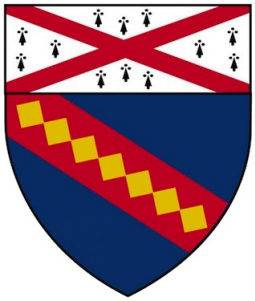
Human Disease Network: A Study Based on Taiwan National Health Insurance Research Database

Xinchun Liu

Thesis submitted to the faculty of the
Yale School of Public Health
in partial fulfillment of the requirements for the degree of
Master of Public Health In Biostatistics



Thesis Advisor: Professor Shuangge Ma

Second Reader: Professor Maria M. Ciarleglio

April 5th, 2017

Acknowledgements

Foremost, I would like to express my deepest gratitude to my thesis advisor: Professor Shuangge Ma for the insights, advices, and guidance he has provided in the creation of this thesis. I could not have imagined having a better advisor and mentor for my master study. I am forever grateful for the support and help he has given to me during my journey at the Yale School of Public Health. His knowledge, patience, and passion added considerably to my experience.

Besides my advisor, I have been very fortunate to have worked with Professor Maria M. Ciarleglio, for her encouragement, insightful comments, and hard questions during completion of this thesis.

My sincere thanks also goes to my fellow labmates Yefei Jiang in Taiwan Fu Jen Catholic University for sharing access to the database and working with me on this stimulating study.

I want to thank all of my families and friends, who always inspire, energize, and embolden me in completing this thesis.

Last but not the least, I would like to thank my parents and my brother for the love they have provided through my entire life.

Abstract

Many of the existing resources from genetic perspective have been constructed to help understand the origins of many diseases. While progress on genetic fronts has been impressive, many of these sources overlook the information we could get from analyzing patient clinical histories. Our primary goal here is to define the human disease network using epidemiological data from a population science perspective and detect pairwise co-morbidity correlations.

Network analysis is increasingly used to explore the co-occurrence of human diseases. We here employed a network analysis approach, called weighted correlation network analysis (WGCNA), to find significant associations among over 600 diseases in over 800,000 patients from Taiwan National Health Insurance Research Database through year 2000 to 2013. The concepts of network construction is straightforward: nodes represent diseases and edges represent the connections between nodes. Nodes are connected if the corresponding diseases has the high possibility of being comorbid. We detect co-morbidity of disease by (1) calculating pairwise connections; (2) identifying diseases with high comorbidity rate; (3) exploring diseases clustering and its changing pattern over fourteen years.

Our findings show that Diabetes mellitus, Disorders of lipoid metabolism and eight other diseases have the highest possibility to co-occur with other diseases. And these ten diseases constantly have high risk of being comorbid with other diseases. Our clustering results show diseases within same category tends to occur together. In addition, diseases share same risk factors but in different categories may also appear simultaneously.

Contents

1 Introduction

Over the past half-decade, studies have been conducted to examine the disease associations based on genes, proteins, and expression patterns [1]. For example, by connecting diseases that have been presented with same genes, Goh et al. created a network of Mendelian gene-disease associations [2]. Their results suggested that disease genes play a central role in the human interactome. Network analysis also has been used in building protein interaction network, like the ones created by Rual et al. and Stelzl et al. The work by Rual et al. described an initial version of a proteome-scale map of human binary protein-protein interactions and revealed more than 300 new connections to over 100 diseases associated with proteins [3]. And the work by Stelzl et al. detected the interacting pairs of human proteins via building a protein-protein interaction maps [4]. Two novel Axin-1 interactions were validated in their study, the results characterizing ANP32A (i.e. Acidic Nuclear Phosphoprotein 32 Family Member A) and CRMP1 (i.e. Collapsin Response Mediator Protein 1) as modulators of Wnt signaling. Both the works did by Rual et al. and Stelzl et al. represent an important step towards a systematic human disease network project. Moreover, in the gene expression front, several studies have used microarray expression profiles and other cellular information to explore networks in brain disease, inflammation and breast cancer. Pujana et al. used a network modeling strategy to identify genes potentially associated with higher risk of breast cancer, and two case-control studies indicated that HMMR locus is associated with higher risk of breast cancer in humans [5]. The analysis conducted by Calvano et al. reveals that, upon acute systemic inflammation, the human blood leukocyte response includes widespread suppression at the transcriptional level of mitochondrial energy production and protein synthesis machinery [6]. Via examining the large-scale organization of gene coexpression networks in human and chimpanzee brains, Oldham et al. provided an integrated view of human brain evolution. In addition, they found module conservation in cerebral cortex is significantly weaker than module conservation in sub-cortical brain regions [7]. Several studies on network analysis methodology provides the scientific references. [8, 9, 10, 11].

The progress in genetic and proteomic fronts of disease studies has been impressive, however, much of the available resources overlook the fact

that we have extensive and continually updated patient clinical histories. In other words, studies based on genetic information may fail to provide an updated description of disease associations. With the consolidation of electronic medical record (EMR) systems in modern healthcare, massive amounts of clinical data and phenotype data are gradually becoming available for researches [12, 13, 14, 15]. The EMR allows clinicians to maintain a problem summary list for each patient to provide a concise overview of significant medical issues and diagnoses. The studies on building diseases network from population-based data start emerging. However, they have often focused on one specific disease. For example, the study conducted by Prather et al. was using the techniques of data mining to search for relationships between several factors and preterm birth in a large clinical database [16]. Hanauer et al. detected significant associations among diagnoses that were not described in the literature before using electronic medical records from hundreds of physicians [17]. In this study, we aim to use patient claim data to provide a systematic overview on human disease associations.

Typically, we say a comorbidity relationship exists between two diseases whenever they affect the same individual substantially more than just chance. With the aim of making available pairwise comorbidity correlations for more than 600 diseases for fourteen years from over 800,000 medical records, this study will build comprehensive and systematic networks for human diseases using population-based datasets. Studying the structure of comorbidities might help understand many clinical questions from a perspective that is complementary to other approaches. In network analysis, a node corresponds to one specific disease and nodes are connected if they have significant pairwise associations. The strength of such connection is measured by edges. In this study, we obtain pairwise disease correlations and capture all diseases in a human disease network. Additionally, we illustrate both diseases within same category and diseases in different categories but share common risk factors tend to co-occur. Finally, with longitudinal data available, we show diseases with significant correlation tend to have constantly high comorbidity rate. We organize the results into (1) detecting disease correlations (2) identifying disease clusters (3) tracking longitudinal trends. Visualization of study results is obtained through Gephi.

This study differs from existing literature in the following points. First, with regard to the total population size of Taiwan (i.e. 23 millions), the

sample size of 1 million for this study is large enough to provide reliable results for future researches. This analysis will provide a comprehensive overview on human disease network, both the severity of diseases and co-occurrence of diseases. Second, using ICD-9 list as the reference for diseases to analyze, our study could provide a systematic overview on population-based disease associations. Few data or research results are currently available in exploring relationships between all diseases at one time. Third, instead of detecting association from a genetic and proteomic perspective, our study will build a human disease network through epidemiological approach. The study results could open new opportunities for public health approaches to diseases. Fourth, with 14 years data available, this study could also define the changing patterns of diseases associations over years.

2 Methods

Network analysis is increasingly used to explore the system-level correlation pattern among diseases. This method focuses on understanding the "system" instead of reporting a list of individual diseases, hence, it replies more heavily on systematic network analysis and other data-mining techniques. Analyzing disease network increases confidence levels for individual interactions and provides a tremendously useful framework for recognizing new disease interactions [18]. Advanced network mapping models to investigate and analyze large and complex patients claim data will provide an accurate and detailed understanding of diseases associations, which will subsequently help health care professionals gain a better understanding of disease tracking and management.

In network construction, nodes represent diseases and nodes are connected if the corresponding diseases are significantly connected. The key steps in network analysis include: (1) measure concordance of diseases with a Pearson correlation; (2) encode the connection strength between each pair of diseases by transforming Pearson correlation into adjacency matrix; (3) identify clusters (modules) of highly correlated diseases. Details will be discussed below.

2.1 Source Data and Study Population

The National Health Insurance Research Database (NHIRD) in Taiwan offer reliable, systematic, and complete data for disease detection [19, 20, 21, 22]. The NHIRD, one of the largest administrative health care databases around the world, has been used widely in academia. The NHIRD studies expanded rapidly in both quantity and quality since the first study was published in 2000 [20]. Taiwan, a 35.8 thousand square kilometers island with more than 23 million people, launched its universal National Health Insurance program on March 1, 1995 and as of 2004, over 99 percent of Taiwan population were enrolled in this single payer system. Each month, Bureau of National Health Insurance (BNHI) collects and sorts several dozens of millions of health service claims from the National Health Insurance program, including registration files and original claim data for reimbursement [23]. After billing process, the National Health Research Institutes (NHRI) compiles data from these insurance claims for research purposes.

The number of NHIRD studies increased rapidly since the first study appeared in 2000. The average annual growth rate of NHIRD studies (i.e.45.8%) over the last decade was considerably higher than all PubMed literature growth rate (i.e.17%) between 2002 and 2006 [24]. Additionally, the NHIRD is widely used in 383 studies by 667 authors and the extensive publication in 210 journals cover 250 study fields by 2009. The numbers of authors, journals and study fields even doubled every 2 years [20]. Besides the rapid growth in quantity, the progress in high quality NHIRD studies is also remarkable. About 92.2% of the NHIRD studies were indexed in the Science Citation Index. Fernandex-Cano et al. revealed that the number of scientific literature normally doubled every 10 to 20 years, while the average time to double the number of SCI indexed NHIRD studies is only 2.49 years [25].

In the datasets we got from NHIRD, each record consists of the date of visit, inpatient hospital stay, number of treatments, costs of each treatment, and a primary diagnosis which is specified by three-digit ICD-9 codes. In total, the ICD-9-CM classification consists of 657 different categories (i.e. 657 diseases). For a detailed list of currently used ICD9 codes see www.icd9data.com. The datasets in our study are from a longitudinal observation study, containing the hospital claims from

2000 to 2013. The effective sample sizes of datasets from 2000 to 2013 are 809015, 826296, 843799, 858705, 888917, 904551, 883616, 873400, 862372, 860250, 855266, 856412, 849989, and 844187 respectively. With the collaborator in Taiwan, we have the access to NHIRD database. It will be cost-effective via using existing database.

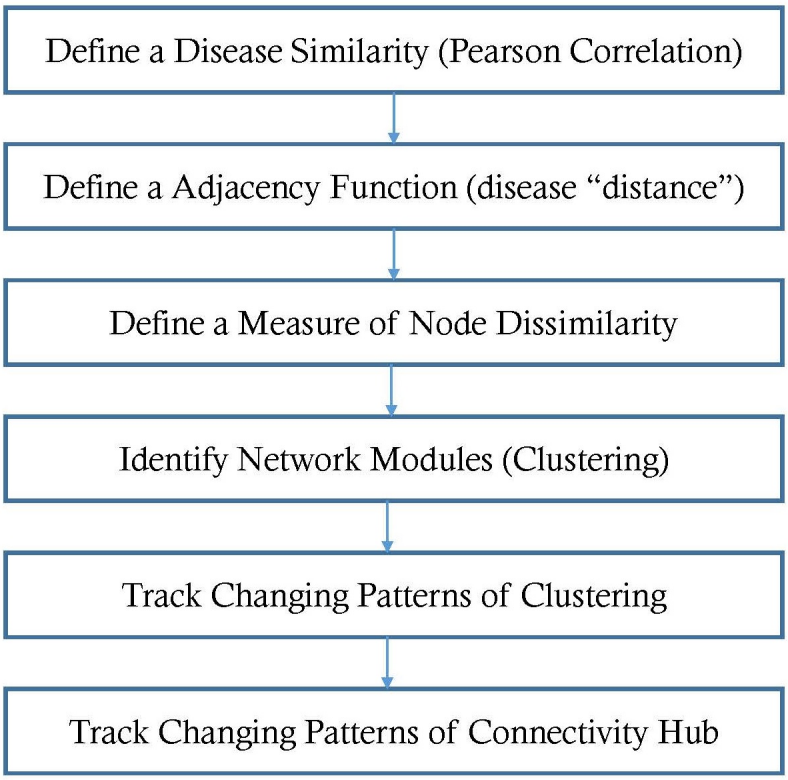


Figure 1: Flowchart and illustration of human disease network analysis

2.2 Steps of the Network Analysis

A flowchart for constructing a human disease network is presented in Figure 1. In this analysis, we will build thirteen networks each year from 2000 to 2013. Each network has its own pairwise disease correlation matrix and corresponding adjacency matrix. In our analysis (i.e. weighted networks), the adjacency matrix encodes the connection

strength between each pair of nodes. While in an unweighted network, the adjacency matrix is composed by binary entries (i.e. 1 or 0) indicating whether or not a pair of nodes is connected. In addition, disease modules are also identified in each year through hierarchical clustering. Finally, the changes in disease clustering over fourteen years are examined and explained, so does the diseases with high comorbidity rate.

2.2.1 Define a Disease Similarity

To begin with, we need to define a measure of similarity between diseases. We here quantify this strength of similarity by introducing a notion of "distance" between two diseases across the population in Taiwan. Specifically, for each pairs of diseases i and j, s_{ij} denotes this similarity measures, which is calculated by Pearson correlation ϕ. The ϕ-correlation, can be expressed mathematically as

$$\phi_{ij} = \frac{C_{ij}N - P_iP_j}{\sqrt{P_jP_j(N-P_i)(N-P_j)}} \tag{1}$$

where C_{ij} is the number of patients affected by both diseases, N is the total number of patients in the population and P_i and P_j are the cases of diseases i and j. We denote the similarity matrix by $S = [s_{ij}]$. The diagonal elements of similarity matrix are set to be one as they represent the correlation between diseases themselves. Note $\phi_{ij} = 1$ has a value from -1 to 1, where 1 is total positive linear correlation, 0 is no linear correlation, and 1 is total negative linear correlation between two diseases.

2.2.2 Define an Adjacency matrix

The $n \times n$ similarity matrix is then transformed into an $n \times n$ adjacency matrix $A = [a_{ij}]$, encoding the strength of correlation between pairs of diseases (i.e. distances between two diseases). The adjacency matrix is the foundation of all subsequent steps. As the networks considered here are undirected, A will be a symmetric matrix with non-negative entries (i.e. $a_{ij} \in [0, 1]$). In addition, the diagonal elements of adjacency are

set to be zero by convention. The adjacency function is a monotonically increasing function that maps from interval [0,1] into [0,1]. The choice of adjacency functions determines whether the resulting network will be weighted or unweighted.

