ORIGINAL ARTICLE
Korean J Intern Med 2012;27:197-202
http://dx.doi.org/10.3904/kjim.2012.27.2.197

pISSN 1226-3303 eISSN 2005-6648
http://www.kjim.or.kr

Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining

Hye Soon Kim[1], A Mi Shin[2], Mi Kyung Kim[1], and Yoon Nyun Kim[1,2]

Departments of [1]Internal Medicine and [2]Medical 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].

Received : April 26, 2011
Revised : June 1, 2011
Accepted: September 14, 2011

Correspondence to Yoon Nyun Kim, M.D.
Department of Internal Medicine, Keimyung University School of Medicine, 1095 Dalgubeol-daero, Dalseo-gu, Daegu 704-701, Korea
Tel: 82-53-250-7432, Fax: 82-53-250-7434, E-mail: endocrine@dsmc.or.kr

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

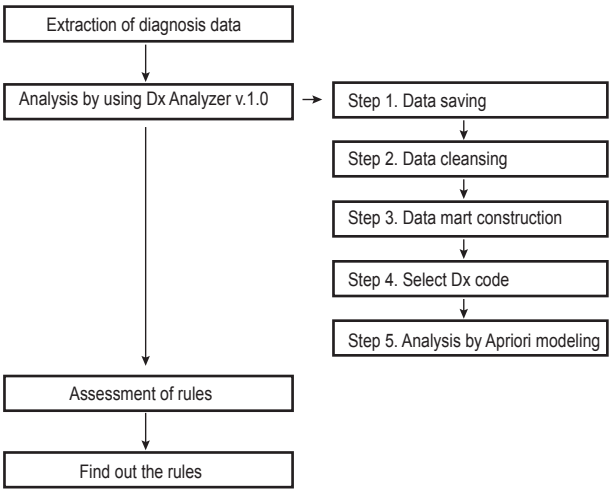

Figure 1. Schematic diagram of the study workflow.

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

$$Confidence\ (\%) = \frac{Number\ of\ disease\ A \cap B}{Number\ of\ disease\ A}$$

for support and confidence have been previously described [4,8,9] and are presented below.

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.

Table 1. High frequency comorbid diseases with type 2 diabetes mellitus (n = 20,314)

Dx code	Dx name	No.	%
I10	Essential (primary) hypertension	7,081	34.86
K29	Gastritis and duodenitis	3,170	15.61
H25	Senile cataract	3,134	15.43
E78	Disorders of lipoprotein metabolism and other lipidemias	2,771	13.64
H36	Retinal disorders in diseases classified elsewhere	2,597	12.78
I63	Cerebral infarction	2,522	12.42
I20	Angina pectoris	2,520	12.41
N18	Chronic renal failure	1,638	8.06
K25	Gastric ulcer	1,617	7.96
M81	Osteoporosis without pathological fracture	1,464	7.21
I50	Heart failure	1,374	6.76
K21	Gastroesophageal reflux disease	1,323	6.51
I21	Acute myocardial infarction	1,192	5.87
H35	Other retinal disorders	1,183	5.82
K76	Other hepatic diseases	1,152	5.67
G63	Polyneuropathy in diseases classified elsewhere	1,082	5.33
J15	Bacterial pneumonia, not elsewhere classified	1,042	5.13
Z03	Medical observation and evaluation for suspected diseases and conditions	1,025	5.05
K74	Hepatic fibrosis and cirrhosis	1,024	5.04

RESULTS

Diseases frequently accompanying T2DM

Diseases that frequently accompany T2DM are summarized in Table 1. The most frequent disease was essential hypertension (34.68% of all subjects), followed by gastritis and duodenitis (15.61%), senile cataract (15.43%), lipidemias and other disorders of lipoprotein metabolism (13.64%), and retinal disease (12.78%).

