7 million patients. The correlation coefficient of every pair of diagnoses d_i and d_j, PCC_{ij}, is calculated as in equation 6.2.

$$PCC_{ij} = \frac{(c_{ij}*N)(c_i*c_j)}{\sqrt{(c_i*c_j)(N-c_i)(N-c_j)}} \quad \textbf{- (6.2)},$$

Next, the T-statistic using PCC_{ij} of every pair of diagnoses is calculated as in equation 6.3. Following the most conservative approach, we use the c_{ij} (minimum of c_{ij}, c_i, and c_j) as the degrees of the freedom. Using the T-statistic, a network at $\alpha=0.01$, $T>2.58$ and $c_{ij}>\sum c_{ij}/p$ (to select pairs occurring more than by chance) is created, where p is maximum number of pairs. In this network, the number of significantly correlated pairs (q) is recorded. Then, the Salton Cosine Index cutoff number is found where number of pairs is equal to q and $c_{ij}>\sum c_{ij}/p$. This cut-off is used to create networks for identifying cliques.

$$T=\frac{PCC_{ij}\sqrt{c_{ij}-2}}{\sqrt{1-PCC_{ij}{}^2}} \quad \textbf{- (6.3)}$$

The network results from the above process is represented by a matrix, D_{nXn} containing strength of connections between diagnoses, where the Salton Cosine Index for a pair (w_{ij}) of diseases indicates the strength between them.

$$D_{nXn} = \begin{bmatrix} 0 & w_{12} & \dots & w_{1n} \\ w_{21} & 0 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & \dots & & 0 \end{bmatrix}$$

6.3.1 DETECTING CLIQUES

A clique is a subset of the network where nodes form a fully connected sub-network. Given the network $C\ (D, E)$ where D represents the nodes, the clique Q ($\subseteq C$) is a sub-network in which all nodes are adjacent to each other. Moreover, a clique is maximal when it is not fully contained in another clique. The problem of clique graph recognition is NP-complete as argued by Alcón, et al. (2009). There are different algorithms created by the researchers to identify cliques and maximal cliques in the graph. However, our aim in this paper is not to identify all the cliques in

the network but to find mortality related maximal cliques in the comorbidity network, given a base diagnosis.

From the matrix resulted in the previous section (D_{nXn}), first, the diagnoses with high mortality risk and then their cliques are identified, if any. From the health records in EMR, mortality risk associated with each diagnosis (d_i) is calculated as proportion of patients deceased having the diagnosis d_i. The diagnoses with mortality risk more than 0.1 are considered as high risk diagnoses. Because the death rate in US is 733.1 per 100,000[14] (in 2015), mortality risk of 0.1 due to a diagnosis is significantly higher than a random death. These diagnoses are considered as the base diagnoses for which cliques are identified.

The concept of a clustering coefficient is adopted to find the maximal clique of a base diagnosis. The clustering coefficient explains the small clusters formed by the nodes in the network. The clustering coefficient of a node explains how well the neighbors of a node are connected (Watts & Strogatz, 1998). With respect to the comorbidity network, clustering coefficient of a disease, d_i, explains how well the direct connections of the disease, d_i, are connected to each other. The clustering coefficient of a node d_i can be mathematically written as

$$t_i = \frac{2l_i}{k_i(k_i-1)}, \qquad \textbf{-(6.4)}$$

where l_i is the number of links among the neighbors of the node i and k_i is the degree of a node i. The number t_i ranges from 0 to 1. At $t_i=0$, none of the direct connections of d_i are connected to each other. On the other hand, at $t_i=1$, all nodes are adjacent to each other making the sub-network a clique comprising d_i and its adjacent diagnoses. The size of the clique is k_i+1 as it includes d_i and its directly connected diagnoses.

[14] https://www.cdc.gov/nchs/data/hus/hus16.pdf#019

The entire step by step process for finding mortality related cliques is presented in Table 6.1. A matrix named A_{4Xn} is used to store mortality risk, high risk indicator and clustering coefficient of each diagnosis. In A_{4Xn}, A_1 contains the name of diagnosis (d_i, where i=1 to n), A_2 represents the mortality risk, A_3 indicates whether the diagnosis is high-risk or not (i.e. mortality risk>.1), and A_4 contains the clustering coefficient of each diagnosis.

First of all, we converted the weighted matrix D_{nXn} into unweighted network by changing the $w_{ij} = 1$ where $w_{ij}>0$ as in Code 6.1 in Table 6.1. In the next step, the diagnoses with high risk of mortality are identified. In addition, clustering coefficient of each diagnosis is calculated as in

Table 6.1. An algorithm to find diagnoses forming cliques with high mortality rate

Input: D_{nXn} matrix of edges with weights w_{ij} $\forall$ $i, j \in$ (1 to n)
In our network, n=1043, the number of distinct diagnoses
Input: A_{4Xn} contains diagnoses with its attributes:
1-Diagnosis, 2-Mortality Risk, 3-High risk indicator, 4-Clustering Coefficient
Let
Numeric *MortalityRate* – It denotes a cutoff to consider mortality risk as higher or lower. In our study, we use 0.1 as the cutoff.

Code 6.1: Convert weighted matrix into unweighted
For i=1 to n
  For j=1 to n
    If $w_{ij} > 0$ **then** w_{ij}=1
  Loop
Loop

Code 6.2: Identify high mortality risk diagnoses forming cliques
For i=1 to n
  If $A_{i2} >$ *MortalityRate* **then** $A_{i3} = 1$
  Else $A_{i3} = 0$
  A_{i4} = Cluster coefficient of A_{i1} calculated as per equation 4.
  If $A_{i3} = 1$ and $A_{i4} = 1$ **then**
    Call Sub-routine *CliqueNodes (i)*
Loop

Code 6.3: A sub routine to direct connections of the diagnoses forming cliques
Sub *CliqueNodes* (i)
  For j=1 to n
    If w_{ij}=1 and $i \neq j$ then
    Write D_i, D_j
  Loop
End

Code 6.2. The diagnoses that are high risk and have maximum clustering coefficient (i.e. 1) are selected. Finally, the direct connections of the identified diagnoses forming cliques are extracted from the matrix of connections as in Code 6.3.

6.4 DATA DESCRIPTION AND PREPARATION

We use an Electronic Medical Record (EMR) containing health records of more than 58 million patients across 662 hospitals in US (2000-2016). This EMR is provided by Cerner Corporation, a major EMR provider. The database includes information about all types of patients (emergency, inpatient, outpatient, etc.), admissions (elective, urgent, new born, etc.), hospitals and payers (Medicaid, Medicare, self-pay and other private payers). For our purpose, we extracted all types of patients and diseases diagnosed across different hospital visits. The diagnoses are recorded according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). We removed hospital visits in which patients were not diagnosed with any type of disease or symptom. After removing such encounters, we had approximately 24.7 million unique patients with sufficient information to perform analysis.

For data analysis, we follow the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines created by Collins et al. (2015). The steps followed in processing the massive data containing medical records of 24.7 million unique patients are presented in a flowchart in Figure 6.2. The flowchart also presents the number of patients and hospital visits at each step in the process. First of all, information about the patients, their hospital visits and diseases diagnosed were merged. An integration of different datasets resulted medical records of 24.7 million patients diagnosed with at least one disease or symptom. The entire dataset was used for find a cut-off for Salton Cosine Index using the *Process 1*. As discussed earlier, we used the relationship between Pearson's correlation coefficient and SCI to find a cut-off for SCI. From the comorbidity network of 24.7 million patients developed using Pearson's Correlations Coefficient significant at $\alpha=0.01$ and cij>Average (cij), the number of

statistically significant edges were equal at the SCI cutoff of 0.04. Therefore, from here onwards, we consistently use SCI cutoff of 0.04 for creating an edge in the comorbidity network.

The large dataset containing health records of 24.7 million patients was divided into two random samples of equal size. The first sample was used to create the comorbidity network, identify cliques related to the high risk diagnoses and calculate the impact of cliques in the patients. Then, the second random sample was used to validate the effect of cliques on mortality on an independent set of patients.

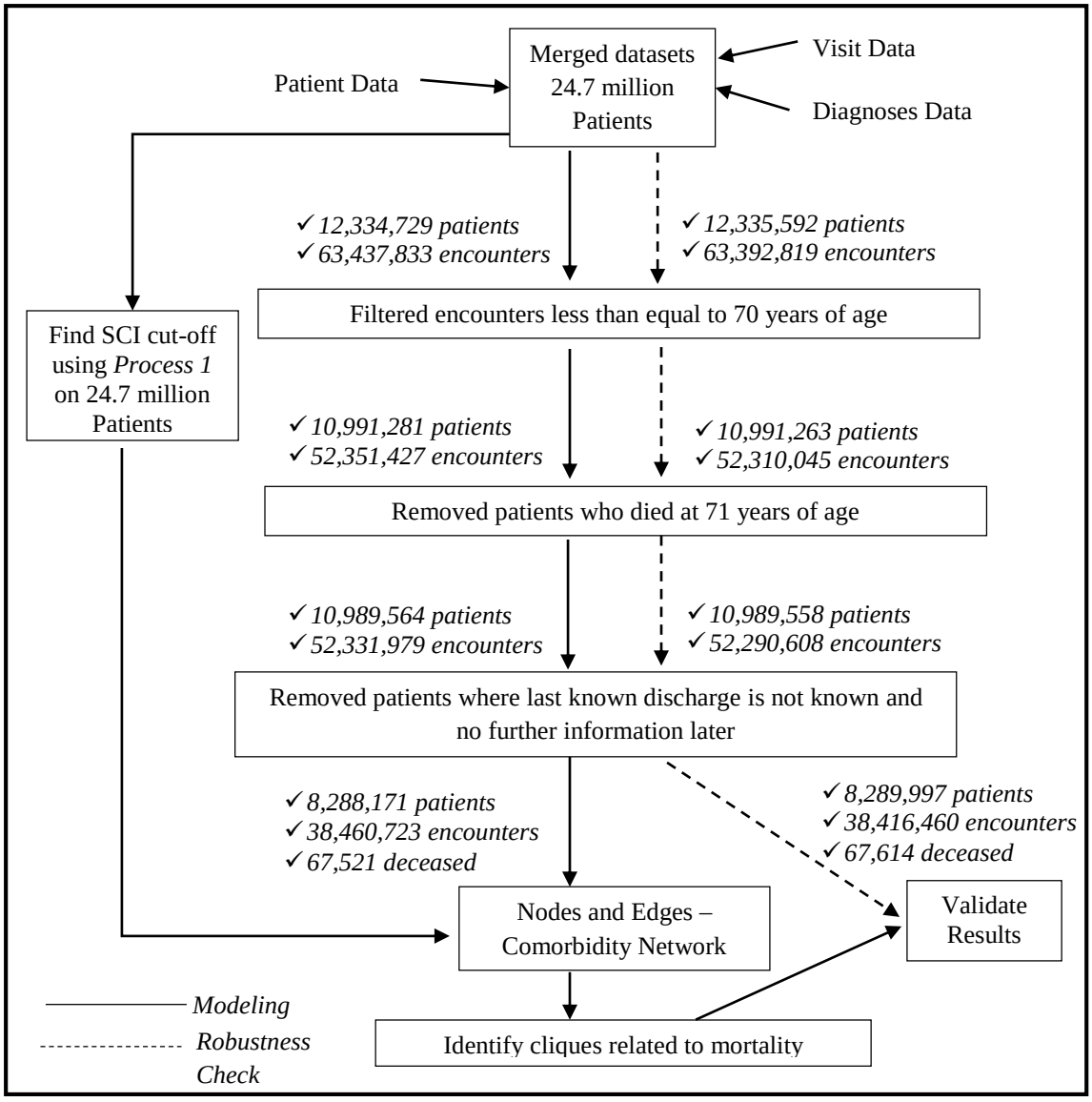


Figure 6.2. Data processing

To reduce some bias related to the age, we restricted our analysis to the patients up to 70 years of age. Moreover, we followed the patient records for one more year and removed those who deceased at 71 years of age so that the number of false positives are minimized. In addition, we deleted the patients from the analysis where the last known discharge of the patients was either not valid or unknown. These filters left us with more than 8.2 million patients in each sample. In both samples, the death rate is around 800 out of 100,000.

