International Journal of Medical Informatics 62 (2001) 103–111

International Journal of Medical Informatics

www.elsevier.com/locate/ijmedinf

Data mining approach to policy analysis in a health insurance domain

Young Moon Chae [a,*], Seung Hee Ho [a], Kyoung Won Cho [a], Dong Ha Lee [b], Sun Ha Ji [a]

[a] *Graduate School of Health Policy and Administration, Yonsei University, CPO Box 8044, Seoul, 120-749, South Korea*
[b] *Department of Computer Science and Engineering, Pohang University of Science and Technology, Pohang, South Korea*

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, CHIAD (Chi-squared Automatic Interaction Detection) and C5.0 (a variant of C4.5) using the test set of 4588 beneficiaries and the training set of 13,689 beneficiaries. Contrary to the previous study, the CHIAD algorithm performed better than the logistic regression in predicting hypertension, and C5.0 had the lowest predictive power. In addition, the CHIAD algorithm and the association rule also provided the segment-specific information for the risk factors and target group that may be used in a policy analysis for hypertension management. © 2001 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Knowledge management; Data mining; Logistic regression; Hypertension; Health insurance

1. Introduction

Healthcare organizations are generally characterized as information-dependent organizations. Knowledge-intensive technology is vital to these information-dependent organizations as knowledge is becoming more important in healthcare organizations, with the impact of technology and the resultant complex, competitive environment. 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

* Corresponding author. Present address: Graduate School of Health Science and Management, Yonsei University, CPO Box 8044, Seoul, 120-749, South Korea. Fax: +82-2-392-7734.

E-mail address: ymchae@yumc.yonsei.ac.kr (Y.M. Chae).

PII: S1386-5056(01)00154-X

civil service workers, teachers, and their dependants. All insured beneficiaries are required to participate in biannual medical examinations performed by KMIC, as a part of nation-wide health-promotion program. A questionnaire was also distributed to all participants 3–4 days before the examination to collect information on perceived health status, tobacco consumption, and exercise habits. While health promotion is a new concept in Korea, recent changes in the cause of death have piqued 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 that a small portion of the medical insurance funds be spent on prevention activities, and requires communities to develop health-promotion plans [2].

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 the information available in the database. Such information is needed for developing policies and inducing business process change. The probable reason is that its ability to collect and store the data has grown proportionately faster than its ability to analyze data from a large temporal database.

This paper presents the knowledge management tool called knowledge discovery and data mining to make effective use of the KMIC database to discover the untapped information using the hypertension-management program as an example of a health-promotion program. 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 a 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 were pathology information systems [4,5]. These systems, however, did not deal explicitly with policy analysis using various data-mining models.

Long et al. [6] 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, CHIAD (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 in a health-insurance domain.

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 a sample for this study, 50% of the total population was randomly selected in the first stage, and 100% of the beneficiaries with hypertension (9,103) and an equal number of beneficiaries without hypertension were randomly selected in the second stage. This sample was further randomly divided into a training set (13,689) and a test set (4,588) (the ratio was approximately 3:1).

Biometric data, including blood pressure, blood glucose, cholesterol, urinary glucose, urinary protein, and height and weight, were collected during the physical examination. Hypertension was defined as systolic blood pressure >140 mmHg or diastolic blood pressure as >90 mmHg, and low risk as values below these. A fasting blood specimen was drawn and analyzed for total cholesterol and blood glucose.

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 the dependent variable. Stepwise selections of the independent variables were made, and the corresponding coefficients were 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 domains. 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. However, 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. In our example, the decision tree categorizes the entire subjects according to whether or not they are likely to have hypertension. CHIAD and C5.0 are two popular decision-tree inducers, based on the ID3 classification algorithm by Quinlan [7].

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. In this paper, the CHIAD 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), a 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 a 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 a 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 [8]. 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 often bought together. In this paper, the association rule was used to identify the occurrence relationship between hypertension and various modifiable risk factors, such as smoking or drinking, in an attempt to develop a hypertension-management program.

An association rule is intended to capture a certain type of dependence among items. Suppose that $i_1 = > i_2$, then

- *support* is a probability that i_1 and i_2 occur together
- *confidence*, of all the baskets containing i_1, is a probability of also containing i_2.

3. Results

3.1. Characteristics of the subjects

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 (0.4%) were current smokers and 22 (0.4%) former smokers. Most of them had moderate weight levels. Complete descriptive statistics for the modifiable risk factors are shown in Table 1.

