Journal of Biomedical Informatics 61 (2016) 276–282

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Network analysis of human diseases using Korean nationwide claims data

Jin Hee Kim [a], Ki Young Son [a], Dong Wook Shin [a], Sang Hyuk Kim [a], Jae Won Yun [b,c], Jung Hyun Shin [a], Mi So Kang [a], Eui Heon Chung [a], Kyoung Hun Yoo [a], Jae Moon Yun [a,*]

[a] Department of Family Medicine & Health Promotion Center, Seoul National University Hospital, Seoul, Republic of Korea
[b] Samsung Genome Institute, Samsung Medical Center, Seoul, Republic of Korea
[c] Department of Molecular Cell Biology, Sungkyunkwan University School of Medicine, Suwon, Republic of Korea

ARTICLE INFO

Article history:
Received 8 March 2016
Revised 22 April 2016
Accepted 10 May 2016
Available online 11 May 2016

Keywords:
Claims data
Network analysis
Disease–disease association
Human disease network
Data-driven analysis

ABSTRACT

Objective: To investigate disease–disease associations by conducting a network analysis using Korean nationwide claims data.

Methods: We used the claims data from the Health Insurance Review and Assessment Service-National Patient Sample for the year 2011. Among the 2049 disease codes in the claims data, 1154 specific disease codes were used and combined into 795 representative disease codes. We analyzed for 381 representative codes, which had a prevalence of >0.1%. For disease code pairs of a combination of 381 representative disease codes, *P* values were calculated by using the χ^2 test and the degrees of associations were expressed as odds ratios (ORs).

Results: For 5515 (7.62%) statistically significant disease–disease associations with a large effect size (OR > 5), we constructed a human disease network consisting of 369 nodes and 5515 edges. The human disease network shows the distribution of diseases in the disease network and the relationships between diseases or disease groups, demonstrating that diseases are associated with each other, forming a complex disease network. We reviewed 5515 disease–disease associations and classified them according to underlying mechanisms. Several disease–disease associations were identified, but the evidence of these associations is not sufficient and the mechanisms underlying these associations have not been clarified yet. Further research studies are needed to investigate these associations and their underlying mechanisms.

Conclusion: Human disease network analysis using claims data enriches the understanding of human diseases and provides new insights into disease–disease associations that can be useful in future research.

1. Introduction

To understand the nature of diseases, many research studies have been conducted to determine the pathogenesis, course, and complications of diseases. Usually, these research studies are based on the discoveries of new evidence or insights about the associations between two disease conditions. So far, investigations on the associations between disease conditions were generally conducted with hypotheses based on up-to-date evidence of two diseases. Then, researchers collected data to support their hypotheses [1–4]. Recently, medical research has undergone significant changes from hypothesis- to data-driven analysis, led by advances in computational technology, development of new analysis methods, and an increase in health and medical data. In data-driven analysis, researchers analyze data first and then search for patterns or obtain insights from data to form hypotheses.

Although claims data or electronic health records are being generated and accumulated rapidly, classical analysis methods have limitations to analyze these large data comprehensively. Advanced computing power, network analysis, and network visualization enable researchers to analyze these data and understand disease–disease associations [5,6].

Recently, several studies were conducted to investigate the disease–disease associations by using large-scale claims data or electronic patient records and by applying new analysis methods. Hidalgo et al. used health insurance claims data of hospitalized individuals to construct a phenotypic disease network that

* Corresponding author at: Department of Family Medicine & Health Promotion Center, Seoul National University Hospital, 101 Daehakro, Jongnogu, Seoul 110-744, Republic of Korea.
E-mail address: jaemoon2@gmail.com (J.M. Yun).

http://dx.doi.org/10.1016/j.jbi.2016.05.002
1532-0464/

summarizes connections between diseases [7]. Roque et al. used text mining from electronic patient records to cluster the patients based on their phenotypic profile similarities and investigate disease comorbidities by identifying co-occurring disease codes in each patient [8].

We constructed a human disease network to investigate disease–disease associations by using Korean nationwide claims data and performing a network analysis.

2. Materials and methods

2.1. Source data and study population

South Korea has a National Health Insurance System (NHIS), which is a compulsory social insurance. NHIS covers approximately 98% of the population living in South Korea, except the low-income population. The Health Insurance Review and Assessment Service (HIRA) conducts reviews and assessments of medical costs and quality of services. The HIRA collects claims data of 46 million patients, which accounts for 90% of the population in South Korea.

The Health Insurance Review and Assessment Service-National Patient Sample (HIRA-NPS) is 3% of the national patient sample data covering all age groups. The sample data were extracted by using a stratified randomized sampling method [9]. We used claims data of 1,375,842 individuals from the HIRA-NPS in the year 2011.

