Epidemiological Characterization of a Directed and Weighted Disease Network Using Data From a Cohort of One Million Patients: Network Analysis

Kyungmin Ko,1,2* MD; Chae Won Lee1,3*, BSc; Sangmin Nam4*, MD, PhD; Song Vogue Ahn2, MD, PhD; Jung Ho Bae6, MD; Chî Yong Ban1,3, BA; Jongman Yoo3,7, MD, PhD; Jungmin Park8, PhD; Hyun Wook Han1,3, MD, PhD

1Department of Biomedical Informatics, CHA University of Medicine, Seongnam, Republic of Korea
2Department of Pathology, Medstar Georgetown University Hospital, Washington, DC, WA, United States
3Institute of Basic Medical Sciences, School of Medicine, CHA University, Seongnam, Republic of Korea
4Department of Ophthalmology, CHA Bundang Medical Center, Seongnam, Republic of Korea
5Department of Health Convergence, Ewha Womans University, Seoul, Republic of Korea
6Department of Internal Medicine, Healthcare Research Institute, Seoul National University Hospital Healthcare System Gangnam Center, Seoul, Republic of Korea
7Department of Microbiology, CHA University School of Medicine, Seongnam, Republic of Korea
8Department of Nursing, School of Nursing, Hanyang University, Seoul, Republic of Korea

*these authors contributed equally

Corresponding Author:
Hyun Wook Han, MD, PhD
Department of Biomedical Informatics
CHA University of Medicine
Pangyo-ro, 335
Seongnam, KS009
Republic of Korea
Phone: 82 31 881 7109
Email: stepano7@gmail.com

Abstract

Background: In the past 20 years, various methods have been introduced to construct disease networks. However, established disease networks have not been clinically useful to date because of differences among demographic factors, as well as the temporal order and intensity among disease-disease associations.

Objective: This study sought to investigate the overall patterns of the associations among diseases; network properties, such as clustering, degree, and strength; and the relationship between the structure of disease networks and demographic factors.

Methods: We used National Health Insurance Service-National Sample Cohort (NHIS-NSC) data from the Republic of Korea, which included the time series insurance information of 1 million out of 50 million Korean (approximately 2%) patients obtained between 2002 and 2013. After setting the observation and outcome periods, we selected only 520 common Korean Classification of Disease, sixth revision codes that were the most prevalent diagnoses, making up approximately 80% of the cases, for statistical validity. Using these data, we constructed a directional and weighted temporal network that considered both demographic factors and network properties.

Results: Our disease network contained 294 nodes and 3085 edges, a relative risk value of more than 4, and a false discovery rate-adjusted P value of <.001. Interestingly, our network presented four large clusters. Analysis of the network topology revealed a stronger correlation between in-strength and out-strength than between in-degree and out-degree. Further, the mean age of each disease population was related to the position along the regression line of the out/in-strength plot. Conversely, clustering analysis suggested that our network boasted four large clusters with different sex, age, and disease categories.

Conclusions: We constructed a directional and weighted disease network visualizing demographic factors. Our proposed disease network model is expected to be a valuable tool for use by early clinical researchers seeking to explore the relationships among diseases in the future.
cohort studies; data science; longitudinal studies; statistical data interpretation; medical informatics

Introduction

Traditionally, clinical researchers have pushed forward to explore a number of risk factors that affect a single disease [1-3], and any diseases previously diagnosed are considered important clinical indicators to predict the disorder under investigation [4,5]. Among various methods for unearthing disease relationships, the concept of network medicine could be better suited to understand health and disease [6-8]. Likewise, a disease network was introduced a decade ago as a useful method to study the complex relationships among diseases [9-17].

Under the assumption that diseases are caused by genetic defects, many disease networks were constructed using genomic data [9,11,14]. For example, Li et al constructed a network to investigate disease relationships according to the genes of their shared pathways [14]. Nonetheless, according to the disease lists of the International Statistical Classification of Diseases, 10th revision (ICD-10), many diseases, such as traumatic bone fracture attributed to a traffic accident, are not related to genetic mutations. As such, genome-based disease networks alone are inevitably limited for accurately representing the complex pathogenesis of the relationships among diseases [18].

