Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining

Hye Soon Kim¹, A Mi Shin², Mi Kyung Kim¹, and Yoon Nyun Kim¹,²

Departments of ¹Internal Medicine and ²Medical Informatics, Keimyung University School of Medicine, Daegu, Korea

Background/Aims: The aim of this study was to analyze comorbidity in patients with type 2 diabetes mellitus (T2DM) by using association rule mining (ARM).

Methods: We used data from patients who visited Keimyung University Dongsan Medical Center from 1996 to 2007. Of 411,414 total patients, T2DM was present in 20,314. The Dx Analyze Tool was developed for data cleansing and data mart construction, and to reveal associations of comorbidity.

Results: Eighteen associations reached threshold (support, ≥ 3%; confidence, ≥ 5%). The highest association was found between T2DM and essential hypertension (support, 17.43%; confidence, 34.86%). Six association rules were found among three comorbid diseases. Among them, essential hypertension was an important node between T2DM and stroke (support, 4.06%; confidence, 8.12%) as well as between T2DM and dyslipidemia (support, 3.44%; confidence, 6.88%).

Conclusions: Essential hypertension plays an important role in the association between T2DM and its comorbid diseases. The Dx Analyze Tool is practical for comorbidity studies that have an enormous clinical database.

Keywords: Diabetes mellitus, type 2; Comorbidity; Data mining

INTRODUCTION

According to national health statistics in Korea, the prevalence of type 2 diabetes mellitus (T2DM) increased from 8.6% in 2001 to 9.5% in 2007, while the prevalence of T2DM in the United States was 10.7% in 2007. Furthermore, the prevalence of T2DM in 2007 in men (11.6%) was higher than in women (7.8%). The prevalence was highest in men aged 60-69 years (26.6%) and in females aged 70-79 years (19.5%) [1].

Patients with T2DM have an increased incidence of disease in several internal organs and tissues. Chronic microvascular and macrovascular diseases have greater influence on the long-term prognosis of patients with T2DM than acute complications [2]. Investigating the associations of these complications with comorbid diseases by using patient diagnostic data is helpful in predicting their incidence and thus more effectively treating patients with T2DM.

Association rule mining (ARM) describes how two items are related using a special method of exploring patterns different from other analysis techniques [3].
The association rule generated from ARM can formulate the relation between X and Y in the form of “X → Y” or “If X..., then Y...,” and analyze it as “If item X exists, item Y coexists” [4]. A rule does not necessarily imply cause and effect. Instead, it identifies simultaneous occurrence between items in antecedent X and consequent Y. ARM makes it possible to analyze the association between not only two diseases, but also among three or more comorbidities that can be calculated from existing statistics. One study revealed the accompanying diseases of attention deficit/hyperactivity disorder by applying ARM to diagnostic data from the National Health Insurance Database of Taiwan [5]. Another study analyzed stroke and its comorbid diseases by ARM [6]. Therefore, the current study was conducted to determine the relations among complications, the various diseases that accompany T2DM, and three or more comorbidities, using ARM based on large amounts of clinical data.

**METHODS**

**Study population**

Data from 411,414 patients examined at the Keimyung University Dongsan Medical Center from 1996 to 2007 were analyzed using the Dx Analyze Tool. Among the patients, 20,314 had T2DM and the total diagnostic data was 145,306. As the control group for the analysis, 20,314 patients without a diagnosis of T2DM were included and the total diagnostic data was 57,379.

**Data collection**

The workflow of the association analysis of T2DM comorbid diseases is shown in Fig. 1. First, data were collected from the database of patients examined at Keimyung University Dongsan Medical Center from 1996 to 2007. Personal information of the subjects such as name, gender, age, and contact details was not collected.

**Analysis method**

For the current study, we developed the Dx Analyze Tool using the Apriori algorithm (C# 2.0, MS Access DB) [4,7] to analyze the association between clinical diagnoses. The Dx Analyze Tool, which refines the data and extracts an association rule between a specific disease and its related diseases, involves five steps: data retention, data cleansing, data mart construction, selection of Dx code, and analysis by the Apriori algorithm. The Apriori algorithm is an ARM technique. The algorithm rules specify when item-set A appears and an item-set B appears with it. The rules are evaluated by support (the number of occurrences of disease A and disease B from all diseases) and confidence (the number of occurrences of disease A co-occurring with disease B). The formulas for support and confidence have been previously described [4,8,9] and are presented below.

\[
\text{Support} (\%) = \frac{\text{Number of disease } A \cap B}{\text{Total number of disease }}
\]

\[
\text{Confidence} (\%) = \frac{\text{Number of disease } A \cap B}{\text{Number of disease } A}
\]

Using SPSS version 18.0 (SPSS Inc., Chicago, IL, USA), the chi-square test was used to review the association rules generated by the Dx Analyze Tool and to discern differences between groups with or without T2DM in the distribution of diseases appearing by the association rule. The results from the Dx Analyze Tool and the chi-square test found that a meaningful association rule exists between T2DM and other diseases.