Association rule resulting from the Apriori algorithm

The association rule between T2DM and comorbid diseases generated by the Apriori algorithm is presented in Table 2. The threshold for values was established as > 3% for support and > 5% for confidence, and 18 rules satisfying these conditions were made. The rule with the highest support and confidence was T2DM→essential hypertension (support, 17.43%; confidence, 34.86%). Other rules with high support and confidence were T2DM→gastritis/duodenitis (support, 7.80%; confidence, 15.61%), T2DM→senile cataract (support, 7.71%; confidence, 15.43%), T2DM→disorders of lipoprotein metabolism and other lipidemias (support, 6.82%; confidence, 13.64%), and T2DM→retinal disease (support, 6.39%; confidence, 12.78%). The rules showing an association for more than three diseases were T2DM→essential hypertension and stroke (support, 4.06; confidence, 8.12%), T2DM→essential hypertension and disorders of lipoprotein metabolism and other lipidemias (support, 3.44%; confidence, 6.88%), and T2DM→senile cataract and retinal disease (support, 3.39%; confidence, 6.78%).

Statistical examination of ARM analysis results

The results of the statistical analysis to determine the distribution of diseases occurring with or without T2DM are summarized in Table 3. Subjects with T2DM were more likely than those without T2DM to have disorders of lipoprotein metabolism and other lipidemias, senile cataract, retinal disorders, essential hypertension, angina pectoris, heart failure, cerebral infarction, gastroesophageal reflux disease, gastric ulcer, gastritis and duodenitis, osteoporosis without pathological frac-

Table 2. Association rules between type 2 diabetes mellitus and comorbid diseases (n = 40,628)

Rule	No.	Support	Confidence
E11 → I10	7,081	17.43	34.86
E11 → K29	3,170	7.80	15.61
E11 → H25	3,134	7.71	15.43
E11 → E78	2,771	6.82	13.64
E11 → H36	2,597	6.39	12.78
E11 → I63	2,522	6.21	12.42
E11 → I20	2,520	6.20	12.41
E11 → I10, I63	1,649	4.06	8.12
E11 → N18	1,638	4.03	8.06
E11 → K25	1,617	3.98	7.96
E11 → M81	1,464	3.60	7.21
E11 → I10, E78	1,398	3.44	6.88
E11 → H25, H36	1,378	3.39	6.78
E11 → I50	1,374	3.38	6.76
E11 → I10, I20	1,342	3.30	6.61
E11 → K21	1,323	3.26	6.51
E11 → I10, K29	1,310	3.22	6.45
E11 → I10, H25	1,263	3.11	6.22

E11, type 2 diabetes mellitus; I10, essential (primary) hypertension; K29, gastritis and duodenitis; H25, senile cataract; E78, disorders of lipoprotein metabolism and other lipidemias; H36, retinal disorders in diseases classified elsewhere; I63, cerebral infarction; I20, angina pectoris; N18, chronic renal failure; K25, gastric ulcer; M81, osteoporosis without pathological fracture; I50, heart failure; K21, gastroesophageal reflux disease.

Table 3. Statistical analysis of the association rule mining results (n = 40,628)

Dx code	E11	Non E11	χ^2	*p* value
E78	2,771 (13.6)	533 (2.6)	1,650.12	0.000
H25	3,134 (15.4)	380 (1.9)	2,362.72	0.000
H36	2,597 (12.8)	21 (0.1)	2,709.25	0.000
I10	7,081 (34.9)	1,186 (5.8)	5,277.43	0.000
I20	2,520 (12.4)	522 (2.6)	1,418.50	0.000
I50	1,374 (6.8)	257 (1.3)	796.97	0.000
I63	2,522 (12.4)	442 (2.2)	1,574.51	0.000
K21	1,323 (6.5)	598 (2.9)	287.20	0.000
K25	1,617 (8.0)	441 (2.2)	707.85	0.000
K29	3,170 (15.6)	1,385 (6.8)	787.82	0.000
M81	1,464 (7.2)	201 (1.0)	999.00	0.000
N18	1,638 (8.1)	167 (0.8)	1,254.54	0.000

Values are presented as number (%).

