Data Mining for Knowledge Management in a Health Insurance Domain

Young Moon Chae* · Seung Hee Ho* · Kyoung Won Cho*,

Dong Ha Lee** · Sun Ha Ji*

Abstract

This study examined the characteristics of the knowledge discovery and data mining algorithms to demonstrate how they can be used to predict health outcomes and provide policy information for hypertension management using the Korea Medical Insurance Corporation database. Specifically, this study validated the predictive power of data mining algorithms by comparing the performance of logistic regression and two decision tree algorithms, CHAID (Chi-squared Automatic Interaction Detection) and C5.0 (a variant of C4.5), since logistic regression has assumed a major position in the healthcare field as a method for predicting or classifying health outcomes based on the specific characteristics of each individual case. This comparison was performed using the test set of 4,588 beneficiaries and the training set of 13,689 beneficiaries that were used to develop the models. On the contrary to the previous study, CHAID algorithm performed better than logistic regression in predicting hypertension, but C5.0 had the lowest predictive power. In addition, CHAID algorithm and association rule also provided the segment characteristics for the risk factors that may be used in developing hypertension management programs. This showed that data mining approach can be a useful analytic tool for predicting and classifying health outcomes data.

Key word: Knowledge Management, Data Mining, Logistic regression, Hypertension, Decision Support System

* Graduate School of Health Policy and Administration, Yonsei University, Korea

** Dept. of Computer Science and Engineering Pohang University of Science and Technology, Korea

1. INTRODUCTION

Healthcare organizations are generally characterized as an information-intense organization. Knowledge-intensive technology is vital to these information-intense organizations as knowledge is becoming more important in healthcare organizations these days, with the impact of technology and the resultant complex, competitive environment. Developing procedures and routines to optimize the creation, flow, learning, and sharing of knowledge and information in the organization become an important

responsibility of management. The process of systematically and actively managing and leveraging the stores of knowledge in an organization is called knowledge management [1]. Information systems can play a valuable role in knowledge management, in which it helps the organization optimize its flow of information and capture its knowledge base.

The Korea Medical Insurance Corporation (KMIC) provides a health insurance to all civil service workers, teachers, and their dependents. All insured workers are required to participate in biannual medical examinations performed by KMIC. A questionnaire was distributed to all participants three to four days before the examination to collect information on perceived health status, tobacco consumption, and exercise habits. The KMIC database is of an enormous value in monitoring health status and developing national health promotion programs because it contains health utilization data as well as risk factors such as demographic data, biomedical data, and lifestyle data for the same beneficiaries over time. Despite its usefulness, KMIC failed to mobilize, exploit, and capitalize on these valuable knowledge database, which is needed for developing policies and inducing business process change. Probable reason is that its ability to collect and store the data has grown proportionally faster than its ability to analyze data from a large temporal database. As a result, the significant untapped knowledge lies hidden in this database.

This paper presents the knowledge management tool called knowledge discovery and data mining tools to make effective use of the KMIC database to discover the untapped knowledge that lies hidden therein using hypertension managemen: program as an example of health promotion program. While health promotion is a new concept in Korea, recent changes in the cause of death have piqued the interest in health promotion. In 1995, a health promotion law was passed that provides funds for health promotion research through a tax on tobacco sales, requires a small portion of the medical insurance fund to be spent on prevention activities, and requires communities to develop health promotion plans [2].

Data mining is a nontrivial process of identifying valid, novel, potentially useful, and an ultimately understandable pattern in data [3]. Typically, the applications involve large-scale information banks such as data warehouse. In healthcare, insurance companies and large hospitals are ideal settings for the application of data mining. Some of the previous applications for data mining in healthcare are pathology information system to discover new patterns that provided new knowledge [4], identifying significant factors influencing prenatal care [5], and automatic detection of hereditary syndromes [6].

These systems, however, did not explicitly deal with policy analysis using various data mining models.