The most widely used adjacency function is to identify a specific threshold parameter τ, which is also known as a unweighted adjacency matrix. To be more specific,

$$a_{ij} = f(S_{ij}, \tau) = \begin{cases} 1 & if \quad s_{ij} \geq \tau \\ 0 & if \quad s_{ij} \leq \tau \end{cases} \tag{2}$$

For instance, if we set τ to be 0.5, then the connection only exists for disease with greater than 0.5 similarity. As a result, these pairs of diseases with a greater than 0.5 similarity measures will be coded as 1 in corresponding adjacency matrix and the remaining disease will be coded as 0. An Unweighted adjacency matrix with binary entries leads to intuitive network concepts, but may also lead to losses of needed information.

To diminish the disadvantages of losing necessary information, we define the adjacency function as follows:

$$a_{ij} = f(S_{ij}, \tau) = \begin{cases} |s_{ij}| & if \quad s_{ij} \geq \tau \\ 0 & if \quad s_{ij} \leq \tau \end{cases} \tag{3}$$

The choice of the parameter τ determines the sensitivity and specificity of pairwise disease distances. Increasing the values of τ results in fewer node connections, which may reduce the noise in the network. However, if τ is defined too high the subsequent network will be too sparse for detecting the disease clustering. In this analysis, τ is defined as 0.05 for all network constructing. Hence, if the absolute value of one entry in the similarity matrix above is greater than 0.05, then its corresponding adjacency value will be the absolute value of its similarity value. Otherwise, the adjacency value will be zero.

2.2.3 Define a Measure of Node Dissimilarity

An important goal of this human disease network analysis is to detect subsets of nodes (modules) that are tightly connected to each other. It is important to mention that authors differ on how they define modules in

network. The existing methods could be classified into three groups. The first group of methods identified modules via three common steps. First, network nodes and edges are annotated with scores quantifying disease connection [26, 27]. Next, a scoring function is formulated to compute an aggregate score for each subnetwork. Subsequently, a search strategy is used to identify subnetworks with high scores, which are corresponding disease modules. The second group of methods emulated the related concepts of diffusion flow and network propagation [28, 29, 30]. The last group of methods employs clustering of network interactions. Clustering based on network connectivity has proven to be instrumental in defining principles of network modules [31, 32, 33]. Here we apply the method used by Zhang and Horvath [8], which uses average linkage hierarchical clustering coupled with the topological overlap dissimilarity measure [33]. A dendrogram will be obtained through hierarchical clustering for each network, we will then choose a height cutoff to arrive at clusters and branches of the dendrogram are the corresponding modules we need.

The topological overlap of two nodes represents their relative inter-connection. The topological overlap matrix (TOM) $\Omega = [\omega_{ij}]$ provides a similarity measure (opposite of dissimilarity), which has been proved to be useful in biological networks [33, 34]. The topological overlap dissimilarity is used as input of hierarchical clustering. And the formula is defined as follows:

$$TOM_{ij} = \omega_{ij} = \frac{l_{ij} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}} \tag{4}$$

$$DistTOM_{ij} = 1 - TOM_{ij} \tag{5}$$

where $l_{ij} = \sum_u a_{iu} a_{uj}$, and $k_i = \sum_u a_{iu}$ is the node connectivity, see equation (6) below. Note $\omega = 1$ if the node satisfies: (a) all of its neighbors are also neighbors of the other node and (b) it is connected to the other node. By contrast, $\omega = 0$ if i and j are unconnected and the two nodes do not share any neighbors. Note in applying this topological overlap dissimlarity to unweighted network, l_{ij} equals the number of connected pair of nodes. Since $l_{ij} \leq min(\sum_{u \neq j} a_{iu}, \sum_{u \neq i} a_{uj})$, it follows that $l_{ij} \leq min(\sum(k_i, k_j) - a_{ij}$. Therefore, $0 \leq a_{ij} \leq 1$ implies $0 \leq \omega_{ij} \leq 1$. The topological overlap matrix $\Omega = [\omega_{ij}]$ is therefore non-negative and symmetric.

2.2.4 Identify Network Clusters

Researchers on network analysis differ on how they define modules. Again, we adopt the definition from Zhang and Horvath that disease modules are groups of diseases highly correlated across the sample patients. We will use the method from Ravaszx et al [33]: modules are groups of nodes with high topological overlap to identify the clusters.

To group diseases with high correlation into modules, we use average linkage hierarchical clustering coupled with TOM-based dissimilarity. Disease modules correspond to branches of the hierarchical tree (known as dendrogram). We need to choose a height cutoff to cut branches off the tree. A common but inflexible method uses a constant height cutoff value, which is called static height cutoff. Large height values lead to big modules, while as small values lead to small but tight modules. The static height cut-off method works quite well at retrieving the true modules, but it misses lots of diseases at the fringes of the modules (e.g. the surrounding diseases in figure 7A) [35]. In other words, the static method has high specificity but low sensitivity. As the aim of this study is to get an overview on human disease co-morbidity, we would adopt static method and set 0.99 as the cutoff point. The resulting branches on color row in the dendrogram represent the disease modules.

The topological overlap matrix plots (known as heatmap) provides a 'reduced' view of the network and allows us to visualize and identify network modules. The TOM plot is a color-coded depiction of the values of the TOM-based dissimilarity. The rows and columns are sorted by the hierarchical clustering tree and each color corresponds to a single disease. An example heatmap plot is shown in Figure 2. In the heatmap, light colors represent low topological overlap while progressively darker orange and red colors correspond higher topological overlap. The corresponding disease dendrograms and module assignments are shown on the left and top of heatmap. Since the topological overlap matrix is symmetric, the heatmap as a result is symmetric as well. The modules are sets of nodes with high topological overlap, and the red squares along the diagonal are corresponding modules (i.e. clusters).

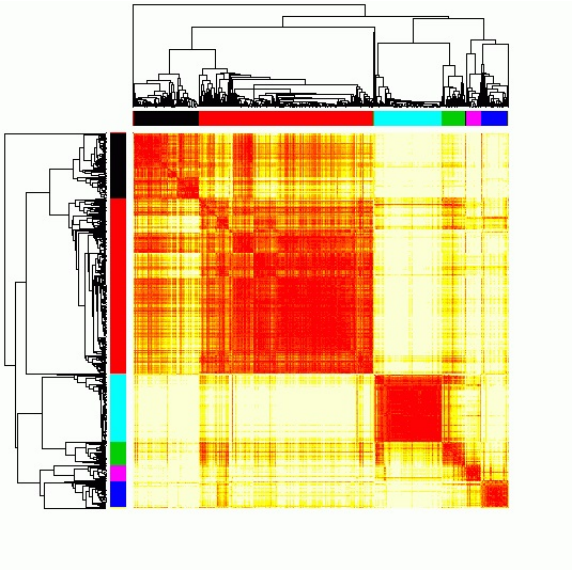


Figure 2: An example of Topological Overlap matrix plot

2.2.5 Track Changing Patterns of Clusters

Since we have data for all fourteen years, we are capable of identifying the network clusters each year. The variations in clusters diagnosis for adjacent years are tracked as well. We record changing patterns of clusters in an $n \times m$ matrix $A = a_{i,j}$. The diagonal values represent the number of diseases stay within the same cluster for both years, while off-diagonal values interpret the number of disease switch from one cluster to the other. An example is shown in Table 1.

		2000				
		1	2	3	4	5
2001	1	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{1,5}$
	2	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{2,4}$	$a_{2,5}$
	3	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{3,4}$	$a_{3,5}$
	4	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$a_{4,4}$	$a_{4,5}$
	5	$a_{5,1}$	$a_{5,2}$	$a_{5,3}$	$a_{5,4}$	$a_{5,5}$

Table 1: An example of cluster pattern tracking

Assume we are tracking the cluster patterns between year 2000 and 2001. And we also hypothesize here that the number of clusters in 2000 and 2001 are both five. So the row will be the clustering information in year 2000 and columns will be the clustering information in year 2001. Hence, $a_{1,1}$ identifies the total number of diseases keep staying in cluster 1 in both year 2000 and 2001. For off-diagonal values, take $a_{2,1}$ as an example, it stands for the total number of diseases classified in cluster 1 in year 2000, but switch to cluster 2 in year 2001.

2.3 Connectivity

2.3.1 Standard Connectivity

For each disease, the connectivity is defined as the sum of connection strengths with the other network genes. In unweighted networks, the connectivity K_i of node i equals the number of its direct connections to other nodes. This could be expressed as follows:

$$K_i = \sum_{j=1}^{n} a_{ij} \qquad i \neq j \tag{6}$$

where a_{ij} are the adjacency matrix defined above. Naturally, we could also use this formula to define the connectivity of node i in weighted networks, which measures how correlated a disease is with all other network diseases. Intuitively, the connectivity measures in a network are the row sums of adjacency matrix.

2.3.2 A TOM-based Connectivity Measure

A key concept of network analysis is node connectivity, also known as centrality. The standard connectivity measure is proposed as equation 6, but many alternatives are available. We here referred to TOM-based measure of connectivity ω_i from Zhang and Horvath:

$$\omega_i = \sum_{j=1}^{n} \omega_{ij} \tag{7}$$

where ω_{ij} is the topological overlap between two nodes i and j. Thus, a node has high TOM-based connectivity ω_i if it has high overlap with many other nodes. A central node (i.e. connectivity hub) is one with many connections to other nodes. In other words, the ones with highest connectivity values are referred as connectivity hubs. The connectivity measures are used to determine the sizes of nodes in our network building also. The higher the connectivity values, the larger the node i, and the greater the comorbidity rate of other diseases will be.

Between year 2002 and 2013, we select ten nodes with highest connectivity and track the changes in these connectivity values over these eleven years. We aim to detect whether hubs have consistently high connectivity values or not.

3 Results

3.1 Source Data and Summary statistics

With the possibility of inaccurate data collection and data entry, there are few diseases not recognized by ICD-9 codes. After excluding these mis-coded ICD-9 diseases, the effective number of diseases from 2000 to 2013 are 642, 642, 640, 642, 638, 632, 629, 625, 627, 623, 626, 625, 623, 624. In order to track variations in clustering over fourteen years, we keep only common diseases occurred every year and the total number is 611 in this study. A disease will be excluded if it is not identified in any given year. There is only one variable used in this study − primary diagnosis, which is a binary variable indicating whether the patient is diagnosed with that disease or not.

The prevalence of 611 diseases are calculated for each year. The top ten diseases with highest prevalence and their connectivity values over fourteen years are listed in Table 2. Most of these diseases are classified into diseases of digestive system and disease of respiratory system, which are common diseases in daily life. The summary statistics and scatter plot for these ten diseases are calculated in Table 3 and Figure 3 also, which provides an overview of connectivity variations between 2000 and 2013.

		2000	2001	2002	2003	2004	2005	2006
465	Acute upper respiratory infections	0.642	0.619	0.601	0.575	0.573	0.566	0.493
523	Gingival and periodontal diseases	0.233	0.244	0.250	0.251	0.273	0.280	0.295
521	Diseases of hard tissues of teeth	0.274	0.283	0.282	0.277	0.289	0.285	0.293
466	Acute bronchitis and bronchiolitis	0.247	0.253	0.255	0.247	0.245	0.263	0.235
372	Disorders of conjunctiva	0.217	0.208	0.236	0.205	0.217	0.216	0.212
460	Acute nasopharyngitis [common cold]	0.247	0.224	0.241	0.232	0.236	0.228	0.197
461	Acute sinusitis	0.143	0.165	0.168	0.166	0.168	0.185	0.179
463	Acute tonsillitis	0.178	0.177	0.177	0.163	0.164	0.180	0.182
692	Contact dermatitis and other eczema	0.128	0.134	0.137	0.141	0.150	0.154	0.150
558	Other and unspecified noninfectious gastroenteritis and colitis	0.140	0.136	0.152	0.131	0.151	0.137	0.158

		2007	2008	2009	2010	2011	2012	2013
465	Acute upper respiratory infections	0.488	0.463	0.473	0.462	0.491	0.460	0.438
523	Gingival and periodontal diseases	0.312	0.328	0.343	0.354	0.358	0.373	0.387
521	Diseases of hard tissues of teeth	0.299	0.309	0.313	0.310	0.304	0.304	0.304
466	Acute bronchitis and bronchiolitis	0.244	0.234	0.238	0.232	0.267	0.244	0.229
372	Disorders of conjunctiva	0.219	0.215	0.214	0.226	0.218	0.215	0.218
460	Acute nasopharyngitis [common cold]	0.201	0.194	0.197	0.197	0.212	0.199	0.194
461	Acute sinusitis	0.187	0.185	0.187	0.192	0.211	0.198	0.197
463	Acute tonsillitis	0.189	0.181	0.185	0.179	0.200	0.180	0.178
692	Contact dermatitis and other eczema	0.150	0.154	0.157	0.160	0.162	0.167	0.173
558	Other and unspecified noninfectious gastroenteritis and colitis	0.143	0.134	0.129	0.145	0.148	0.154	0.131

Table 2: Prevalence of Top Ten Diseases

ICD9	Disease Name	Mean	Median	Min	Max	STD
465	Acute upper respiratory infections	0.525	0.492	0.438	0.642	0.066
523	Gingival and periodontal diseases	0.306	0.303	0.233	0.387	0.050
521	Diseases of hard tissues of teeth	0.295	0.296	0.274	0.313	0.012
466	Acute bronchitis and bronchiolitis	0.245	0.244	0.229	0.267	0.011
372	Disorders of conjunctiva	0.217	0.216	0.205	0.236	0.007
460	Acute nasopharyngitis [common cold]	0.214	0.206	0.194	0.247	0.019
461	Acute sinusitis	0.181	0.185	0.143	0.211	0.017
463	Acute tonsillitis	0.179	0.179	0.163	0.200	0.009
692	Contact dermatitis and other eczema	0.151	0.152	0.128	0.173	0.012
558	Other and unspecified noninfectious gastroenteritis and colitis	0.142	0.141	0.129	0.158	0.009

Table 3: Summary Statistics of Prevalence for Top Ten Diseases

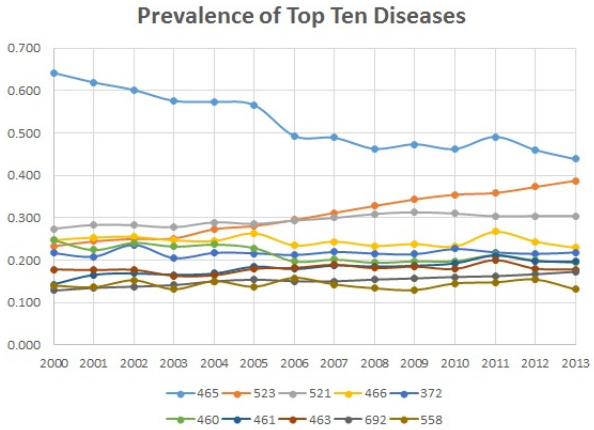


Figure 3: A Scatter Plot of Prevalence for Top Ten Diseases between 2000 and 2013

Two diseases with the highest mean prevalence (i.e. Acute upper respiratory infections and Gingival and periodontal diseases) also have the highest variations over fourteen years, while the remaining eight diseases have relative consistent prevalence rates. Specifically, acute upper respiratory infections has a consistently uprising trend while Gingival and periodontal diseases pose a sustained downward trend between 2000 and 2013. Most upper respiratory infections are self-diagnosed and self-treated at home. Patients who present with upper respiratory infections often benefit from reassurance, education and instructions for appropriate home treatment. In Taiwan, with the development of healthcare

system, medical education is more accessible for residents. Prevalence of upper respiratory infections may decrease with the improved home treatment knowledge. From National Institute of Health, the risk factors of Gingival and periodontal diseases include illness (e.g. cancer, diabetes), bad habits (e.g. smoking) and poor oral hygiene habits. With greater number of smokers and diabetic patients, the prevalence of gingival and periodontal diseases also increases.