Thus, disease networks were later constructed using shared clinical information, such as symptoms and comorbidities [12,17]. Zhou et al generated a symptom-based network of human diseases that was based on the similarity of symptoms [17], whereas Hidalgo et al and Barabási et al constructed a comorbidity network using the Medicare database [7,12]. Because these efforts were focused on demonstrating the relationships among shared diseases or symptoms occurring or present at a single point in time, the networks did not take into account investigations of the temporal order of disease manifestations [19].

Recently, researchers have suggested that disease networks should consider temporal directionality when exploring the connections among diseases [13]. For instance, Jensen et al analyzed temporal disease progression patterns according to disease trajectory using the Danish National Patient Registry. In this study, we constructed a directional and weighted disease network visualizing the effects of demographic factors, such as sex, age, and disease outbreak size, according to the relative risk (RR) among diseases using the National Health Insurance Service-National Sample Cohort (NHIS-NSC) of South Korea, which includes epidemiological time series data of 12 years for approximately 1 million patients.

Finally, we investigated the overall patterns of the associations among diseases; network properties, such as clustering, degree, and strength; and the relationship between the structure of the disease network and demographic factors.

Methods

Construction and Visualization of the Disease Network

South Korea is a representative country implementing national health insurance services. The NHIS-NSC contains time insurance information of 1 million out of 50 million Korean (approximately 2%) patients, which was collected between 2002 and 2013. Thus, clinical information can be tracked for 12 years for every patient.

To examine the risk factors for diseases that a patient already had at the beginning of the cohort study, we needed to set an initial period before the main study period to serve as the medical history period. For most chronic diseases, the recommended follow-up interval rarely exceeds 2 years. Therefore, we set the observation period as 2002 through 2003 and the outcome period as 2004 through 2013.

From the sample of 1,016,580 patients who were eligible for National Health Insurance in 2004, we selected 885,125 patients who had at least one record of a medical visit during the aforementioned observation period. We defined this group of patients as the sample cohort. In South Korea, diagnoses are coded in the Korean Classification of Diseases sixth revision (KCD-6), an extension of the ICD-10. The only difference between the KCD-6 and ICD-10 is that the diagnosis codes for Korean medicine are included in the KCD-6 using U20-U99 codes.

To simplify the study, we truncated the KCD-6 codes beyond their third digit, in effect, grouping subcategories of conditions together. In total, the KCD-6, when used between 2002 and 2013, consisted of 2,097 unique diagnoses at the third digit level, and of these, 1,971 diagnoses were included in our data. Ultimately, we chose only 520 common KCD-6 codes that were the most prevalent diagnoses, covering approximately 80% of the cases for statistical validity.

Support Offered by the Clinical Evidence From Relationships Among Diseases

All statistical analyses and visualizations were performed using the R package “igraph” (version 3.4.4) and Cytoscape. For calculation of the RR, we sought to obtain P values against the null hypothesis, which states that any two diseases present occur independently of one another in the sample cohort. False-discovery rate (FDR) corrections were performed using the Bonferroni method.

Clusters of associated diseases were identified using the random walktrap community detection algorithm [20,21]. This method detects clusters purely according to connectivity (unless specified to use weights) using random walks along edges. The demographic profiling of disease clusters was carried out by pooling the patients identified with at least one of the diagnoses in the cluster.
As a result, patient pools for each cluster are not exclusive but instead overlap somewhat with other clusters. The age distribution of the patient pools was calculated at the beginning of the observation period. An enrichment analysis of the clusters for the KCD categories was performed using the Fisher exact test for adjusted P values <.05.

**Topological Characteristics of the Disease Network**

In graph theory, the degree of a node is the total number of connections with other nodes. In a directed network, the out-degree of a node is the number of connections with that node as the source, whereas the in-degree of a node is the number of connections with that node as the target. Hence, the degree can be thought of as a measure of the level of disease risk in our network.

In contrast, the strength of a node is the sum of the RRs to achieve connections with other nodes. For example, the out-strength and in-strength of node i are defined, respectively, as follows:

\[ s_{out}(i) = \sum_j RR_{ij} \]  
\[ s_{in}(i) = \sum_j RR_{ji} \]

where \( RR_{ij} \) is the weight of the edge from node i to node j, and \( RR_{ji} \) is the weight of the edge from node j to node i. The out-strength is a measure of the magnitude of disease morbidity, whereas the in-strength is a measure of the magnitude of a disease’s tendency to follow from other diseases.

