Clinical Decision Support System for Hypertension Management

Young M Chae[a], Seung Y Cho[b], Seok M Kang[b], Hwa Y Kim[a], Tae S Kim[a], Jin Lee[a], Seung H Ho[a]

[b.] *Department of Health Informatics, Graduate School of Health Science and Management, Yonsei University, C.P.O. Box 8044, Seoul Korea, e-mail:ymchae.yumc.yonsei.ac.kr*
[b.] *Department of Cardiology, Yonsei University College of Medicine, Seoul, Korea*

Abstract
This study examined the predictive power of data mining algorithms by comparing the performance of a logistic regression algorithm with that of a decision tree algorithm called CHAID (Chi-squared Automatic Interaction Detection). Unlike in the previous studies, the decision tree performed better than logistic regression. This paper also demonstrated how another data mining model, called association rule, could be used in validating clinical decision rules for hypertension management. We also developed a CDSS (Clinical Decision Support System) with three modules (doctor, nurse, and patient) based on data warehouse architecture. This system can help improve doctors' decision- and can make educational material more accessible to patients.

Keywords:
Clinical decision support system, data mining, hypertension, decision tree, association rule

Introduction
Over the last several years, there has been increasing pressure from the public to improve quality of healthcare service and to reduce the total cost of healthcare. Since hospitals are generally characterized as information-dependent organizations, information systems have played an important role in meeting this demand. In hospitals, operational systems such as physician order entry system (POES) has been designed to efficiently store medical data and rapidly process individual patient transactions.

Although POES decrease outpatient waiting time and inpatient length of stay by speeding up the information flow, they have little effect on clinical decision-making. Clinical decision support systems (CDSS) based on operational systems have been designed to automate alerts and warnings, offer physicians instantaneous access to reference materials and standards of care, assist the physician perform compliance checking, and maintain a complete, accurate patient medical records [1].

CDSS can be considered a knowledge management system (KMS) because it systematically captures and processes the knowledge of physicians and other health professionals to support their decision-making. Various knowledge-intensive technologies including data mining have been used to develop KMS. Data mining is a nontrivial process of identifying valid, novel, potentially useful, and an ultimately understandable pattern in data [2]. 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.

Hypertension is a major contributor to coronary heart disease and the leading cause of death in Korea. Various factors have been implicated in the pathogenesis of hypertension, although the exact cause of hypertension is still unknown.

This paper presents the CDSS for hypertension management using data mining approach. Specifically, the performances of logistic regression and decision tree algorithm, CHAID (Chi-squared Automatic Interaction Detection), in predicting the prognosis of hypertension treatment were compared. This paper also demonstrated how another data mining models, called the clustering and the association rule, could be used in segmenting a diverse patient group into a number if similar subgroups and validating decision rules for hypertension management. In addition, CDSS with three modules (doctor, nurse, and patient) was developed based on data warehouse architecture to collect and integrate various relevant data from hospital operational systems.

Subjects and Methods
1. Subjects
The subjects of the study were 2,507 inpatients and 2,470 outpatients treated from the Severance Hospital between October 1999 and May 2002. Biometric data, including blood pressure, blood glucose, cholesterol, urinary glucose urinary protein, and height and weight, were collected during patients' physical examination. A questionnaire was distributed to all patients to collect information on perceived health status, lifestyle factors (smoking, exercise, and drinking), and demographics (gender, age, job status).

2. Methods
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 by using patient characteristics, history, lifestyle, and test results as independent variables and the hypertension status as dependent variable. The independent variables were stepwise selected, and the corresponding coefficients were computed.

Decision Tree
Decision trees are known as effective classifiers in a variety of domains. In our example, the decision tree categorizes all subjects according to whether or not they are likely to have hypertension. 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.

Clustering

Clustering is the task of segmenting a diverse group into a number of more similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. The records are grouped together on the basis of self-similarity.

Association rule

An association rule gives an occurrence relationship among factors. In this paper, the association rule was used to identify occurrence relationship between hypertension and various modifiable risk factors, such as smoking or drinking, to develop a hypertension management program. An association rule is intended to capture a certain type of dependence among items. Suppose that $i_1 \Rightarrow i_2$, then

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

Results

1. Clustering of Treatment Outcome

We performed a clustering of treatment outcome to examine the differences in the distribution of variables between the controlled group and the uncontrolled group. Table 1 shows the variables that were significantly different between the two groups based on the t-test at 5% level for inpatients and outpatients.