3.2 Disease Correlation Matrix

As mentioned in method part, the first step of constructing a disease network is to define the disease correlation matrix. It is standard to use the Pearson correlation coefficient ϕ as a co-occurrence measure in network analysis [36, 37, 38, 39]. The values in similarity matrix $S = [s_{ij}]$ is between -1 and 1, with 1 indicates total positive linear correlation, 0 represents no linear correlation and -1 corresponds total negative linear correlation. Since we restrict our analysis to one-year period, the prevalence of each disease may not be very high. As a result, the absolute values of similarity matrix are relatively small. In the meanwhile, it is less likely to have two diseases with high incidence rate simultaneously. The summary statistics for off-diagonal similarity measures in each year is summarized in Table 4 and are listed separately for positive and negative correlations.

	For Sij>0			For Sij<0		
Year	Min	Median	Max	Min	Median	Max
2000	7.840E-08	3.277E-03	4.336E-01	-6.508E-02	-2.237E-04	-1.720E-08
2001	4.798E-08	3.253E-03	5.000E-01	-6.351E-02	-2.216E-04	-2.666E-09
2002	1.386E-07	3.270E-03	4.228E-01	-6.104E-02	-1.735E-04	-3.243E-08
2003	8.638E-08	3.225E-03	4.292E-01	-6.899E-02	-2.194E-04	-3.541E-08
2004	1.685E-07	3.275E-03	5.000E-01	-6.606E-02	-2.368E-04	-2.457E-07
2005	2.414E-07	3.441E-03	4.973E-01	-6.902E-02	-2.072E-04	-1.617E-07
2006	3.489E-08	3.464E-03	5.040E-01	-7.073E-02	-2.229E-04	-3.089E-08
2007	2.641E-07	3.353E-03	4.638E-01	-7.411E-02	-2.627E-04	-6.852E-08
2008	7.889E-08	3.340E-03	4.507E-01	-7.901E-02	-2.495E-04	-6.287E-08
2009	1.416E-07	3.338E-03	4.570E-01	-8.189E-02	-2.569E-04	-3.263E-08
2010	4.812E-08	3.353E-03	4.601E-01	-8.455E-02	-2.505E-04	-6.202E-08
2011	1.985E-08	3.343E-03	4.795E-01	-8.877E-02	-2.537E-04	-1.129E-08
2012	2.942E-08	3.302E-03	4.904E-01	-9.273E-02	-2.529E-04	-1.287E-09
2013	2.747E-08	3.292E-03	4.968E-01	-9.641E-02	-2.523E-04	-4.492E-08

Table 4: Range of Similarity Measures between 2000 and 2013

Our results in Table 4 does show that the absolute values in similarity matrices are relatively small. The range of similarity over fourteen years are constant. The medians of similarity measures indicate over 50 percent of pairwise disease associations are insignificant.

3.3 Disease Adjacency Matrix

The $n \times n$ similarity matrix is then transformed into an $n \times n$ adjacency matrix $A = [a_{ij}]$, demonstrating the strength of correlation between pair of diseases. The adjacency measures will be used as edges (i.e. distances) of two nodes (i.e. diseases) when we build disease networks later. Adjacency matrices are symmetric with non-negative entries (i.e. $a_{ij} \in [0, 1]$) and diagonal elements are set to be zero by convention. In this study, we are using a weighted adjacency matrix, as defined in equation 3. As we could see from similarity matrix, there exists plenty of tiny correlations between two diseases (i.e. $7.84E-08$), indicating insignificant pairwise correlations. We would treat them as noises in the network instead of true correlations, an threshold τ is then defined to squeeze out those noises. Several groups have suggested it may be necessary to threshold the Pearson correlation coefficient in constructing a network [40, 41, 42]. The choice of the parameter τ determines the sensitivity and specificity of pairwise edges (i.e. disease distances). Greater τ means fewer node connections, which may reduce the noises. However, the network will be too sparse for further investigation if τ is defined too small. In this study, τ is defined as 0.05 for all fourteen years network constructions. Hence, in order to get adjacency matrices $A = [a_{ij}]$, entries in similarity matrices with values lower than 0.05 are transformed into 0, and the remaining entries are converted into their absolute values. In an adjacency matrix, zeros represent no pairwise connections between two diseases, while non-zeros display the strength of disease relations. Table 5 below outlined the summary statistics of adjacency matrices.

The mean values and standard deviations of adjacency measures are relatively constant throughout fourteen years, with an increase between 2003 and 2004. By the end of year 2004, over 99 percent of the population in Taiwan are covered by National Health Insurance (NHI). The system promises equal access to healthcare for Taiwan residents. People then have improved accessibility to medical treatment, encouraging treatment

Year	Min	Max	Median	Mean	STD
2000	0	0.43	0	7.49E-04	9.13E-03
2001	0	0.42	0	7.25E-04	9.03E-03
2002	0	0.42	0	7.65E-04	9.21E-03
2003	0	0.43	0	7.76E-04	9.20E-03
2004	0	0.50	0	9.02E-04	1.00E-02
2005	0	0.50	0	9.71E-04	1.04E-02
2006	0	0.50	0	9.96E-04	1.05E-02
2007	0	0.46	0	1.00E-03	1.06E-02
2008	0	0.45	0	9.89E-04	1.05E-02
2009	0	0.46	0	1.01E-03	1.07E-02
2010	0	0.46	0	1.01E-03	1.06E-02
2011	0	0.48	0	1.01E-03	1.06E-02
2012	0	0.49	0	1.01E-03	1.06E-02
2013	0	0.50	0	9.94E-04	1.05E-02

Table 5: Summary Statistics of Adjacency Measures between 2000 and 2013

seeking behavior on diseases covered by NHI. Hence, more diseases covered by NHI are detected and reported, resulting an increased pairwise disease correlations on average. After diminishing network noises, the effective numbers of edge between 2000 and 2013 are 1725, 1662, 1764, 1820, 2072, 2174, 2219, 2203, 2201, 2201, 2227, 2207, 2184, 2169.

3.4 Connectivity

Besides the disease distances (i.e. edges), the second key concept of network construction is node connectivity, also known as centrality. Refer to TOM-based measure of connectivity $_i$ in equation 7, the connectivity of disease i equals to the summation of topological overlaps between i and all other diseases. The connectivity identifies the sizes of nodes within each network as well, so the greater the connectivity measures, the bigger the nodes will be. The co-occurrence rate of other diseases will ascend with the increase of node size. Nodes with high connectivity values are classified as connectivity hubs. With longitudinal data available in this study, we select top ten connectivity hubs and track their variations in connectivity measures. The ten connectivity hubs include Diabetes mellitus, Disorders of lipoid metabolism, Anxiety, dissociative and somatoform disorders, Cataract, Essential hypertension, Hypertensive heart disease, Chronic ischemic heart disease, Heart failure, Osteoarthrosis and allied disorders, and Spondylosis and allied disorders.

These diseases are classified into five categories in ICD-9 − endocrine, nutritional and metabolic diseases; mental disorder; diseases of the nervous system and sense organs; diseases of circulatory system; diseases of musculoskeletal system and connective tissue.

Our results are consistent with some existing studies. The outcomes of a cohort study conducted by Luijks et al. showed the prevalence of comorbidity in patients with type 2 diabetes pose a wide range [43]. Metra et al. indicated in their study a broad spectrum of concomitant disorders complicate heart failure. Comorbidities could be subdividied into cardiovascular (e.g. hypertension, coronary artery disease) and noncardiovascular (e.g. respiratory, renal, metabolic conditions)[44]. Leite et al. evaluated the frequency of comorbidities in osteoarthritis patients, metabolic syndrome, hypertension, dyslipidemis and obesity are shown to have high incidences. The coexistence of some chronic diseases is common among patients with osteoarthritis [45].

Between year 2002 and 2013, the variations in connectivity measures for connectivity hubs are minor. The diseases with high connectivity values tend to keep having a high co-occurrence with other diseases throughout the study period. We notice that diabetes have a particularly high connectivity value, which may be due to the fact that diabetes has been linked to multiple complications. The scatter plot for top ten diseases is shown in Figure 4.

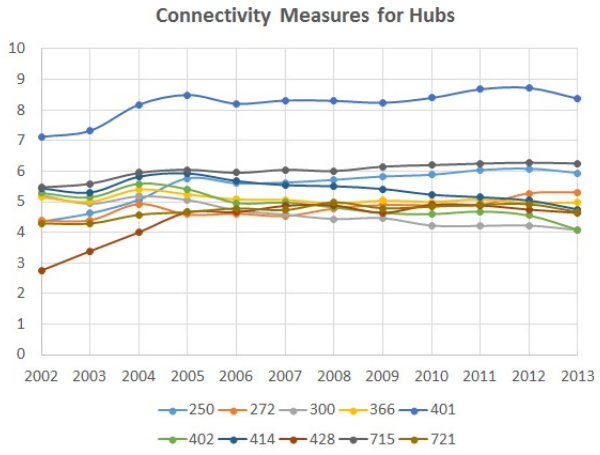


Figure 4: A Scatter Plot of Connectivity Measures for hubs between 2002 and 2013

3.5 Disease Dissimilarity and Modules

3.5.1 TOM Dissimilarity

An important aim of this network analysis is to detect the subsets of diseases that are tightly connected to each other. Refer to Zhang and Horvath, we apply average linkage hierarchical clustering coupled with topological overlap similarity measures. Through hierarchical clustering, a dendrogram is obtained each year, and we choose a constant height cut-off of 0.99 to arrive at disease clusters. The branches of the dendrogram are the corresponding disease subsets. The color row under dendrogram visualizing the disesae modules via assigning unique color to each module. The topological overlap of two nodes identify their inter-connection. The topological overlap dissimilarity (i.e. opposite of topological overlap) $DistTOM_{ij}$ is used as input of hierarchical clustering, defined in equation 5. Note $DistTOM_{ij} \in [0, 1]$ and $DistTOM_{ij} = 1$ if node i and node j are unconnected and the two nodes do not share any neighbors. And lower $DistTOM_{ij}$ means higher inter-connection between disease i and disease j. Summary statistics of topological overlap dissimilarity matrices for fourteen years are in Table 6.

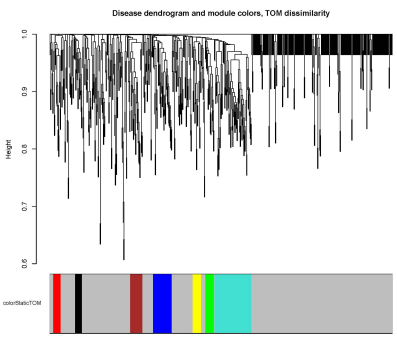
Year	Min	Max	Median	Mean	STD
2000	0.6425	1	1	0.9989	0.0090
2001	0.5236	1	1	0.9990	0.0088
2002	0.6146	1	1	0.9989	0.0090
2003	0.6506	1	1	0.9989	0.0089
2004	0.5000	1	1	0.9987	0.0097
2005	0.5625	1	1	0.9985	0.0101
2006	0.5528	1	1	0.9985	0.0101
2007	0.6028	1	1	0.9985	0.0101
2008	0.5825	1	1	0.9985	0.0100
2009	0.5683	1	1	0.9985	0.0102
2010	0.5728	1	1	0.9985	0.0107
2011	0.5981	1	1	0.9985	0.0102
2012	0.6294	1	1	0.9984	0.0102
2013	0.6270	1	1	0.9985	0.0101

Table 6: Summary Statistics of TOM Dissimilarity Measures between 2000 and 2013

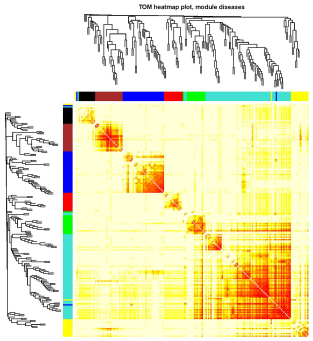
The average TOM dissimilarity measures are constant throughout fourteen years, and the medians indicate that over 50 percent of the dissimilarity measures equal to 1 in each year. For the same reasons as we define disease correlations − the study period for each disease network is short and possibility of having two diseases with high morbidity simultaneously is low, the dissimilarity measures are large and most entries in dissimilarity matrices are close to 1.

