International Journal of Medical Informatics 108 (2017) 22–28

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Research Paper

# Examining health disparities by gender: A multimorbidity network analysis of electronic medical record

Pankush Kalgotra[a,*], Ramesh Sharda[b], Julie M. Croff[c]

[a] Graduate School of Management, Clark University, 950 Main Street (Carlson Hall), Worcester, 01610 MA, USA
[b] Spears School of Business, Oklahoma State University, 102 Gunderson, Stillwater, 74078 OK, USA
[c] Oklahoma State University, 429 Willard Hall, Stillwater, 74078 OK, USA

ARTICLE INFO

*Keywords:*
Gender disparity
Multimorbidity
Network analysis

ABSTRACT

*Problem:* Multimorbidity health disparities have not been well examined by gender. Co-occurring diseases may be mutually deleterious, co-occurring independently, or co-occurring from a common antecedent. Diseases linked by a common antecedent may be caused by biological, behavioral, social, or environmental factors. This paper aims to address the co-occurrences of diseases using network analysis.

*Methods:* In this study, we identify these multi-morbidities from a large electronic medical record (EMR) containing diagnoses, symptoms and treatment data on more than 22.1 million patients. We create multimorbidity networks from males and females medical records and compare their structural properties.

*Results:* Our macro analysis at the organ-level indicates that females have a stronger multimorbidity network than males. For example, the female multimorbidity network includes six linkages to mental health, wherein the male multimorbidity network includes only two linkages to mental health. The strength of some disease associations between lipid metabolism and chronic heart disorders is stronger in males than females.

*Conclusion:* Our multimorbidity network analysis by gender identifies specific differences in disease diagnosis by gender, and presents questions for biological, behavioral, clinical, and policy research.

## 1. Introduction

Multiple ecological levels interact to influence disparities in health and health outcomes by gender. Health disparities observed between genders are caused by genetic, hormonal, physiological, behavioral, and sociocultural factors. Life expectancy at birth is notably longer for females at 81.4 years compared to males at 76.4 years [1]. During this longer lifetime, females are more likely to visit the hospital or health care provider, but less likely to die [2]. Notably this male-female health-survival paradox is explained by chronic diseases which are most prevalent by gender: females are more likely to experience pain, reproductive cancers, and depression, while males are more likely to experience cardiovascular disease and diabetes [3]. Additionally, when males and females are compared on the same chronic diseases, males may experience severe cases of chronic disease. Previous epidemiological studies of health disparities address individual diseases experienced by gender; however, most patients are diagnosed with multiple diseases. The goal of this paper is to explore disparities among males and females diagnosed with more than one disease, and present research and policy implications.

Two terms are often used to discuss the presence of more than one disease in a patient: comorbidity and multimorbidity. Comorbidity is a condition when an additional disease is diagnosed in presence of an index disease [4]. Multimorbidity is defined as the coexistence of multiple chronic diseases and conditions in a patient [5,6]. Throughout this manuscript we will use these terms interchangeably to denote co-occurrence of diseases, unless we need to specifically highlight the differences between comorbidity and multimorbidity. Previous studies on comorbidities have controlled for gender but rarely focused and reported differences in genders explicitly as pointed out by Short et al. [7]. Further examination of comorbidities by gender may be critically important for treatment of disease, and in identifying contraindications of common pharmaceuticals. The availability of large medical records affords the opportunity to study all possible disease relationships as observed in practice.

We adapt a network approach to model the multimorbidities [8]. Networks are formed from the interactions between the elements or nodes. Network analysis has been used in health and medical literature to understand the interaction of genes [9,10], molecular involvement in disease [11], drug trials [12], and historical epidemiological data on

* Corresponding author.
*E-mail addresses:* pkalgotra@clarku.edu (P. Kalgotra), ramesh.sharda@okstate.edu (R. Sharda), julie.croff@okstate.edu (J.M. Croff).

http://dx.doi.org/10.1016/j.ijmedinf.2017.09.014
Received 8 June 2017; Received in revised form 19 September 2017; Accepted 24 September 2017

disease phenotypes [13,14]. Tai and Chiu [15] applied association rule mining to create comorbidity network in ADHD patients using clinical database. Similarly, Chmiel et al. [16] applied network approach to study the prevalence of different cluster of diseases over lifetime of genders. However, to the best of our knowledge, no one has applied this approach to study multimorbidity by gender in order to better understand health disparities.