2.2. Definition of diseases

The diseases are defined by using the disease codes in the claims data. The disease codes in the claims data are encoded according to the *Korean Standard Classification of Diseases-6* (*KCD-6*), which is a modification of the *International Statistical Classification of Disease and Related Health Problems, 10th Revision* (*ICD-10*), suited to Korean medical circumstances. Because *KCD-6* includes several disease codes for oriental medicine, we excluded these disease codes, as they are not included in the *ICD-10*. The *ICD-10* consists of 5 digits, and we used the first 3 digits, which comprise the main disease category. Among the 2049 *ICD-10* codes, those that belong to the categories "Symptoms, signs, and abnormal clinical and laboratory findings (R00–R99)," "Injury, poisoning, and certain other consequences of external causes (S00–T98)," "External causes of morbidity and mortality (V01–Y98)," "Factors influencing health status and contact with health services (Z00–Z99)," and "Codes for special purposes (U00–U85)" were excluded because they are not codes for specific diseases. Codes in the categories "Certain conditions originating in the perinatal period (P00–P96)" and "Congenital malformations, deformations, and chromosomal abnormalities (Q00–Q99)" were excluded because of their low prevalence. The remaining 1154 disease codes in the "A–O" disease categories were reviewed, and similar disease codes or disease codes that can be categorized into one broader disease category were combined into one representative disease code so that clinically important rare diseases, such as cancers, were not excluded. We considered the prevalence and clinical importance of the individual diseases when combining them. In particular, we took into consideration the lower prevalence of specific disease subtypes when combining diseases into broader representative disease codes, since their low prevalence means they can be excluded. For example, we used the bacterial intestinal infection representative disease code (A04) for cholera (A00), typhoid and paratyphoid fevers (A01), *Salmonella* infections (A02), and shigellosis (A03). Anemia (D64) was used as a representative disease code for several subtypes of anemia, including iron deficiency anemia (D50), vitamin B12 deficiency anemia (D51), and thalassemia (D56). We also combined disease codes according to similar anatomical localizations; for example, malignant neoplasm of colon (C18), rectosigmoid junction (C19), rectum (C20), and the anus and anal canal (C21) were combined into the representative disease code for colorectal cancer (C18). However, clinically important diseases or those with a high prevalence, such as acute hepatitis A (B15) and B (B16), or malignant neoplasm of stomach (C16) or esophagus (C15), were not combined into representative disease codes. As a result, 1154 disease codes were combined into 795 representative disease codes, and among these, we analyzed 381 representative codes that had a prevalence of more than 0.1% (Fig. 1).

2.3. Statistical analysis

P values were calculated for a combination of 381 disease codes using the χ^2 test with Bonferroni correction; the cutoff was set to $P < 1.38 \times 10^{-7}$. The degrees of the associations were expressed in odds ratios (ORs).

$$\mathrm{OR} = \frac{A_B/NA_B}{A_{NB}/NA_{NB}} \quad (1)$$

A_B: Population having disease codes A and B.
A_{NB}: Population having disease code A but not B.
NA_B: Population having disease code B but not A.
NA_{NB}: Population not having disease code A nor B.

Statistical analyses were performed by using Stata version 14.0 (StataCorp, Texas, USA).

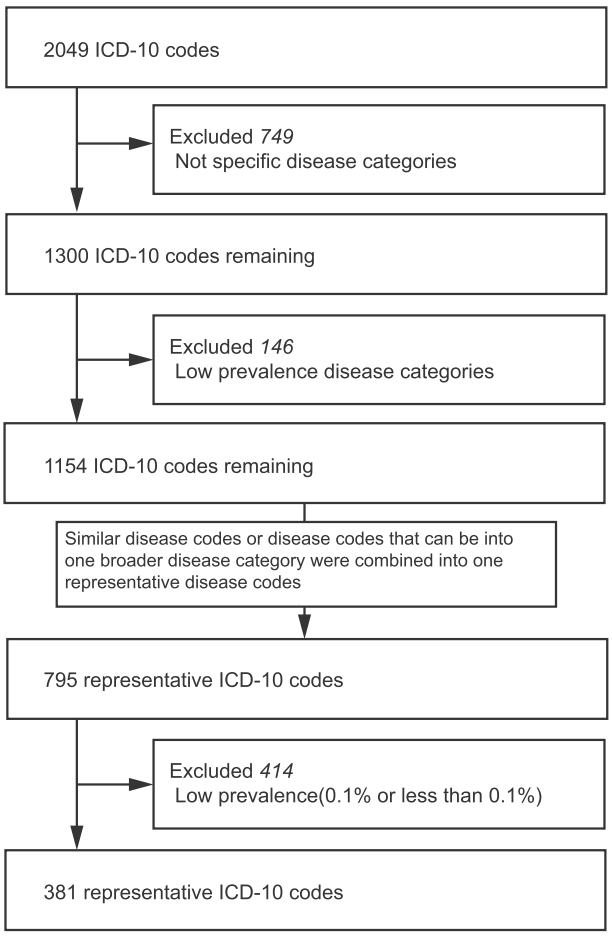

Fig. 1. Flow diagram of the representative disease codes selected.