Original Article

Prediction of Hospital Charges for the Cancer Patients with Data Mining Techniques

Jin Oh Kang[1], Suk-Hoon Chung[2], Yong-Moo Suh[2]

Dept. of Radiation Oncology, School of Medicine, Kyung Hee Univ., Seoul, Korea[1],
Graduate School of Business, Korea Univ., Seoul, Korea[2]

Abstract

Objective: Predictions of hospital charges for cancer patients are very important, because they provide a basis for allocating medical resources in the hospital and for establishing national medical policies. But previous studies to predict hospital charges were mainly based on statistical analysis, which has used only a small aspect among huge medical data so that the prediction power was limited. Thus we developed four data mining models, including two artificial neural network (ANN) models and two classification and regression tree (CART) models, to predict both the total amount of hospital charges and the amount paid by the insurance of cancer patients and compared their efficacies. **Methods:** The data was generated from 400,625 medical records of 1,605 cancer patients who had been hospitalized to Kyung Hee University Hospital from March 1, 2003 to February 29, 2004. Clementine 8.1 program was used to build four data mining prediction models, two for the total amount and two for the amount paid by insurance. The variables included all of the data fields of standard medical record form of Korea. The neural network model used feed-forward back propagation method, which had 2 hidden layers. For decision tree model, RELIEFF method was used and the maximum tree depth was set to 30. We divided the dataset into 67% of training dataset and 33% of test dataset, using stratified sampling. Linear correlation coefficient and gain chart were compared. **Results:** The ANN models showed better linear correlation coefficient than the CART models in predicting both the total amount (0.824 vs. 0.791) and the amount paid by insurance (0.838 vs. 0.699). The estimated accuracy of ANN model was more than 98% to predict both total amount and amount paid by insurance. The CART model for total amount showed that the relative importance of the variables were duration of admission(0.073), number of consultation(0.061), and treatment group 16(0.06). The CART model for the amount paid by insurance showed that the relative importance of the cariables were duration of admission (0.09), number of ICU admission (0.063), and number of consultations (0.062). The percent gain of ANN model shows better %gain than CART to predict total amount but to predict amount paid by insurance, ANN showed similar pattern to CART **Conclusion:** The ANN models showed better prediction accuracy than CART models. However, the CART models, which serve different information from ANN model, can be used to allocate limited medical resources effectively and efficiently. For the purpose of establishing medical policies and strategies, using those models together is warranted. ***(Journal of Korean Society of Medical Informatics 15-1, 13-23, 2009)***

Key words: Cost, Cancer, Data Mining, Neural Network Models, Decision Tree Models

Received for review: June 2, 2008; **Accepted for publication:** January 12, 2009.

Corresponding Author: Jin Oh Kang, M.D., Department of Radiation Oncology, School of Medicine, KyungHee University,
Hoiki-dong, Dongdaemun-gu, Seoul, Korea
Tel: +82-2-958-8664, **Fax:** +82-2-962-3002, **E-mail:** kangjino@khmc.or.kr

* This research was supported by the Kyung Hee University Research Fund in 2008. (KHU-20080383)

Ⅰ. Introduction

According to the 2007 report of Health Insurance Review Agency[1], the estimated number of medical insurance claim was 967,735,494 and the total amount of expense was 32,258,975 million Won. The estimated number of claims associated with malignant neoplasm was 707,478 and the total amount of expense for that was 1,604,788 million Won. Although the number of claims related with cancer occupies only 0.07% of all the claims, the amount of expenses of that reaches about 4.97%. Moreover, the average expense per medical claim related to malignant neoplasm was 2,268,323 Won, which is 68 times greater than average expense 33,334 Won of all the medical claims. Accordingly, the hospital charges related to cancer show huge expansion[2]. Therefore, it has become very important to predict the hospital charge related to the cancer for the proper allocation of medical resources and establishment of medical policies in hospitals.

Meanwhile, the medical data is difficult to analyze because of its characteristics such as huge volume and heterogeneity, temporality of the data, and high frequency of null values. It is not uncommon that a patient data exceed more than 1000 fields when the patients' data collection is longer than one year[3]. In addition, medical data consists of various types of data such as image, numbers, videos and electrical signals (EKG, EEG), it is more difficult to analyze than other domains. Moreover, the medical data collection occurs very irregularily because the disease breaks out unexpectedly. The irregularity of medical data leads to many null values in the aspect of time flow. And the null values influence negatively to build proper prediction models. For the reasons mentioned above, these researches have shown very low prediction accuracy, for they could not help using very limited parts of a huge amount of medical data which consists of various data types.

But, researches related to hospital expense are relatively small. Especially, the researches about the prediction of hospital expense using data mining techniques are not easy to find which are known to be better prediction results than other methods[4]. Thus we built prediction models to predict the expenses for the cancer patients using artificial neural network and decision trees methods and compared their efficacies.