3.5.2 Disease Modules

Adopting the definition by Zhang and Horvath, disease modules are groups of diseases highly correlated across the sample patients. After applying average linkage hierarchical clustering, dendrogram and topological overlap matrix plots (i.e. heatmap) is attained for each year. The branches of the dendrogram are the corresponding disease modules. Each modules are specified by unique color, which are shown on color row under the dendrogram. Note the background color of the color row − grey is reserved to color diseases that are not part of any module, while the remaining colors represent the true disease subsets defined by hierarchical clustering. Light colors represent low topological overlap while progressively darker colors means higher topological overlap. Heatmap is a color-coded depiction of the values of the TOM-based dissimilarity, which allow us to visualize and identify network modules. As the modules are defined to be groups of nodes with high topological overlap, and the red/dark orange squares along the diagonal are corresponding modules (i.e. clusters).



(a) 2000: Dendrogram



(b) 2000: TOM plot

Figure 5: Dendrograms and heatmaps of diseases in 2000: unique colors on color row correspond to clusters; light colors in heatmap represent low topological overlap and progressively darker orange and red colors represent higher topological overlap.

Take year 2000 as an example (Figure 5), the color row in dendrogram (a) has seven unique colors − red, black, brown, blue, yellow, green, turquoise − indicating seven disease modules. Seven modules relating to same anatomical area or type of disorder are identified along the diagonal of the heatmap (i.e. red squares) as well. Table 7 gives the full list of disease in each cluster. Most diseases (ordered by color shown in color row) can be classified as (1) oral malignant neoplasm; (2) skin and subcutanrous tissues disorders; (3) gynecological disorders; (4) otolaryngological and respiratory disorders; (5) genitourinary disorders; (6) ophthalmological disorders; and (7) endocrine, metabolic diseases and immunity disorder; mental disorders; nervous system and sense organs disorders; circulatory system; digestive system; and diseases of musculoskeletal system and connective tissues. Dendrograms and heatmaps in other years are shown in Appendix.

Table 7: Disease modules in 2000

Seven clusters in 2000 range from trivial correlations (e.g. different mallignant oplasm in lip, oral cavity and pharynx.), to correlations of cause and effect codes (e.g. diseases of nervous and mental disorders). A number of different factors determine the overall health of populations and individuals, ranging from genetic and biologic characteristics of the individual to social context. They all play an important role in the etiology of any particular disease, as a result, they are expected to play a role in co-occurring diseases. Several studies have investigated comorbidity patterns of diseases [46, 47, 48, 49, 50]. Intuitively, groups of diseases cluster in one patient can be explained by two reasons [51]. First, directly shared biological factors such as common disease genes can cause comorbid diseases, which explains the clustering of diseases within the same category. Second, comorbid diseases can occur together since they share a common pattern of influences such as correlated risk factors, which may be the reason of co-occurrence of diseases classified in similar but different categories. But other reasons may explain these clustering as well. Refer to Valderas et al., different diseases may be found in the same individual by chance and selection bias, or by one or more types of causal association [50]. Four models of genuine etiological association between conditions are described in Figure 6. With limited information, this study only focuses on co-occurrence of diseases instead of causality.

In the direct causation model, the presence of one disease is directly responsible for another. This model may explain the clustering of complications, such as diabetes and cardiovascular diseases; diabetes and joint disorders; glaucoma and cataract; glaucoma and retinal detachments in our results. In the associated risk factors model, the risk factors for one disease are correlated with the risk factor for another disease. For example, smoking and alcohol consumption are correlated; smoking is a risk factor for heart failure while alcohol consumption is a risk factor for chronic liver diseases, making these two disease be more likely to occur simultaneously. By contrast, in the heterogeneity model, disease risk factors are not correlated but each one of them can cause either disease, also increasing the likelihood of disease co-occurrence. Taking hypertension and liver disease in our results as an example. Both alcohol consumption and age are risk factors of hypertension and risk factors of liver diseases, but they are independent risk factors. In the last model (Independence) − the simultaneous presence of two diseases may correspond to a third

distinct disease. In our results, for example, depressive disorder, Parkinson's disease and dementia are diseases occur in the same cluster. But the co-occurrence of depressive disorder and dementia might both be due to Parkinson's disease.

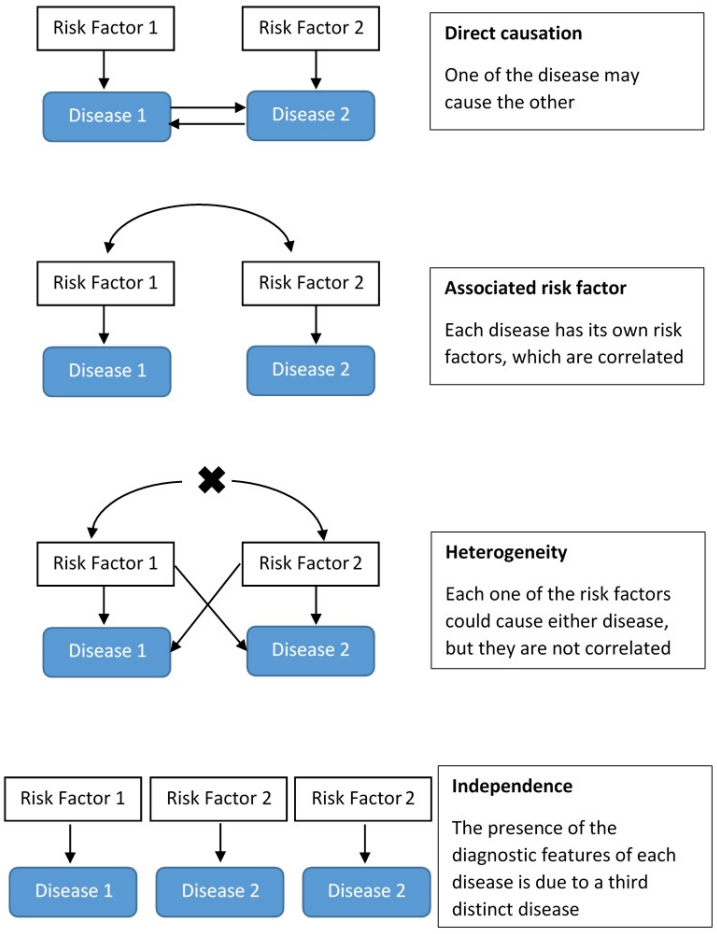


Figure 6: Four Models of Genuine Etiological Association

Note the four models here are not necessary mutually exclusive and have not to be applied extensively in comorbidity researches. However, all models here have been successfully tested and proved empirically valid in selected comorbidities [52]. We then use Gephi to visualize disease network, which is shown in Figure 7A and 7B.

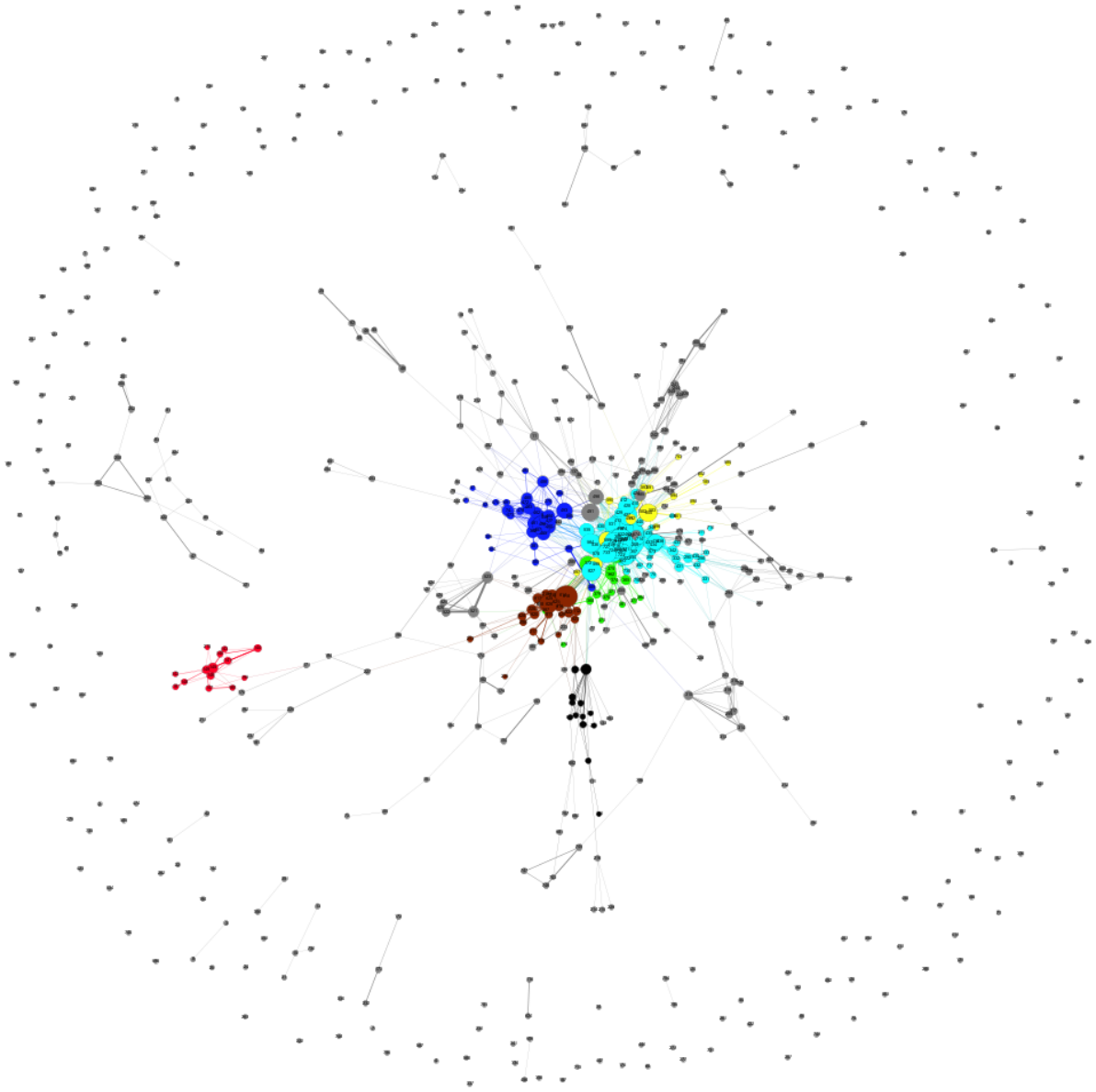


Figure 7A: Full Disease Network in 2000

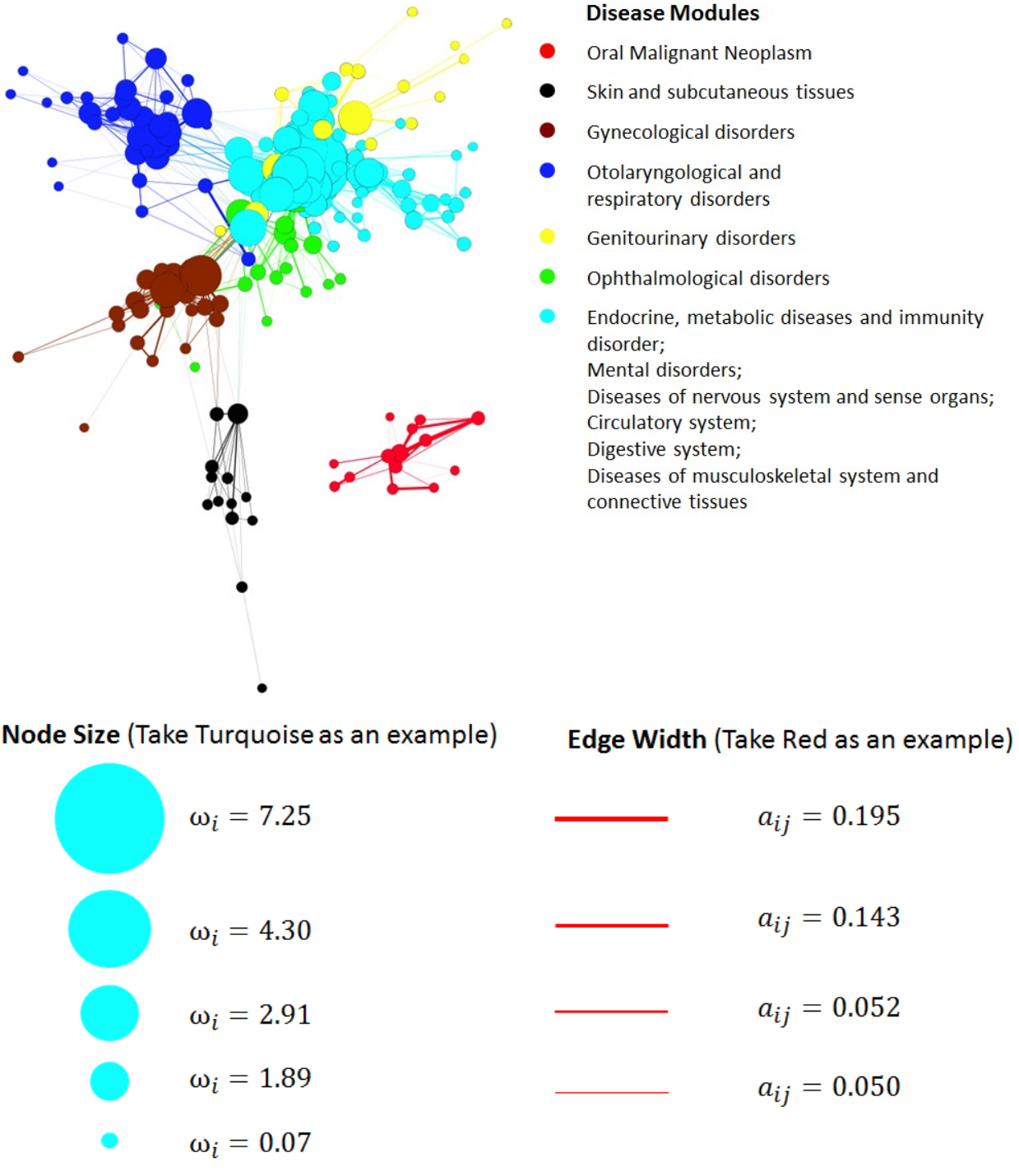


Figure 7B: Network of Modules in 2000

Node size represents the connectivity of each diseases, with larger node means higher connectivity value. Edges represent association between nodes. The thicker the edges, the more significant the association between nodes. We apply the results from hierarchical clustering to coloring nodes. Hence, the nodes within the same disease module share same color. Figure 7A displays the whole network while Figure 7B displays the network of seven modules produced by hierarchical clustering in 2000. The corresponding disease categories are listed also. As we notice in Figure 7A, there exist small nodes and corresponding edges surrounding significant disease modules. And these diseases with insignificant pairwise correlation are classified in neither disease modules.