6.5 RESULTS

In the dataset of 12 million patients, we identify eighteen base diagnoses where the mortality risk of having that specific diagnoses is more than 0.1. The list including the ICD-9 codes is given in Table 6.2. The table also includes the size of clique, which is the number of diagnoses in a clique including the base diagnosis. The list includes cancers of plasma cells, lymphatic tissue, pancreas, stomach, esophagus, head, face, neck, thorax, abdomen, pelvis, limbs, ovary and other uterine adnexa. Other diagnoses include nutritional marasmus, portal vein thrombosis, acute and subacute endocarditis, subarachnoid hemorrhage, mycoses, abscess of lung and mediastinum, alveolar and parietoalveolar pneumonopathy, pneumococcal pneumonia and dementias. The encounter for dialysis and dialysis catheter care also forms a clique with its neighbors.

The comparison of the mortality rate among patients with and without a specific clique, given the base diagnosis, is performed. Since our focus is on the base disease, we perform the comparison of the mortality risk in the patients with a particular base disease. For instance, Figure 6.3 presents the clique of Portal vein thrombosis (ICD9-452) as the base diagnosis. The size of this clique is three, which includes chronic liver disease and cirrhosis (ICD9-571) and liver abscess and sequelae of chronic liver disease (ICD9-572). There were 1,327 portal vein thrombosis patients without the clique having the mortality rate of 18.3%. However, there were 871 portal vein thrombosis patients

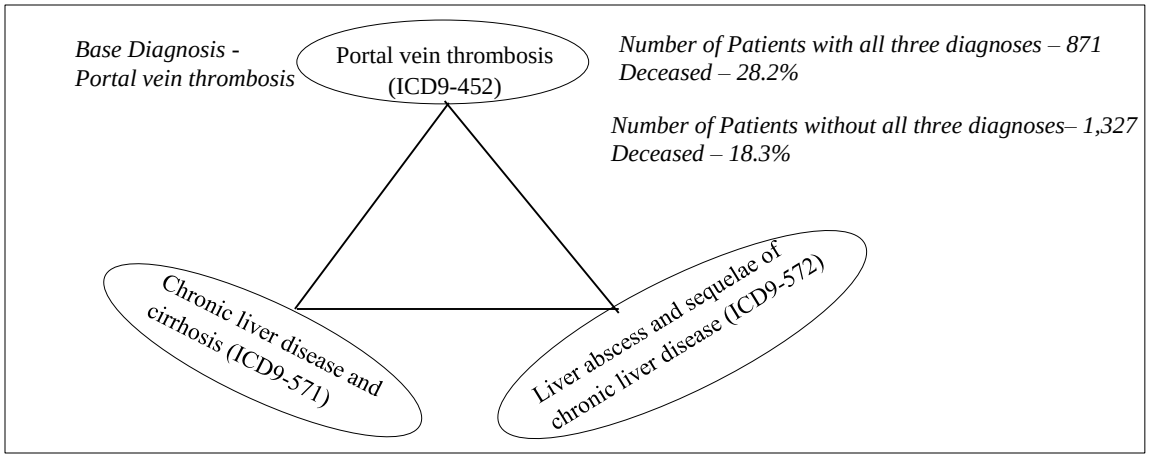


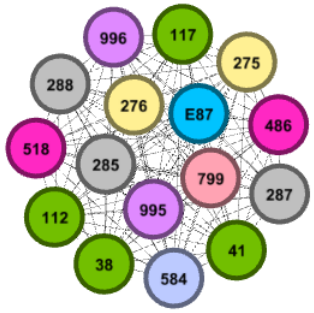
Figure 6.3. A clique/triangle of three diseases with their joint impact on mortality

diagnosed with all three diseases with mortality rate of 28.2%. It clearly indicates that clique increases the mortality risk significantly ($p<0.05$). The same comparison is done for all the base diagnoses. The description of all cliques including diagnoses can be found in Figures 6.4a to 6.4r.

The increase in risk of mortality due to the clique formation within the patient is listed in Table 6.2 and presented in Figure 6.5. Except for the cancer of esophagus, the presence of all other cliques have significant effect on the mortality risk. There are some diagnoses in the table where the number of patients with a clique is small such as patients diagnosed with nutritional marasmus, malignant neoplasm of pancreas, acute and subacute endocarditis, mycoses, malignant neoplasm of head, face, neck, thorax, abdomen, pelvis and limbs, and multiple myeloma and immunoproliferative neoplasms. Although the effect of clique on mortality in these patients was large, the issue of small sample size still exists. However, in other categories such as patients with portal vein thrombosis, stomach cancer, ovary and other uterine adnexa cancer, plasma cell cancer, abscess of lung and mediastinum, pneumococcal pneumonia and dementias, the number of patients deceased is significantly large. The rise in mortality risk due to clique diagnoses is confirmed from these types of patients. Since the mortality risk increases when cliques are present, it should obviously alarm the physicians.

Table 6.2. Mortality rate with and without cliques						
ICD-9	**Description**	**Clique Size**	**Patient count w/o Clique**	**Mortality rate w/o a clique (%)**	**Patient count with Clique**	**Mortality rate with clique (%)**
261*	Nutritional marasmus	13	3,725	32.0	30	60
452*	Portal vein thrombosis	3	1,327	18.3	871	28.2
157*	Malignant neoplasm of pancreas	7	5,808	21.5	73	43.8
421*	Acute and subacute endocarditis	15	3,760	21.2	67	41.8
151*	Malignant neoplasm of stomach	3	2,968	17.1	427	42.9
150	Malignant neoplasm of esophagus	3	2,772	18.8	905	19.0
430*	Subarachnoid hemorrhage	4	5,520	16.0	192	24.0
117*	Mycoses	16	6,558	15.6	21	33.3
513*	Abscess of lung and mediastinum	3	1,513	9.8	659	22.2
516*	Alveolar and parietoalveolar pneumonopathy	6	4,013	12.0	201	30.8
510*	Empyema	9	3,985	12.1	151	30.5
195*	Malignant neoplasm of head, face, neck, thorax, abdomen, pelvis and limbs	12	6,096	12.6	14	71.4
481*	Pneumococcal pneumonia	8	4,098	11.6	160	26.3
V56*	Encounter for dialysis and dialysis catheter care	9	2,074	9.5	628	18.2
200*	Lymphosarcoma and reticulosarcoma	5	3,845	10.5	375	18.1
203*	Multiple myeloma and immunoproliferative neoplasms	12	5,843	10.8	19	47.4
183*	Malignant neoplasm of ovary and other uterine adnexa	7	6,929	10.2	142	40.1
290*	Dementias	8	4,305	9.5	152	28.3

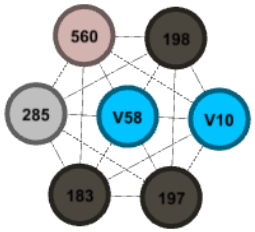
** The proportions are statistically different ($p<0.05$)*



Base Diagnosis – Mycoses (117)

117- Mycoses
287- Purpura and other hemorrhagic conditions
584- Acute kidney failure
995- Certain adverse effects not elsewhere classified
518- Other diseases of lung
275- Disorders of mineral metabolism
486- Pneumonia, organism unspecified
288- Diseases of white blood cells
799- Ill-defined causes of morbidity and mortality
276- Disorders of fluid electrolyte and acid-base balance
996- Complications peculiar to certain procedures
285- Other and unspecified anemias
E87- Accident during medical care
038- Septicemia
112- Candidiasis
041- Bacterial infection

Figure 6.4a. Clique 1



Base Diagnosis – Malignant neoplasm of ovary and other uterine adnexa (183)

183- Malignant neoplasm of ovary and other uterine adnexa
560- Intestinal obstruction without mention of hernia
198- Secondary malignant neoplasm of other specified sites
V58- Encounter for other and unspecified procedures and aftercare
V10- Personal history of malignant neoplasm
285- Other and unspecified anemias
197- Secondary malignant neoplasm of respiratory and digestive systems

Figure 6.4b. Clique 2

150- Malignant neoplasm of esophagus
V10- Personal history of malignant neoplasm
530- Diseases of esophagus

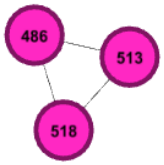
Base Diagnosis - Malignant neoplasm of esophagus (150)

Figure 6.4c. Clique 3

151- Malignant neoplasm of stomach
197- Secondary malignant neoplasm of respiratory and digestive systems
V10- Personal history of malignant neoplasm

Base Diagnosis - Malignant neoplasm of stomach (151)

Figure 6.4d. Clique 4



513- Abscess of lung and mediastinum
486- Pneumonia, organism unspecified
518- Other diseases of lung

Base Diagnosis - Abscess of lung and mediastinum (513)

Figure 6.4e. Clique 5

430- Subarachnoid hemorrhage
437- Other and ill-defined cerebrovascular disease
518- Other diseases of lung
348- Other conditions of brain

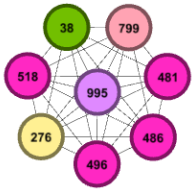
Base Diagnosis - Subarachnoid hemorrhage (430)

Figure 6.4f. Clique 6

200- Lymphosarcoma, reticulosarcoma and malignant tumors of lymphatic tissue
288- Diseases of white blood cells
V58- Encounter for other and unspecified procedures and aftercare
V10- Personal history of malignant neoplasm
202- Other malignant neoplasms of lymphoid and histiocytic tissue