3.2. Comparison of predictive rates between logistic regression and decision tree algorithms

The results of the logistic regression show that the biomedical variables were excellent predictors of hypertension (Table 2). While four biomedical variables (BMI, urinary protein, blood glucose, and cholesterol) were significant predictors, none of the lifestyle factors predicted hypertension, and age was the only significant predictor among demographic factors.

A comparison of the sensitivity, specificity, and overall predictive rate for the three models is shown in Table 3. The CHAID algorithm had the best overall predictive rate (64.06%), followed by logistic regression (63.84%) and C5.0 (59.22%). CHIAD 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%).

3.3. Policy analysis for hypertension management with data mining

3.3.1. Decision tree

As shown in Fig. 1, the decision tree has 75 leaf nodes. Each node depicted in the decision tree can be expressed in terms of an ‘if–then’ rule, as follows:

/* Node 75 */

If ((gender = male) and BMI = obesity and $(44 < \text{age} < = 50)$ and (family history of stroke = Yes) and $(0 < \text{Urinary PH} < = 60)$ and $(92 < \text{blood glucose} < = 571)$), then hypertension proportion = 77.65%

The gains chart produced by the decision tree can be used for a policy analysis for hypertension management. There are two parts to the gains chart: node-by-node statistics and cumulative statistics (Table 4). In the gains chart, nodes were 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 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 shows 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 also provides valuable information about which segments to target

Table 1
Descriptive statistics for the study sample

Category	Measure	Value	Men ($n = 12{,}077$)		Women (6200)	
			Count	Percentage	Count	Percentage
Lifestyle variables	Diet habits	Irregular	447	3.7	376	6.1
	Salt intake	Salted	2085	17.3	651	10.5
	Preferred food	Meat	709	5.9	304	4.9
	Alcohol	Heavy	4718	39.1	36	0.6
	Tobacco	Former	1994	16.5	22	0.4
		Current	5762	47.7	22	0.4
	Exercise	No	4900	40.6	4237	68.3
Biometric variables	BMI	Overweight	3807	31.5	1770	28.5
		Obesity	3059	25.3	1813	29.2
	Disease	Hypertension	6752	55.9	2350	37.9
	U. sugar	Positive	11,389	94.3	6100	98.4
	U. protein	Positive	11,558	95.7	6019	98.5
	U. RBC	Positive	11,573	95.8	5532	89.2
	Blood glucose	High (>126 mg/dl)	1031	8.5	199	3.2
	Total cholesterol	High (>240 mg/dl)	1537	12.7	1022	16.5
Demographic variables	Age group	40–44	1467	12.1	990	16
		45–49	3088	25.6	2180	35.2
		50–54	3061	25.3	1504	24.3
		55–59	2988	24.7	1041	16.8
		60–69	1473	12.2	476	7.7
		H.D.	478	4	94	1.5
	Past history	Stroke	56	0.5	51	0.8
		D.M.	1055	8.7	473	7.6
		Hypertension	6971	57.7	3491	56.3
	Family history	Stroke	7189	59.5	3801	61.3
		H.D.	7485	62	3912	63.1
		D.M.	7249	60	3761	60.7

H.D.: heart disease; D.M.: diabetes mellitus.

Table 2
Results of logistic regression

Category	Variable	Odds ratio	Para. est.	$\Pr > \chi^2$
Lifestyle variables	Salt intake	1.02	0.02	0.45
	Alcohol	1.00	0.00	1.00
	Tobacco	1.00	−0.00	0.93
	Exercise	1.00	−0.00	0.94
	Meat	0.97	−0.03	0.39
Biometric variables	BMI	1.58	0.46	0.00
	U. protein	1.45	0.37	0.00
	Blood glucose	1.01	0.01	0.00
	T. cholesterol	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 H.D.	1.16	0.15	0.16
	PHx of stroke	1.08	0.08	0.75
	Age	1.08	0.07	0.00
	FHx of H.D.	1.04	0.04	0.45
	FHx of D.M.	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 D.M.	0.93	−0.07	0.26

PHx: past history; FHx: family history; H.D: heart disease; D.M: diabetes mellitus.

Table 3
Comparison of the predictive rates for logistic regression, CHAID, and C5.0

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

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. This segment-specific information can be used for planning for a hypertension-management program.

3.3.2. *Association rule*

The association rules provide specific information about risk factors based on the following generalized rule induction (Table 5).