Table 1. Result of clustering of treatment outcome for inpatients

Variables	Controlled		Uncontrolled		t-value*
	Mean	S.D	Mean	S.D	
Change in BP	33.8	12.9	2.9	10.4	49.3
Initial SBP	169.2	16.7	144.7	10.1	34.9
Initial DBP	96.9	12.4	86.4	8.3	19.2
Diabetes	0.2	0.4	0.2	0.4	-1.9
Glucose AC	0.1	0.2	0.1	0.3	-2.1
HDL	0.0	0.1	0.0	0.0	2.8
Exercise	0.5	0.5	0.6	0.5	-2.4
Smoking & drinking	0.8	0.4	0.8	0.4	-2.9

* $P<0.05$

Initial blood pressures (for both systolic blood pressure and diastolic blood pressure) were most significantly different variables for inpatients. This shows that the patients with high blood pressures before the treatment responded better to the treatment than those with low blood pressure. These variables were also used to build a decision tree and logistics regression model as the input variables.

2. Comparison of Decision Tree and Logistic Regression

To obtain misclassification rates, the predictive value decided by the experimental result from each knowledge model was compared with the actual value decided by physicians. The comparison of root ASE (adjusted standard error) and the misclassification (or error) rates for the logistic regression (LR) and decision tree is shown in Table 2. Comparing the ASE and misclassification rates of the LR classification to the decision tree classifications, it is clear that decision tree performs better on both outpatients and inpatients.

3. Validation of Clinical Decision Rules with the Association rule

Table 2. Comparison of the error rates between logistic regression and decision tree

Patient Type	Model	Root ASE	Misclassification Rate
Outpatients	Logistic Reg.	0.3957	0.1942
	Decision Tree	0.3757	0.1867
Inpatients	Logistic Reg.	0.4398	0.2613
	Decision Tree	0.3916	0.2222

We introduced the association rule mining in order to examine how prescription orders, expressed in terms of clinical decision rules, were effective in control of hypertension. If-part of the rule was expressed in terms of characteristics of patients and relevant prescriptions, whereas then-part of the rule was either controlled or uncontrolled. A total of 50 rules were used in the system.

Table 3 shows three accuracy measures for the validation of clinical decision rules, sorted by the support. Rules based on high support and confidence factors represent a higher degree of relevance than rules with low support and confidence factors.

The performance improvement of rules is measured for lift by stating the percentage of the study population for which the prediction will be used and the lift for that subset. In order to show the effectiveness of the clinical decision rule, its lift score should be greater than 1. As seen Table 3 rule 1, rule 12, rule 13, rule 15, and rule 16 had the lift score greater than 1 and were considered as effective rules in controlling high blood pressure.

Table 3. Accuracy measures for validation of clinical decision rules

Rules	Lift	Support (%)	Confidence (%)
Rule 1	1.21	21.2	85.52
Rule 5	0.89	18.29	62.94
Rule 10	0.80	6.32	56.06
Rule 16	1.20	5.47	84.21
Rule 2	1.00	4.44	70.27
Rule 20	0.71	3.08	50.00
Rule 13	1.25	2.56	88.24
Rule 15	1.24	2.39	87.50
Rule 12	1.09	2.22	76.47

4. CDSS for hypertension management

As shown in Figure 2, the CDSS for hypertension management consists of data warehouse and three modules (doctor, nurse, and patient). Data warehouse collects and integrates relevant information from various databases from the hospital information system (HIS).

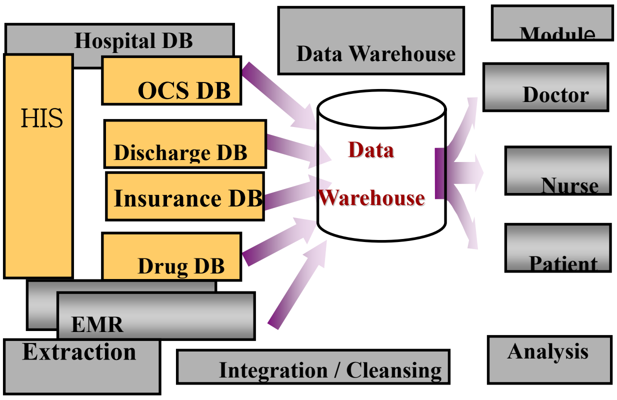

Figure 1. Structure of CDSS for hypertension management

The doctor module provides diagnostic as well as therapeutic information to support a doctor's decision making; the nurse module deals with key entries of patient lifestyle information; and the patient module provides educational information to patient. Doctor module provides diagnostic information based on the clinical practice guidelines from WHO and JNC (Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure). In addition, this module also provides drug recommendation according to priority order, depicted in color, to support decision making for doctors, as seen in Figure 2.