Base Diagnosis - Lymphosarcoma, reticulosarcoma and malignant tumors of lymphatic tissue (200)

Figure 6.4g. Clique 7



481- Pneumococcal pneumonia
038- Septicemia
799- Other ill-defined and unknown causes of morbidity and mortality
995- Certain adverse effects not elsewhere classified
486- Pneumonia, organism unspecified
496- Chronic airway obstruction, not elsewhere classified
276- Disorders of fluid electrolyte and acid-base balance
518- Other diseases of lung

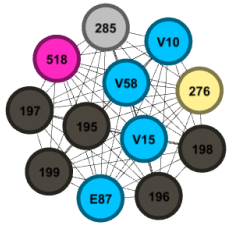
Base Diagnosis - Pneumococcal pneumonia (481)

Figure 6.4h. Clique 8

452- Portal vein thrombosis
572- Liver abscess and sequelae of chronic liver disease
571- Chronic liver disease and cirrhosis

Base Diagnosis - Portal vein thrombosis (452)

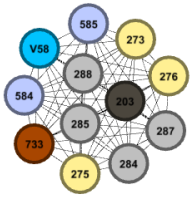
Figure 6.4i. Clique 9



285- Other and unspecified anemias
V10- Personal history of malignant neoplasm
V58- Encounter for other and unspecified procedures and aftercare
V15- Other personal history presenting hazards to health
E87- Accident during medical care
276- Disorders of fluid electrolyte and acid-base balance
198- Secondary malignant neoplasm of other specified sites
196- Secondary and unspecified malignant neoplasm of lymph nodes
195- Malignant neoplasm of other and ill-defined sites
199- Malignant neoplasm without specification of site
197- Secondary malignant neoplasm of respiratory and digestive systems
518- Other diseases of lung

Base Diagnosis - Malignant neoplasm of other and ill-defined sites (195)

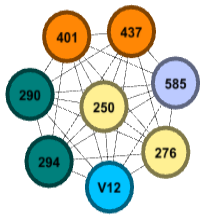
Figure 6.4j. Clique 10



203- Multiple myeloma and immunoproliferative neoplasms
585- Chronic kidney disease (ckd)
273- Disorders of plasma protein metabolism
276- Disorders of fluid electrolyte and acid-base balance
288- Diseases of white blood cells
285- Other and unspecified anemias
275- Disorders of mineral metabolism
284- Aplastic anemia and other bone marrow failure syndromes
287- Purpura and other hemorrhagic conditions
V58- Encounter for other and unspecified procedures and aftercare
584- Acute kidney failure
733- Other disorders of bone and cartilage

Base Diagnosis - Multiple myeloma & immunoproliferative neoplasms (203)

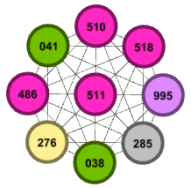
Figure 6.4k. Clique 11



Base Diagnosis – Dementias (290)

290- Dementias
401-Essential hypertension
437- Ill-defined cerebrovascular disease
250- Diabetes mellitus
294- Persistent mental disorders
276- Disorders of fluid electrolyte and acid-base balance
585- Chronic kidney disease (ckd)
V12- Personal history of certain other diseases

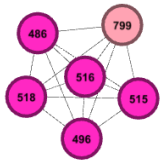
Figure 6.4l. Clique 12



Base Diagnosis – Empyema (510)

510- Empyema
518- Other diseases of lung
511- Pleurisy
995- Certain adverse effects not elsewhere classified
285- Other and unspecified anemias
038- Septicemia
276- Disorders of fluid electrolyte and acid-base balance
486- Pneumonia, organism unspecified
041- Bacterial infection

Figure 6.4m. Clique 13



Base Diagnosis - Alveolar & parietoalveolar pneumonopathy (516)

486- Pneumonia, organism unspecified
496- Chronic airway obstruction
515- Postinflammatory pulmonary fibrosis
516- Other alveolar and parietoalveolar pneumonopathy
518- Other diseases of lung
799- Ill-defined causes of morbidity and mortality

Figure 6.4n. Clique 14

Base Diagnosis – Encounter for dialysis and dialysis catheter care (V56)

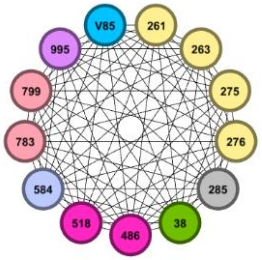
V56- Encounter for dialysis and dialysis catheter care
588- Disorders resulting from impaired renal function
585- Chronic kidney disease (ckd)
V45- Other postprocedural states
E87- Accident during medical care
276- Disorders of fluid electrolyte and acid-base balance
285- Other and unspecified anemias
403- Hypertensive chronic kidney disease
996- Complications peculiar to certain procedures

Figure 6.4o. Clique 15

Base Diagnosis – Malignant neoplasm of pancreas (157)

157- Malignant neoplasm of pancreas
V10- Personal history of malignant neoplasm
197- Secondary malignant neoplasm of respiratory and digestive systems
276- Disorders of fluid electrolyte and acid-base balance
285- Other and unspecified anemias
576- Other disorders of biliary tract
577- Diseases of pancreas

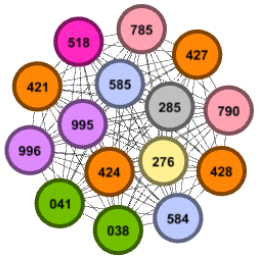
Figure 6.4p. Clique 16



Base Diagnosis – Nutritional marasmus (261)

261- Nutritional marasmus
V85- Body mass index (bmi)
263- Other and unspecified protein-calorie malnutrition
275- Disorders of mineral metabolism
276- Disorders of fluid electrolyte and acid-base balance
285- Other and unspecified anemias
038- Septicemia
486- Pneumonia, organism unspecified
518- Other diseases of lung
584- Acute kidney failure
783- Symptoms concerning nutrition metabolism and development
799- Ill-defined causes of morbidity and mortality
995- Certain adverse effects not elsewhere classified

Figure 6.4q. Clique 17



Base Diagnosis – Acute and subacute endocarditis (421)

421- Acute and subacute endocarditis
518- Other diseases of lung
785- Symptoms involving cardiovascular system
427- Cardiac dysrhythmias
790- Nonspecific findings on examination of blood
428- Heart failure
584- Pneumonia in infectious diseases
038- Septicemia
041- Bacterial infection
996- Complications peculiar to certain specified procedures
585- Chronic kidney disease (ckd)
285- Other and unspecified anemias
276- Disorders of fluid electrolyte and acid-base balance
424- Other diseases of endocardium
995- Certain adverse effects not elsewhere classified

Figure 6.4r. Clique 18

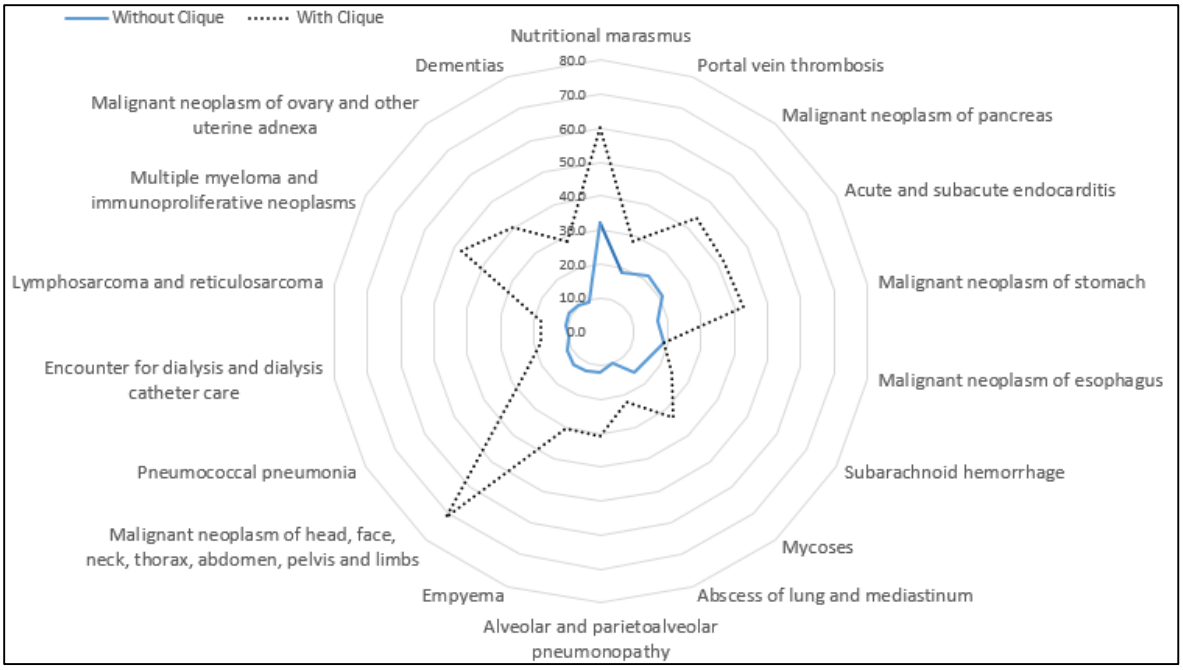


Figure 6.5. Mortality rate with and without clique

Moreover, we validated the effect of cliques on mortality by performing a similar comparison on an independent dataset. We found similar results with cliques having more risk of mortality. The results on the validation dataset are listed in Table 6.3. This validation on an independent dataset confirms the robustness of our analysis and results.

Although the method and findings from this study are unique, we relate these back to the literature of comorbidity. For instance, we found several cancer-related cliques indicating comorbidities impacting the mortality rate. Our finding is supported by the previous researchers. For example, Sarfati, Koczwara and Jackson (2016) argued that comorbidities among cancer patients lead to poorer survival and quality of life. Similarly, empyema is known to be related to multiple other diseases like pneumonia, septicemia, lung diseases as discussed by Li and Gates (2008), which are related to mortality. Adding to the above literature of comorbidities, using a novel approach, we are able to show the risk of collection of diagnoses in terms of cliques.