Ⅱ. Materials and Methods

In general, appropriate feature selection subset improves accuracy than using total feature set. The authors used RELIEFF suggested by Demsar[5], which is an extension of RELIEF as a feature selection method. Further, a domain expert verified selected features. The features were selected for split criteria in decision trees and in naïve Bayes classifier for building models.

The dataset is based on the records of cancer patients who have been treated in Kyung Hee University Hospital from March 1, 2003 to February 29, 2004. The hospital had more than 130,000 admissions, 4,000,000 out-patients' visits and 5,000 newly diagnosed cancer patients during the period. Among them, the data from 1622 patients who had been hospitalized at least once for the treatment of cancer were enrolled. Data from 17 patients who have no personal identification were excluded. Finally, 400,625 records from 1605 patients were used for the analysis. The variables included all the fields based on the standard medical record form of Korea. Dataset were prepared totally with 66 (65 input variables and 1 output variable) variables (Table 1).

The output variable was set once to predict 'the total amount of hospital charge' and then to predict 'the amount paid by insurance'.

We removed null values, and performed variables selection using RELIEFF algorithm with the help of medical domain experts because building models with a subset of appropriate variables results in better accuracy than with a total set[5]. For example, original fields, 'operation_1' to 'operation_10', consist of two-digit code to identify surgeon and two-digit code to identify operation numbers. They had many null values, because a patient rarely receives more than five operations during an admission period. So, we derived a new field, *no. of operations*, which simply stores the number of times of operations performed on a patient, thereby reducing both the number of null values and the number of fields. And disease codes other than cancer were divided into 19 fields,

Table 1. Input Variables Used for Analysis

No.	Variables	No.	Variables
1	Age	19	Thyroid cancer
2	Sex	20	Leukemia
3	Duration of admission	21	Bladder cancer
4	Department at discharge	22	Pancreatic cancer
5	Doctor ID	23	Prostate cancer
6	Resident ID	24	Other cancer
7	Intern ID	25	Benign tumor
8	Readmission	26-44	Disease group 1-19†
9-12	Patient group 1-4*	45-60	Treatment group 1-16†
14	Lung cancer	61	No. of ICU admissions
15	Hepatoma	62	No. of transfers to other Dept.
16	Colorectal cancer	63	No. of consultations
17	Breast cancer	64	No. of operations
18	Uterine cervix cancer	65	No. of hospital infections

* Patient group were classified by payment methods. Group 1 has insurance, group 2 has government warrant, group 3 has no insurance and group 4 has private insurance.
† Disease group and treatment group denotes the number of diseases and treatments which belong to a patient according to the Korean Standard Classification of Disease-4 and ICD-9CM.

each of which denotes the number of diseases in each disease group. As a consequence, 65 input fields were created in total. 19 disease groups were generated according to the Korean Classification of Diseases and 16 treatment groups were also generated according to the ICD-9CM classification. Each kind of cancer was stored into one of the twelve fields.

Clementine 7.0 (SPSS, Chicago Illinois, USA) program was used to build data mining models. Having carried feature selection using the well-known RELIEFF method, we build models. A feed-forward back-propagation method was used to build neural network models. 70% of original dataset were set to be training dataset and the rest to be test dataset. Two neural network models were created using the training dataset: one to predict the total amount of hospital charge and another to predict the amount paid by insurance. Similarly, two CART models were built using the same input variables selected from the RELIEFF method, as were used to build neural network models. For the CART models, we set maximum tree depth to 30. Same training and test datasets were used as were used when building neural network models. Gini index which indicates a level of impurity of a node is used as a basis for splitting nodes. All the models were built using Clementine 8.1.

Ⅲ. Results

To predict total amount, ANN model was created with 55 input neurons in its 3 hidden layers (Fig. 1). To predict total mount, the ANN model with feature selection showed better linear correlations than without feature selection. The linear correlation coefficient of ANN models with or without feature selection were 0.824 and 0.794, respectively. To predict the amount paid by insurance, ANN model was created with 53 input neurons in its 3 hidden layers. Also, the ANN model with feature selection showed better linear correlations than without feature selection. The linear correlation coefficient of ANN models with or without feature selection were 0.838 and 0.82, respectively (Table 2). The estimated accuracy of neural network model for total amount and the amount paid by insurance was 98.3% and 98.7% respectively. The relative weights of factors that affect hospital charge were analyzed. The most important factors in predicting the total amount were *duration of admission* (0.074), *number of consultations* (0.062) and *treatment group 16* (0.061) (Table 3). *Treatment group 16* is designated as the miscellaneous diagnostic and therapeutic procedures. The most important factors in predicting the amount paid by insurance were *duration of admission*

(0.091), *the number of ICU admission* (0.063) and *the number of consultation* (0.063). Among the variables, physician relative variables such as Doctor ID did not influence on the relative importances.