**Characterization of Large Clusters Throughout Computational Clustering**

To calculate the risk ratio from a risk disease \( D_1 \) to an outcome disease \( D_2 \) (\( D_1 \rightarrow D_2 \)), we need to first identify the group of patients at risk of acquiring \( D_2 \). We regarded a patient as being at risk of disease \( D_2 \) if that patient had no record of being diagnosed with \( D_2 \) during the observation period. Patients were considered to be exposed if they had been diagnosed at least once with disease \( D_1 \) during the observation period. The RR of \( D_1 \rightarrow D_2 \) was defined using the following formula:

\[ RR = \frac{a / (a + b)}{c / (c + d)} \]

where \( a \) is the number of patients exposed to \( D_1 \) in the initial period and \( D_2 \) in the outcome period; \( b \) is the number of patients exposed to \( D_1 \) in the initial period but not exposed to \( D_2 \) in the outcome period; \( c \) is the number of patients not exposed to \( D_1 \) in the initial period but exposed to \( D_2 \) in the outcome period; and \( d \) is the number of patients not exposed to either \( D_1 \) in the initial period or \( D_2 \) in the outcome period (Table 1).

Since a single misdiagnosis can cause a very large error in the RR value if the numbers in the contingency table are small, we established a minimum size of 947 patients for each group. For example, the diagnosis with the highest prevalence in the initial period was “J20: acute bronchitis,” with 355,045 patients diagnosed at least once in the observation period.

The lowest diagnosis was “R80: isolated proteinuria,” with 947 patients diagnosed during the observation period. Consequently, the at-risk group sizes ranged from 530,080 (885,125 – 355,045) for acute bronchitis to 884,178 (885,125 – 947) for isolated proteinuria.

To select the cut-off value for the RR, we chose the closest integer to the top percentile (ie, the closest integer to x where \( P \left[ RR > x \right] = .01 \)), which was 4. Therefore, we selected disease relationships with an RR of more than 4 and an FDR-corrected P value of <.001 to construct our final network.

Accordingly, the prevalence and at-risk group sizes were large enough to accurately determine the RR. Since the self-interaction in this study was not the subject, the total number of theoretical interactions of a total of 520 nodes was found to be 269,880.

**Results**

**Construction and Visualization of the Disease Network**

Initially, for the construction and visualization of our final disease network, we selected an RR of more than 4 and an FDR-adjusted \( P \) value of <.001. As a result, we were able to obtain a disease network with four clusters, 294 nodes, and 3085 edges (Figure 1). For better clinically intuitive visualization, we designed a visualization scheme such that the color of the disease node would reflect the age of the patient affected with the disease and that the outbreak size would reflect the relative number of patients. The shape of the node was indicated by a rectangle. Node widths represented the number of male patients, whereas the heights represented the number of female patients. For node colors, the intensity of the red channel was proportional to the ratio of patients younger than 30 years, the intensity of the green channel was proportional to the ratio of patients more than 65 years, and the intensity of the blue channel was proportional to the remaining patients.
channel was proportional to the ratio of patients aged between 30 and 59 years, and the intensity of the blue channel was proportional to the ratio of patients aged 60 years or older.

Meanwhile, to indicate the directionality and weight of the node, the edges were represented by arrows and the relative thickness was represented in gray. The number of patients in each sex and age group for each disease was calculated at the beginning of the outcome period based on the included patients’ histories during the observation period.

**Figure 1.** Visualization of the disease network. The network is constructed using the prefuse force-directed layout.

---

**Support Offered by the Clinical Evidence From Relationships Among Diseases**

To determine whether the correlations among inferred diseases in our disease network model were clinically meaningful, we investigated the available literature concerning the top six disease-disease associations with the highest RR values from the established disease network.

Most results appeared in agreement with previously known associations among diseases (Table 2). A substantial nosologic and biologic overlap exists between bipolar disorder and schizophrenia [22,23]. Further, long-standing hypertension is known to be an important cause and consequence of chronic kidney disease [24]. It is also well-known that anemia develops into chronic kidney disease and portends an unfavorable prognosis [25].