Table 6.3. Robustness Check on Validation Dataset					
ICD-9	**Description**	**Patient count w/o Clique**	**Mortality rate w/o clique (%)**	**Patient count with Clique**	**Mortality rate with clique (%)**
261	Nutritional marasmus	3,514	31.5	36	41.7
452*	Portal vein thrombosis	1,220	18.3	902	29.6
157*	Malignant neoplasm of pancreas	5,756	21.7	54	51.9
421*	Acute and subacute endocarditis	3,792	19.2	60	50.0
151*	Malignant neoplasm of stomach	3,042	16.9	436	37.8
150	Malignant neoplasm of esophagus	2,679	17.6	928	19.4
430*	Subarachnoid hemorrhage	5,621	16.3	172	28.5
117	Mycoses	6,645	15.7	26	23.1
513*	Abscess of lung and mediastinum	1,468	10.5	673	22.3
516*	Alveolar and parietoalveolar pneumonopathy	4,015	12.9	198	33.3
510*	Empyema	4,145	12.2	121	33.9
195*	Malignant neoplasm of head, face, neck, thorax, abdomen, pelvis and limbs	6,045	12.2	12	50.0
481*	Pneumococcal pneumonia	4,194	11.8	140	33.6
V56*	Encounter for dialysis and dialysis catheter care	2,061	10.0	604	18.9
200*	Lymphosarcoma and reticulosarcoma	3,770	9.2	340	16.8
203*	Multiple myeloma and immunoproliferative neoplasms	5,694	10.7	15	26.7
183*	Malignant neoplasm of ovary and other uterine adnexa	6,902	10.0	150	38.7
290*	Dementias	4,342	10.1	148	29.7

** The proportions are statistically different ($p<0.05$)*

6.6 CONCLUSIONS

We showed structure properties such as cliques of a network inferred from the co-occurrences effect health outcomes of the patients. We discovered underlying interaction between different diagnoses related to the mortality. Our study has several practical implications. Clique being a complete subnetwork indicates a trap of diseases. Since the mortality rate increases if a clique is developed, physicians should take measures to avoid such trap of diagnoses. Moreover, the presence of a clique in a patient should alarm the physician to take preemptive action. Studying the impact of multimorbidity on mortality from the clique point of view definitely increases the understanding about the possible impact of multiple diagnoses on mortality.

We recognize few limitations of our approach. First, we identified maximal cliques related to the base diagnosis. However, it is possible that some diagnoses are redundant and may not be related to mortality. In the future, we will find the high risk sub-network of the clique by eliminating the redundant diagnoses. Second, we acknowledge that there are several diagnoses that are highly related to mortality but not forming the cliques. We did not include such diagnoses in our analysis because the main focus was on the clique property.

Notwithstanding the limitations, the approach presented in this paper is generalizable to multiple other business and social problems where a network can be inferred from the actions of the users. For example, a clique of products in the product purchase network can help vendors to create bundles of products to sell. In this case, the cliques of products can be identified based on the profitability, analogous to mortality in our case. A similar idea can be applied to find cliques of mobile apps in the mobile-apps network, which can be inferred from the usage patterns of users. Such cliques of mobile apps can be bundled together on the online store. Another application of our approach is in finding cliques of locations in the network of locations inferred from the travel patterns of the tourists. Some tourist spots may emerge as profitable cliques and therefore, travel agents can create holiday packages comprising clique locations.

CHAPTER VII

CONCLUSIONS

7.1. CONTRIBUTIONS AND GENERALIZABILITY

Our network is one of a category that emerges without the intentions of its source, often referred to as unintentional, inferred or implicit. Patients are the source, but the network is formed implicitly unknowingly. This class differs from traditional or explicit networks, which are developed from the intentional actions of its members. The method and applications presented in this dissertation has practical, methodological and theoretical contributions.

Inferring network itself is a challenge. Therefore, the contribution of Chapter 2 is significant where we proved that a cosine index is a better choice than a correlation coefficient. We showed that selecting an index independent of the sample size provides valid network comprising true relationships. This contribution can be generalizable to all implicit networks inferred from the sample. For example, affinity between products in a market basket of buyers in a grocery store should be computed using an index independent of the number of buyers.

Moreover, in Chapter 2, we proved that the comorbidity network follows the small-world topology. This is a significant contribution to the comorbidity and medical literature. The small-world property of the network has several practical implications for the providers. For example, the clusters in the network can be used to lay out departments. Moreover, clusters of diseases can guide pharmaceutical companies to understand the side-effects of a medicine. Generally, the

network properties can help the physicians prioritize the diagnoses based on the position of a disease in the network.

A network emerged from the unintended actions can be different in groups because groups might behave differently in different situations. Like in Chapter 3 and 4, we discovered different behavior of diseases in different population groups based on gender, race, and insurance, this is true for other inferred networks as well. For instance, the co-purchasing pattern of males and females may form two different structures. Therefore, the approach of analyzing different network is different population groups is generalizable to other networks to gain more insights about the phenomenon.

The network method can also help create models in high dimensional space by summarizing the relationships between different components of a system. A network can be a basis for the algorithms to predict future outcomes. In Chapters 5 and 6, we described two algorithms using the network properties. The algorithms were used to create models of different levels i.e. i.e. at the network level and the performance outcome level. Therefore, this dissertation has algorithmic contributions. These algorithms can be applied in other problems domains where inferred network can be used to build models at different levels.

Because a network can handle high dimensional space, a scale incorporating high dimensionality to quantify any phenomenon can be created. In Chapter 5, we used network properties to devise a measure to compute comorbidity at a patient level. This is another contribution of this dissertation like Shmueli and Koppius (2011) explained in their commentary that a measure development is one of the roles of analytics in scientific research.

The measure developed from the network was used to predict exogenous outcome i.e. length of stay. Here, we take the application of implicit networks to next level, exploring how these impact the uncontrollable performance of the external source. Similarly, in other inferred networks,

structural properties can be used to explain and predict outcomes external to the network. The use of network variables for modeling is itself a major contribution of this dissertation. This approach provides researchers a new dimension to understand the phenomenon.

A network embeds several inherent structural properties resulting from the interactions between the nodes. Understanding these properties theoretically is important. In Chapter 6 of this dissertation, we theorized a very critical inherent property of the comorbidity network i.e. cliques. A clique was theorized as a trap state from where the exit is difficult. The cliques identified were related to high mortality risk. To the best of our knowledge, no one in the past has theorized a clique property of the comorbidity network in the form of a trap state. Moreover, the method of finding cliques related to an exogenous property (mortality in our case) can be applied in other networks. Finding Clique is a never ending problem and a simpler algorithm described in Chapter 6 to find these based on prevalent outcomes is a methodological contribution of this dissertation.

7.2. FUTURE WORK

We identified comorbidity differences across population groups in Chapter 3 and 4. To increase the usefulness of our results for practical purposes, we are creating a website describing the comorbidity differences interactively. Physicians will be able to use our website to predict future diagnoses and understand the state of a patient.

In addition, we primarily focused on the length of stay and mortality as the health outcomes. However, in future, we will study how network properties influence other health outcomes of patients such as readmission probability. In addition, we will study how a network can be used to design new interventions.

The comorbidity networks created in this dissertation were undirected with no directions between the diagnoses. However, if we consider the time of a diagnosis, a directed network with direction

between the diagnoses can be created. It will explain how one disease leads to another disease. This will help explain the causal relationship between diagnoses.

As a part of our future work, we will generalize our methodology and approach to other domains. We will extend our method to the retail industry problems. One future project is to create a network of products based on shopping cart of customers in a grocery store. The structural properties of product network will be utilized to study performance outcomes of the customers. For example, the structural properties of the products in current shopping cart might predict the future cart and future spending.

Our approach can also be applied to understand technology related behavioral outcomes. For instance, an implicit network formed from technology use might explain performance outcomes such as technology addiction, satisfaction, etc. An example of such an implicit network is the smartphone app network formed from the use of multiple apps one after the other by users. Similarly, we will explore different hidden networks in various domains.

REFERENCES

Adler, P. S., & Kwon, S. W. (2002). Social capital: Prospects for a new concept. *Academy of management review*, *27*(1), 17-40.

Agarwal, R., & Dhar, V. (2014). Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research. *Information systems research*, *25*(3), 443-448.

Agarwal, R., & Lucas Jr, H. C. (2005). The information systems identity crisis: Focusing on high-visibility and high-impact research. *MIS Quarterly*, 381-398.

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, *54*(6), 550-560.

Ahluwalia, S. C., Gross, C. P., Chaudhry, S. I., Ning, Y. M., Leo-Summers, L., Van Ness, P. H., & Fried, T. R. (2012). Impact of comorbidity on mortality among older persons with advanced heart failure. *Journal of general internal medicine*, *27*(5), 513-519.

Alcón, L., Faria, L., de Figueiredo, C. M., & Gutierrez, M. (2009). The complexity of clique graph recognition. *Theoretical Computer Science*, *410*(21-23), 2072-2083.

Azevedo, A. I. R. L., & Santos, M. F. (2008). Kdd, semma crisp-dm: a parallel overviee. *IADS-DM*.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, *12*(1), 56-68.

Bardhan, I., Oh, J. H., Zheng, Z., & Kirksey, K. (2014). Predictive analytics for readmission of patients with congestive heart failure. *Information Systems Research*, *26*(1), 19-39.

Bassett, D. S., & Bullmore, E. (2006). Small-World Brain Networks. *The Neuroscientist*, *12*(6), 512-523. doi:10.1177/1073858406293182

Bauer-Mehren, A., Bundschus, M., Rautschka, M., Mayer, M. A., Sanz, F., & Furlong, L. I. (2011). Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PloS one*, *6*(6), e20284.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, *77*(2), 81-97.

Blazer, D. G., Moody-Ayers, S., Craft-Morgan, J., & Burchett, B. (2002). Depression in diabetes and obesity: racial/ethnic/gender issues in older adults. *Journal of Psychosomatic Research*, *53*(4), 913-916.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 1170-1182.

Bonacich, P., & Liggett, T. M. (2003). Asymptotics of a matrix valued Markov chain arising in sociology. *Stochastic Processes and their Applications*, *104*(1), 155-171.

Braunstein, J. B., Anderson, G. F., Gerstenblith, G., Weller, W., Niefeld, M., Herbert, R., & Wu, A. W. (2003). Noncardiac comorbidity increases preventable hospitalizations and mortality among Medicare beneficiaries with chronic heart failure. *Journal of the American College of Cardiology*, *42*(7), 1226-1233.

Bresnahan, M., Begg, M. D., Brown, A., Schaefer, C., Sohler, N., Insel, B., ... & Susser, E. (2007). Race and risk of schizophrenia in a US birth cohort: another example of health disparity?. *International journal of epidemiology*, *36*(4), 751-758.

Butte, A. J., & Kohane, I. S. (2006). Creation and implications of a phenome-genome network. *Nature biotechnology*, *24*(1), 55-62.

Carter, E. M., & Potts, H. W. (2014). Predicting length of stay from an electronic patient record system: a primary total knee replacement example. *BMC medical informatics and decision making*, *14*(1), 26.

Case, A. C. & Paxson, C. (2004). Sex Differences in Morbidity and Mortality. *Demography 42*(2), 189-214.

Chapman, E. N., Kaatz, A., & Carnes, M. (2013). Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. *Journal of general internal medicine*, *28*(11), 1504-1510.

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

Chen, Y., Zhang, X., Zhang, G. Q., & Xu, R. (2015). Comparative analysis of a novel disease phenotype network based on clinical manifestations. *Journal of biomedical informatics*, *53*, 113-120.