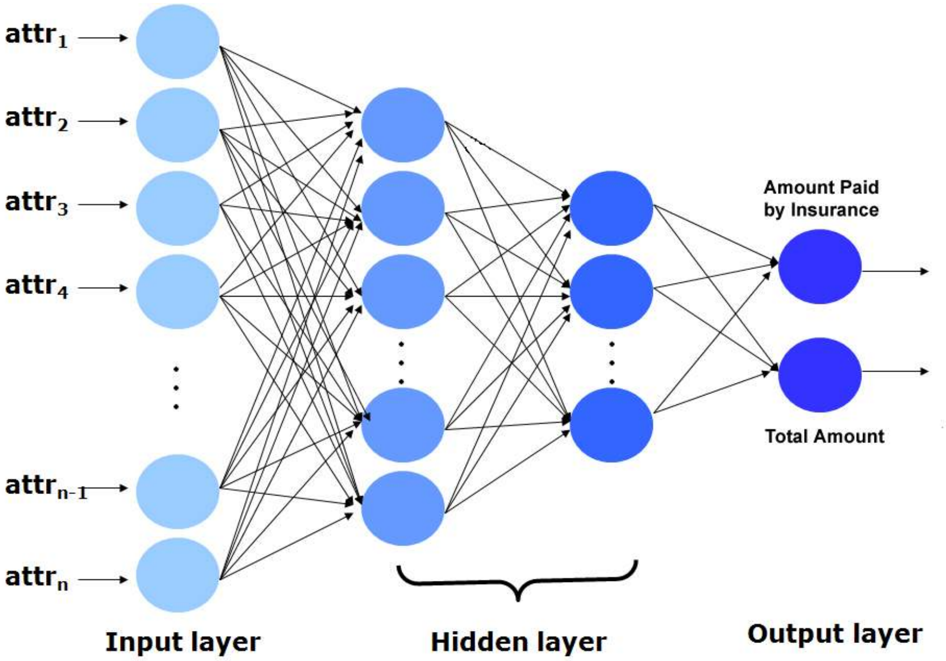

Figure 1. Structure of Artificial Neural Network. The ANN model had two hidden layers. The ANN model for total amount included 56 input neurons and the model for amount paid by insurance included 53 neurons.

Table 2. Linear Correlations of Each Data Mining Models.

Total Amount						
Used Method	Linear Correlation				Occurrences	
	Without FS*		With FS			
ANN	Whole	0.794	Whole	0.824	Whole	3654
	Training	0.819	Training	0.823	70%	2558
	Test	0.803	Test	0.776	30%	1096
CART	Whole	0.791	Whole	0.791	Whole	3654
	Training	0.755	Training	0.755	70%	2558
	Test	0.734	Test	0.734	30%	1096
The Amount Paid by Insurance						
Used Method	Linear Correlation				Occurrences	
	Without FS		With FS			
ANN	Whole	0.82	Whole	0.838	Whole	3654
	Training	0.786	Training	0.899	70%	2558
	Test	0.79	Test	0.816	30%	1096
CART	Whole	0.699	Whole	0.699	Whole	3654
	Training	0.687	Training	0.687	70%	2558
	Test	0.73	Test	0.73	30%	1096

*FS: Feature Selection

Table 3. Relative Importance of Each Input Variables for the Amount

Total Amount			Amount Paid by Insurance		
Rank	Feature	Relative Importance	Rank	Feature	Relative Importance
1	Duration of admission	0.073684654	1	Duration of Admission	0.09091958
2	No. of consultation	0.061788311	2	No. of ICU admission	0.06318038
3	TG16	0.060594119	3	No. of consultation	0.06276517
4	No. of ICU admission	0.050140664	4	TG16	0.06241662
5	Nosocomial infection	0.036508323	5	Nosocomial infection	0.04538375
6	No. of Transfer to other dept.	0.035945577	6	DG18	0.03961705
7	No. of Operations	0.032656611	7	DG1	0.03544701
8	DG18	0.027907454	8	DG8	0.0342562
9	DG1	0.02764464	9	No. of transfer to other dept.	0.03349989
10	TG9	0.022238082	10	DG10	0.03144369
11	DG8	0.022163646	11	TG7	0.02843765
12	DG10	0.021860789	12	DG9	0.02302769
13	TG7	0.021783816	13	No. of Operations	0.02292118
14	DG9	0.02153655	14	Age	0.02270374
15	Age	0.020800857	15	DG3	0.0224453
16	DG3	0.019286175	16	TG8	0.0199803
17	TG8	0.017990889	17	TG9	0.01832424
18	TG1	0.015112707	18	DG17	0.01692561
19	DG7	0.015038423	19	DG7	0.01648804
20	DG17	0.013499681	20	DG2	0.01465006