We also calculate the summary statistics of adjacency values within each cluster (Figure 8). The bar plot is showing average adjacency values within each disease module. The brown (i.e. 0.0433) and green (i.e. 0.0436) clusters tend to have higher mean adjacency values, indicating higher pairwise comorbidiy rate of diseases within these two clusters on average. To be more specific, gynecological disorders are more likely to occur simultaneously, so does the opthalmological disorders. This may due to the highly correlated risk factors the diseases share within these two categories. We also notice turquoise cluster shows the lowest average adjacency values, which may result from wide range of disease categories in this cluster. Diseases range from Endocrine, metabolic diseases and immunity disorders, to mental disorders, to diseases of nervous system and sense organs, to circulatory disorders, to digestive disorders, to diseases of musculoskeletal system. Previous study showed that the pairwise associations of diseases within the same category are more significant than diseases between different categories. The study also indicated diseases classified as "Metabolic/immunity disorders", "Circulatory system", and "Nervous system and sense organs" had prevalent association patterns with other diseases. Disease belonging to "Musculoskeletal system", and "Digestive system" also showed similar patterns [51]. This could also explain the wide range of diseases categories in cluster turquoise in our results.

Summary Statistics of Adjacency Values

Cluster	Min	Max	Median	Mean	STD
	0	0.1424	0	0.0265	0.0396
	0	0.199	0	0.0348	0.0545
	0	0.3572	0	0.0433	0.0643
	0	0.2637	0	0.036	0.0552
	0	0.2419	0	0.0293	0.0510
	0	0.3994	0	0.0223	0.0467
	0	0.2341	0	0.0436	0.0577

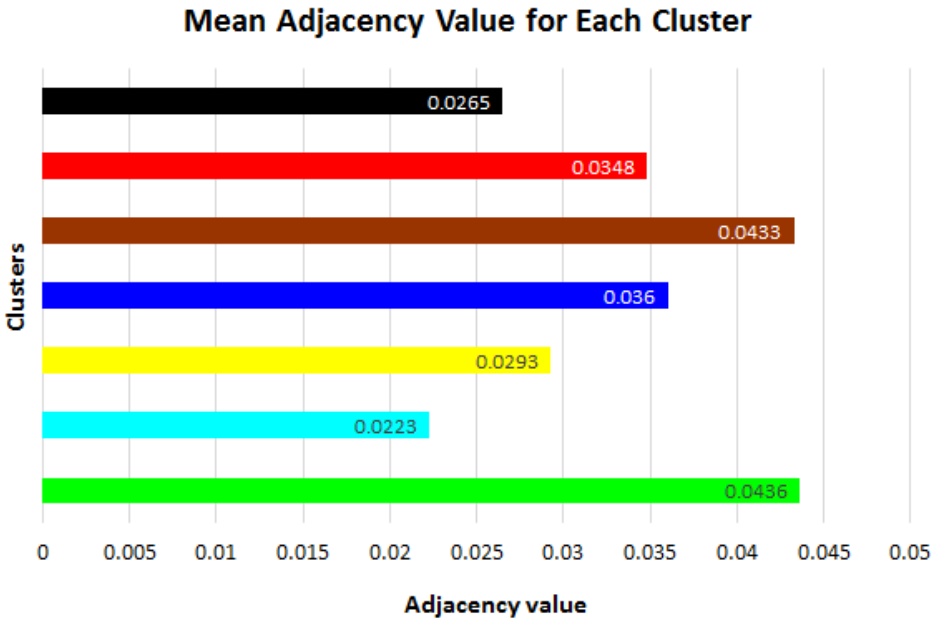
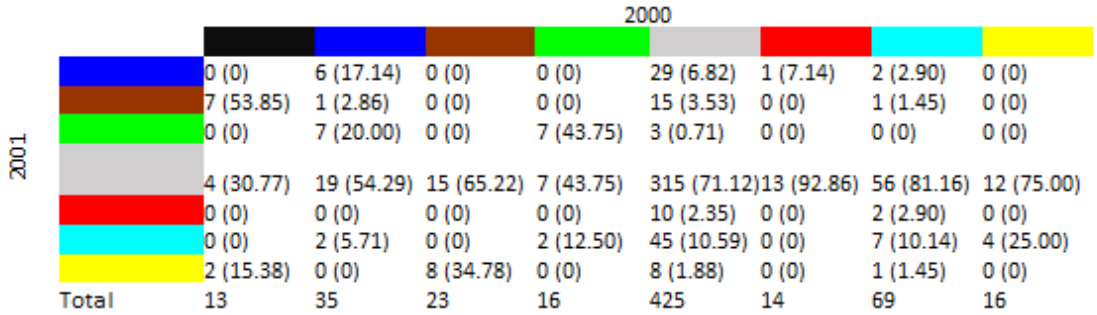


Figure 8: Summary Statistics of Adjacency for Disease Modules

3.5.3 Changing Patterns of Disease Modules

With the longitudinal data for fourteen years, we are able to detect the changes of clusters over this time period. As mentioned in 3.1.5, we track the changing patterns in each two adjacent years in an $n \times m$ matrix. The diagonal values indicate the number of disease stay constant within the same cluster for these two adjacent years, while off-diagonal values track the number of diseases switch between modules. The number in parenthesis indicates the corresponding percentage. Results for every two adjacent years are shown in Table 8. Also note that the diseases are classified into grey if they belong to neither modules, hence, the grey is the background color instead of the true disease module. Take the matrix of changes between 2003 and 2004 as an example, 90.63 percent of the diseases in blue cluster in 2003 stay within that cluster in 2004 also, while the percentage for green cluster is 93.75. The 6.25 percent of diseases in cluster green in 2003 switch to brown cluster in 2004. Observing from Table 8, the clusters tends to fluctuate between 2000 to 2003, and become relatively stable since 2003, and become erratic again starting 2010. Moreover, we notice diseases within the clusters with higher adjacency values such as brown and blue tend to be more stable than other clusters with lower adjacency values. The diseases classified into gynecological disorders and respiratory disorders have their unique etiological pathways, hence, they are more likely to be assigned in the same cluster as original one throughout years.

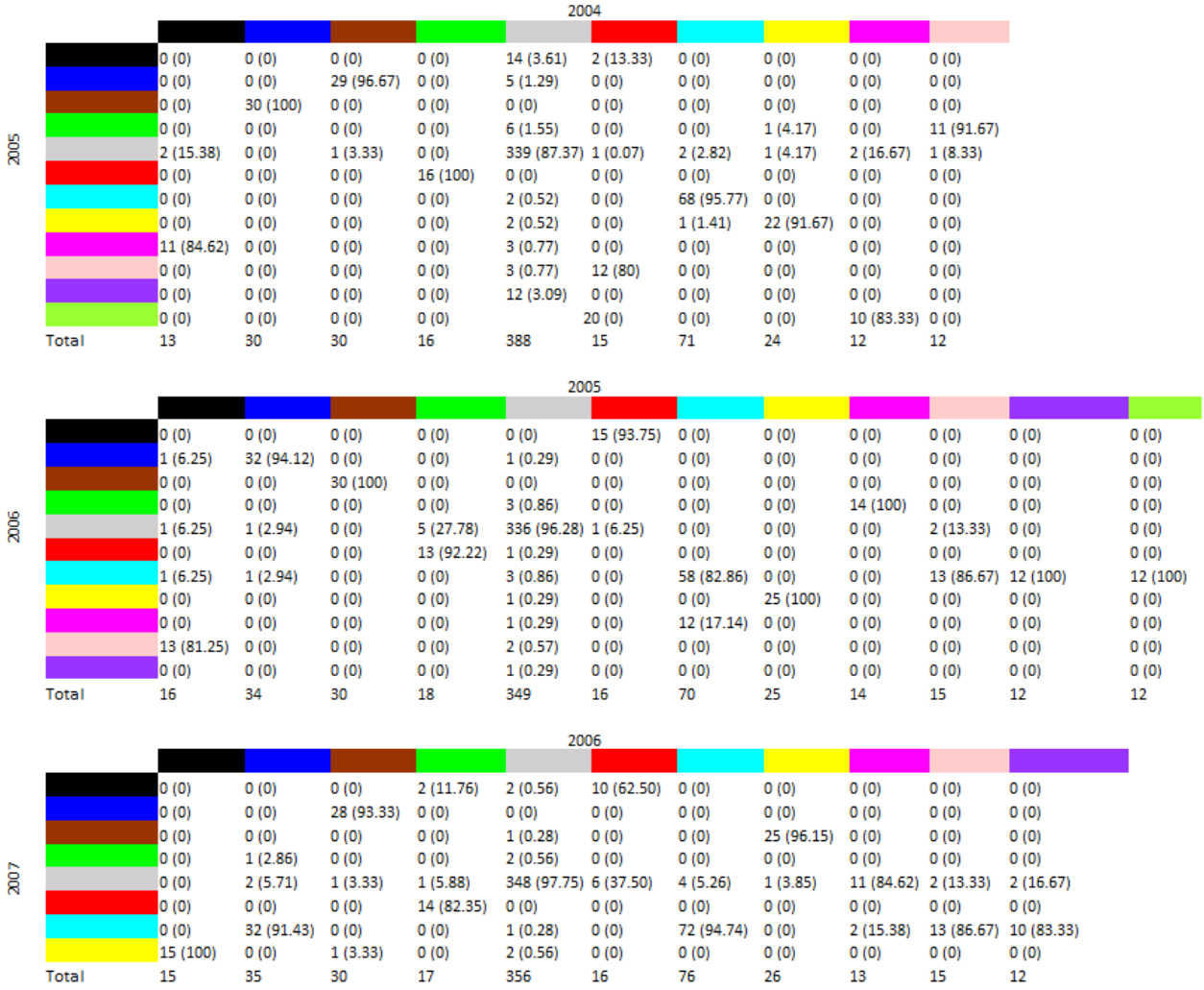


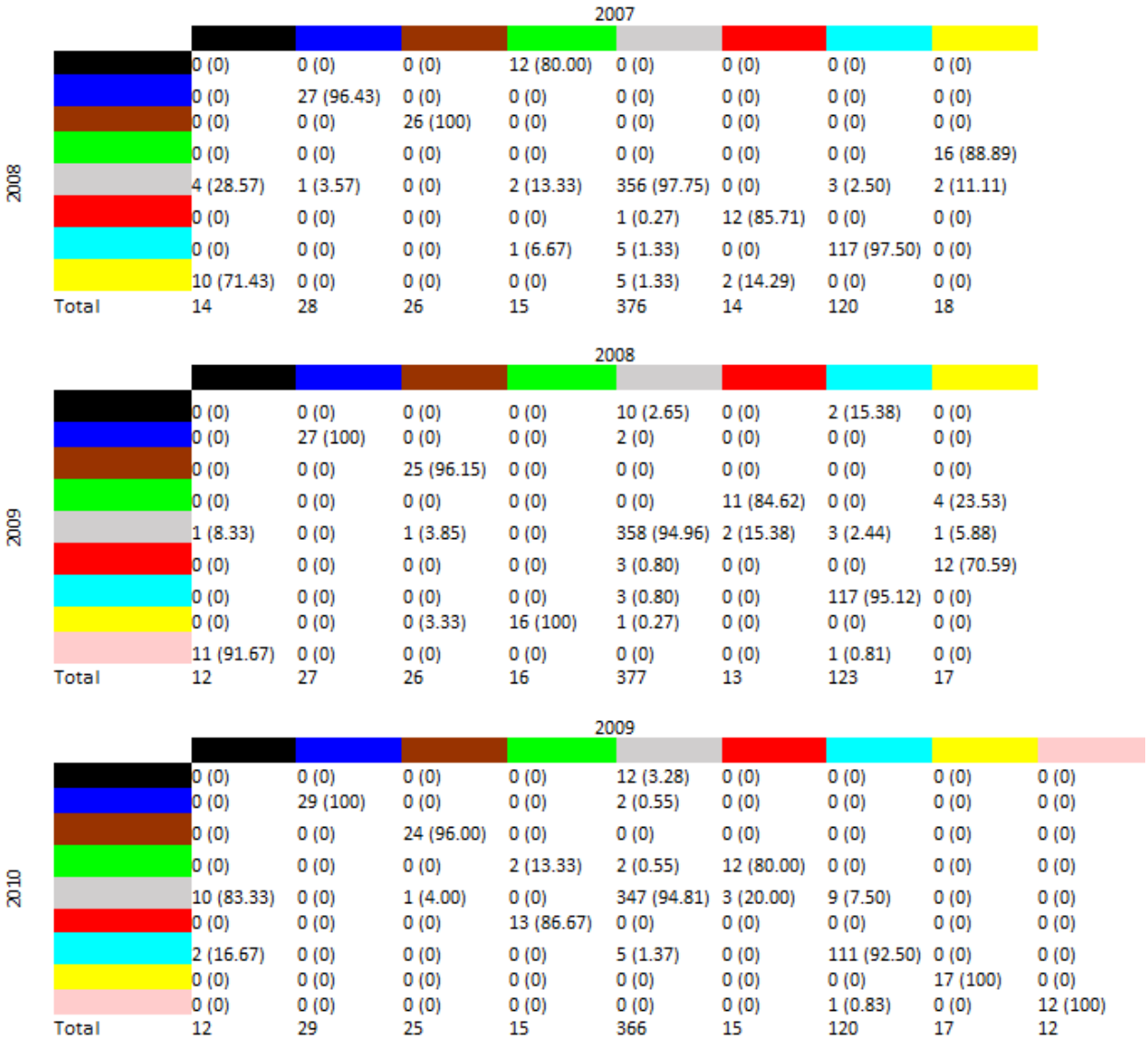
2001 \ 2000								
	0 (0)	6 (17.14)	0 (0)	0 (0)	29 (6.82)	1 (7.14)	2 (2.90)	0 (0)
	7 (53.85)	1 (2.86)	0 (0)	0 (0)	15 (3.53)	0 (0)	1 (1.45)	0 (0)
	0 (0)	7 (20.00)	0 (0)	7 (43.75)	3 (0.71)	0 (0)	0 (0)	0 (0)
	4 (30.77)	19 (54.29)	15 (65.22)	7 (43.75)	315 (71.12)	13 (92.86)	56 (81.16)	12 (75.00)
	0 (0)	0 (0)	0 (0)	0 (0)	10 (2.35)	0 (0)	2 (2.90)	0 (0)
	0 (0)	2 (5.71)	0 (0)	2 (12.50)	45 (10.59)	0 (0)	7 (10.14)	4 (25.00)
	2 (15.38)	0 (0)	8 (34.78)	0 (0)	8 (1.88)	0 (0)	1 (1.45)	0 (0)
Total	13	35	23	16	425	14	69	16