Chertow, G. M., Burdick, E., Honour, M., Bonventre, J. V., & Bates, D. W. (2005). Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *Journal of the American Society of Nephrology*, *16*(11), 3365-3370.

Chmiel, A., Klimek, P., & Thurner, S. (2014). Spreading of diseases through comorbidity networks across life and gender. *New Journal of Physics*, *16*(11), 115013.

Clague, J. E., Craddock, E., Andrew, G., Horan, M. A., & Pendleton, N. (2002). Predictors of outcome following hip fracture. Admission time predicts length of stay and in-hospital mortality. *Injury*, *33*(1), 1-6.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American journal of sociology*, S95-S120.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC medicine, 13*(1), 1.

Corrigan, P. (2004). How stigma interferes with mental health care. *American psychologist, 59*(7), 614.

de Groot, V., Beckerman, H., Lankhorst, G. J., & Bouter, L. M. (2003). How to measure comorbidity: a critical review of available methods. *Journal of clinical epidemiology, 56(*3), 221-229.

Dhar, V., Geva, T., Oestreicher-Singer, G., & Sundararajan, A. (2014). Prediction in economic networks. *Information Systems Research, 25*(2), 264-284.

D'hoore, W., Sicotte, C., & Tilquin, C. (1993). Risk adjustment in outcome assessment: the Charlson comorbidity index. *Methods of information in medicine*, *32*(5), 382-387.

Divo, M. J., Casanova, C., Marin, J. M., Pinto-Plata, V. M., de-Torres, J. P., Zulueta, J. J., . . . Berto, J. (2015). Chronic obstructive pulmonary disease comorbidities network. *European Respiratory Journal*, ERJ-01716-02014.

Egghe, L., & Leydesdorff, L. (2009). The relation between Pearson's correlation coefficient r and Salton's cosine measure. *Journal of the American Society for Information Science and Technology, 60*(5), 1027-1036.

Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical care*, *36*(1), 8-27.

Erdős, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen, 6*, 290-297.

Erdos, P., & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci, 5*(1), 17-60.

Euler, L. (1953). Leonhard Euler and the Königsberg bridges. *Scientific American, 189*, 66-70.

Feinstein, A. R. (1970). The pre-therapeutic classification of co-morbidity in chronic disease. *Journal of chronic diseases, 23*(7), 455-468.

Ferrazzi, F., Magni, P., Sacchi, L., Nuzzo, A., Petrovič, U., & Bellazzi, R. (2007). Inferring gene regulatory networks by integrating static and dynamic data. *International Journal of Medical Informatics, 76*, S462-S475.

Fine, M. J., Ibrahim, S. A., & Thomas, S. B. (2005). The Role of Race and Genetics in Health Disparities Research. *American Journal of Public Health, 95*(12), 2125-2128.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks, 1*(3), 215-239.

Funke, J. (1991). Solving complex problems: Exploration and control of complex systems. *Complex problem solving: Principles and mechanisms*, 185-222.

Furlanetto, L. M., & da Silva, R. V. (2003). The impact of psychiatric comorbidity on length of stay of medical inpatients. *General hospital psychiatry, 25*(1), 14-19.

Ganesh, A., Massoulié, L., & Towsley, D. (2005, March). The effect of network topology on the spread of epidemics. In *INFOCOM 2005. Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies* (Vol. 2, pp. 1455-1466).

Glaser, B. G., & Strauss, A. L. (2009). *The discovery of grounded theory: Strategies for qualitative research*: Transaction publishers.

Glaser, B. G., Strauss, A. L., & Strutzel, E. (1968). The Discovery of Grounded Theory; Strategies for Qualitative Research. *Nursing Research, 17*(4), 364.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences, 104*(21), 8685-8690.

Hachesu, P. R., Ahmadi, M., Alizadeh, S., & Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare informatics research, 19*(2), 121-129.

Haslam, B., & Perez-Breva, L. (2016). Learning disease relationships from clinical drug trials. *Journal of the American Medical Informatics Association, 24*(1), 13-23.

Hidalgo, C. A., Blumm, N., Barabási, A.-L., & Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol, 5*(4), e1000353.

Holguin, F., Folch, E., Redd, S. C., & Mannino, D. M. (2005). Comorbidity and mortality in COPD-related hospitalizations in the United States, 1979 to 2001. *Chest Journal, 128*(4).

Hong, W., Chan, F. K., Thong, J. Y., Chasalow, L. C., & Dhillon, G. (2013). A framework and guidelines for context-specific theorizing in information systems research. *Information systems research, 25*(1), 111-136.

Huntley, D. A., Cho, D. W., Christman, J., & Csernansky, J. G. (1998). Predicting length of stay in an acute psychiatric hospital. *Psychiatric services, 49*(8), 1049-1053.

Jakovljevic, M., & Ostojic, L. (2013). Comorbidity and multimorbidity in medicine today: challenges and opportunities for bringing separated branches of medicine closer to each other. *Psychiatr Danub, 25*(Suppl 1), 18-28.

Johnson, P., Fitzgerald, T., Salganicoff, A., Wood, S. F., & Goldstein, J. M. (2014). Sex-specific medical research: why women's health can't wait. *A report of the Mary Horrigan Connors Center for Women's Health & Gender Biology at Brigham and Women's Hospital. Brigham and Women's Hospital*.

Kalyani, R. R., Saudek, C. D., Brancati, F. L., & Selvin, E. (2010). Association of diabetes, comorbidities, and A1c with functional disability in older adults. *Diabetes care, 33*(5), 1055-1060.

Konstas, I., Stathopoulos, V., & Jose, J. M. (2009, July). On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 195-202). ACM.

Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., ... & Moreau, Y. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, *25*(3), 309-316.

Lee, D. S., Park, J., Kay, K. A., Christakis, N. A., Oltvai, Z. N., & Barabási, A. L. (2008). The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, *105*(29), 9880-9885.

Leontiadis, G. I., Molloy-Bland, M., Moayyedi, P., & Howden, C. W. (2013). Effect of comorbidity on mortality in patients with peptic ulcer bleeding: systematic review and meta-analysis. *The American journal of gastroenterology*, *108*(3), 331-345.

Li, S. T. T., & Gates, R. L. (2008). Primary operative management for pediatric empyema: decreases in hospital length of stay and charges in a national sample. *Archives of pediatrics & adolescent medicine*, *162*(1), 44-48.

Librero, J., Peiró, S., & Ordiñana, R. (1999). Chronic comorbidity and outcomes of hospital care: length of stay, mortality, and readmission at 30 and 365 days. *Journal of clinical epidemiology*, *52*(3), 171-179.

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary-too big to fail: large samples and the p-value problem. *Information systems research*, *24*(4), 906-917.

Liu, H. (2009). Statistical properties of Chinese semantic networks. *Chinese Science Bulletin*, *54*(16), 2781-2785.

Liu, V., Kipnis, P., Gould, M. K., & Escobar, G. J. (2010). Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Medical care*, *48*(8), 739-744.

Lopez-Gonzalez, L., Pickens, G. T., Washington, R., Weiss, A.J. (October 2014) Characteristics of Medicaid and Uninsured Hospitalizations, 2012. HCUP Statistical Brief #182. Agency for Healthcare Research and Quality, Rockville, MD. http://www.hcup-us.ahrq.gov/reports/statbriefs/sb182-Medicaid-Uninsured-Hospitalizations-2012.pdf

Lowell, W. E., & Davis, G. E. (1994). Predicting length of stay for psychiatric diagnosis-related groups using neural networks. *Journal of the American Medical Informatics Association*, *1*(6), 459-466.

Lyketsos, C. G., Dunn, G., Kaminsky, M. J., & Breakey, W. R. (2002). Medical comorbidity in psychiatric inpatients: relation to clinical outcomes and hospital length of stay. *Psychosomatics*, *43*(1), 24-30.

Mainous, A. G., Diaz, V. A., Everett, C. J., & Knoll, M. E. (2011). Impact of insurance and hospital ownership on hospital length of stay among patients with ambulatory care–sensitive conditions. *The Annals of Family Medicine*, *9*(6), 489-495.

Marazzi, A., Paccaud, F., Ruffieux, C., & Beguin, C. (1998). Fitting the distributions of length of stay by parametric models. *Medical care*, *36*(6), 915-927.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, *15*(4), 251-266.

Markides, K. S., & Eschbach, K. (2005). Aging, migration, and mortality: current status of research on the Hispanic paradox. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *60*(Special_Issue_2), S68-S75.

Markus, M. L., & Mao, J.-Y. (2004). Participation in development and implementation-updating an old, tired concept for today's IS contexts. *Journal of the Association for Information Systems*, *5*(11), 14.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American society for information science*, *41*(6), 433.

Marrie, R. A., Elliott, L., Marriott, J., Cossoy, M., Blanchard, J., Leung, S., & Yu, N. (2015). Effect of comorbidity on mortality in multiple sclerosis. *Neurology*, *85*(3), 240-247.

Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., & O'Connor, P. J. (2014). A machine learning approach to improving dynamic decision making. *Information systems research*, *25*(2), 239-263.

Moller-Leimkuhler, A. (2007). Gender differences in cardiovascular disease and comorbid depression. *Dialogues in clinical neuroscience, 9*(1), 71.

Müller, O., Junglas, I., vom Brocke, J., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems*.

National Center for Health Statistics (2016). Health, United States, 2015: With Special Feature on Racial and Ethnic Health Disparities. Hyattsville, MD.

Newman, M. E. (2005). A measure of betweenness centrality based on random walks. *Social networks*, *27*(1), 39-54.

Oh, S. S., Galanter, J., Thakur, N., Pino-Yanes, M., Barcelo, N. E., White, M. J., . . . Wu, A. H. (2015). Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS Med, 12*(12), e1001918.

Oksuzyan, A., Juel, K., Vaupel, J. W., & Christensen, K. (2008). Men: good health and high mortality. Sex differences in health and aging. *Aging clinical and experimental research, 20*(2), 91

Olson, S. H., Atoria, C. L., Cote, M. L., Cook, L. S., Rastogi, R., Soslow, R. A., . . . Elkin, E. B. (2012). The impact of race and comorbidity on survival in endometrial cancer. *Cancer Epidemiology Biomarkers & Prevention, 21*(5), 753-760.

Ottman, R., Lipton, R. B., Ettinger, A. B., Cramer, J. A., Reed, M. L., Morrison, A., & Wan, G. J. (2011). Comorbidities of epilepsy: results from the Epilepsy Comorbidities and Health (EPIC) survey. *Epilepsia, 52*(2), 308-315.

Ovseiko, P. V., Greenhalgh, T., Adam, P., Grant, J., Hinrichs-Krapels, S., Graham, K. E., ... & Al Rahbi, I. S. (2016). A global call for action to include gender in research impact assessment. *Health Research Policy and Systems, 14*(1), 50.