Interestingly, the association between diabetes mellitus in pregnancy and neonatal jaundice was also very high, despite the fact that neonates are never pregnant. This outcome is possible because diagnoses for infants who are not yet in the national registry are filled out under the mother’s account for insurance [26-28]. Another interesting aspect of our results was the fact that neonatal jaundice and diaper dermatitis were strongly associated with one another, which was not observed in previous epidemiological studies.
Table 2. Top relative risk values.

| Risk factor disease                          | Outcome disease                  | RR\(^a\) | References |
|---------------------------------------------|----------------------------------|----------|------------|
| Bipolar affective disorder                  | Schizophrenia                    | 34.4     | [22,23]    |
| Chronic kidney disease                      | Hypertensive renal disease       | 31.9     | [24]       |
| Diabetes mellitus in pregnancy              | Neonatal jaundice                | 29.1     | [26-28]    |
| Neonatal jaundice                           | Diaper dermatitis                | 28.1     | N/A\(^b\) |
| Chronic kidney disease                      | Anemia in chronic disease        | 27.4     | [25]       |
| Hemorrhage in early pregnancy               | Neonatal jaundice                | 26.1     | [27]       |

\(^a\) RR: relative risk.

\(^b\) N/A: not applicable.

**Topological Characteristics of the Disease Network**

We investigated the in- and out-degree distributions of our constructed network. Like many other networks, the in- and out-degrees of our network followed a power-law distribution with a long tail [29] (Figure 2). However, in contrast with the degrees, neither in- nor out-strength followed the power-law distribution (Figure 3).

**Figure 2.** Distribution of the network for in- and out-degrees.

**Figure 3.** Distribution of the network for in- and out-strength.
Table 3 shows the top six diseases with the highest out-degree, in-degree, out-strength, and in-strength results. The top out-degree diseases included diseases that are known to affect many other conditions such as chronic kidney disease and essential hypertension.

The top in-degree diseases are known to be associated with long-term hospitalization or immunocompromise, which are statuses that can arise from various diseases. The top out-degree and top out-strength diseases had considerable overlap, with polyneuropathy, senile cataract, and retinal disorders all being both high out-degree and high out-strength diseases. Patients with these diseases may be at greater risk for developing multiple comorbidities.

The top in-degree and top in-strength diseases included Parkinson disease, chronic kidney disease, anemia in chronic disease, and osteoporosis with pathological fracture. This suggests that many different diseases can have a strong tendency for coverage onto these diseases. Subsequently, we explored the relationships between out-degree and in-degree and between out-strength and in-strength results.

The correlation between the out-strength and in-strength findings (Pearson correlation coefficient: 0.72) was stronger than that between the out-degree and in-degree findings (Pearson correlation coefficient: 0.57) (Figure 4). This means that diseases show strong tendencies to develop from other diseases. For better characterization, we color-coded the diseases in the out-/in-strength plot according to the age composition of the patients (Figure 5). This revealed that mean age was related with positioning along the regression line of the out-/in-strength plot.

Table 3. Top out/in-degree diseases and top out/in-strength diagnoses.

| KCD\(^a\) code and disease                      | Degree |
|------------------------------------------------|--------|
| **Top out-degree diseases**                    |        |
| G63: polyneuropathy                           | 43     |
| C61: malignant neoplasm of the prostate       | 43     |
| H25: senile cataract                           | 43     |
| H36: retinal disorders                        | 42     |
| N18: chronic kidney disease                   | 42     |
| I10: essential hypertension                   | 39     |
| **Top in-degree diseases**                    |        |
| G20: Parkinson disease                        | 82     |
| M80: osteoporosis with pathological fracture  | 64     |
| N18: chronic kidney disease                   | 62     |
| D63: anemia in chronic diseases               | 61     |
| A41: sepsis                                   | 59     |
| **Top out-strength diseases**                 |        |
| G63: polyneuropathy                           | 1057   |
| H36: retinal disorders                        | 998    |
| M48: spondylopathies                          | 992    |
| H25: senile cataract                           | 992    |
| M81: osteoporosis without pathological fracture| 981    |
| M17: arthrosis of the knee                    | 979    |
| **Top in-strength diseases**                  |        |
| G20: Parkinson disease                        | 1197   |
| N18: chronic kidney disease                   | 1135   |
| D63: anemia in chronic diseases               | 1123   |
| M80: osteoporosis with pathological fracture  | 1120   |
| I12: hypertensive renal disease               | 1100   |
| H27: disorder of the lens                     | 1089   |

\(^{a}\)KCD: Korean Classification of Diseases.
Characterization of Large Clusters Throughout Computational Clustering

To confirm whether the visually observed clustering in Figure 1 was an artifact of the layout algorithm, we employed a random walktrap algorithm for network clustering [20,21]. A total of 19 clusters were detected, including four large clusters of a size greater than 38 and 15 small clusters of a size less than 13. When we color-coded the network using these four major clusters, we could see that the top right and top left clusters were almost exactly as visualized, but the largest cluster was detected as two large subclusters (Figure 6).