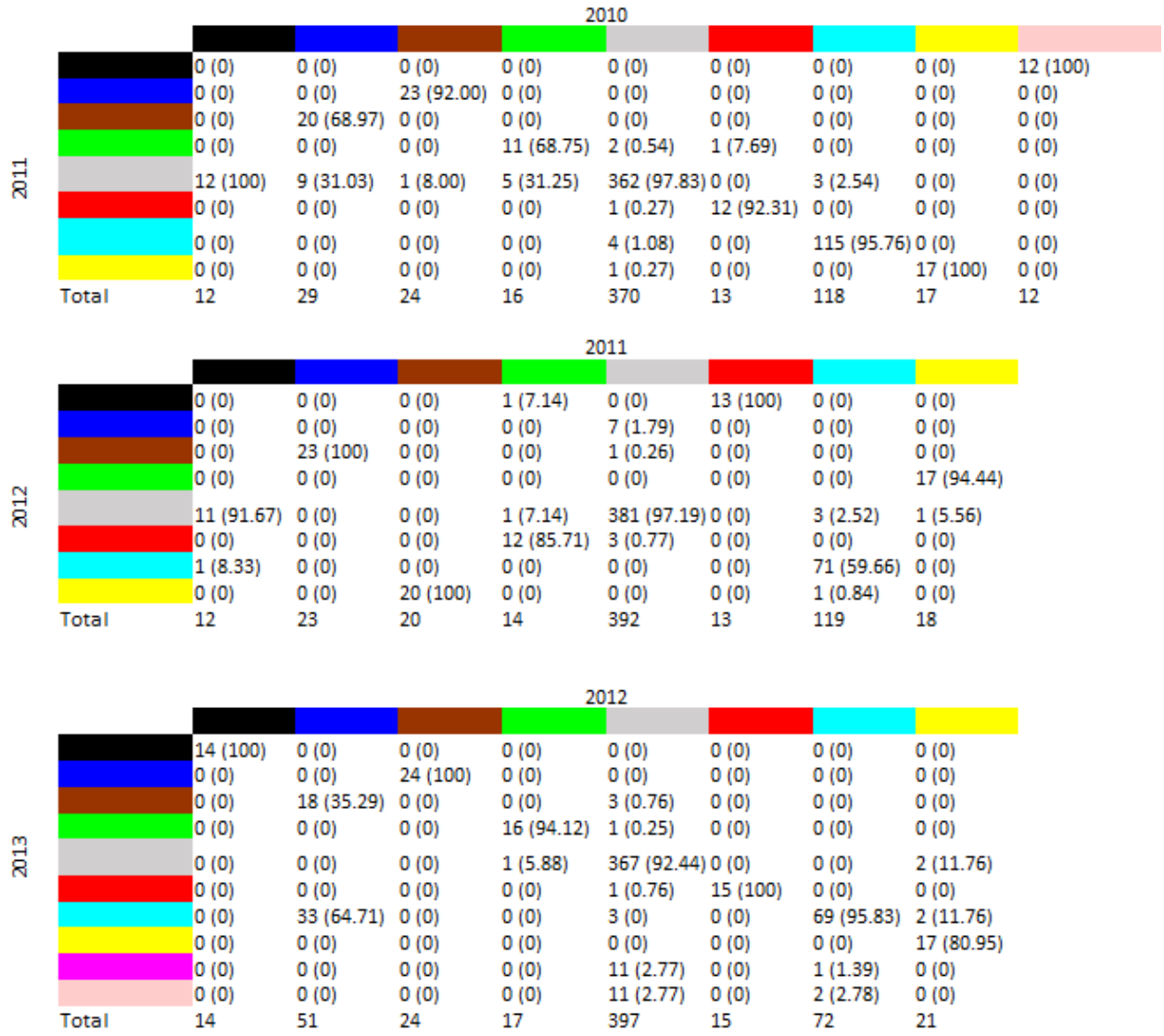
2002 \ 2001							
	6 (15.79)	1 (4.17)	6 (35.29)	18 (4.08)	0 (0)	1 (1.67)	0 (0)
	0 (0)	0 (0)	0 (0)	14 (3.17)	0 (0)	0 (0)	8 (42.11)
	0 (0)	0 (0)	7 (41.18)	8 (1.81)	0 (0)	2 (3.33)	0 (0)
	28 (73.68)	14 (58.33)	4 (23.53)	330 (74.83)	10 (83.33)	49 (81.67)	9 (47.37)
	0 (0)	8 (33.33)	0 (0)	4 (0.91)	0 (0)	0 (0)	1 (5.26)
	4 (10.53)	1 (4.17)	0 (0)	50 (11.34)	2 (16.67)	6 (10.00)	1 (5.26)
	0 (0)	0 (0)	0 (0)	17 (3.85)	0 (0)	2 (3.33)	0 (0)
Total	38	24	17	441	12	60	19

2003 \ 2002							
	31 (96.88)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	0 (0)	22 (100)	0 (0)	2 (0.43)	0 (0)	1 (1.56)	0 (0)
	0 (0)	0 (0)	0 (0)	4 (0.85)	12 (92.31)	0 (0)	0 (0)
	1 (3.13)	0 (0)	0 (0)	418 (89.32)	1 (7.69)	8 (12.5)	2 (10.53)
	0 (0)	0 (0)	0 (0)	12 (2.56)	0 (0)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	8 (1.71)	0 (0)	55 (85.94)	17 (89.47)
	0 (0)	0 (0)	17 (100)	0 (0)	0 (0)	0 (0)	0 (0)
Total	32	22	17	468	13	64	19

2004 \ 2003							
	0 (0)	0 (0)	0 (0)	5 (1.16)	0 (0)	8 (10.00)	0 (0)
	29 (90.63)	0 (0)	0 (0)	1 (0.23)	0 (0)	0 (0)	0 (0)
	1 (3.13)	0 (0)	1 (6.25)	11 (2.56)	0 (0)	17 (21.25)	0 (0)
	0 (0)	0 (0)	15 (93.75)	1 (0.23)	0 (0)	0 (0)	0 (0)
	1 (3.13)	2 (8.00)	0 (0)	369 (85.81)	12 (100)	2 (2.5)	2 (11.76)
	0 (0)	0 (0)	0 (0)	15 (3.49)	0 (0)	0 (0)	0 (0)
	0 (0)	1 (4.00)	0 (0)	2 (0.47)	0 (0)	53 (66.25)	15 (88.24)
	0 (0)	22 (88.00)	0 (0)	2 (0.47)	0 (0)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	12 (2.79)	0 (0)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	12 (2.79)	0 (0)	0 (0)	0 (0)
Total	31	25	16	430	12	80	17


2004
2005
0 (0) 0 (0) 0 (0) 0 (0) 14 (3.61) 2 (13.33) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 29 (96.67) 0 (0) 5 (1.29) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 30 (100) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 6 (1.55) 0 (0) 0 (0) 1 (4.17) 0 (0) 11 (91.67)
2 (15.38) 0 (0) 1 (3.33) 0 (0) 339 (87.37) 1 (0.07) 2 (2.82) 1 (4.17) 2 (16.67) 1 (8.33)
0 (0) 0 (0) 0 (0) 16 (100) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 2 (0.52) 0 (0) 68 (95.77) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 2 (0.52) 0 (0) 1 (1.41) 22 (91.67) 0 (0) 0 (0)
11 (84.62) 0 (0) 0 (0) 0 (0) 3 (0.77) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 3 (0.77) 12 (80) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 12 (3.09) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 20 (0) 0 (0) 0 (0) 10 (83.33) 0 (0)
Total 13 30 30 16 388 15 71 24 12 12
2006
2005
0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 15 (93.75) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
1 (6.25) 32 (94.12) 0 (0) 0 (0) 1 (0.29) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 30 (100) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 3 (0.86) 0 (0) 0 (0) 0 (0) 14 (100) 0 (0) 0 (0) 0 (0)
1 (6.25) 1 (2.94) 0 (0) 5 (27.78) 336 (96.28) 1 (6.25) 0 (0) 0 (0) 0 (0) 2 (13.33) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 13 (92.22) 1 (0.29) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
1 (6.25) 1 (2.94) 0 (0) 0 (0) 3 (0.86) 0 (0) 58 (82.86) 0 (0) 0 (0) 13 (86.67) 12 (100) 12 (100)
0 (0) 0 (0) 0 (0) 0 (0) 1 (0.29) 0 (0) 0 (0) 25 (100) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 1 (0.29) 0 (0) 12 (17.14) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
13 (81.25) 0 (0) 0 (0) 0 (0) 2 (0.57) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 1 (0.29) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
Total 16 34 30 18 349 16 70 25 14 15 12 12
2007
2006
0 (0) 0 (0) 0 (0) 2 (11.76) 2 (0.56) 10 (62.50) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 28 (93.33) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 1 (0.28) 0 (0) 0 (0) 25 (96.15) 0 (0) 0 (0) 0 (0)
0 (0) 1 (2.86) 0 (0) 0 (0) 2 (0.56) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 2 (5.71) 1 (3.33) 1 (5.88) 348 (97.75) 6 (37.50) 4 (5.26) 1 (3.85) 11 (84.62) 2 (13.33) 2 (16.67)
0 (0) 0 (0) 0 (0) 14 (82.35) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 32 (91.43) 0 (0) 0 (0) 1 (0.28) 0 (0) 72 (94.74) 0 (0) 2 (15.38) 13 (86.67) 10 (83.33)
15 (100) 0 (0) 1 (3.33) 0 (0) 2 (0.56) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
Total 15 35 30 17 356 16 76 26 13 15 12


2007
2008
0 (0) 0 (0) 0 (0) 12 (80.00) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 27 (96.43) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 26 (100) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 16 (88.89)
4 (28.57) 1 (3.57) 0 (0) 2 (13.33) 356 (97.75) 0 (0) 3 (2.50) 2 (11.11)
0 (0) 0 (0) 0 (0) 0 (0) 1 (0.27) 12 (85.71) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 1 (6.67) 5 (1.33) 0 (0) 117 (97.50) 0 (0)
10 (71.43) 0 (0) 0 (0) 0 (0) 5 (1.33) 2 (14.29) 0 (0) 0 (0)
Total 14 28 26 15 376 14 120 18
2008
2009
0 (0) 0 (0) 0 (0) 0 (0) 10 (2.65) 0 (0) 2 (15.38) 0 (0)
0 (0) 27 (100) 0 (0) 0 (0) 2 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 25 (96.15) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 11 (84.62) 0 (0) 4 (23.53)
1 (8.33) 0 (0) 1 (3.85) 0 (0) 358 (94.96) 2 (15.38) 3 (2.44) 1 (5.88)
0 (0) 0 (0) 0 (0) 0 (0) 3 (0.80) 0 (0) 0 (0) 12 (70.59)
0 (0) 0 (0) 0 (0) 0 (0) 3 (0.80) 0 (0) 117 (95.12) 0 (0)
0 (0) 0 (0) 0 (3.33) 16 (100) 1 (0.27) 0 (0) 0 (0) 0 (0)
11 (91.67) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 1 (0.81) 0 (0)
Total 12 27 26 16 377 13 123 17
2009
2010
0 (0) 0 (0) 0 (0) 0 (0) 12 (3.28) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 29 (100) 0 (0) 0 (0) 2 (0.55) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 24 (96.00) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 2 (13.33) 2 (0.55) 12 (80.00) 0 (0) 0 (0) 0 (0)
10 (83.33) 0 (0) 1 (4.00) 0 (0) 347 (94.81) 3 (20.00) 9 (7.50) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 13 (86.67) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0)
2 (16.67) 0 (0) 0 (0) 0 (0) 5 (1.37) 0 (0) 111 (92.50) 0 (0) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 17 (100) 0 (0)
0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 0 (0) 1 (0.83) 0 (0) 12 (100)
Total 12 29 25 15 366 15 120 17 12



2011 \ 2010

	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	12 (100)
	0 (0)	0 (0)	23 (92.00)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	0 (0)	20 (68.97)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	11 (68.75)	2 (0.54)	1 (7.69)	0 (0)	0 (0)	0 (0)
	12 (100)	9 (31.03)	1 (8.00)	5 (31.25)	362 (97.83)	0 (0)	3 (2.54)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.27)	12 (92.31)	0 (0)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	0 (0)	4 (1.08)	0 (0)	115 (95.76)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.27)	0 (0)	0 (0)	17 (100)	0 (0)
Total	12	29	24	16	370	13	118	17	12

2012 \ 2011

	0 (0)	0 (0)	0 (0)	1 (7.14)	0 (0)	13 (100)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	0 (0)	7 (1.79)	0 (0)	0 (0)	0 (0)
	0 (0)	23 (100)	0 (0)	0 (0)	1 (0.26)	0 (0)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	17 (94.44)
	11 (91.67)	0 (0)	0 (0)	1 (7.14)	381 (97.19)	0 (0)	3 (2.52)	1 (5.56)
	0 (0)	0 (0)	0 (0)	12 (85.71)	3 (0.77)	0 (0)	0 (0)	0 (0)
	1 (8.33)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	71 (59.66)	0 (0)
	0 (0)	0 (0)	20 (100)	0 (0)	0 (0)	0 (0)	1 (0.84)	0 (0)
Total	12	23	20	14	392	13	119	18

2013 \ 2012

	14 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	0 (0)	0 (0)	24 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	0 (0)	18 (35.29)	0 (0)	0 (0)	3 (0.76)	0 (0)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	16 (94.12)	1 (0.25)	0 (0)	0 (0)	0 (0)
	0 (0)	0 (0)	0 (0)	1 (5.88)	367 (92.44)	0 (0)	0 (0)	2 (11.76)
	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.76)	15 (100)	0 (0)	0 (0)
	0 (0)	33 (64.71)	0 (0)	0 (0)	3 (0)	0 (0)	69 (95.83)	2 (11.76)
	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	17 (80.95)
	0 (0)	0 (0)	0 (0)	0 (0)	11 (2.77)	0 (0)	1 (1.39)	0 (0)
	0 (0)	0 (0)	0 (0)	0 (0)	11 (2.77)	0 (0)	2 (2.78)	0 (0)
Total	14	51	24	17	397	15	72	21

Table 8: Changing Patterns of Clustering

4 Discussion

4.1 Study Limitation

This study has several limitations. Exploring disease associations do not imply causation and with limited information, we did not take into account the temporal sequence of the comorbid diseases. In addition, simply because a significant connection exists between two diseases does not directly imply valid clinical relationship. In this study, we set a threshold (i.e. 0.05) for detecting significant associations since some of the weaker ones may simply due to chance given the large sample size. Nevertheless, some real but less significant associations may be overlooked with such methodologies.