ICU: Intensive Care Unit

DG: Disease group according to the Korean Standard Classification of Disease (Appendix 1)

TG: Treatment group according to the ICD-9 (Appendix2)

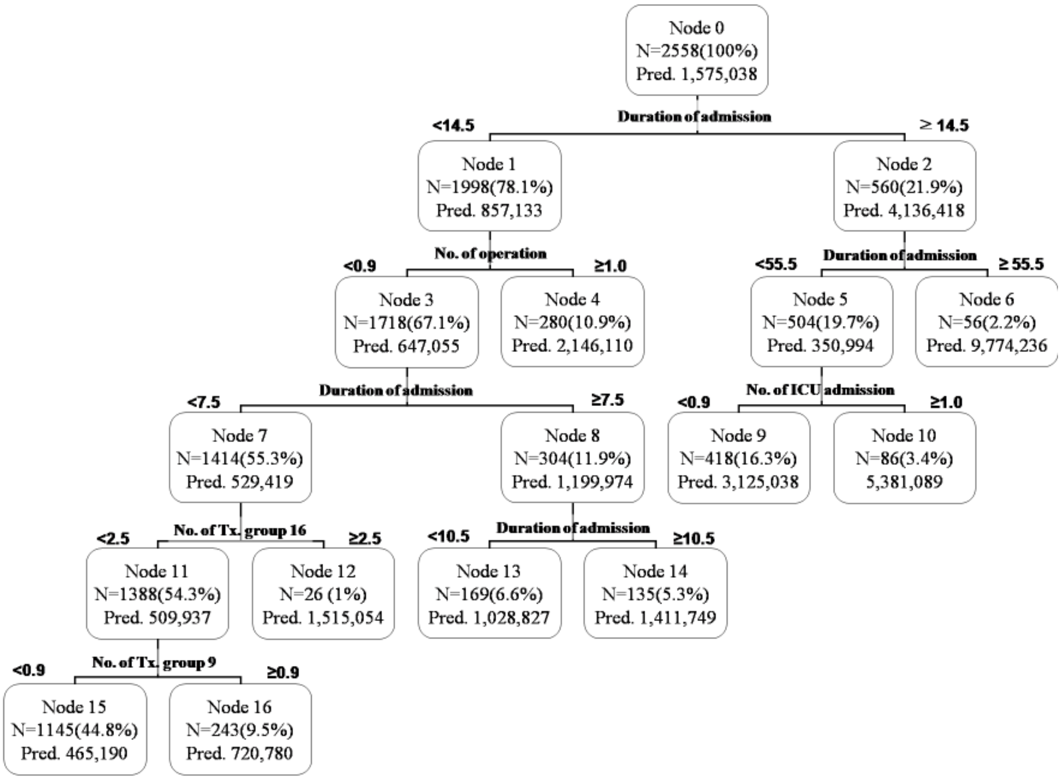

Figure 2. Decision Tree Model for Total Amount. The duration of admission was the most important variable to split.

The most important variable was the duration of admission, where the first branch was split at the point of 14.5 days of admission. The second important variable is the *number of operations* at the left branch of the tree. Then, the nodes at other levels were split based on *number of operations*, *treatment group16*, and *treatment group 9*. The number of rules in the resulting rule set was eleven and these rules classified the part of high hospital expense well. For example, consider these rules: IF "(1) days of admission ≥14.5 and (2) days of admission <55.5 and (3) the number of ICU admission <0.5" THEN "3,125,038". The correlation coefficients of the CART models were 0.791 for the total amount of hospital charge and 0.699 for the amount paid by insurance regardless of feature selection. In the CART model for amount paid by insurance, the *duration of admission* was most important variable also but instead of the *number of operation*, *department* was demonstrated as second important variable. The other variables were number of *ICU admission* and *treatment group 16*(Fig. 3).

The percent gain of ANN model shows better %gain than CART to predict total amount but to predict amount paid by insurance, ANN showed similar pattern to CART (Fig. 4).

Ⅳ. Discussion

With the development of information technologies, it becomes possible to record and search the historical states of patients through database. As a result, tremendous medical data of various types have been accumulated into a database of a medical information system. But because of the complexity, the medical data is difficult to analyze. It frequently occurs that if the period of data collection is longer than one year, the record of such a patient has more than 1000 fields[3]. In addition, medical data consists of various types of data such as image, numbers, video, etc and