This confirmed that disease associations grouped diseases into a few distinct clusters and that this occurred independently of the prefuse force-directed layout. Interestingly, the modularity score for the random walktrap algorithm (0.53) was more than twice the score for the KCD categories (0.24). To see whether the four major clusters actually had the characteristics that we noticed in the visualization, we profiled the clusters with respect to the age distribution and sex ratio of the affected patients (Figure 7).

Patients diagnosed with diseases in clusters 1 and 3 were relatively older (mean age of 47.4 [SD 18.22] years and 48.19 [SD 18.66] years, respectively). The diseases in cluster 2 were...
dominated by women of reproductive age (the ratio of males to females was 1:18.67; mean age: 39.38 [SD 13.08] years). Cluster 4 included patients who were relatively young, with slightly more females (the ratio of males to females was 1:1.22; mean age: 31.7 [SD 21.56] years).

We profiled the KCD classes of each cluster and performed an enrichment analysis to investigate the types of diseases that were enriched in each cluster (P=.05). Although every cluster contained its own disease groups (Multimedia Appendix 1), the enrichment analysis revealed that each of the four major clusters was enriched with nonoverlapping sets of KCD categories (Multimedia Appendix 2). Since each cluster had distinct characteristics, we labeled the major clusters from 1 to 4, according to their most prominent features, as “chronic debilitation,” “women’s disease,” “hemato-oncology,” and “infectious disease” clusters, respectively.

**Figure 6.** Four major clusters of the network.

Figure 7. Age distribution and sex ratio for the four major clusters.

**Discussion**

**Principal Findings**

In this study, we proposed a comprehensive method for modeling a disease network with directionality and weight of edges using medical claims data. We selected only the most common diagnoses to avoid an overestimation of RR among rare diseases. The ϕ correlation is also useful to avoid overestimation of associations among rare diseases [12], but it is clinically less intuitive and unnecessary for the purpose of studying the overall pattern of common disease associations. Epidemiological factors, such as age and sex, are important inducers of disease development [30-33]; they are, in effect, the most critical clinical factors affecting the prevalence and classification of diseases.

Another purpose of this study was to dissolve these various factors in the disease network and to see how various factors
affect the structure and dynamics of the disease network. In addition, these factors were reflected in the visualization of the disease network. In our disease network model, we proposed an intuitive visualization method that maximizes clinical usability.

Nodes indicate the patient outbreak size, and at the same time, represent the relative proportion of width (women) and height (men) in a rectangle. In addition, each node is divided into red for young patients, green for middle-aged patients, and blue for old patients. Conversely, the RR and direction among diseases can be intuitively grasped through the arrow and the thickness of the edge.

As a result, our visualization method of a disease network can help to intuitively identify the direction and RR among diseases and can help to effectively understand the age distribution, sex ratio, and disease outbreak size. The directionality of the disease relationship is a consequence of the study design being longitudinal with a chronological order. Strong RR values support disease association in a chronological order, which is a prerequisite for causality among diseases in clinical research [34]. Because of this, our network can be a starting point to investigate causality among diseases. Here, we examined the literature on disease relationships with high RRs.

Limitations

The NHIS-NSC includes the proportional stratified sampling data of 1,025,340 patients from among 47,851,928 patients. These patients were randomly extracted by age group, sex, eligibility status, and income level using a proportional stratified sampling method [35]. In general, NHIS-NSC data are representative, but some rare diseases may lose their representativeness owing to the difficulty in obtaining statistical significance. Owing to these limitations, this study excluded disease groups with a small sample size, and rare diseases that have not been assessed in this study are expected to be evaluated in future studies.