In the statistics and epidemiology community, logistic regression has assumed a major position as a method for predicting or classifying outcomes based on the specific features of an individual case. Long et al. [7] compared the performance of logistic regression to a popular data mining model, called C4.5 decision tree induction, in classifying patients as having acute cardiac ischemia, and found that logistic regression performed better than C4.5. In this paper, the performance of logistic regression and two decision tree algorithms, CHAID (Chi-squared Automatic Interaction Detection) and C5.0 (a variant of C4.5), in predicting hypertension were compared. In addition, this paper demonstrated how the decision tree algorithm and another data mining model, called association rule, could be used in a policy analysis for hypertension management.

2. METHODS

2.1 Subjects

The subjects were randomly selected from a population of 127,886 beneficiaries who participated in a biannual medical examination conducted by KMIC in 1998. In selecting the sample for this study, 50% of the total population were randomly selected in the first stage, and 100% of the beneficiaries with hypertension (9,103) and an equal number of the beneficiaries without hypertension were randomly selected in the second stage. This sample is further randomly divided into a training set (13,689) and a test set (4,588), whose ratio is approximately three to one.

Biometric data, including blood pressure, blood glucose, cholesterol, urinary glucose urinary protein, and height and weight, was drawn collected during the physical examination. Hypertension was defined as systolic blood pressure >140 or diastolic blood pressure as >90 mmHg, and low risk as values below those. A fasting blood specimen was drawn and analyzed for total cholesterol and blood glucose. A questionnaire was distributed to all participants three to four days before the examination to collect information on perceived health status, lifestyle factors (smoking, exercise, and drinking), and demographics (gender, age, blue collar job status and income level).

2.2 Knowledge models

2.2.1 Logistic regression

Logistic regression is a nonlinear regression method for predicting a dichotomous dependent variable. Logistic regression was performed to identify risk factors for hypertension using patient characteristics, history, lifestyle, and test results as independent variables and the hypertension status as dependent variable. Stepwise selections of the independent variables can be made and the corresponding coefficients computed. In producing the logistic regression equation, the maximum-likelihood ratio was used to determine the statistical significance of the variables. Logistic regression has proven to be very robust in a number of medical domains and proves an effective way of estimating probabilities from dichotomous variables.

2.2.2 Decision Tree

Decision trees are known as effective classifiers in a variety of domain. In our example, the decision tree categorizes the entire subjects according to whether or not they are likely to have hypertension. Most of the decision tree algorithms use a standard top-down approach to building trees. CHAID and C5.0 are two popular decision tree inducers, based on the ID3 classification algorithm by Quinlan [8].

A CHAID tree is a decision tree that is constructed by splitting subsets of the space into two or more child nodes repeatedly, beginning with the entire data set. To determine the best split at any node, any allowable pair of categories of the predictor variables is merged until there is no statistically significant difference within the pair with respect to the target variable. The process is repeated until no insignificant pair is found. The resulting set of categories of the predictor variable is the best split with respect to that predictor variable [9]. In this paper, the CHAID algorithm with growing criteria of the likelihood ratio chi-square statistic was used for building the tree and evaluating splits because most of our variables were ordinal and discretized continuous variables. To identify nodes of interest (that is, nodes with a relatively high probability), gains chart was used. The gains chart shows the nodes sorted by the number of cases in the target category for each node.

C5.0 uses pruning strategy where a branch is pruned when the introduced error is one standard error of the existing errors adjusted for the continuity correction. C5.0 also uses boosting technique to generate and combine multiple classifiers to give improved predictive accuracy. Compared with C4.5, the error rate of C5.0's boosted classifiers is about one-third of the error rate of C4.5's single classifiers [10]. Both algorithms also provide classification rules for identifying risk factors that may be used to develop programs for hypertension management.

2.2.3 Association rule

An association rule gives an occurrence relationship among factors. In a well-known market basket problem, the association rule has been used to discover buying patterns such as two or more items that are bought together often. In this paper, the association rule was used to identify occurrence relationship between hypertension and various modifiable risk factors, such as smoking or drinking, in an attempt to develop hypertension management program.