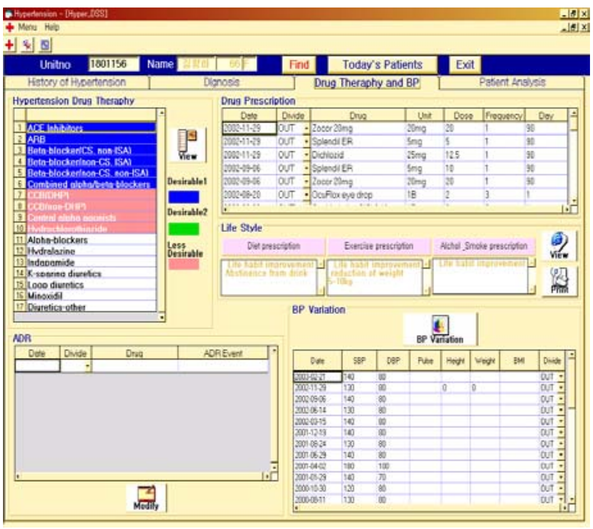

Figure 2. Screen of doctor module in CDSS

Discussion

Diagnosis and therapeutic planning of hypertension are a multi-factorial process, in which symptoms, severity, personal characteristics, lifestyle, past and family history, laboratory tests, and physical examination are assessed. Many studies have identified the relationship among these various factors to improve effectiveness in hypertension management. This study examined the predictive power of data mining algorithms by comparing the performance of logistic regression and decision tree algorithm, called CHAID, since logistic regression has assumed a major position in the healthcare field as a method for predicting or classifying health outcomes based on specific characteristics of each individual case. Comparing the ASE and misclassification rates of the LR classification to the decision tree classifications, unlike in the previous studies, decision tree performs better on both outpatients and inpatients. Long et al. [3] found that the logistic regression performed better than the decision tree in classifying patients as having acute cardiac ischemia. Similarly, Chae et al. [4] found that the discriminant analysis performed better than other data mining methods, neural network and case-based reasoning.

In addition, we presented the association rules in order to show the effectiveness of the clinical decision rules using lift, support, and confidence factor. For example, the association rule showed that five clinical decision rules were effective in controlling high blood pressure. This shows that the association rule can be used as a useful validation tool in developing knowledge-base systems.

Study limitations include the low specificity for the CHAID algorithm and the low confidence measures for the association rule. Other limitations are weak measures for exercise behavior, and no measures for nutrition, stress, and depression. Therefore, future analyses will improve the 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. Since most hypertension patients are required to visit hospitals on a long term basis and their biomedical as well as lifestyle data are well maintained in a temporal database at the hypertension clinic, the sequence rule can be effectively applied to predict treatment outcomes by using the trends data. For example, the sequence rule states that the blood pressure increases if the BMI and cholesterol levels both increase at two consecutive medical examinations.

We have also developed a CDSS with three modules (doctor, nurse, and patient) by using data warehouse architecture. Data warehouse allows the information to flow smoothly from the operational system and enables easy integration with other decision support systems (e.g. case management, cost analysis, etc.).

References

[1] Ball MJ, Lillis J. E-health: transforming the physician: patient relationship, *Int J of Med Informatics* 2001: 61, pp. 1–10.

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

[3] Long WJ, Griffith JL, Selker, HP. A comparison of logistic regression to decision-tree induction in a

medical domain, *Computers and Biomed Research* 1993: 26; 74-97
[4] Chae YM, Lee SH Ho SH, Bae MY, Ohrr HC. Medical decision support system for the management of hypertension, *Informatica 1997:* 21; 219-25.

Address for correspondence
Young Moon Chae, Ph.D. Professor, Graduate School of Public Health, Yonsei University,
C.P.O. Box 8044, Seoul, 120-749, Korea
Email: ymchae@yumc.yonsei.ac.kr