As a result, we have shown that our network can provide clues to reveal the causal relationships among diseases. In our network, neonatal jaundice and diaper dermatitis presented a statistically meaningful association (RR=28.1, P<.001), but we did not find other supporting evidence of such an association during the literature search. Nonetheless, this does not mean that our network had an incorrect result. Rather, it suggests the possibility of associations that researchers have not yet discovered. In a strict sense, it can be difficult to say that this is a causal relationship because directionality can only be thought of as the natural progression of a disease, the outcome of a treatment, or the process of making a diagnosis.

Conclusions

During our research, we investigated network topologies, such as degree and strength. Both in- and out-degrees followed power-law distribution like other biological networks; however, strength distributions did not. Since the RR values did not indicate causality, we cannot say that a certain disease is the cause of many other diseases by only looking at out-degrees and out-strengths. Despite this, patients with high out-degree and high out-strength diseases (eg, polyneuropathy, retinal disorders, and senile cataract) are worthy of special attention for secondary prevention purposes. Similarly, diseases with high in-degree and high in-strength findings, such as Parkinson disease, osteoporosis with pathological fracture, and chronic kidney disease, can be seen as common comorbidities of many different diseases.

We found stronger correlations between in- and out-strengths than between in- and out-degrees. Moreover, a stronger risk associated with a disease tended to be related to older affected patients. The association between age and strength suggested that the previously discovered correlation between disease connectivity and mortality could be explained by the phenomenon of increased risk strength.

Through clustering of the network, we found four major disease clusters with distinct demographic characteristics. Interestingly, each cluster was exclusively enriched in KCD categories and had a different mean age and sex ratio. The clustering patterns analyzed using our network suggest that KCD categories, age, and sex have strong influences on disease associations and highlight the importance of demographic factors. Since patients with diseases within a cluster tend to acquire other diseases within the same cluster, we may be able to minimize the onset of comorbidities through patient care by configuring specialty clinics to cater to clusters or subclusters of associated diseases, as is the case with obstetrics and gynecology.

In this regard, our proposed disease network model will likely serve as a valuable tool for early clinical researchers seeking to further explore the relationships of diseases in the future.

For future study attempts, we will take into account the dynamicity of network-considered time order and assess the network collapse point that can affect the overall network structure.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1E1A1A03070934) and the Ministry of Science & ICT (NRF-2019M3C7A1032262).

Authors’ Contributions

HWH conceived the project; HWH, SVA, JHB, and NSM designed the research; KMK and JMP performed the research; KMK and JMY analyzed the data; and CWL, CYB, and HWH wrote the paper.
Conflicts of Interest
None declared.

Multimedia Appendix 1

Korean Classification of Diseases composition of disease nodes in each cluster. The composition of each disease cluster in terms of disease categories is shown in this bar chart.
[DOCX File, 238 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Enrichment analyses revealed that each of the four major communities was enriched with nonoverlapping sets of Korean Classification of Diseases categories.
[DOCX File, 14 KB-Multimedia Appendix 2]

References

1. Mooe T, Björklund F, Graipe A, Huber D, Jakobsson S, Kajermo U, et al. The Nurse-Based Age Independent Intervention to Limit Evolution of Disease After Acute Coronary Syndrome (NAILED ACS) Risk Factor Trial: Protocol for a Randomized Controlled Trial. JMIR Protoc 2014 Aug 15;3(3):e42 [FREE Full text] [doi: 10.2196/protoc.3466] [Medline: 25131960]

2. Peng S, Shen F, Wen A, Wang L, Fan Y, Liu X, et al. Detecting Lifestyle Risk Factors for Chronic Kidney Disease With Comorbidities: Association Rule Mining Analysis of Web-Based Survey Data. J Med Internet Res 2019 Dec 10;21(12):e14204 [FREE Full text] [doi: 10.2196/14204] [Medline: 31821152]

3. Tluway F, Urio F, Mmbando B, Sangeda RZ, Makubi A, Makani J. Possible Risk Factors for Severe Anemia in Hospitalized Sickle Cell Patients at Muhimbili National Hospital, Tanzania: Protocol for a Cross-Sectional Study. JMRI Res Protoc 2018 Feb 28;7(2):e46 [FREE Full text] [doi: 10.2196/protoc.7349] [Medline: 29490896]

4. Linet M, Vajdic C, Morton L, de Roos AJ, Skibola C, Boffetta P, et al. Medical history, lifestyle, family history, and occupational risk factors for follicular lymphoma: the InterLymph Non-Hodgkin Lymphoma Subtypes Project. J Natl Cancer Inst Monogr 2014 Aug;2014(48):26-40 [FREE Full text] [doi: 10.1093/jncimonographs/lgu006] [Medline: 25174024]

