International Journal of Medical Informatics 115 (2018) 1–9

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression

Arif Khan[a,b,*], Shahadat Uddin[a], Uma Srinivasan[b]

[a] Complex Systems Research Group, Project Management Program, The University of Sydney, Sydney, NSW, Australia
[b] Health Market Quality Research Stream, Capital Markets CRC, Level 3, 55 Harrington Street, Sydney, NSW, Australia

ARTICLE INFO

Keywords:
Health informatics
Chronic disease
Comorbidity
Diabetes
Type 2 diabetes
Network theory
Social network analysis
Electronic health records
Knowledge management
Data mining
Attribution

ABSTRACT

Background: Chronic diseases management outside expensive hospital settings has become a major target for governments, funders and healthcare service providers. It is well known that chronic diseases such as Type 2 Diabetes (T2D) do not occur in isolation, and has a shared aetiology common to many other diseases and disorders. Diabetes Australia reports that it is associated with a myriad of complications, which affect the feet, eyes, kidneys, and cardiovascular health. For instance, nerve damage in the lower limbs affects around 13% of Australians with diabetes, diabetic retinopathy occurs in over 15% of Australians with diabetes, and diabetes is now the leading cause of end-stage kidney disease. Our research focus is therefore to understand the comorbidity pattern, which in turn can enhance our understanding of the multifactorial risk factors of chronic diseases like Type 2 Diabetes.

Our research approach is based on utilising valuable indicators present in pre-existing administrative healthcare data, which are routinely collected but often neglected in health research. One such administrative healthcare data is the hospital admission and discharge data that carries information about diagnoses, which are represented in the form of ICD-10 diagnosis codes. Analysis of diagnoses codes and their relationships helps us construct comorbidity networks which can provide insights that can be used to understand chronic disease progression pattern and comorbidity network at a population level. This understanding can subsequently enable healthcare providers to formulate appropriate preventive health policies targeted to address high-risk chronic conditions.

Methods and findings: The research utilises network theory principles applied to administrative healthcare data. Given the high rate of prevalence, we selected Type 2 Diabetes as the exemplar chronic disease. We have developed a research framework to understand and represent the progression of Type 2 diabetes, utilising graph theory and social network analysis techniques. We propose the concept of a '*comorbidity network*' that can effectively model chronic disease comorbidities and their transition patterns, thereby representing the chronic disease progression. We further take the attribution effect of the comorbidities into account while generating the network; that is, we not only look at the pattern of disease in chronic disease patients, but also compare the disease pattern with that of non-chronic patients, to understand which comorbidities have a higher influence on the chronic disease pathway.

The research framework enables us to construct a *baseline comorbidity network* for each of the two cohorts. It then compares and merges these two networks into single *comorbidity network* to discover the comorbidities that are exclusive to diabetic patients. This framework was applied on administrative data drawn from the Australian healthcare context. The overall dataset contained approximately 1.4 million admission records from 0.75 million patients, from which we filtered and sampled the records of 2300 diabetics and 2300 non-diabetic patients. We found significant difference in the health trajectory of diabetic and non-diabetic cohorts. The diabetic cohort exhibited more comorbidity prevalence and denser network properties. For example, in the diabetic cohort, heart and liver-related disorders, cataract etc. were more prevalent. Over time, the prevalence of diseases in the health trajectory of diabetic cohorts were almost double of the prevalence in the non-diabetic cohort, indicating entirely different ways of disease progression.

Conclusions: The paper presents a research framework based on network theory to understand chronic disease progression along with associated comorbidities that manifest over time. The analysis methods provide insights that can enable healthcare providers to develop targeted preventive health management programs to reduce

* Corresponding author at: Complex Systems Research Group, Project Management Program, The University of Sydney, Sydney, NSW, Australia.
E-mail address: arif.khan@sydney.edu.au (A. Khan).

https://doi.org/10.1016/j.ijmedinf.2018.04.001
Received 24 November 2017; Received in revised form 1 April 2018; Accepted 7 April 2018

hospital admissions and associated high costs. The baseline *comorbidity network* has the potential to be used as the basis to develop a chronic disease risk prediction model.

1. Introduction and background

Chronic diseases, such as diabetes, chronic obstructive pulmonary disease, cancer, are the group of diseases that tend to be long-lasting and have persistent effects [1], either over the lifetime or through recurrent episodes of relapse and remission. They often lead to hospitalisation that are potentially preventable. In many cases, patients are unaware of their chronic conditions [2,3], until they are admitted to a hospital and subsequently when these chronic conditions are discovered as a secondary diagnosis [4]. This results in further complications, longer length of stay and an increased burden on limited healthcare resources [5–10]. Had the chronic conditions been diagnosed earlier, many such complications could have been averted or managed outside the hospital setup. However, traditional methods of conducting intervention studies and regular monitoring of a large population are often resource-intensive in terms of available clinical resources and economic viability. A potential alternative data resource is available in the form of hospital admission and discharge data [11] that carry a substantial amount of health information in the form of standardised ICD (International Classification of Diseases) codes [12]. This provides a snapshot of the patient's overall health condition during a hospital admission episode. Analysis of this vast amount of systematically generated data using advanced data mining and network analysis techniques can help us understand the disease footprints left by chronic patients. This understanding can then be utilised to evaluate the health status of a larger population cohort.