Pemantle, R., & Skyrms, B. (2004). Network formation by reinforcement learning: the long and medium run. *Mathematical Social Sciences, 48*(3), 315-327.

Pfuntner A, Wier LM, Steiner C. (December 2013) Costs for Hospital Stays in the United States, 2011. HCUP Statistical Brief #168. Agency for Healthcare Research and Quality, , Rockville, MD. http://europepmc.org/books/NBK179289

Provan, K. G., & Sebastian, J. G. (1998). Networks within networks: Service link overlap, organizational cliques, and network effectiveness. *Academy of Management journal*, *41*(4), 453-463.

Redelmeier, D. A., Tan, S. H., & Booth, G. L. (1998). The treatment of unrelated disorders in patients with chronic medical diseases. *New England Journal of Medicine*, *338*(21), 1516-1520.

Roca, C. P., Lozano, S., Arenas, A., & Sánchez, A. (2010). Topological traps control flow on real networks: The case of coordination failures. *PLoS One*, *5*(12), e15210.

Rochon, P. A., Katz, J. N., Morrow, L. A., McGlinchey-Berroth, R., Ahlquist, M. M., Sarkarati, M., & Minaker, K. L. (1996). Comorbid illness is associated with survival and length of hospital stay in patients with chronic disability: a prospective comparison of three comorbidity indices. *Medical care*, *34*(11), 1093-1101.

Roth, M., Ben-David, A., Deutscher, D., Flysher, G., Horn, I., Leichtberg, A., ... & Merom, R. (2010, July). Suggesting friends using the implicit social graph. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 233-242). ACM.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York, NY: McGraw-Hill, Inc.

Sarfati, D., Koczwara, B., & Jackson, C. (2016). The impact of comorbidity on cancer and its treatment. *CA: a cancer journal for clinicians*.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, *5*(4), 13-22.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, 289-310.

Shmueli, G., & Koppius, O. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly, 35*(3), 553-572.

Short, S. E., Yang, Y. C., & Jenkins, T. M. (2013). Sex, gender, genetics, and health. *American journal of public health, 103*(S1), S93-S101.

Simon, H. A. (1973). The structure of ill structured problems. *Artificial intelligence, 4*(3-4), 181-201.

Singer, M. (1996). A dose of drugs, a touch of violence, a case of AIDS: conceptualizing the SAVA syndemic. *Free Inquiry in Creative Sociology, 24*(2), 99-110.

Solé, R. V., Corominas-Murtra, B., Valverde, S., & Steels, L. (2010). Language networks: Their structure, function, and evolution. *Complexity, 15*(6), 20-26.

Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. (2001). Social networks and the performance of individuals and groups. *Academy of management journal, 44*(2), 316-325.

Spratt, N., Wang, Y., Levi, C., Ng, K., Evans, M., & Fisher, J. (2003). A prospective study of predictors of prolonged hospital stay and disability after stroke. *Journal of Clinical Neuroscience, 10*(6), 665-669.

Strauss, A., & Corbin, J. (1967). Discovery of grounded theory.

Struijs, J. N., Baan, C. A., Schellevis, F. G., Westert, G. P., & van den Bos, G. A. (2006). Comorbidity in patients with diabetes mellitus: impact on medical health care utilization. *BMC health services research, 6*(1), 84.

Tai, Y. M., & Chiu, H. W. (2009). Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan. *International journal of medical informatics, 78*(12), e75-e83.

Teng, C.-Y., Lin, Y.-R., & Adamic, L. A. (2012). *Recipe recommendation using ingredient networks.* Paper presented at the Proceedings of the 4th Annual ACM Web Science Conference.

Thomas, J. W., Guire, K. E., & Horvat, G. G. (1997). Is patient length of stay related to quality of care?. *Journal of Healthcare Management, 42*(4), 489.

Thombs, B. D., Singh, V. A., Halonen, J., Diallo, A., & Milner, S. M. (2007). The effects of preexisting medical comorbidities on mortality and length of hospital stay in acute burn injury: evidence from a national sample of 31,338 adult patients. *Annals of surgery, 245*(4), 629-634.

Thompson, J. W., Ryan, K. W., Pinidiya, S. D., & Bost, J. E. (2003). Quality of care for children in commercial and Medicaid managed care. *Jama, 290*(11), 1486-1493.

Tierney, W. M. (2001). Improving clinical decisions and outcomes with information: a review. *International journal of medical informatics, 62*(1), 1-9.

Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., & Roland, M. (2009). Defining comorbidity: implications for understanding health and health services. *The Annals of Family Medicine, 7*(4), 357-363.

Valente, T. W., Coronges, K., Lakon, C., & Costenbader, E. (2008). How correlated are network centrality measures?. *Connections (Toronto, Ont.), 28*(1), 16.

van den Akker, M., Buntinx, F., & Knottnerus, J. A. (1996). Comorbidity or multimorbidity: what's in a name? A review of literature. *The European Journal of General Practice, 2*(2), 65-70.

van den Akker, M., Buntinx, F., Metsemakers, J. F., Roos, S., & Knottnerus, J. A. (1998). Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. *Journal of clinical epidemiology, 51*(5), 367-375.

Van Den Heuvel, M. P., & Pol, H. E. H. (2010). Exploring the brain network: a review on resting-state fMRI functional connectivity. *European neuropsychopharmacology, 20*(8), 519-534.

Van Eck, N. J., & Waltman, L. (2008). Appropriate similarity measures for author co-citation analysis. *Journal of the American Society for Information Science and Technology, 59*(10), 1653-1661.

Von Alan, R. H., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature, 393*(6684), 440-442.

Weston, J., Elisseeff, A., Zhou, D., Leslie, C. S., & Noble, W. S. (2004). Protein ranking: from local to global structure in the protein similarity network. *Proceedings of the National Academy of Sciences of the United States of America, 101*(17), 6559-6563.

Wolda, H. (1981). Similarity indices, sample size and diversity. *Oecologia, 50*(3), 296-302.

Woolf, A. D., & Pfleger, B. (2003). Burden of major musculoskeletal conditions. *Bulletin of the World Health Organization, 81*(9), 646-656.

Yang, Y., Yang, K. S., Hsann, Y. M., Lim, V., & Ong, B. C. (2010). The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. *Journal of critical care, 25*(3), 398-405.

Yazdany, J., Feldman, C. H., Liu, J., Ward, M. M., Fischer, M. A., & Costenbader, K. H. (2014). Quality of care for incident lupus nephritis among Medicaid beneficiaries in the United States. *Arthritis care & research, 66*(4), 617-624.

Zhou, X., Menche, J., Barabási, A. L., & Sharma, A. (2014). Human symptoms–disease network. *Nature communications, 5*.

Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems, 74*, 150-161.

APPENDICES

Appendix A. Performance of the hospital length of stay models for patients with a specific primary diagnosis

ICD	Name	N	Mean	Baseline 4 R-square	COM 3 R-square
002	Typhoid and paratyphoid fevers	4389	1.2422	0.318	0.496
003	Other salmonella infections	6617	2.3350	0.231	0.254
005	Other food poisoning (bacterial)	1794	1.3016	0.370	0.381
008	Intestinal infections due to other organisms	3671	2.1370	0.287	0.382
009	Ill-defined intestinal infections	572	2.3794	0.391	0.661
034	Streptococcal sore throat and scarlet fever	6871	1.0971	0.189	0.234
038	Septicemia	4669	5.1527	0.101	0.207
041	Bacterial infection in conditions classified elsewhere	2153	2.5235	0.319	0.602
053	Herpes zoster	1224	1.4232	0.418	0.532
054	Herpes simplex	804	1.7674	0.343	0.469
057	Other viral exanthemata	1334	1.0997	0.244	0.309
074	Specific diseases due to coxsackie virus	1542	1.0759	0.282	0.405
079	Viral and chlamydial infection in conditions classified elsewhere	18690	1.1546	0.296	0.395
110	Dermatophytosis	2586	1.1257	0.283	0.575
112	Candidiasis	3199	1.2813	0.624	0.780
133	Acariasis	1764	1.0884	0.251	0.320
153	Malignant neoplasm of colon	530	5.2792	0.109	0.205
649	Other conditions or status of the mother complicating pregnancy, childbirth	2043	2.4258	0.095	0.215
650	Normal delivery	10455	2.2477	0.045	0.115
651	Multiple gestation	578	3.3789	0.112	0.225
652	Malposition and malpresentation of fetus	2164	3.0231	0.047	0.117
654	Abnormality of organs and soft tissues of pelvis	9271	2.7025	0.059	0.118
655	Known or suspected fetal abnormality affecting management of mother	917	2.4089	0.265	0.325
656	Other fetal and placental problems affecting management of mother	2796	2.6692	0.042	0.136
658	Other problems associated with amniotic cavity and membranes	3823	3.4057	0.030	0.078
659	Other indications for care or intervention related to labor	6071	2.6655	0.045	0.108
660	Obstructed labor	959	2.9124	0.061	0.142
661	Abnormality of forces of labor	2432	2.9416	0.122	0.216
663	Umbilical cord complications during labor and delivery	3649	2.2710	0.049	0.138
664	Trauma to perineum and vulva during delivery	10292	2.1785	0.038	0.071
669	Other complications of labor and delivery not elsewhere classified	2016	2.9122	0.098	0.210
681	Cellulitis and abscess of finger and toe	2955	1.4115	0.428	0.547
682	Other cellulitis and abscess	39685	1.6068	0.419	0.451
684	Impetigo	1335	1.1363	0.313	0.382