Even though all patient claims are collected each year by National Health Insurance program, they are actually entered at the discretion of the clinicians. The strictness of classification criteria used is unknown. It has been shown in our dataset that some disease diagnose code are unrecognized, indicating some data entries may be inaccurate. It has been shown in Rhodes et al. study that coded diagnoses from billing data can often be inaccurate [53]. Moreover, one study based on Veterans health care system specified that clinicians may also fail to enter patients' full information [54].

The datasets we could access are in ICD-9-CM formation, which are in three digits format and designed mainly for insurance claim. Potential errors for using the ICD classification scheme at data collection and coding in general have been noted. For example, at the 5-digit ICD-9 classification, there are 33 diagnoses associated with hypertension, which are only five at the 3-digit level [55].

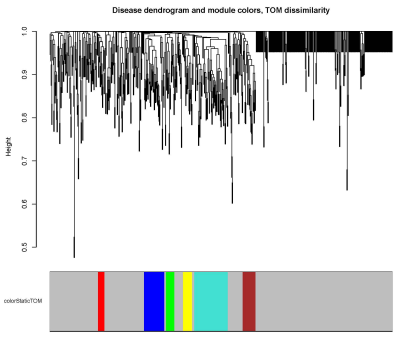
4.2 Impacts on Disease Management

The importance of recognizing comorbid health problems in terms of their relevance to clinical management was shown in previous research [56]. Various classification systems have been developed by Kaplan et al., Angold et al., and Piette and Kerr [57, 58, 59], which are widely reflected under current clinical care practice. For example, ischemic heart disease,

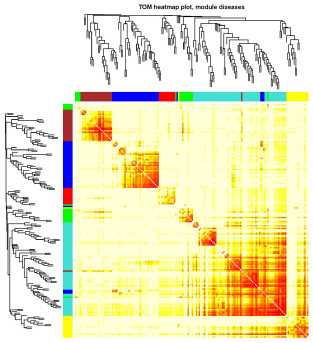
cardiovascular diseases (e.g. hypertension), and diabetes are commonly managed within the same clinics in primary care since they share important aspects of disease management. Our study results also show they belong to the same cluster (i.e. cluster turquoise in 2000). Drawing together patients with similar needs in clinical management will be efficient. Understanding the likelihood of getting comorbid diseases is also a key point in preventing worsening of current medical conditions. Future work could involve implementing this disease network in a clinical care setting to provide real-time suggestions to clinicians in disease prevention and management.

Detecting co-occurred diseases from human disease network could be useful for hypothesis generation and the confirmation in disease associations in existing literature. Comparing our results with those of other studies would help to support or refute some findings uncovered in our analysis. It also provides an useful clue for further research on exploring detailed disease comorbidity. Further work in the laboratory to elucidate biological mechanisms and improve the understanding of interactions among comorbid diseases is still needed [18].

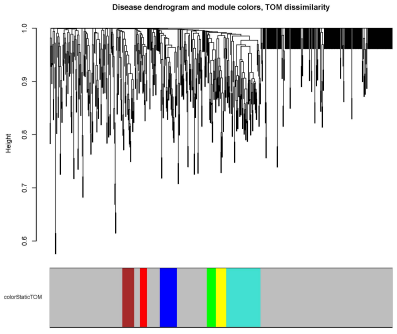
5 Appendix



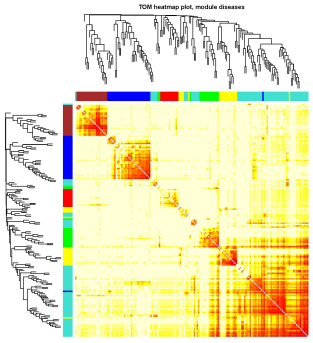
(a) 2001: Dendrogram



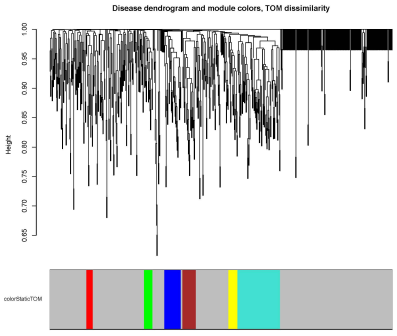
(b) 2001: TOM plot



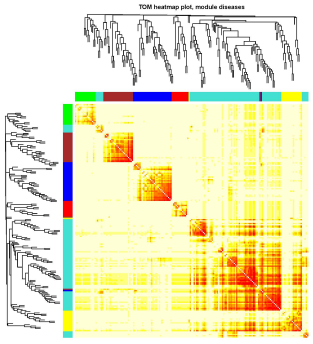
(c) 2002: Dendrogram



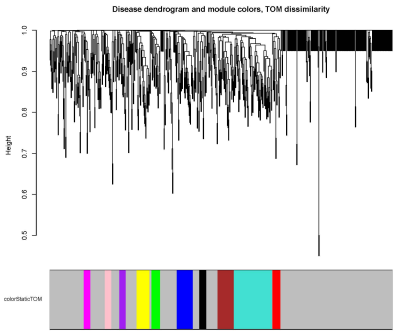
(d) 2002: TOM plot



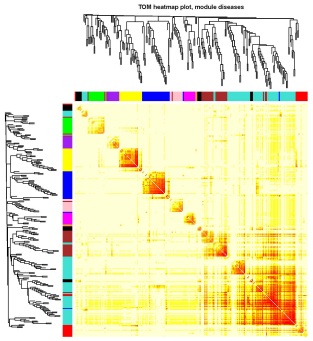
(e) 2003: Dendrogram



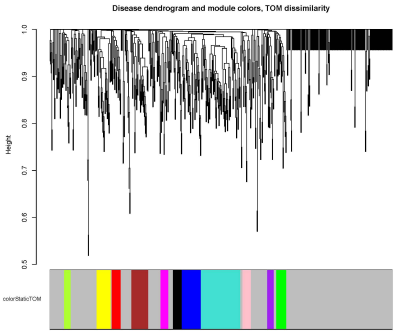
(f) 2003: TOM plot



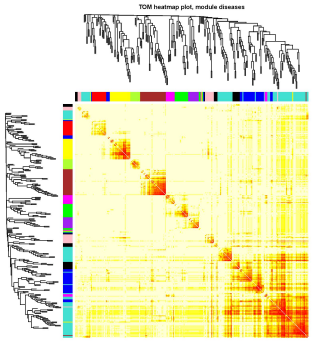
(g) 2004: Dendrogram



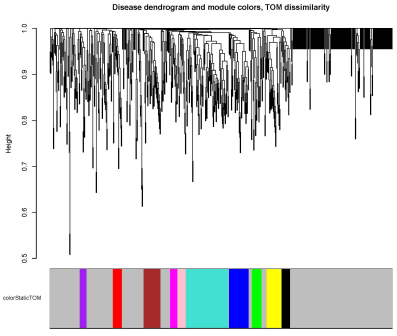
(h) 2004: TOM plot



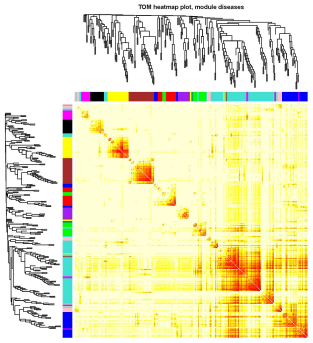
(i) 2005: Dendrogram



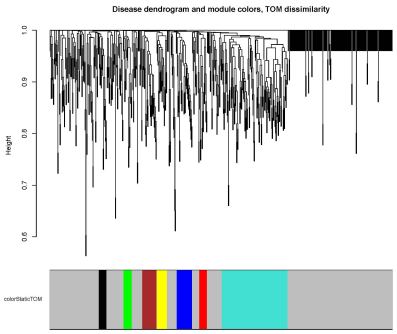
(j) 2005: TOM plot



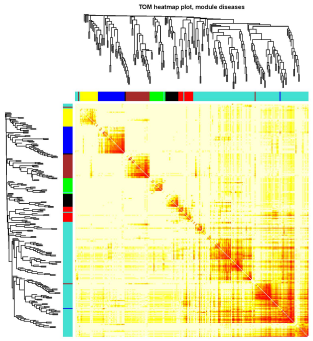
(k) 2006: Dendrogram



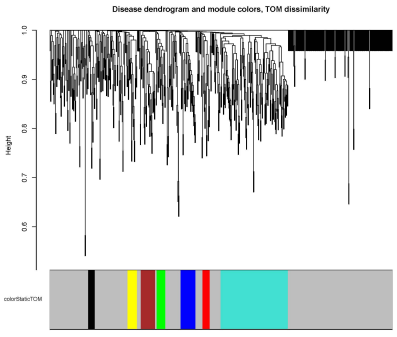
(l) 2006: TOM plot



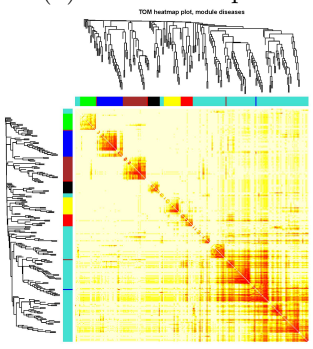
(m) 2007: Dendrogram



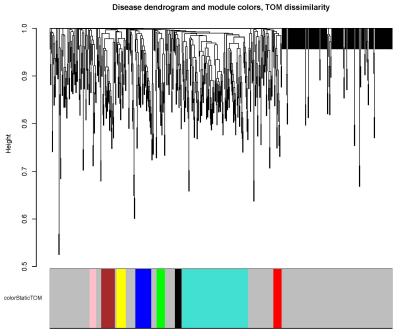
(n) 2007: TOM plot



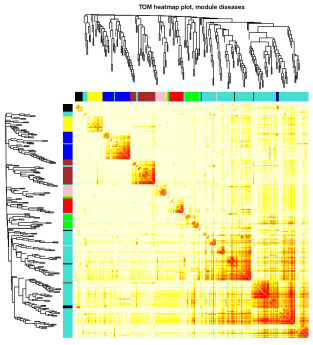
(o) 2008: Dendrogram



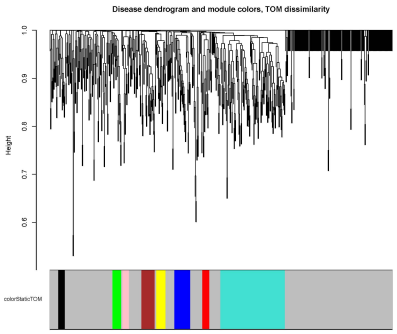
(p) 2008: TOM plot



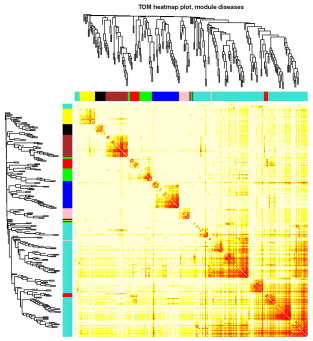
(q) 2009: Dendrogram



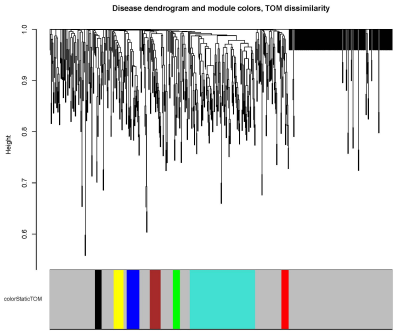
(r) 2009: TOM plot



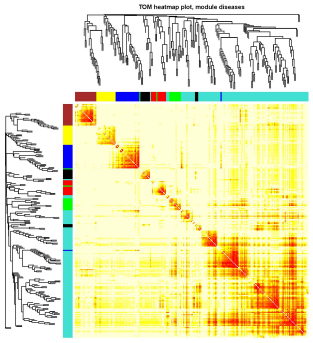
(s) 2010: Dendrogram



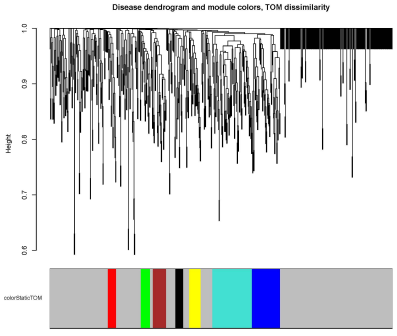
(t) 2010: TOM plot



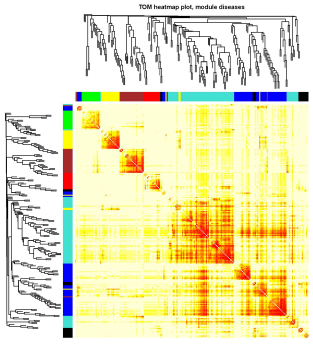
(u) 2011: Dendrogram



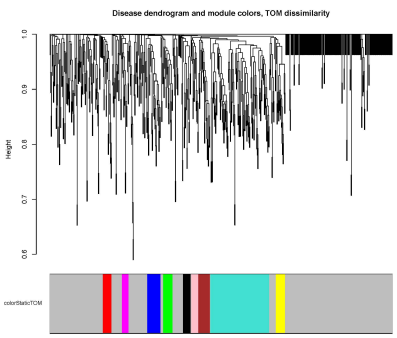
(v) 2011: TOM plot



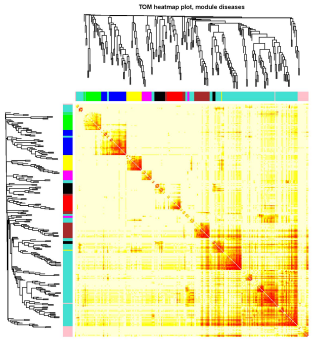
(w) 2012: Dendrogram



(x) 2012: TOM plot



(y) 2013: Dendrogram



(z) 2013: TOM plot

Figure 9: Dendrograms and heatmaps of diseases between 2001 to 2013: unique colors on color row correspond to clusters; light colors in heatmap represent low topological overlap and progressively darker orange and red colors represent higher topological overlap.