An association rule is intended to capture a certain type of dependence among items represented in the database B [11]. Specifically, we say that $i_1 => i_2$

- i_1 and i_2 occur together in at least s % of the n baskets (s % is called the support);
- of all the baskets containing i_1, at least c % also contain i_2 (c % is called the confidence)

3. RESULTS

The mean age of the study participants was 52.1 years among men and 51.4 among women. Among the 12,077 men, 5,762 (47.7%) were current smokers and 1,994 (16.5%) were ex-smokers. However, among the 6,200 women, only 22 (.4%) were current smokers and 22 (.4%) former smokers. Most had moderate weight levels. Complete descriptive statistics for the modifiable risk factors are shown in <Table 1>.

[Table 1] Descriptive statistics for the study sample

			Men (n=12,077)		Women (6,200)	
Cate-gory	Measure	Value	Count	Percent	Count	Percent
Lifestyle factor variables	Diet habits	Irregular	447	3.7%	376	6.1%
	Salt Intake	Salted	2,085	17.3%	651	10.5%
	Liking Intake	Meat	709	5.9%	304	4.9%
	Alcohol consumption	Heavy	4,718	39.1%	36	0.6%
	Tobacco user	Former	1,994	16.5%	22	0.4%
		Current	5,762	47.7%	22	0.4%
	Exercise Habits	No	4,900	40.6%	4,237	68.3%
Biometric variables	BMI	Overweight	3,807	31.5%	1,770	28.5%
		Obesity	3,059	25.3%	1,813	29.2%
	Dx	Hypertension	6,752	55.9%	2,350	37.9%
	Urinary sugar	Positive	11,389	94.3%	6,100	98.4%
	Urinary protein	Positive	11,558	95.7%	6,019	98.5%
	Urinary RBC	Positive	11,573	95.8%	5,532	89.2%
	Blood Glucose	High (>126mg/dl)	1,031	8.5%	199	3.2%
	Total Cholesterol	High (>240mg/dl)	1,537	12.7%	1,022	16.5%
Demo -graphic variables	Age Group	40-44	1,467	12.1%	990	16%
		45-49	3,088	25.6%	2,180	35.2%
		50-54	3,061	25.3%	1,504	24.3%
		55-59	2,988	24.7%	1,041	16.8%
		60-64	1,340	11.1%	439	7.1%
		65-69	133	1.1%	37	0.6%
	Past History	Heart Disease	478	4%	94	1.5%
		Stroke	56	0.5%	51	0.8%
		Diabetes Mellitus	1,055	8.7%	473	7.6%
	PHx	Yes	1,539	12.7%	608	9.8%
	Family History	Hypertension	6,971	57.7%	3,491	56.3%
		Stroke	7,189	59.5%	3,801	61.3%
		Heart Disease	7,485	62%	3,912	63.1%
		Diabetes Mellitus	7,249	60%	3,761	60.7%
	FHx	Yes	2,553	21.1%	1,777	28.7%

PHx : Past History
FHx : Family History

3.1 Comparison of logistic regression to decision tree algorithms

The comparison of the sensitivity, specificity, and overall predictive rate for the three models is shown in [Table 2] The CHAID algorithm had the best overall predictive rate (64.06%), followed by logistic regression (63.84%) and C5.0 (59.22%). CHAID algorithm also had the best sensitivity (76.3%), followed by logistic regression (64.36%) and C5.0 (59.34%). However, CHAID had the lowest specificity (52.3%) and the logistic regression had the best specificity (63.33%).

[Table 2] Comparison of the predictive rates for logistic regression, CHAID, and C5.0

	Sensitivity(%)	Specificity(%)	PredictiveRate(%)
Logistic Regression	64.36	63.33	63.84
CHAID	76.30	52.3	64.06
C 5.0	59.34	59.1	59.22

3.2 Identifying segment characteristics for risk factors

As shown in Table 3, biomedical variables were excellent predictors of hypertension. Four biomedical variables (BMI, urinary protein, blood glucose, and cholesterol) were significant predictors. However, none of the lifestyle factors predicted hypertension, and age was the only significant predictor among demographic factors.

The decision tree has 75 leaf nodes. As shown in figure 1, the induced rules for nodes 65, 75, and 40 are depicted in the decision tree.