E11, type 2 diabetes mellitus; E78, disorders of lipoprotein metabolism and other lipidemias; H25, senile cataract; H36, retinal disorders in diseases classified elsewhere; I10, essential (primary) hypertension; I20, angina pectoris; I50, heart failure; I63, cerebral infarction; K21, gastroesophageal reflux disease; K25, gastric ulcer; K29, gastritis and duodenitis; M81, osteoporosis without pathological fracture; N18, chronic renal failure.

ture, and chronic renal failure ($p < 0.05$).

DISCUSSION

This study was conducted to analyze the association between T2DM and comorbid diseases. Prior to this study, a pilot study was performed, in which comorbidity of cerebral infarction patients [6] and essential hypertension patients [10] were analyzed by ARM. On the basis of the pilot study, the present study constructed a data mart by refining diagnostic data extracted from patients of our medical center. The association rule related to more than three diseases comorbid with T2DM was ascertained by developing a program to generate the association rule by applying the ARM Apriori algorithm.

T2DM is frequently accompanied by one or more components of metabolic syndrome such as obesity, dyslipidemia, and hypertension. A patient with hypertension is 2.4 times more likely to develop cerebrovascular disease [11]. A study that examined diabetic complications in 5,652 patients with diabetes from 13 university hospitals in Korea reported that hypertension and dyslipidemia are accompanying comorbid conditions in 60.4% and 44.1%, respectively, of these patients. Additionally, 38.4% and 44.7% of patients had retinopathy and neuropathy, respectively [2]. Another study [12] reported that 77.9% of 4,240 patients with T2DM from 13 university hospitals in Korea had metabolic syndrome, with the prevalence of each component of metabolic syndrome being 56.8% for central obesity, 42.0% for hypertriglyceridemia, 65.1% for low high-density lipoprotein cholesterol, and 74.9% for hypertension. Despite different research methods, the results of the present study agree with previous studies and link T2DM with essential hypertension, disorders of lipoprotein metabolism and other lipidemias, retinal disease, cerebral infarction, and angina pectoris. Specifically, T2DM and essential hypertension had the highest association, and this association produced the following association rules: T2DM→essential hypertension and cerebral infarction, T2DM→essential hypertension and disorders of lipoprotein metabolism and other lipidemias, and T2DM→essential hypertension and angina pectoris. A previous comorbidity study on cerebral infarction revealed disorders of lipoprotein metabolism and essential hypertension→cerebral infarction by the Apriori algorithm, as well as an association of T2DM and essential hypertension→cerebral infarction [5].

Patients with T2DM often have irregular diet patterns, which deleteriously influences glucose control, lipid metabolism, and micronutrient intake [13]. In addition, T2DM is progressive and generally incurable, precluding several complications related to poor glucose regulation [14]. The use of medications to counteract the complications of diet and disease itself can cause and exacerbate gastric disorders. This was recently shown by the link between T2DM and gastroesophageal reflux disease, gastric ulcer, and gastritis and duodenitis.

Fasting glucose and diabetes correlate with the occurrence of cataracts, and metabolic disorders of the body increase the risk of the occurrence of cataracts. Specifically, the risk of cataracts increases in low levels of high-density lipoprotein cholesterol, hypertension, and high fasting glucose [15]. The present data also support an association between T2DM and senile cataract and essential hypertension. However, an association with dyslipidemia was not found and this requires further study.

Although the present study showed that T2DM is associated with heart failure and chronic renal failure, other studies on T2DM did not show such results [2,11,14]. Park et al. [16] investigated the cause of death in 680 patients with T2DM and reported that cerebrovascular disease (15.0%), ischemic heart disease (15.6%), infectious disease (25.3%), cancer (21.9%), congestive heart failure (7.1%), kidney disease (4.7%), and other diseases are major causes of death, which offers support for an association rule for T2DM, congestive heart failure, and chronic kidney disease.

In the present study, 7.21% (1,464 patients) of the patients with T2DM displayed accompanying osteoporosis without pathological fracture, and the association rule of T2DM→osteoporosis without pathological fracture was generated. Patients with T2DM were found to have more concurrent osteologic diseases than nondiabetic patients, suggesting that patients with T2DM may have decreased bone density [17].