Bibliography

[1] Hidalgo CA, Blumm N, Barabasi AL, and Christakis NA. "A Dynamic Network Approach for the study of Human Phenotypes". In: *PLoS Comput Biol* 5.4 (2009). DOI: 10.1371/journal.pcbi.1000353.

[2] Goh K-I, Gusick ME, Valle D, Childs B, and Vidal M. "The human disease network". In: *Proc Natl Acad Sci USA* 104.21 (2007), pp. 8685–8690. DOI: 10.1073/pnas.0701361104.

[3] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawwa T, and Dricot A. "Towards a proteome-scale map of the human protein-protein interaction network". In: *Nature* 437.7062 (2005), pp. 1173–1178. DOI: 10.1038/nature04209.

[4] Stelzl U, Worm U, Lalowski M, Haenig C, and Brembeck FH é. "A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome". In: *Cell* 122.6 (2005), pp. 957–968. DOI: http://dx.doi.org/10.1016/j.cell.2005.08.029.

[5] Pujana MA et al. "Network modeling links breast cancer susceptibility and centrosome dysfunction". In: *Nat Genet* 39.11 (2007), pp. 1338–1349. DOI: 10.1038/ng.2007.2.

[6] Calvano SE, Xiao WE, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, and Brownstein BH. "A network-based analysis of systemic inflammation in humans". In: *Nature* 437 (2005), pp. 1032–1037. DOI: 10.1038/nature03985.

[7] Oldham MC, Horvath S, and Geschwind DH. "Conservation and evolution of gene coexpression networks in human and chimpanzee brains". In: *PNAS* 103.47 (2006), pp. 17973–17978. DOI: 10.1073/pnas.0605938103.

[8] Zhang B and Horvath S. "A General Framework for Weighted Gene Co-Expression Network Analysis". In: *Statistical Applications in Genetics and Molecular Biology* 4.17 (2005). DOI: 10.2202/1544-6115.1128.

[9] Dong J and Horvath S. "Understanding Network Concepts in Modules". In: *BMC Systems Biology* 1.24 (2007). DOI: 10.1186/1752-0509-1-24.

[10] Dong J and Horvath S. "Geometric Interpretation of Gene Co-expression Network Analysis". In: *PLoS Comput Biol* 4.8 (2008). DOI: 10.1371/journal.pcbi.1000117.

[11] Yip A and Horvath S. "Gene network interconnectedness and the generalized topological overlap measure". In: *BMC Bioinformatics* 8.22 (2007). DOI: 10.1186/1471-2105-8-22.

[12] Haux R. "Health care in the information society: what should be the role of medical informatics?" In: *Methods Inf Med* 41.1 (2002), pp. 31–35.

[13] Prokosch HU and Ganslandt T. "Perspectives for medical informatics. Reusing the electronic medical record for clinical research". In: *Methods Inf Med* 48.1 (2009), pp. 38–44.

[14] DesRoches CM, Campbell EG Rao SR, Donelan K, Ferris TG, Jha A, Kaushal R, Levy DE, Rosenbaum S, Shields AE, and Blumenthal D. "Electronic Health Records in Ambulatory Care — A National Survey of Physicians". In: *N Engl J* 359 (2008), pp. 50–60. DOI: 10.1056/NEJMsa0802005.

[15] Hoffman S. "Electronic health records and research: privacy versus scientific priorities". In: *Am J Bioeth* 10 (9), pp. 19–20. DOI: 10.1080/15265161.2010.492894.

[16] Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, and et al. "Medical data mining: knowledge discovery in a clinical data warehouse". In: *Proc AMIA Annu Fall Symp* (1997), pp. 101–105.

[17] Hanauer DA, Rhodes DR, and Chinnaiyan AM. "Exploring Clinical Association Using '-Omics' Based Enrichment Analyses". In: *PLoS ONE* 4.4 (2009). DOI: 10.1371/journal.pone.0005203.

[18] Kwoh CK and Ng PY. "Network analysis approach for biology". In: *Cell. Mol. Life Sci.* 64 (2007), pp. 1739–1751. DOI: 10.1007/s00018-007-7053-7.

[19] Hsiao FY, Yang CL, Huang YT, and Huang WF. "Using Taiwan's National Health Insurance Research Databases for Pharmacoepidemiology Research". In: *Journal of Food and Drug Analysis* 15.2 (2007), pp. 99–108.

[20] Chen YC, Yun YH, Wu JC, Haschler I, Chen TJ, and Wetter T. "Taiwan's National Health Insurance Research Database: administrative health care database as study object in bibliometrics". In: *Scientometrics* 86.2 (2011), pp. 365–380. DOI: 10.1007/s11192-010-0289-2.

[21] Harpe SE. "Using secondary data sources for pharmaco-epidemiology and outcomes research". In: *Pharmacotherapy* 29.2 (2009), pp. 138–153. DOI: 10.1592/phco.29.2.138.

[22] Tseng CH. "Diabetes and risk of bladder cancer: a study using the National Health Insurance database in Taiwan". In: *Diabetologia* 54.8 (2011), pp. 2009–2015. DOI: 10.1007/s00125-011-2171-z.

[23] National Health Research Institutes. *National Health Insurance Research Database, Taiwan*. URL: http://nhird.nhri.org.tw/en/index.htm.

[24] DeShazo JP, LaVallie DL, and Wolf FM. "Publication trends in the medical informatics literature: 20 years of "Medical Informatics" in MeSH". In: *BMC Medical Informatics and Decision Making* 9.7 (2009). DOI: 10.1186/1472-6947-9-7.

[25] Fernandez-Cano A, Torralbo M, and Vallejo M. "Reconsidering Price's model of scientific growth: An overview". In: *Scientometrics* 61.3 (2004), pp. 301–321. DOI: 10.1023/B:SCIE.0000045112.11562.11.

[26] Zien A, Küffner R, Zimmer R, and Lengauer T. "Analysis of gene expression data with pathway scores". In: *Proc Int Conf Intell Syst Mol Biol* 8 (2000), pp. 407–417.

[27] Kurhekar MP, Adak S, Jhunjhunwala S, and Raghupathy K. "Genome-wide pathway analysis and visualization using gene expression data". In: *Pac Symp Biocomput* (2002), pp. 462–473.

[28] Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, and Habegger L. "Personal omics profiling reveals dynamic molecular and medical phenotypes". In: *Cell* 148.6 (2012), pp. 1293–1307. DOI: 10.1016/j.cell.2012.02.009.

[29] Leiserson MD, Blokh D, Sharan R, and Raphael BJ. "Simultaneous identification of multiple driver pathways in cancer". In: *PLoS Comput Biol* 9.5 (2013). DOI: 10.1371/journal.pcbi.1003054.

[30] Qiu YQ, Zhang S, Zhang XS, and Chen L. "Detecting disease associated modules and prioritizing active genes based on high throughput data". In: *BMC Bioinformatics* 11.26 (2010). DOI: 10.1186/1471-2105-11-26.

[31] Spirin V and Mirny LA. "Protein complexes and functional modules in molecular networks". In: *Proc Natl Acad Sci USA* 100.21 (2003), pp. 12123–12128.

[32] Rives AW and Galitski T. "Modular organization of cellular networks". In: *Proc Natl Acad Sci USA* 100.3 (2003), pp. 1128–1133. DOI: 10.1073/pnas.0237338100.

[33] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, and Barabasi AL. "Hierarchical organization of modularity in metabolic networks". In: *Science* 297.5586 (2002), pp. 1551–1555. DOI: 10.1126/science.1073374.

[34] Ye Y and Godzik A. "Comparative analysis of protein domain organization". In: *Genome Biology* 14.3 (2004), pp. 343–353. DOI: 10.1101/gr.1610504.

[35] Horvath S and Langfelder P. *Tutorials for the WGCNA package*. URL: https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/.

[36] Eisen M, Spellman P, Brown P, and Botstein D. "Cluster analysis and display of genome-wide expression patterns". In: *Proc Natl Acad Sci U S A* 95.25 (1998), pp. 14863–14868.

[37] Qin J, Lewis D, and Noble W. "Kernel Hierarchical Gene Clustering from Microarray Expression Data". In: *Bioinformatics* 19.16 (2003), pp. 2097–2104.

[38] Gat-Vilks I, Sharan R, and Shamir R. "Scoring Clustering Solutions by their Biological Relevance". In: *Bioinformatics* 19.18 (2003), pp. 2381–2389.

[39] Bar-Joseph Z, Demaine E, Gifford D, Srebro N, Hamel A, and Jaakkola T. "K -ary Clustering with Optimal Leaf Ordering for Gene Expression Data". In: *Bioinformatics* 19.9 (2003), pp. 1070–1078.

[40] Cheng CL, Kao YH, Lin SJ, Lee CH, and Lai ML. "Validation of the National Health Insurance Research Database with ischemic stroke cases in Taiwan". In: *Pharmacoepidemiol Drug Saf* 20.3 (2011), pp. 236–242. DOI: 10.1002/pds.2087.

[41] Cheng CL, Kao YH, Lin SJ, Lee CH, and Lai ML. "Validation of acute myocardial infarction cases in the national health insurance research database in Taiwan". In: *J Epidemiol* 24.6 (2014), pp. 500–507.

[42] Chen TJ, Chen YC, Hwang SJ, and Chou LF. "International collaboration of clinical medicine research in Taiwan, 1990–2004: A bibliometric analysis". In: *Journal of the Chinese Medical Association* 70.3 (2007), pp. 110–116.

[43] Luijks H, Schermer T, Bor H, Biermans M Weel CV Janssen TL, and Grauw WD. "Prevalence and incidence density rates of chronic comorbidity in type 2 diabetes patients: an exploratory cohort study". In: *BMC Medicine* 10.128 (2012). DOI: 10.1186/1741-7015-10-128.

[44] Metra M, Zacà V, Parati G, Agostoni P, Bonadies M, Ciccone M, Cas AD, and Iacoviello M et al. "Cardiovascular and noncardiovascular comorbidities in patients with chronic heart failure". In: *J Cardiovasc Med (Hagerstown)* 12.2 (2011). DOI: 10.2459/JCM.0b013e32834058d1.

[45] Leite AA, Costa AJ, Lima BA, Padilha AV, Albuquerque E, and Marques CD. "Comorbidities in patients with osteoarthritis: frequency and impact on pain and physical function". In: *Rev Bras Reumatol* 51.2 (2011).

[46] Cramer AO, Waldorp LJ, van der Maas HL, and Borsboom D. "Comorbidity: a network perspective". In: *Behav Brain Sci* 32.2 (2010), pp. 137–150. DOI: 10.1017/S0140525X09991567.

[47] Melamed RD, Emmett KJ, Madubata C, Rzhetsky A, and Rabadan R. "Genetic similarity between cancers and comorbid Mendelian diseases identifies candidate driver genes". In: *Nat Commun* 6.7033 (2015). DOI: 10.1038/ncomms8033.

[48] Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, and Barabási AL. "The implications of human metabolic network topology for disease comorbidity". In: *Proc Natl Acad Sci U S A* 105.29 (2008). DOI: 10.1073/pnas.0802208105.

[49] Zhou X, Menche J, Barabási AL, and Sharma A. "Human symptoms-disease network". In: *Nat Commun* 5.4212 (2014). DOI: 10.1038/ncomms5212.

[50] Valderas JM, Starfield B, Sibbald B, Salisbury C, and Roland M. "Defining comorbidity: implications for understanding health and health services". In: *Ann Fam Med* 7.4 (2009). DOI: 10.1370/afm.983.

[51] Younhee Ko, Minah Cho, Jin-Sung Lee, and Jaebum Kima. "Identification of disease comorbidity through hidden molecular mechanisms". In: *Sci Rep* 6.39433 (2016). DOI: 10.1038/srep39433.

[52] Johnson EO, Rhee SH, Chase GA, and Breslau N. "Comorbidity of depression with levels of smoking: an exploration of the shared familial risk hypothesis". In: *Nicotine Tob Res* 6.6 (2004).

[53] Rhodes ET, Laffel LM, Gonzalez TV, and Ludwig DS. "Accuracy of administrative coding for type 2 diabetes in children, adolescents, and young adults". In: *Diabetes Care* 30.1 (2007), pp. 141–143. DOI: 10.2337/dc06-1142.

[54] Williams C, Mosley-Williams A, and McDonald C. "Accuracy of provider generated computerized problem lists in the Veterans Administration". In: *AMIA Annu Symp Proc* 1155 (2007).

[55] Surján G. "Questions on validity of International Classification of Diseases-coded diagnoses". In: *Int J Med Inform* 54.2 (1999), pp. 77–95. DOI: http://dx.doi.org/10.1016/S1386-5056(98)00171-3.

[56] Feinstein AR. "The pre-therapeutic classification of co-morbidity in chronic disease". In: *J Chronic Dis* 23.7 (1970), pp. 455–468.

[57] Kaplan MH and Feinstein AR. "The importance of classifying initial co-morbidity in evaluatin the outcome of diabetes mellitus". In: *J Chronic Dis* 27.7 (1974), pp. 387–404. DOI: http://dx.doi.org/10.1016/0021-9681(74)90017-4.

[58] Angold A, Costello EJ, and Erkanli A. "Comorbidity". In: *J Child Psychol Psychiatry* 40.1 (1999), pp. 57–87.

[59] Piette JD and Kerr EA. "The impact of comorbid chronic conditions on diabetes care". In: *Diabetes Care* 29.3 (2006), pp. 725–731. DOI: https://doi.org/10.2337/diacare.29.03.06.dc05-2078.