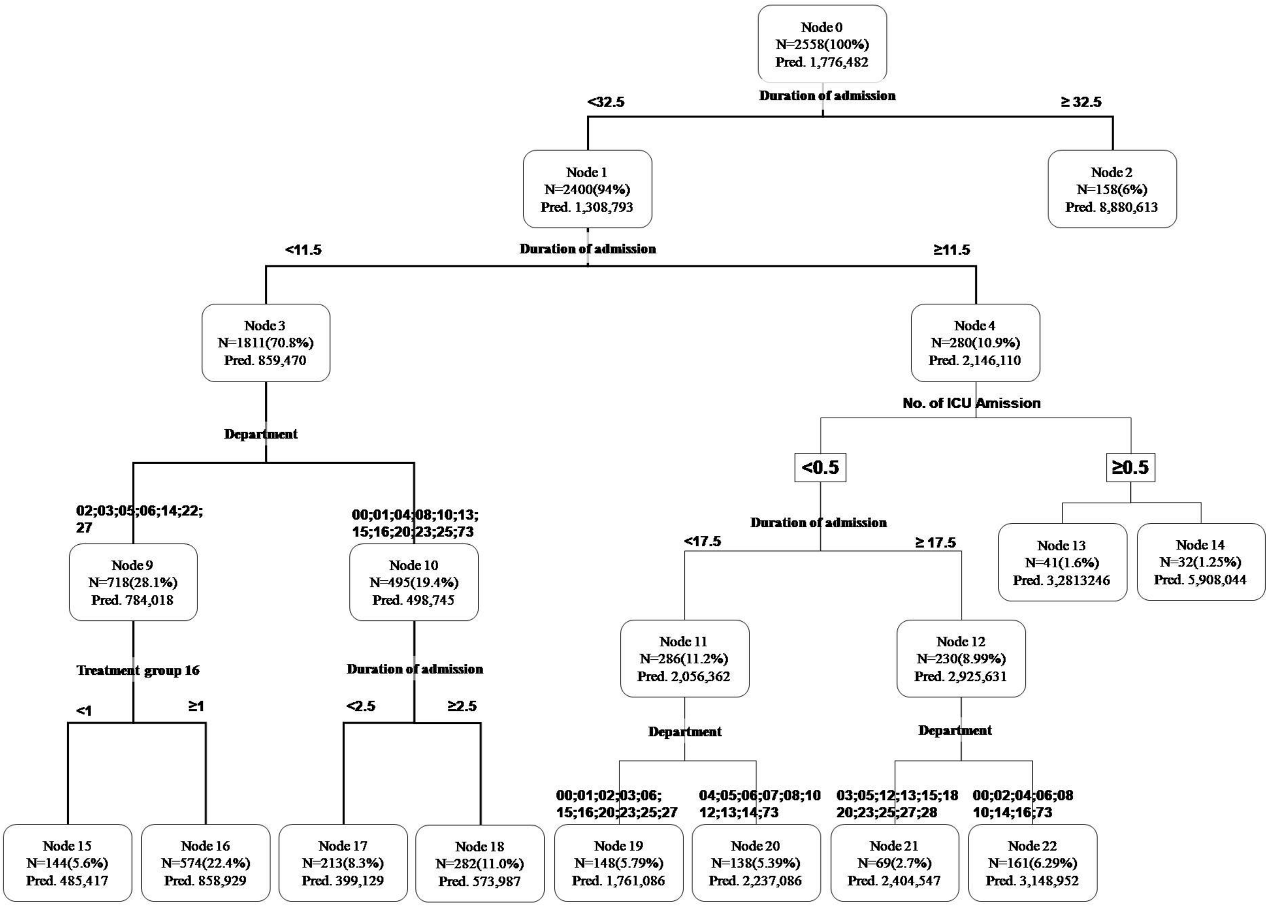

Figure 3. Decision Tree Model for Amount Paid by Insurance. The duration of admission was the most important variable to split.