5. Silverman DT, Schiffman M, Everhart J, Goldstein A, Lillemoe KD, Swanson GM, et al. Diabetes mellitus, other medical conditions and familial history of cancer as risk factors for pancreatic cancer. Br J Cancer 1999 Aug;80(11):1830-1837 [FREE Full text] [doi: 10.1038/sj.bjc.6690607] [Medline: 10468306]

6. Tillmann T, Gibson AR, Scott G, Harrison O, Dominiczak A, Hanlon P. Systems Medicine 2.0: potential benefits of combining electronic health care records with systems science models. J Med Internet Res 2015 Mar 23;17(3):e64 [FREE Full text] [doi: 10.2196/jmir.3082] [Medline: 25831125]

7. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011 Jan;12(1):56-68 [FREE Full text] [doi: 10.1038/nrg2918] [Medline: 21645525]

8. Loscalzo J, Barabasi A. Systems biology and the future of medicine. Wiley Interdiscip Rev Syst Biol Med 2011;3(6):619-627 [FREE Full text] [doi: 10.1002/wsbm.144] [Medline: 21928407]

9. Al-Harazi O, Al Insai S, Al-Aljan MA, Kaya N, Dzimir N, Colak D. Integrated Genomic and Network-Based Analyses of Complex Diseases and Human Disease Network. J Genet Genomics 2016 Jun 20;43(6):349-367. [doi: 10.1016/j.jgg.2015.11.002] [Medline: 27318646]

10. Davis DA, Chawla NV. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. PLoS One 2011;6(7):e22670 [FREE Full text] [doi: 10.1371/journal.pone.0022670] [Medline: 21829475]

11. Goh K, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci U S A 2007 May 22;104(10):5685-5690 [FREE Full text] [doi: 10.1073/pnas.0701361104] [Medline: 17502601]

12. Hidalgo CA, Blumm N, Barabási AL, Christakis NA. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol 2009 Apr;5(4):e1000353 [FREE Full text] [doi: 10.1371/journal.pcbi.1000353] [Medline: 19360091]

13. Jensen AB, Moseley PL, Oprea TI, Ellesæe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun 2014 Jun 24;5:4022 [FREE Full text] [doi: 10.1038/ncomms5022] [Medline: 24959948]

14. Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. PLoS One 2009;4(2):e4346 [FREE Full text] [doi: 10.1371/journal.pone.0004346] [Medline: 19194489]

15. Park J, Lee D, Christakis NA, Barabási AL. The impact of cellular networks on disease comorbidity. Mol Syst Biol 2009;5:262 [FREE Full text] [doi: 10.1038/msb.2009.16] [Medline: 19357641]

16. Zhang X, Zhang R, Jiang Y, Sun P, Tang G, Wang X, et al. The expanded human disease network combining protein-protein interaction information. Eur J Hum Genet 2011 Jul;19(7):783-788 [FREE Full text] [doi: 10.1038/ejhg.2011.30] [Medline: 21836875]

17. Zhou X, Menche J, Barabási AL, Sharma A. Human symptoms-disease network. Nat Commun 2014 Jun 26;5:4212. [doi: 10.1038/ncomms5212] [Medline: 24967666]
18. Roberts NJ, Vogelstein JT, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE. The predictive capacity of personal genome sequencing. Sci Transl Med 2012 May 09;4(133):133ra58 [FREE Full text] [doi: 10.1126/scitranslmed.3003380] [Medline: 22472521]

19. Gross F. What systems biology can tell us about disease. Hist Philos Life Sci 2011;33(4):477-496. [Medline: 22662506]

20. Pons P, Latapy M. Computing communities in large networks using random walks. In: Computer and Information Sciences. Berlin, Heidelberg: Springer; 2005 Presented at: 20th International Conference on Computer and Information Sciences; October 26-28, Istanbul, Turkey p. 284-293. [doi: 10.1007/11569596_31]

21. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci U S A 2008 Jan 29;105(4):1118-1123 [FREE Full text] [doi: 10.1073/pnas.0706851105] [Medline: 18216267]