A substantial amount of work has been done in the related field of understanding disease comorbidities, i.e., any two or more diseases that occur in one person at the same time [13]. Many rule-based scoring models are based on the clinical and empirical understanding of symptoms, disease comorbidity and prevalence. In addition to the presence of comorbidity to assess a patient's health condition, they also assign scores to various physiologically observable symptoms and demographic information. For example, Charlson Comorbidity Index [14] was proposed as early as 1987 which predicts the ten-year mortality for a patient by ranking a range of demographic and comorbid conditions like heart disease, cancer, AIDS etc. Elixhauser index [15] shows slightly better prediction performance [16,17], especially when predicting the mortality beyond 30 days. Similar models such as APACHE-II (Acute Physiology and Chronic Health Evaluation-II), SAPS (Simplified Acute Physiology Score), MPM (Mortality Probability Model) [18,19] are also used to assess intensive care unit (ICU) patients' health condition to determine the aggressiveness of treatment. While these scoring models work well in specific healthcare setting, they are usually derived from rigid empirical observation and do not scale up for a large diverse population with multiple co-morbidities. It is well known that chronic or non-communicable diseases, in reality, do not occur in isolation [20]. They often share a common risk factor, which can be genetic, environmental, behaviour etc. These risk factors have a synergistic effect [21,22] on the health outcome which are often hard to understand if considered in isolation.

Therefore, we use a network-based approach on longitudinal hospital admission and discharge data to understand the disease progression by considering the comorbidities that occur over a period of time. Network-based approaches have been used, especially in the biomedical domain (21). For instance, it has been used to understand disease pathogenesis by mapping gene expression and the associated protein that act on the same pathway [23]. A few data mining based method like the collaborative filtering [24,25] were proposed to understand disease progression and prediction that specifically use administrative data. Other related research involving administrative healthcare data includes comorbidity assessment utilising Charlson index [26], identifying particular group of patient (e.g., with lower gastrointestinal bleeding) using classification algorithms [27] etc. More recently, social network analysis (SNA) based approaches have gained attention among researchers [28] as the data elements in hospital data are inherently linked. For example, effective coordination for patient transfers between hospitals or related entities has been studied using administrative data and SNA [29]. SNA and traditional network analysis were also applied [30] on electronic health record and administrative data of congestive heart failure patients to understand the patterns of service delivery for effective care coordination. The ICD codes [12] contained in administrative datasets have tremendous potential in understanding the nature of comorbidities [31], disease prediction [32,33], mortality prediction [34,35] etc. However, there is very little work that explicitly uses the network theories and administrative data to systematically develop a research framework to model chronic disease progression in terms of comorbidity, which is the primary goal of the research presented in this paper. We chose type 2 diabetes (T2D) as the example chronic disease to explore.

2. Methods

This section focuses on the methods to create a network-based view of chronic disease progression in terms of comorbidity, i.e., the *Comorbidity Network* creation. This network is designed to represent the health trajectory of patients who have a certain chronic disease in common; for our research context, this will be T2D.

2.1. Data selection

Data used in the research protocol is obtained from the administrative health data. It contained hospital admission information for about 749,000 de-identified patients who were members of private healthcare funds based in Australia. The effective time frame represented in the cohort was for a period of six years between September 2009 and March 2015. The health condition was represented in the form of ICD codes present in the admission and discharge data of each admission episode. Both T2D and non-diabetic patient data were analysed over the full period of the research dataset. Required human ethics approval was obtained from the Human Research Ethics Committee of The University of Sydney under the project number 2015/824.

In order to identify the research dataset from the original administrative data, we followed a systematic process of cleaning, filtering and appropriate cohort selection that can enable us to draw conclusions that have statistical significance and validity. Some of the filtering methods followed include - 1) selecting patients having at least two admissions with valid ICD codes over the dataset period, 2) excluding patients requiring frequent pre-scheduled admissions such as for chemotherapy or dialysis, 3) excluding irrelevant ICD codes such as those related to physical injuries, fever, vertigo, vomiting etc.

In the next section, we present some graph theory and social network analysis based definitions that will act as a basis for some of the methods used in the framework. After that, we will introduce the subsequent functional concepts of the framework.

2.2. Graph theory-based definitions used in the research framework

We use several concepts from graph theory to formally represent the