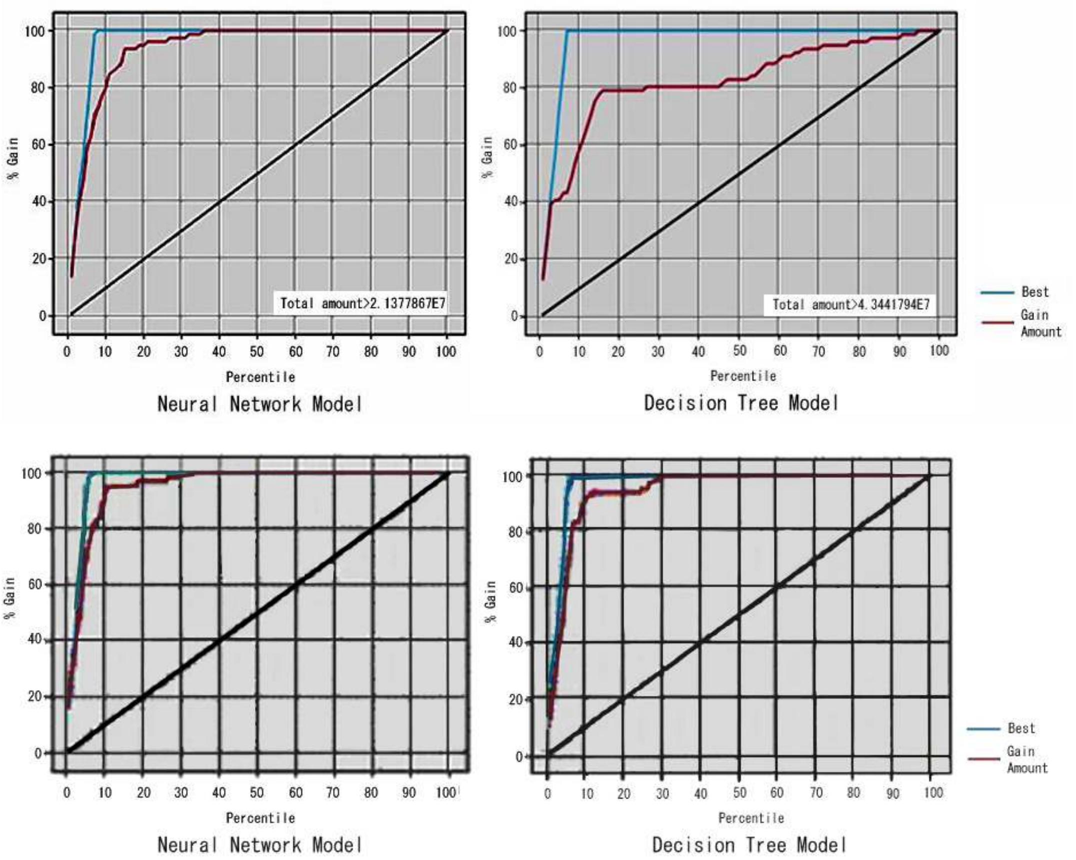

Figure 4. The y-axis Shows the Percentage of Gain. The x-axis shows the percentage of samples selected based on the data mining model, which is a fraction of total samples selected. ANN model shows better %gain than CART to predict total amount (upper). But ANN and CART showed similar pattern to predict amount paid by insurance (lower).

thus it is more difficult to analyze than simple data collected in other domains. The characteristics of medical data include 1) huge volume, 2) heterogeneity, 3) temporal (historical) data, and 4) relatively high frequency of null value.

There have been several researches related to the prediction of hospital charges of cancer patients using statistical analysis such as regression or ANOVA[6-8]. Since most of these researches were based on a small number of variables among many affecting the hospital charge, their prediction accuracy was not satisfactory. Therefore, these regression models can be hardly used for the prediction of hospital expense. In this aspect, data mining has emerged as an analytical method which can discover interesting knowledge from tremendous data from diverse domains using various techniques such as pattern recognition, statistics, database, machine learning and so forth[9]. Data mining can discover interesting knowledge from a large amount of data in the form of rules, patterns or trends, which may be difficult to obtain using traditional statistical methods[10]. In the medical field, data mining techniques such as association rules, artificial neural network, decision tree and genetic algorithm have been used to achieve various objectives and several data mining studies concerning medical cost were performed. Marshall et al. [11] built conditional phase-type distribution model to predict elderly patient's outcome and duration of stay. In the research, they were able to identify that there is a strong relationship between Barthel grade, patient outcome and length of stay. Chae et al. 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[12]. They built logistic regression, CHAID and C5.0 models from a dataset related to hypertensive and non-hypertensive and compared their performance one another. They reported that the CHIAD algorithm performed better than the logistic regression in predicting hypertension, and C5.0 had the lowest predictive power. These researches are pioneers to introduce data mining models to predict medical costs. But still, data mining models for cancer patients' cost are rarely found.

Thus, we aimed the objective of this study to build prediction models for the hospital charge of cancer patients because there are very limited researches concerning for the cost of cancer patient in Korea, where the medical insurance system is very unique and governed by the governance. The current research to build data mining models to predict cost of whole cancer patients may be the first in Korea. Although the authors have reported a data mining models to predict cost of cancer patients, it was limited only for the colorectal cancer[13]. However, our current research has some limitations which should be taken into account in later researches. The dataset we have used to build the predictive models did not have records of all treatments and examinations which a patient has experienced, because they were not digitalized at the time of data collection. Also, we could not make use of the information indicating the staging of cancer, the use of which may enable us to build a more exact predictive model in various aspects. If we have more digitalized records in a few years to come as we plan, we will make more accurate predictive models of the hospital expense. Further, as more cases of cancer patients are accumulated into the medical information systems, we may apply other data mining technique such as case-based reasoning to the medical data to get the similar results which may be better.