/* Node 65 */

if ((gender=male) and BMI=obesity and (50<age<=55) and (103<blood glucose<=571)), then hypertension proportion =79.12

[Table 3] Results of logistic regression

Category	Variable	Odds Ratio	Para-meter Estimate	Pr > Chi-Square
Lifestyle factor variables	Salt Intake	1.02	0.02	0.45
	Alchol Consumption	1.00	0.00	1.00
	Tabacco User	1.00	-0.00	0.93
	Exercise habits	1.00	-0.00	0.94
	Meet	0.97	-0.03	0.39
Bio-metric variables	BMI	1.58	0.46	0.00
	Urinary Protein	1.45	0.37	0.00
	Blood Glucose	1.01	0.01	0.00
	T.Chole-sterol	1.00	0.00	0.00
	U_RBC	1.00	-0.00	0.90
	U_sugar	0.94	-0.06	0.13
Demographic variables	PHx of Heart Disease	1.16	0.15	0.16
	PHx of Stroke	1.08	0.08	0.75
	Age	1.08	0.07	0.00
	FHx of Heart Disease	1.04	0.04	0.45
	FHx of Diabets Mellitus	1.01	0.01	0.87
	FHx of Stroke	0.99	-0.01	0.87
	FHx of Hypertension	0.99	-0.01	0.82
	PHx of Diabetes Mellitus	0.93	-0.07	0.26

PHx : Past History

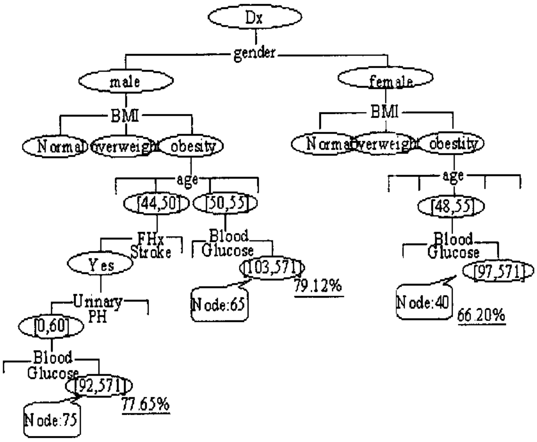
FHx : Family History

/* Node 75 */

if ((gender=male) and BMI=obesity and (44<age<=50) and (family history of stroke=Yes) and (0<Urinary PH<=60) and (92<blood glucose<=571)), then hypertension proportion=77.65

/* Node 40 */

if ((gender=female) and BMI=obesity and (48<age<=55) and (97<Blood glucose<=571)), then hypertension proportion=66.20



[Figure 1] Decision tree by CHAID algorithm

As shown in [Table 4], there are two parts to the gains chart: node-by-node statistics and cumulative statistics. The gains chart shows the nodes sorted by the number of cases in the target category for each node. The first node in the table, node 67, contains 271 hypertensive cases out of 333 subjects, or 81.38% hypertension rate. For this type of gains chart, with a categorical target variable, the gain score

[Table 4] Gains chart by CHAID algorithm

	Node-by-Node						Cumulative			
Node	Node : n	Node : %	Resp : n	Resp : %	Gain (%)	Index (%)	Node : %	Resp : %	Gain (%)	Index (%)
67	333	2.4	271	4.0	81.4	162.5	2.4	4.0	81.4	162.5
21	137	1.0	110	1.6	80.3	160.4	3.4	5.6	81.1	161.9
65	182	1.3	144	2.1	79.1	158.0	4.7	7.7	80.5	160.8
75	179	1.3	139	2.0	77.7	155.1	6.1	9.7	79.9	159.6
52	165	1.2	126	1.8	76.4	152.5	7.3	11.5	79.3	158.4
58	554	4.0	396	5.8	71.5	142.8	11.3	17.3	76.5	152.8
66	388	2.8	277	4.0	71.4	142.6	14.1	21.3	75.5	150.8
56	215	1.6	152	2.2	70.7	141.2	15.7	23.5	75.0	149.8
64	140	1.0	96	1.4	68.6	136.9	16.7	24.9	74.6	149.0
40	142	1.0	94	1.4	66.2	132.2	17.8	26.3	74.1	148.0
.	.	.	.	.	.	.	.	.	.	.
39	349	2.6	176	2.6	50.4	100.7	62.5	76.9	61.6	123.1
60	152	1.1	70	1.0	46.1	92.0	63.6	77.9	61.4	122.5
18	308	2.3	137	2.0	44.5	88.8	65.8	79.9	60.8	121.4
.	.	.	.	.	.	.	.	.	.	.

