J Korean Med Sci 2004; 19: 677-81
ISSN 1011-8934

Comparison of Hospital Charge Prediction Models for Colorectal Cancer Patients: Neural Network vs. Decision Tree Models

Analysis and prediction of the care charges related to colorectal cancer in Korea are important for the allocation of medical resources and the establishment of medical policies because the incidence and the hospital charges for colorectal cancer are rapidly increasing. But the previous studies based on statistical analysis to predict the hospital charges for patients did not show satisfactory results. Recently, data mining emerges as a new technique to extract knowledge from the huge and diverse medical data. Thus, we built models using data mining techniques to predict hospital charge for the patients. A total of 1,022 admission records with 154 variables of 492 patients were used to build prediction models who had been treated from 1999 to 2002 in the Kyung Hee University Hospital. We built an artificial neural network (ANN) model and a classification and regression tree (CART) model, and compared their prediction accuracy. Linear correlation coefficients were high in both models and the mean absolute errors were similar. But ANN models showed a better linear correlation than CART model (0.813 vs. 0.713 for the hospital charge paid by insurance and 0.746 vs. 0.720 for the hospital charge paid by patients). We suggest that ANN model has a better performance to predict charges of colorectal cancer patients.

Key Words : *Hospital Charges; Neural Networks (Computer); Decision Trees; Data Mining; Colorectal Neoplasms*

Seung-Mi Lee[*,†], **Jin-Oh Kang**[‡], **Yong-Moo Suh**[*]

Graduate School of Business, Korea University*, Seoul; Medical Cell, Samsung SDS, Gyeonggido†; Medical College, Kyung Hee University‡, Seoul, Korea

Received : 27 February 2004
Accepted : 20 May 2004

Address for correspondence
Jin Oh Kang, M.D.
Medical College, Kyung Hee University, 1 Hoiki-dong, Dongdaemun-gu, Seoul 130-701, Korea
Tel : +82.2-958-8664, Fax : +82.2-962-3002
E-mail : kangjino@khmc.or.kr

INTRODUCTION

The incidence of colorectal cancer is rapidly increasing in Korea as the life style and the dietary pattern change. The incidence has been increased about 200% in recent 15 yr occupying more than 10% of total cancer incidence in Korea (1). Accordingly, the hospital charges related to colorectal cancer show huge expansion (2). Therefore, it has become very important to predict the hospital charges related to the colorectal cancer for the allocation of medical resources and the establishment of medical policies in Korea.

There are several researches related to the prediction of hospital charges of cancer patients using a statistical analysis such as regression or aualysis of variance (3-5). Since most of these researches were based on a small number of variables among many affecting the hospital charge, their prediction accuracy was less satisfactory than expected. For example, Penberthy et al. developed the prediction models for hospital expense of elderly patients with breast, colorectal, lung or prostate cancers but the R-square values were only 38-49% (4). In this aspect, better prediction models for cancer care costs are warranted. In the meantime, data mining has emerged as an analytical method which can discover interesting knowledge from tremendous data using various technologies from diverse domains such as pattern recognition, statistics, database, machine learning and so forth (6). To achieve various objectives, data mining techniques such as association rules, cluster analysis, classification, artificial neural network (ANN), decision tree, genetic algorithm are used. Among these, we built hospital charge prediction models using ANN and decision trees because these two methods are able to use more various types of data than statistical methods.

MATERIALS AND METHODS

Dataset

The dataset is based on the digitalized records of colorectal cancer patients who have been treated in Kyung Hee University Hospital from January 1999 to December 2002. This 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, 492 patient's 1,022 admission records with 154 variables were used to build prediction models. The median number of admission for a patient is 2.1 (ranges 1-14) and the average days of admission were 11.38 days.