In this study, we included the all kinds of cancers as an input, so that prediction of hospital charge of cancer patients could be made, independently of the type of cancer. And we used artificial neural networks and decision trees to build prediction models and compare their prediction accuracy because those two models are most commonly used data mining tools. Although, our results showed that both models are efficient to predict cancer patient's hospital charge, the prediction accuracy of ANN model was slightly higher than that of the CART model and the ANN model shows better percent gain for predicting total amount (Fig. 3). Generally, ANN models have shown higher sensitivity, specificity and prediction accuracy than other data mining techniques. In the previous research, the authors have compared the performance of ANN model and that of CART model in predicting hospital charges of colorectal cancer patients[13]. The result showed that ANN model showed better performance than CART model. The current results demonstrate that with the complicated database, the ANN model shows better prediction than other models. Chien et al.[14] applied three data mining techniques to improve prediction of post-operative complication of gastric cancer. The data mining techniques included Artificial Neural Networks (ANN), Decision Tree (DT) and Logistic Regression (LR). The results indicated that ANN was a better technique than DT and LR in predicting post-operative complication. Goss et al. [15] compared traditional decision support system such as Binary Logit Regression (BLR) and non-parametric methodologies such as neural network (NN) model to provide objective measures of the likelihood of Intensive Care Unit (ICU) recovery. The study showed that the NN technique predicts mortality rates more correctly than BLR, and offers a promising non-parametric alternative to the parametric methodologies in hospital settings. For the cancer patients, Fogel et al. were first to apply neural networks and linear classifiers to breast cancer patients' dataset[16]. They used cross-validation to estimate error rate and relied on evolutionary computation to mitigate the black-box problem. However, the fact that neural network models could not give an adequate explanation of their results to doctors is very fatal because they hardly accept the result of 'black box' classifiers unless their performances overwhelm other classifiers[17]. To overcome such limitations and adjust weights in neural networks, genetic algorithm has been used. Bojarczuk et al. [18] proposed a new constrained-syntax genetic programming (GP) algorithm for discovering classification rules in five medical data sets: chest pain, Ljubljana breast cancer, dermatology, Wisconsin breast cancer, and pediatric adrenocortical tumor. The proposed GP algorithm obtained good results with respect to predictive accuracy and rule comprehensibility, by comparison with C4.5 and Boolean inputs (BGP).

Meanwhile, decision tree is one of the most frequently used techniques for classification and prediction tasks not only in medical data mining area but also in other areas. This enables one to predict prognoses and diagnoses using tree-structured models and to identify useful features which play an important role in making such predictions. Demšar et al. [5] built models which can be used to predict whether a severe trauma patient would survive or not. They found out that features selected as split criteria in the decision tree corresponded to factors which other researchers found to have an influence on the survival of a patient who suffered from severe trauma. But the size of their dataset was so small (68 cases) that their models could not be used as a prediction model for severe trauma patient's survival. Breault et al.[19] have analyzed diabetes patients' data with CART and discovered that a patient's age rather than whether one has other diseases or not has an association with adjustment of blood sugar.

Although the ANN model showed better results to predict cost, the fact that neural network models could not give an adequate explanation for the result of 'black box' classifiers, CART have their own advantages and unique use so that both models are needed to build proper strategies of hospital and national policies.

REFERENCE

1. Korea HIRaAS. National Health Insurance Statistics 2007. 2008 [updated 2008; cited 2008]; Available from: http://www.hira.or.kr/cms/rd/rdi_statistics/morgue/1188982_5295.html.
2. Yoon SJ, Lee H, Shin Y, Kim YI, Kim CY, Chang H. Estimation of the burden of major cancers in Korea. J Korean Med Sci. 2002 Oct;17(5):604-10.
3. Hirano S, Tsumoto S. Multiscale analysis of long time-series medical databases. AMIA Annu Symp Proc. 2003:289-93.
4. Ismael MB, Eisenstein EL, Hammond WE. A comparison of neural network models for the prediction of the cost of care for acute coronary syndrome patients. Proc AMIA Symp. 1998:533-7.
5. Demsar J, Zupan B, Aoki N, Wall MJ, Granchi TH, Robert Beck J. Feature mining and predictive model construction from severe trauma patient's data. Int J Med Inform. 2001 Sep;63(1-2):41-50.
6. Brooks SE, Ahn J, Mullins CD, Baquet CR, D'Andrea A. Health care cost and utilization project analysis of comorbid illness and complications for patients undergoing hysterectomy for endometrial carcinoma. Cancer. 2001 Aug 15;92(4):950-8.
7. Penberthy L, Retchin SM, McDonald MK, McClish DK, Desch CE, Riley GF, et al. Predictors of Medicare costs in elderly beneficiaries with breast, colorectal, lung, or prostate cancer. Health Care Manag Sci. 1999 Jul;2(3): 149-60.
8. Tollestrup K, Frost FJ, Stidley CA, Bedrick E, McMillan G, Kunde T, et al. The excess costs of breast cancer health care in Hispanic and non-Hispanic female members of a managed care organization. Breast Cancer Res Treat. 2001 Mar;66(1):25-31.
9. Dayhoff JE, DeLeo JM. Artificial neural networks: opening the black box. Cancer. 2001 Apr 15;91(8 Suppl):1615-35.
10. Goss E, Vozikis G. Improving Health Care Organizational Management Through Neural Network Learning. Health Care Manag Sci. 2002;5(3):221-7.
11. Marshall AH, McClean SI, Millard PH. Addressing bed costs for the elderly: a new methodology for modelling patient outcomes and length of stay. Health Care Manag Sci. 2004 Feb;7(1):27-33.
12. Chae YM, Ho SH, Cho KW, Lee DH, Ji SH. Data mining approach to policy analysis in a health insurance domain. Int J Med Inform. 2001 Jul;62(2-3):103-11.
13. Lee SM, Kang JO, Suh YM. Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs. decision tree models. J Korean Med Sci. 2004 Oct;19(5):677-81.
14. Chien CW, Lee YC, Ma T, Lee TS, Lin YC, Wang W, et al. The application of artificial neural networks and decision tree model in predicting post-operative complication for gastric cancer patients. Hepatogastroenterology. 2008 May-Jun;55(84):1140-5.
15. Goss EP, Vozikis GS. Improving health care organizational management through neural network learning. Health Care Manag Sci. 2002 Aug;5(3):221-7.
16. Fogel DB, Wasson EC, 3rd, Boughton EM, Porto VW. Evolving artificial neural networks for screening