equals the percentage of cases with the target category –in this case, hypertension for the node. The Index score shows how the proportion of hypertension for this particular node compares to the overall proportion of hypertension. For node 65, the Index score is about 158.0%, meaning that the proportion of respondents for this node is about 1.6 times the hypertension rate for the overall sample.

The cumulative statistics can show us how well we do at finding hypertensive cases by taking the best segments of the sample. If we only take the best node (node 67), we reach 3.95% of hypertensive cases by targeting only 2.43% of the sample. If we include the next best node as well (node 21), then we get 5.55% of the hypertensive cases from only 3.43% of the sample. Including the node 65 increases those values to 7.7% of hypertensive cases from 4.7% of the sample. At this stage, we are at the crossover point described above, where we start to see diminishing returns. Note what happens if we include the next node (node 40)--we get 26.3% of hypertensive cases, but we must contact 17.8% of the sample to get them.

The gains chart can also provide valuable information about which segments to target and which to avoid. We might base the decision on the number of prospects we want, the desired hypertension rate for the target sample, or the desired proportion of all potential hypertension cases we want to contact. In this example, suppose we want an estimated hypertension rate of at least 80%. To achieve this, we would target the first three nodes, nodes 67, 21, and 65.

3.3 Association rule

[Table 5] shows an example of the association rules based on the generalized rule induction. The rules provide specific information about risk factors as follows:

If exercise = no and (43<age<48) and gender = female, then Dx = hypertension with probability of 21%.

(1.8% of the sample has the above risk factors and hypertension)

If exercise = no and (43<age<48) and smoke = yes and gender = female, then Dx = hypertension with probability of 21%.

If exercise = no and (43<age<48) and gender = female and past history of heart disease = no, then Dx = hypertension with probability of 21%.

If exercise = no and drink = no and (43<age <48) and gender = female, then Dx = hypertension with probability of 20%.

If exercise = no and smoke =yes and (43<age< 48) and gender = female, then Dx = hypertension with probability of 22%.

If exercise = no and smoke =yes and smoke = yes, then Dx = hypertension with probability of 26%.

This shows that an existence of all three modifiable risk factors significantly increases the probability of hypertension regardless of gender.

〔Table 5〕 Example of an association rule

Gender	Age	Phx of Heart Disease	BMI	Smoke	Drink	Exercise	Support (%)	Confidence (%)
Female	43<Age<48	*	*	*	*	No	1.8	21.0
Female	43<Age<48	*	*	Yes	*	No	1.8	21.0
Female	43<Age<48	Yes	*	*	*	No	1.8	21.0
Female	43<Age<48	*	*	*	Yes	No	1.6	20.0
Female	43<Age<48	*	*	Yes	*	No	2.3	22.0
*	43<Age<48	*	*	Yes	Yes	No	2.9	26.0

* non-significant

4. DISCUSSION

This study examined the characteristics of the knowledge discovery and data mining models to demonstrate how they can be used to predict health outcomes and provide policy information from the Korea Medical Insurance Corporation database using hypertension management program as an example. First, this study validated the predictive power of data mining algorithms by comparing the performance of logistic regression and two decision tree algorithms, CHAID and C5.0, since logistic regression has assumed a major position in the healthcare field as a method for predicting or classifying health outcomes based on the specific characteristics of each individual case.