22. Laursen TM, Agerbo E, Pedersen CB. Bipolar disorder, schizoaffective disorder, and schizophrenia overlap: a new comorbidity index. J Clin Psychiatry 2009 Oct;70(10):1432-1438. [doi: 10.4088/JCP.08m04807] [Medline: 19538905]

23. Pearlson GD. Etiologic, phenomenologic, and endophenotypic overlap of schizophrenia and bipolar disorder. Annu Rev Clin Psychol 2015;11:251-281. [doi: 10.1146/annurev-clinpsy-032814-112915] [Medline: 25581236]

24. Gargiulo R, Suhail F, Lerma EV. Hypertension and chronic kidney disease. Dis Mon 2015 Sep;61(9):387-395. [doi: 10.1016/j.disamonth.2015.07.003] [Medline: 26328515]

25. Iseki K, Kohagura K. Anemia as a risk factor for chronic kidney disease. Kidney Int Suppl 2007 Nov(107):S4-S9 [FREE Full text] [doi: 10.1038/sj.ki.5002481] [Medline: 17943141]

26. Alam M, Raza SJ, Sherali AR, Akhtar SM. Neonatal complications in infants born to diabetic mothers. J Coll Physicians Surg Pak 2006 Mar;16(3):212-215. [Medline: 16542603]

27. Wang Z, Kanguru L, Hussein J, Fitzmaurice A, Ritchie K. Incidence of adverse outcomes associated with gestational diabetes mellitus in low- and middle-income countries. Int J Gynaecol Obstet 2013 Apr;121(1):14-19. [doi: 10.1016/j.ijgo.2012.10.032] [Medline: 23321368]

28. Yang J, Cummings EA, O'connell C, Jangaard K. Fetal and neonatal outcomes of diabetic pregnancies. Obstet Gynecol 2006 Sep;108(3 Pt 1):644-650. [doi: 10.1097/01.AOG.0000231688.08263.47] [Medline: 16946226]

29. Albert R. Scale-free networks in cell biology. J Cell Sci 2005 Nov 01;118(Pt 21):4947-4957 [FREE Full text] [doi: 10.1242/jcs.02714] [Medline: 16254242]

30. Cohen P, Cohen J, Kasen S, Veliz CN, Hartmark C, Johnson J, et al. An epidemiological study of disorders in late childhood and adolescence--I. Age- and gender-specific prevalence. J Child Psychol Psychiatry 1993 Sep;34(6):851-867. [doi: 10.1111/j.1469-7610.1993.tb01094.x] [Medline: 8408371]

31. Farrer L, Cupples L, Haines J, Hyman B, Kukull W, Mayeux R, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. JAMA 1997;278(16):1349-1356. [Medline: 9343467]

32. Meisinger C, Thorand B, Schneider A, Stieber J, Döring A, Löwel H. Sex differences in risk factors for incident type 2 diabetes mellitus: the MONICA Augsburg cohort study. Arch Intern Med 2002 Jan 14;162(1):82-89. [doi: 10.1001/archinte.162.1.82] [Medline: 11784224]

33. Piccinelli M, Wilkinson G. Gender differences in depression. Critical review. Br J Psychiatry 2000 Dec;177:486-492. [doi: 10.1192/bjp.177.6.486] [Medline: 11102321]

34. Bang S, Kim J, Shin H. Causality modeling for directed disease network. Bioinformatics 2016 Sep 01;32(17):i437-i444. [doi: 10.1093/bioinformatics/btw439] [Medline: 27587660]

35. Lee J, Lee JS, Park S, Shin SA, Kim K. Cohort Profile: The National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. Int J Epidemiol 2017 Apr 01;46(2):e15. [doi: 10.1093/ije/dyw319] [Medline: 26822938]

Abbreviations

| Abbreviation | Description |
|--------------|-------------|
| FDR | false-discovery rate |
| ICD-10 | International Statistical Classification of Diseases, 10th version |
| KCD-6 | Korean Classification of Diseases, sixth revision |
| NHIS-NSC | National Health Insurance Service-National Sample Cohort |
| RR | relative risk |
©Kyungmin Ko, Chae Won Lee, Sangmin Nam, Song Vogue Ahn, Jung Ho Bae, Chi Yong Ban, Jongman Yoo, Jungmin Park, Hyun Wook Han. Originally published in the Journal of Medical Internet Research (http://www.jmir.org), 09.04.2020. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on http://www.jmir.org/, as well as this copyright and license information must be included.