features from mammograms. Artif Intell Med. 1998 Nov;14(3):317-26.

17. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med. 2001 Aug;23(1):89-109.
18. Bojarczuk CC, Lopes HS, Freitas AA, Michalkiewicz EL. A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets. Artif Intell Med. 2004 Jan;30(1): 27-48.
19. Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artif Intell Med. 2002 Sep-Oct; 26(1-2):37-54.

Appendix 1. Disease Group according to KSD-4

Group	Disease
Group 1	I. Certain infectious and parasitic diseases (A00-B99)
Cancers	II. Neoplasms (C00-D48)
Group 2	III. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (D50-D89)
Group 3	IV. Endocrine, nutritional and metabolic disease(E00-E90)
Group 4	V. Mental and behavioral disorders(F00-F99)
Group 5	VI. Diseases of the nervous system (G00-G99)
Group 6	VII. Diseases of the eye and adnexa(H00-H59)
	VIII. Diseases of the ear and mastoid process(H60-H95)
Group 7	IX. Diseases of the circulatory system (I00-I99)
Group 8	X. Diseases of th respiratory system (J00-J99)
Group 9	XI. Diseases of the digestive system (K00-K93)
Group 10	XII. Diseases of the skin and subcutaneous tissue (L00-L99)
Group 11	XIII. Diseases of the musculo-skeletal system and connective tissue(M00-M99)
Group 12	XIV. Diseases of The genitoruinary system (N00-N99)
Group 13	XV. Pregnancy, childbirth and the puerperium(O00-O99)
Group 14	XVI. Certain conditions originating in the perinatal period(P00-P96)
Group 15	XVII. Congenital malformations, deformations and chromosomal abnormalities(Q00-Q99)
Group 16	XVIII. Symptoms, signs and abnormal clinical and laboratory findings, NEC (R00-R99)
Group 17	XIX. Injury, poisoning and certain other consequences of external causes (S00-T98)
Group 18	XX. External causes of morbidity and mortality(V01-Y98)
Group 19	XXI. Factors influencing health status and contact with health servisces (Z00-Z99)

Appendix 2. Classification of Treatment Group according to ICD-9

Group	Classification
Group 1	Operations on the nervous system
Group 2	Operations on the endocrine system
Group 3	Operations on the eye
Group 4	Operations on the ear
Group 5	Operations on the nose, mouth, and pharynx
Group 6	Operations on the respiratory system
Group 7	Operations on the cardiac system
Group 8	Operations on the hemic and lymphatic system
Group 9	Operations on the digestive system
Group 10	Operations on the urinary system
Group 11	Operations on the male genital organs
Group 12	Operations on the female genital organs
Group 13	Obstetrical procedures
Group 14	Operations on the musculoskeletal system
Group 15	Operations on the integumentary system
Group 16	Miscellaneous diagnostic and therapeutic procedures

The CART models for total amount and amount paid by insurance were generated. Figure 3 and 4 represents the decision tree of CART for total amount and amount paid by insurance, respectively. The CART model showed same linear correlation coefficient regardless of feature selection. (Table 2) The CART model for total amount had 6 layers and the impurity estimation was measured using Gini index. Unlike the ANN model, CART model was not influence by feature selection. In the resulting decision tree for total amount, the root node was spilt based on the value of *duration of admission* (Fig. 2).