Logistic regression and decision tree induction have different underlying assumptions. For logistic regression, it is assumed that the influence of a variable on the outcome is uniform across all subjects unless specific interactions with other variables are included. On the other hand, the decision tree assumes that the effect of a variable in the subset is unrelated to the effect of the variable in other subsets of subjects. This comparison was performed using the test set of 4,588 subjects and the training set of 13,689 subjects that were used to develop the models. Similar to the study by Long et al. [7], logistic regression performed better than C5.0. Unexpectedly, the CHAID algorithm (64.06%) performed slightly better than logistic regression (63.84%) in predicting overall hypertension, and it provided much higher sensitivity (76.3%) than logistic regression (64.36%). This shows that the CHAID algorithm is capable of predicting health outcomes from a large database. In a similar study by Chae et al. [12], statistical method (discriminant analysis) performed better than another data mining methods, neural network and case-based reasoning.

Second, we demonstrated how CHAID could be used in a policy analysis for hypertension management. While logistic regression provides risk factors for hypertension, it does not provide information about the segment characteristics of age or risk factors that may be useful for policy analysis. The CHAID algorithm provided cumulative statistics that shows how well we do at finding the hypertensive cases by taking the best segments of the sample. The gains chart also provided valuable information about which segments to target and which to avoid.

In addition, we presented the association rules that provided an occurrence relationship among risk factors. For example, the association rule showed that an existence of all three modifiable risk factors (smoking, drinking and exercise) significantly increased the probability of hypertension regardless of gender. Such information can be used in examining the effects of individual (modifiable) risk factors on the specific segment of target population.

Study limitations include a low specificity for the CHAID algorithm and a low confidence measures for the association rule. Another limitations are weak measures for exercise behavior, absence of measures for nutrition, stress, and depression. In addition, the population was limited to teachers and civil servants, and was thus was biased from the perspective of affluence.

Future analyses will include an improvement of decision tree algorithm and association rule. Another area of improvement in data mining is an application of a sequence rule. The sequence rule gives a temporal relationship among factors [10]. Since all insured workers in Korea are required to participate in a biannual medical examinations performed by KMIC and their biomedical data as well as lifestyle data are well maintained in a temporal database at the KMIC, the sequence rule can effectively be applied to predict health outcomes based on the trends data. For example, the sequence rule provides information that blood pressure goes up if the BMI and cholesterol level both go up at two consecutive biannual medical examinations. Finally, medical care costs will be incorporated into data mining algorithm based on diagnostic codes for each of the risk factors in order to provide an estimated budget for health promotion programs.

REFERENCES

[1] Laudon KC, Laudon JP. Management information systems, fifth edition, Prentice Hall, 1998, p.553

[2] Ministry of Health and Welfare. Health Promotion Law. Republic of Korea, 1995

[3] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery; an overview. In: Fayyad U, Piatetsky-Shapiro G, Smyth, eds. Advances in Knowledge Discovery and Data Mining. MIT Press; 1996: 1-34

[4] McDonald JM, Brossette S, Moser SA. Pathology information systems: data mining leads to knowledge discovery. Arch Pathol Lab Med, May 1998; 122: 409-411

[5] Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. Medical data mining: knowledge discovery in c clinical data warehouse. Proceedings of Annual Conference of AMIA, 1997; 101-105

[6] Evans S, Lemon S, Deters CA, Fusaro RM, Lynch HT. Automated detection of hereditary syndromes using data mining. Computers and Biomedical Research, 1997; 30: 337-348

[7] Long WJ, Griffith JL, Selker HP, DAgostino RB. A comparison of logistic regression to decision-tree induction in a medical domain. Computers and Biomedical Research, 1993; 26: 74-97

[8] Quinlan JR. C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo, CA, 1993

[9] Biggs DB, de Ville B, Suen E. A method of choosing multiway partitions for classification and decision trees. J of Applied Statistics, 1991; 18: 49-62

[10] Arawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In Proceeding of the ACM SIGMOD. International Conference on the Management of Data, 1993: 207-216

[11] Rulequest. C5.0 better than C4.5. http://www.rulequest.com/see5-comparison.htma,1999

[12] Chae YM, Lee SH, Ho SH, Bae MY, Ohrr HC. Medical decision support system for the management of hypertension. Informatica, 1997; 21: 219-25