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

## 2. Method and analysis

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

### 2.1. Data description

We obtained data from the Oklahoma State University Center for Health Systems Innovation (CHSI), which houses HIPAA compliant patient data provided by Cerner Corporation, a major Electronic Medical Record (EMR) provider. The data warehouse contains an EMR on the visits of 58 million unique patients across 662 US hospitals (2000–2016). We used information about the demographics of the patients, hospitals and disease diagnoses coded by ICD-9 system.[1] We removed several hospital visits in which patients were either not diagnosed with a disease or were marked only for symptoms. After data preprocessing, we had approximately 22.1 million unique patients with the sufficient information to perform analysis.

We extracted medical records for males and females in two different datasets from this pseudo-population dataset for comparing comorbidities by gender. The datasets were further cleaned based on the detected anomalies in particular category. For example, there were a few patients who were coded as a male during one visit and a female or null in another. Although males can also have breast diseases biologically, we removed the male patients diagnosed with such diseases with a suspicion that these are erroneously coded (ICD9: 610–612).[2] We also removed males who were diagnosed with diseases such as inflammatory diseases of female pelvic organs (ICD9: 614–616)[3], and complications of pregnancy, childbirth, and the puerperium (ICD9: 630–679).[4] Similarly, we removed female patients diagnosed with diseases of male genital organs (ICD9: 600–608).[5] After cleaning the data, we had records of 12 million female patients and 9.9 million male patients. From the two samples, networks were created, one each for males and females.

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

[2] There were 38,980 male patients with ICD-9 codes 610–612, which is 0.34% of the male database.

[3] 1594 patients.

[4] 20,009 patients.

[5] 8627 patients.

### 2.2. Measuring multimorbidity

In the past, comorbidity and multimorbidity were largely defined at the cross-sectional level [4,17]. The chronic diseases, which we would not expect to go away in one hospital visit, could be overestimated from the medical records because they are recorded multiple times in an EMR. However, we delineate multimorbidity considering the lifetime history of a patient rather than a single hospital visit. We measure multimorbidity as the presence of multiple diseases in the lifetime history of a patient. This measurement has two advantages over previous definitions. First, the EMR recording of a disease over multiple hospitals visits is only considered once. Considering the same disease as different across hospital visits can overestimate its presence and bias the analysis and conclusions. Second, our definition considers the impact of a disease in one visit on subsequent visits. Therefore, it incorporates a wider span of disease developments. However, there is a concern of taking into account the association between diseases diagnosed across hospital visits occurring after long period of time. Given the relatively short time span of the database (17 years), short average length between first and last hospital visit in the database (527 days), average number of hospital visits of a patient being 5.1 (all types of visits including inpatient, outpatient, etc.) and statistical analysis on millions of patients, we mitigate the concern of false positives.

### 2.3. Multimorbidity network

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

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

Historically, associations between diagnoses or comorbidities were modeled using a simple Pearson's correlation coefficient [18,13]. However, number of significant correlations is directly proportional to the number of observations used, and thus affected by the sample size. Power to detect rare comorbidities is low because of the rareness of events. Therefore, to establish the right measure to model a comorbidity, we use a cosine index known as Salton Cosine index [19]. SCI is immune to the total number of observations used [20] and measures the prevalence of a relationship between two diseases considering their individual prevalence. Salton Cosine Index, *SCI*, is calculated as in Eq. (1), where $c_{ij}$ is the number of co-occurrences of diseases $i$ and $j$, $i_c$ is the prevalence of disease $i$ and $j_c$ is the prevalence of disease $j$. The cosine similarity has been used in the past to find phenotype overlaps [21,22]. We propose this as an appropriate measure for finding the strength of a comorbidity.

$$SCI_{ij} = \frac{(c_{ij})}{\sqrt{(i_c * j_c)}} \tag{1}$$

Statistical significance of SCI was determined by assessing the relationship between correlation and SCI, because this approach has been