This study determined comorbidities using the association rules generated for the diagnosis data of patients with T2DM by applying ARM from previous

studies. While the possibility exists that doctors added diagnoses excessively to increase prescriptions or that comorbidities were found but not recorded, the majority of cases were diagnosed accurately, and the few inaccuracies were filtered by using large amounts of clinical data.

This study was significant because it was based on a large amount of data generated using electronic medical records in clinical use, a constructed data mart, and analysis of the comorbidity of DM using a program that automates the determination of the Apriori algorithm. However, a limitation of the present study is that the data came from a single medical institution. Data from other medical facilities should be collected and analyzed to demonstrate the relevance of the program and its results. Furthermore, the Apriori algorithm is limited in determining precedence or causality of disease. Therefore, future studies to identify the temporal complications of diseases considering chronology (e.g., the sequential pattern of disease occurrence) should be conducted.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by a grant from the Regional Technology Innovation Program of the Ministry of Knowledge Economy (MKE) (RTI04-01-01).

REFERENCES

1. Ministry of Health & Welfare; Korean Centers for Disease Control & Prevention. 2007 National Health Statistics: National Health and Nutrition Examination Survey 4th. Seoul: Korean Centers for Disease Control & Prevention, 2008.
2. Lim S, Kim DJ, Jeong IK, et al. A nationwide survey about the current status of glycemic control and complications in diabetic patients in 2006: The Committee of the Korean Diabetes Association on the Epidemiology of Diabetes Mellitus. Korean Diabetes J 2009;33:48-57.
3. Bae HS, Cho DH, Suk KH, et al. Data Mining Using SAS Enterprise Miner. 2nd ed. Seoul: Kyowoosa, 2008.
4. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data; 1993 May 26-28; Washington, DC. New York: ACM, 1993: 207-216.
5. Tai YM, Chiu HW. Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan. Int J Med Inform 2009;78:e75-e83.
6. Lee IH, Shin AM, Son CS, et al. Association analysis of comorbidity of cerebral infarction using data mining. J Korean Soc Phys Ther 2010;22:75-81.
7. Wikipedia. Association rule learning [Internet]. San Francisco (CA): Wikimedia Foundation Inc., 2012 [cited 2012 Mar 30]. Available from: http://en.wikipedia.org/wiki/Association_rule_learning.
8. Kang HC, Han ST, Choi JH, Kim ES, Kim MK. Data Mining with SAS Enterprise Miner 4.0: Methodology and Application. 3rd ed. Seoul: Jayuacademi, 2002.
9. Heo MH, Lee YG. Data Mining Modeling and Case. 2nd ed. Seoul: Hannarae, 2008.
10. Shin AM, Lee IH, Lee GH, et al. Diagnostic analysis of patients with essential hypertension using association rule mining. Healthc Inform Res 2010;16:77-81.
11. Chung HS, Seo JA, Kim SG, et al. Relationship between metabolic syndrome and risk of chronic complications in Koreans with type 2 diabetes. Korean Diabetes J 2009;33:392-400.
12. Kim TH, Kim DJ, Lim S, et al. Prevalence of the metabolic syndrome in type 2 diabetic patients. Korean Diabetes J 2009;33:40-47.
13. Ahn HJ, Han KA, Koo BK, et al. Analysis of meal habits from the viewpoint of regularity in Korean type 2 diabetic patients. Korean Diabetes J 2008;32:68-76.
14. Kim SG, Choi DS. The present state of diabetes mellitus in Korea. J Korean Med Assoc 2008;51:791-798.
15. Park SS, Lee EH. Relations of cataract to metabolic syndrome and its components: based on the KNHANES 2005, 2007. J Korean Ophthalmic Opt Soc 2009;14:103-108.
16. Park SK, Park MK, Suk JH, et al. Cause-of-death trends for diabetes mellitus over 10 years. Korean Diabetes J 2009;33:65-72.
17. Lipscombe LL, Jamal SA, Booth GL, Hawker GA. The risk of hip fractures in older individuals with diabetes: a population-based study. Diabetes Care 2007;30:835-841.