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

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

…

If exercise = no and smoke = yes and

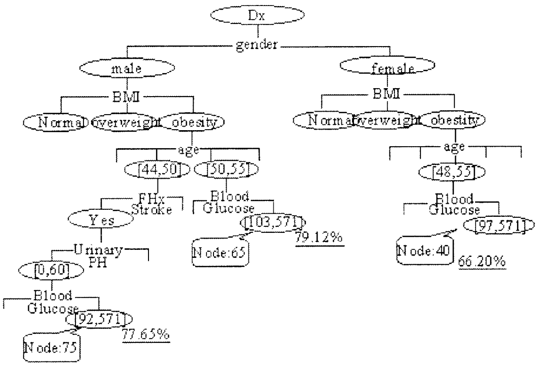

Fig. 1. Decision tree by the CHAID algorithm.

(43 < age < 48) and gender = female, then a probability of hypertension is 22%.

If exercise = no and smoke = yes and drink = yes, then a probability of hypertension is 26%.

This shows that the existence of all three modifiable risk factors significantly increases the probability of hypertension (from 22% to 26%) regardless of gender.

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 a hypertension-management program as an example. First, this study compared the performance of logistic regression and two decision-tree algorithms, CHIAD 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. This comparison was performed using the test set of 4588 subjects and

Table 4
Gains chart by CHIAD algorithm

Node	Node-by-node						Cumulative			
	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
…	…	…	…	…	…	…	…	…	…	…

Resp.: Respondents.

Table 5
Example of an association rule

Gender	Age	Phx of heart disease	BMI	Smoke	Drink	Exercise	Support (%)	Confidence (%)
Female	$43<\text{age}<48$	*	*	*	*	No	1.8	21.0
Female	$43<\text{age}<48$	*	*	Yes	*	No	1.8	21.0
Female	$43<\text{age}<48$	Yes	*	*	*	No	1.8	21.0
Female	$43<\text{age}<48$	*	*	*	Yes	No	1.6	20.0
Female	$43<\text{age}<48$	*	*	Yes	*	No	2.3	22.0
*	$43<\text{age}<48$	*	*	Yes	Yes	No	2.9	26.0

*Non-significant.

the training set of 13,689 subjects that were used to develop the models. Contrary to the study by Long et al. [6], the CHIAD algorithm (64.06%) performed better than logistic regression (63.84%) in predicting overall hypertension, and it provided a much higher sensitivity (76.3%) than logistic regression (64.36%), but logistic regression performed better than C5.0. In a similar study by Chae et al. [10], discriminant analysis performed better than other data-mining methods, neural network and case-based reasoning.

Second, we demonstrated how CHIAD could be used in a policy analysis for hypertension management. While logistic regression provides risk factors for hypertension, it does not provide specific information about the segment characteristics of age or risk factors that may be useful for policy analysis. The CHAID algorithm provided cumulative statistics that show 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 the existence of all three modifiable risk factors (smoking, drinking and exercise) significantly increased the probability of hypertension regardless of gender. Such information, which could not be obtained from the logistic regression, 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 low confidence measures for the association rule. The other 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 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 [9]. Since all insured workers in Korea are required to participate in biannual medical examinations performed by KMIC, and their biomedical as well as lifestyle data are well maintained in a temporal database at the KMIC, the sequence rule can be effectively 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, cost information will be incorporated into a data-mining algorithm for each of the risk factors in order to estimate the budgets for providing hypertension management services for the specific target population.

References

[1] K.C. Laudon, J.P. Laudon, Management Information Systems, Prentice Hall, Englewood Cliffs, NJ, 1998, p. 553.

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

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery; an overview, in: U. Fayyad, G. Piatetsky-Shapiro (Eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, Boston, MA, 1996, pp. 1–34.

[4] J.M. McDonald, S. Brossette, S.A. Moser, Pathology information systems: data mining leads to knowledge discovery, Arch. Pathol. Lab. Med. 122 (1998) 409–411.

[5] S. Evans, S. Lemon, C.A. Deters, R.M. Fusaro, H.T. Lynch, Automated detection of hereditary syndromes using data mining, Comput. Biomed. Res. 30 (1997) 337–348.

[6] W.J. Long, J.L. Griffith, H.P. Selker, R.B. D'Agostino, A comparison of logistic regression to decision-tree induction in a medical domain, Comput. Biomed. Res. 26 (1993) 74–97.

[7] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

[8] D.B. Biggs, B. de Ville, E. Suen, A method of choosing multi-way partitions for classification and decision trees, J. Appl. Stat. 8 (1991) 49–62.

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

[10] Y.M. Chae, S.H. Lee, S.H. Ho, M.Y. Bae, H.C. Ohrr, Medical decision support system for the management of hypertension, Informatica 21 (1997) 219–225.