162	Malignant neoplasm of trachea bronchus and lung	547	5.0494	0.136	0.262
174	Malignant neoplasm of female breast	960	2.0479	0.150	0.327
185	Malignant neoplasm of prostate	1698	1.8893	0.122	0.362
218	Uterine leiomyoma	2988	2.0576	0.159	0.235
244	Acquired hypothyroidism	870	2.5667	0.149	0.558
250	Diabetes mellitus	13397	2.3259	0.227	0.342
251	Other disorders of pancreatic internal secretion	1165	2.0077	0.282	0.504
272	Disorders of lipoid metabolism	2214	2.6500	0.061	0.373
274	Gout	1214	1.4185	0.454	0.759
276	Disorders of fluid electrolyte and acid-base balance	14732	2.1792	0.257	0.338
278	Overweight, obesity and other hyperalimentation	3499	2.3172	0.150	0.442
280	Iron deficiency anemias	840	2.5405	0.170	0.478
282	Hereditary hemolytic anemias	2911	2.5737	0.306	0.412
285	Other and unspecified anemias	4126	2.8323	0.178	0.350
287	Purpura and other hemorrhagic conditions	1078	2.6503	0.241	0.589
288	Diseases of white blood cells	1533	3.2857	0.212	0.442
289	Other diseases of blood and blood-forming organs	1422	1.8432	0.402	0.620
291	Alcohol-induced mental disorders	2552	3.2692	0.271	0.370
292	Drug-induced mental disorders	2091	2.2975	0.391	0.509
294	Persistent mental disorders due to conditions classified elsewhere	510	3.5137	0.436	0.560
295	Schizophrenic disorders	4978	6.5749	0.302	0.346
296	Episodic mood disorders	13858	4.8611	0.256	0.287
298	Other nonorganic psychoses	3691	3.9225	0.376	0.452
300	Anxiety, dissociative and somatoform disorders	13556	1.6713	0.409	0.474
303	Alcohol dependence syndrome	5750	1.6957	0.445	0.503
304	Drug dependence	775	4.7871	0.383	0.430
305	Nondependent abuse of drugs	17013	1.9214	0.321	0.419
307	Special symptoms or syndromes not elsewhere classified	2477	2.8292	0.674	0.708
309	Adjustment reaction	2536	2.7165	0.338	0.443
686	Other local infections of skin and subcutaneous tissue	1230	1.3846	0.371	0.693
691	Atopic dermatitis and related conditions	2141	1.2644	0.294	0.565
692	Contact dermatitis and other eczema	7949	1.1250	0.266	0.492
698	Pruritus and related conditions	1448	1.1250	0.378	0.663
704	Diseases of hair and hair follicles	1188	1.1052	0.338	0.462
708	Urticaria	5151	1.0641	0.181	0.248
709	Other disorders of skin and subcutaneous tissue	2914	1.2172	0.237	0.379
715	Osteoarthrosis and allied disorders	6552	2.3288	0.175	0.307
716	Other and unspecified arthropathies	2080	1.3620	0.365	0.653
719	Other and unspecified disorders of joint	36318	1.1642	0.244	0.355
721	Spondylosis and allied disorders	1739	1.8735	0.142	0.277
722	Intervertebral disc disorders	5082	1.8851	0.127	0.201
723	Other disorders of cervical region	11664	1.1526	0.210	0.293
724	Other and unspecified disorders of back	42330	1.1911	0.269	0.340
726	Peripheral enthesopathies and allied syndromes	2118	1.2899	0.426	0.505
727	Other disorders of synovium tendon and bursa	1712	1.3271	0.408	0.597
728	Disorders of muscle ligament and fascia	2460	1.9459	0.313	0.569
729	Other disorders of soft tissues	42639	1.2080	0.337	0.397
733	Other disorders of bone and cartilage	3514	1.5677	0.316	0.505
737	Curvature of spine	591	4.4196	0.278	0.409
745	Bulbus cordis anomalies and anomalies of cardiac septal closure	1048	5.7424	0.194	0.330
765	Disorders relating to short gestation and low birthweight	2635	6.4159	0.174	0.294
766	Disorders relating to long gestation and high birthweight	1173	2.4859	0.062	0.132
770	Other respiratory conditions of fetus and newborn	1637	5.4319	0.197	0.360
774	Other perinatal jaundice	2953	2.5645	0.158	0.291
778	Integument and temperature regulation of fetus and newborn	595	2.7193	0.144	0.506
779	Other and ill-defined conditions originating in the perinatal period	2718	3.2826	0.239	0.377
780	General symptoms	124364	1.4193	0.317	0.364
781	Symptoms involving nervous and musculoskeletal systems	2793	1.7333	0.245	0.575

311	Depressive disorder, not elsewhere classified	8074	2.7524	0.363	0.419
312	Disturbance of conduct not elsewhere classified	1483	2.1908	0.458	0.547
314	Hyperkinetic syndrome of childhood	532	4.2519	0.418	0.513
338	Pain, not elsewhere classified	3415	1.5356	0.310	0.573
345	Epilepsy and recurrent seizures	6650	2.0329	0.242	0.332
346	Migraine	9547	1.2151	0.315	0.467
348	Other conditions of brain	1183	3.4725	0.176	0.486
351	Facial nerve disorders	1036	1.1931	0.458	0.701
362	Other retinal disorders	1300	1.2500	0.435	0.829
366	Cataract	527	1.1006	0.543	0.724
368	Visual disturbances	1649	1.3869	0.301	0.555
372	Disorders of conjunctiva	9741	1.0813	0.135	0.234
373	Inflammation of eyelids	2340	1.1440	0.371	0.457
379	Other disorders of eye	8543	1.0912	0.266	0.367
380	Disorders of external ear	4340	1.1187	0.295	0.483
381	Nonsuppurative otitis media and eustachian tube disorders	555	1.1135	0.356	0.404
382	Suppurative and unspecified otitis media	25208	1.0948	0.171	0.210
386	Vertiginous syndromes and other disorders of vestibular system	753	1.3997	0.347	0.420
388	Other disorders of ear	11850	1.1097	0.164	0.198
401	Essential hypertension	16603	1.8777	0.260	0.394
403	Hypertensive chronic kidney disease	512	3.4785	0.230	0.401
410	Acute myocardial infarction	7737	3.4349	0.036	0.092
411	Other acute and subacute forms of ischemic heart disease	1413	2.5718	0.075	0.174
414	Other forms of chronic ischemic heart disease	6974	2.9604	0.030	0.096
415	Acute pulmonary heart disease	1911	4.2575	0.121	0.303
424	Other diseases of endocardium	521	5.3647	0.129	0.303
427	Cardiac dysrhythmias	11823	2.7515	0.149	0.234
428	Heart failure	6959	3.8817	0.133	0.197
433	Occlusion and stenosis of precerebral arteries	1909	1.9460	0.266	0.391
434	Occlusion of cerebral arteries	3384	3.3502	0.085	0.252
435	Transient cerebral ischemia	2381	2.1277	0.115	0.285

782	Symptoms involving skin and other integumentary tissue	30349	1.2404	0.292	0.399
783	Symptoms concerning nutrition metabolism and development	1724	3.6990	0.230	0.407
784	Symptoms involving head and neck	56072	1.1691	0.316	0.384
785	Symptoms involving cardiovascular system	9361	1.4525	0.319	0.490
786	Symptoms involving respiratory system and other chest symptoms	137777	1.4710	0.317	0.359
787	Symptoms involving digestive system	54882	1.3368	0.346	0.393
788	Symptoms involving urinary system	13048	1.1914	0.296	0.516
789	Other symptoms involving abdomen and pelvis	122241	1.3435	0.387	0.422
790	Nonspecific findings on examination of blood	4868	2.1781	0.328	0.463
794	Nonspecific abnormal results of function studies	559	2.8068	0.182	0.446
796	Other nonspecific abnormal findings	1296	1.4915	0.239	0.528
799	Other ill-defined and unknown causes of morbidity and mortality	3047	2.6754	0.201	0.428
802	Fracture of face bones	2943	1.5912	0.245	0.355
805	Fracture of vertebral column without mention of spinal cord injury	1867	2.6133	0.239	0.389
807	Fracture of rib(s) sternum larynx and trachea	1871	2.0208	0.257	0.509
810	Fracture of clavicle	1858	1.2374	0.391	0.572
812	Fracture of humerus	3943	1.3779	0.225	0.380
813	Fracture of radius and ulna	8166	1.2103	0.201	0.299
814	Fracture of carpal bone(s)	528	1.1553	0.226	0.310
815	Fracture of metacarpal bone(s)	1851	1.0929	0.195	0.256
816	Fracture of one or more phalanges of hand	3526	1.1075	0.116	0.239
821	Fracture of other and unspecified parts of femur	691	3.0564	0.241	0.399
823	Fracture of tibia and fibula	2644	2.1048	0.277	0.365
824	Fracture of ankle	4045	1.5773	0.241	0.359
825	Fracture of one or more tarsal and metatarsal bones	2123	1.3118	0.366	0.562
826	Fracture of one or more phalanges of foot	1345	1.0900	0.202	0.594
829	Fracture of unspecified bones	758	1.4024	0.323	0.466
831	Dislocation of shoulder	1255	1.0988	0.260	0.347
832	Dislocation of elbow	1994	1.0507	0.112	0.226
840	Sprains and strains of shoulder and upper arm	3944	1.0649	0.097	0.198
841	Sprains and strains of elbow and forearm	515	1.1262	0.462	0.850

440	Atherosclerosis	523	3.1243	0.202	0.339
441	Aortic aneurysm and dissection	678	3.2478	0.301	0.409
453	Other venous embolism and thrombosis	2341	3.0880	0.286	0.430
455	Hemorrhoids	2020	1.3619	0.352	0.598
458	Hypotension	1927	2.5817	0.231	0.484
461	Acute sinusitis	2857	1.1491	0.345	0.429
462	Acute pharyngitis	22449	1.0923	0.137	0.286
463	Acute tonsillitis	2499	1.1493	0.256	0.343
464	Acute laryngitis and tracheitis	5039	1.1861	0.192	0.354
465	Acute upper respiratory infections of multiple or unspecified sites	43320	1.0932	0.233	0.331
466	Acute bronchitis and bronchiolitis	22947	1.5328	0.326	0.365
473	Chronic sinusitis	4275	1.1539	0.313	0.550
477	Allergic rhinitis	1996	1.1728	0.284	0.488
478	Other diseases of upper respiratory tract	4415	1.3905	0.260	0.483
482	Other bacterial pneumonia	1392	3.8111	0.239	0.371
486	Pneumonia, organism unspecified	17078	2.5750	0.319	0.371
487	Influenza	6241	1.3259	0.380	0.535
490	Bronchitis, not specified as acute or chronic	7927	1.1629	0.349	0.412
491	Chronic bronchitis	6218	3.0317	0.276	0.322
493	Asthma	28397	1.5484	0.324	0.390
496	Chronic airway obstruction, not elsewhere classified	2295	2.6078	0.273	0.382
511	Pleurisy	1825	3.0104	0.386	0.508
512	Pneumothorax and air leak	588	4.6497	0.119	0.267
518	Other diseases of lung	3533	4.7220	0.137	0.269
519	Other diseases of respiratory system	1636	1.5348	0.258	0.558
521	Diseases of hard tissues of teeth	3880	1.0907	0.333	0.855
522	Diseases of pulp and periapical tissues	5741	1.1312	0.419	0.471
525	Other diseases and conditions of the teeth and supporting structures	13928	1.0556	0.097	0.177
528	Oral soft tissues excluding lesions specific for gingiva and tongue	3220	1.2891	0.346	0.501
530	Diseases of esophagus	6458	1.8741	0.239	0.405
535	Gastritis and duodenitis	6808	1.3819	0.397	0.544
536	Disorders of function of stomach	1549	2.0575	0.348	0.469

842	Sprains and strains of wrist and hand	6249	1.0570	0.110	0.128
843	Sprains and strains of hip and thigh	1347	1.0995	0.228	0.308
844	Sprains and strains of knee and leg	5855	1.0828	0.149	0.199
845	Sprains and strains of ankle and foot	14357	1.0499	0.058	0.198
846	Sprains and strains of sacroiliac region	518	1.1293	0.168	0.238
847	Sprains and strains of other and unspecified parts of back	22641	1.0509	0.079	0.157
848	Other and ill-defined sprains and strains	2903	1.0568	0.242	0.419
850	Concussion	6848	1.1513	0.193	0.245
852	Subarachnoid subdural and extradural hemorrhage following injury	594	3.3754	0.107	0.331
870	Open wound of ocular adnexa	560	1.0911	0.312	0.542
872	Open wound of ear	593	1.0742	0.237	0.282
873	Other open wound of head	31786	1.0788	0.174	0.250
879	Open wound of other and unspecified sites except limbs	2870	1.4160	0.357	0.522
881	Open wound of elbow forearm and wrist	4093	1.1481	0.266	0.486
882	Open wound of hand except finger(s) alone	6850	1.0778	0.190	0.265
883	Open wound of finger(s)	14455	1.0706	0.153	0.187
884	Multiple and unspecified open wound of upper limb	984	1.1047	0.242	0.310
890	Open wound of hip and thigh	641	1.2559	0.346	0.575
891	Open wound of knee leg (except thigh) and ankle	5015	1.1741	0.279	0.357
892	Open wound of foot except toe(s) alone	3579	1.1260	0.264	0.583
893	Open wound of toe(s)	1258	1.0588	0.148	0.259
910	Superficial injury of face neck and scalp except eye	4553	1.0995	0.279	0.525
911	Superficial injury of trunk	1488	1.1559	0.240	0.721
913	Superficial injury of elbow forearm and wrist	1624	1.1305	0.232	0.805
916	Superficial injury of hip thigh leg and ankle	2574	1.1076	0.269	0.362
917	Superficial injury of foot and toe(s)	575	1.0974	0.217	0.306
918	Superficial injury of eye and adnexa	3266	1.0582	0.134	0.515
919	Superficial injury of other multiple and unspecified sites	6477	1.0871	0.300	0.464
920	Contusion of face, scalp, and neck except eye(s)	11273	1.0931	0.151	0.296
921	Contusion of eye and adnexa	625	1.1488	0.505	0.566
922	Contusion of trunk	7317	1.1152	0.148	0.266
923	Contusion of upper limb	10982	1.0533	0.112	0.195

540	Acute appendicitis	5994	2.8223	0.035	0.107
541	Appendicitis, unqualified	606	2.7508	0.079	0.187
550	Inguinal hernia	1243	1.5414	0.226	0.339
553	Other hernia of abdominal cavity without mention of obstruction or gangrene	2398	2.4854	0.245	0.368
555	Regional enteritis	1047	3.5244	0.232	0.376
558	Other and unspecified noninfectious gastroenteritis and colitis	20523	1.2614	0.365	0.443
560	Intestinal obstruction without mention of hernia	4558	4.1057	0.136	0.196
562	Diverticula of intestine	5127	3.0971	0.256	0.309
564	Functional digestive disorders not elsewhere classified	13421	1.2843	0.362	0.514
566	Abscess of anal and rectal regions	593	1.9224	0.316	0.423
569	Other disorders of intestine	5600	1.9546	0.368	0.434
571	Chronic liver disease and cirrhosis	763	4.3172	0.190	0.302
574	Cholelithiasis	6227	2.5405	0.267	0.325
575	Other disorders of gallbladder	1741	3.0689	0.136	0.265
577	Diseases of pancreas	4637	3.9198	0.145	0.213
578	Gastrointestinal hemorrhage	5587	2.8341	0.260	0.346
584	Acute kidney failure	3318	4.5069	0.070	0.177
585	Chronic kidney disease (ckd)	1196	3.6455	0.214	0.410
590	Infections of kidney	4924	1.9175	0.400	0.473
592	Calculus of kidney and ureter	8884	1.3540	0.256	0.415
593	Other disorders of kidney and ureter	2005	2.5416	0.222	0.419
595	Cystitis	1471	1.1924	0.408	0.562
597	Urethritis not sexually transmitted and urethral syndrome	1328	1.0497	0.238	0.264
599	Other disorders of urethra and urinary tract	29431	1.4133	0.379	0.472
600	Hyperplasia of prostate	853	2.1290	0.176	0.370
604	Orchitis and epididymitis	1310	1.2382	0.533	0.627
605	Redundant prepuce and phimosis	1057	2.1116	0.246	0.548
607	Disorders of penis	1662	1.2196	0.248	0.651

924	Contusion of lower limb and of other and unspecified sites	15534	1.0811	0.233	0.565
930	Foreign body on external eye	1287	1.0645	0.063	0.092
931	Foreign body in ear	1983	1.0756	0.153	0.173
932	Foreign body in nose	1439	1.0737	0.110	0.198
935	Foreign body in mouth esophagus and stomach	1371	1.1758	0.159	0.420
938	Foreign body in digestive system, unspecified	1641	1.1395	0.254	0.469
941	Burn of face head and neck	523	1.8260	0.248	0.484
943	Burn of upper limb except wrist and hand	501	1.6946	0.387	0.453
944	Burn of wrist(s) and hand(s)	2099	1.3149	0.385	0.501
945	Burn of lower limb(s)	1128	1.9025	0.419	0.598
949	Burn unspecified site	1233	1.5750	0.322	0.377
959	Injury other and unspecified	57468	1.1375	0.225	0.306
965	Poisoning by analgesics antipyretics and antirheumatics	1680	1.8268	0.261	0.456
969	Poisoning by psychotropic agents	1228	1.9992	0.234	0.408
977	Poisoning by other and unspecified drugs and medicinal substances	1602	1.4313	0.324	0.434
989	Toxic effect of other substances chiefly nonmedicinal as to source	4021	1.1450	0.285	0.420
995	Certain adverse effects not elsewhere classified	12151	1.3123	0.333	0.547
996	Complications peculiar to certain specified procedures	4480	2.9498	0.209	0.302
997	Complications affecting specified body system not elsewhere classified	1616	3.3923	0.198	0.357
998	Other complications of procedures not elsewhere classified	4459	2.3290	0.317	0.431
E81	Motor vehicle traffic accident	3118	1.2017	0.381	0.532
E84	Vehicle accidents not elsewhere classifiable	531	1.1921	0.553	0.880
E88	Accidental fall	3438	1.3048	0.274	0.414
E90	Accident due to weather	1186	1.2057	0.397	0.747
E91	Accidents caused by submersion, suffocation, and foreign bodies	570	1.2035	0.388	0.510
E92	Late effects of accidental injury	580	1.3259	0.250	0.473
E96	Homicide and injury purposely inflicted by other persons	687	1.9563	0.285	0.544
V01	Contact with or exposure to communicable diseases	1500	1.1520	0.458	0.583

608	Other disorders of male genital organs	3254	1.3190	0.334	0.406
611	Other disorders of breast	2009	1.4191	0.251	0.396
614	Inflammatory disease of ovary fallopian tube pelvic cellular tissue and peritoneum	1576	1.8433	0.383	0.588
616	Inflammatory disease of cervix vagina and vulva	3966	1.1377	0.386	0.444
617	Endometriosis	561	2.0321	0.189	0.270
618	Genital prolapse	1274	1.5785	0.032	0.139
620	Noninflammatory disorders of ovary fallopian tube and broad ligament	3658	1.3857	0.266	0.317
623	Noninflammatory disorders of vagina	9226	1.0942	0.135	0.247
625	Pain and other symptoms associated with female genital organs	7314	1.2177	0.269	0.485
626	Disorders of menstruation and other abnormal bleeding from female genital tract	3548	1.4594	0.263	0.355
632	Missed abortion	550	1.3927	0.198	0.277
633	Ectopic pregnancy	769	1.5475	0.166	0.220
634	Spontaneous abortion	1497	1.1643	0.162	0.269
637	Unspecified abortion	520	1.1327	0.332	0.699
640	Hemorrhage in early pregnancy	6315	1.0515	0.135	0.221
641	Antepartum hemorrhage abruptio placentae and placenta previa	1262	2.6006	0.163	0.190
642	Hypertension complicating pregnancy childbirth and the puerperium	4815	3.4636	0.070	0.155
643	Excessive vomiting in pregnancy	2595	1.3145	0.392	0.442
644	Early or threatened labor	5106	2.5805	0.112	0.176
645	Late pregnancy	5648	2.6084	0.070	0.149
646	Other complications of pregnancy not elsewhere classified	6245	1.3894	0.313	0.366
647	Infective and parasitic conditions in the mother classifiable	689	2.3237	0.210	0.567
648	Other current conditions in the mother classifiable elsewhere	15413	1.9507	0.257	0.303

V15	Other personal history presenting hazards to health	1755	1.6171	0.404	0.725
V20	Health supervision of infant or child	1355	1.7063	0.565	0.650
V22	Normal pregnancy	20230	2.4254	0.116	0.190
V23	Supervision of high-risk pregnancy	1348	2.9206	0.075	0.243
V24	Postpartum care and examination	2278	2.7428	0.055	0.162
V27	Outcome of delivery	1322	2.3215	0.123	0.356
V30	Single liveborn	103027	2.5700	0.022	0.069
V31	Twin birth mate liveborn	2630	5.9167	0.061	0.121
V39	Liveborn unspecified whether single twin or multiple	1598	2.3949	0.098	0.340
V45	Other postprocedural states	1910	2.8215	0.110	0.312
V55	Attention to artificial openings	1153	2.9809	0.407	0.565
V57	Care involving use of rehabilitation procedures	1890	10.214	0.057	0.193
V58	Encounter for other and unspecified procedures and aftercare	9968	1.5058	0.349	0.532
V62	Other psychosocial circumstances	3227	2.5696	0.394	0.453
V64	Persons encountering health services for specific procedures not carried out	1798	1.1107	0.140	0.223
V65	Other persons seeking consultation	1581	1.2739	0.096	0.205
V67	Follow-up examination	1144	1.3706	0.341	0.589
V68	Encounters for administrative purposes	5862	1.1682	0.164	0.227
V70	General medical examination	3294	1.8352	0.360	0.513
V71	Observation and evaluation for suspected conditions not found	8451	1.1619	0.241	0.501
V72	Special investigations and examinations	770	2.0792	0.246	0.583
V82	Special screening for other conditions	1417	1.0607	0.120	0.202

VITA

Pankush Kalgotra

Candidate for the Degree of

Doctor of Philosophy

Thesis: IMPACT OF INFERRED COMORBIDITY NETWORKS ON HEALTH OUTCOMES

Major Field: Management Science and Information Systems

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Management Science and Information Systems at Oklahoma State University, Stillwater, Oklahoma in July, 2017.

Completed the requirements for the Master of Science in Management Information Systems at Oklahoma State University, Stillwater, Oklahoma in May, 2013.

Completed the requirements for the Bachelor of Technology in Information Technology at National Institute of Technology, Raipur, Chhattisgarh, India in 2011.