Korea has a single national health insurance system for all the Korean people by the national policy. The health insurance system provides only a portion of total charges of the

patients. For example, the medical charges are composed of two parts, one is paid by health insurance and the other is paid by patients. The charges for hospital diet, the charges for superior class room, charges for assigning a specified doctors, and the charges for some expensive radiological examinations such as MRI or ultrasonography are 100% paid by patients. The other parts are paid 20% by patients and 80% by health insurance. So, we have analyzed both the charges paid by insurance (charge A) and the charges paid by patients (charge B) to reveal actual financial burden of the patients.

Preprocessing of raw data

The initial dataset cannot be used without preprocessing because it has so many variables, some of which has quite a few null values. So variable selection is one of the most careful steps since the prediction accuracy depends on the set of variables used for the analysis. In general, building models with a subset of appropriate variables results better accuracy than with a total set (7). Thus we have performed variable selection and null value processing with the help of medical domain experts. For example, fields 'operation_1' to 'operation_10' each consisting of a two-digit doctor code and a two-digit operation number have many null values, because it is rare that a patient receives more than ten operations during an admission period. So, we derived a new field 'operation_count', which simply stores the sum of the numbers of operations, thereby reducing both the number of null values and the number of fields. Similarly, fields 'diagnosis (Dx)_1' to 'diagnosis_12' store all the disease codes of a patient. We categorized these fields into 6 'cancer_code' fields and 12 'other_Dx' fields and additionally created 'other_Dx_count' field to store the number of other diagnosis. Each 'cancer_code' field denotes the location of primary cancer according to International Classification of Diseases for Oncology-3 (ICD-O-3) and each 'other-Dx' field stores disease codes other than cancer. We further divided 'other_Dx' into 23 groups, creating 'other_ Dx_group_1' to 'other_Dx_group_23', each of which represents a disease group according to the Korean Standard Classification of Disease. ICD-9CM procedure codes stored in 'Mx_1' to 'Mx_10' fields were categorized into 4 groups according to the treatment taken to each patient, such as OPx, RTx, RDx and CTx, related to operational treatments (OPx), radiological tests (RDx), radiation therapy (RTx) and chemotherapy (CTx), respectively. As a consequence, 61 new fields were derived and among these 38 fields were selected as input variables. Then, we divided the dataset into a training dataset (681 records: 67%) and a test dataset (340 records: 33%), using stratified sampling method. We used Clementine 7.1 program to build ANN and classification and regression tree (CART) models.

Table 1. The descriptions of input variables

Variables	Explanation
Hospital_stay	the days of hospital admissions
ICU_count	the number of admissions to intensive care unit (ICU)
Transfer_count	the number of transfers to other department
Consult_count	the number of consults to other doctors
Operation_count	the number of surgical operations
Other_Dx_count	the number of diagnosis other than cancer
OPx	the number of operational treatments
RDx	the number of radiological tests
RTx	the number of radiotherapy treatments
CTx	the number of chemotherapy treatments
Age	patient age in Korean age
Sex	male or female
Main_diagnosis	one of C18, C19, C20 and other
Hospital_infection	nosocomial infection or not
Patient_diff	patient classification according to insurance status
Other_Dx_group 1-Other_Dx_group23	non-cancer disease code grouped according to the Korean Standard Classification of Disease

Artificial neural network model

Artificial neural network models are created using a training dataset. After many attempts to train neural networks, two ANN models were created heuristically using training dataset. We named models for hospital charge A and hospital charge B as NN-A and NN-B respectively. Test dataset was used to test the models' prediction accuracy. 38 input variables are listed in Table 1.

Classification and regression tree model

Similarly, we built two CART models, CART-A for charge A and CART-B for charge B. We used the same 38 input variables which were used for ANN models. When creating CART models, the Gini index which represents a level of impurity of a node is used as a basis for splitting the node. Training dataset and test dataset were identical to those for building and testing the ANN models.

Table 2. Prediction errors for artificial neural network models

	NN-A		NN-B	
	Training dataset	Test dataset	Training dataset	Test dataset
Minimum error	-3608365	-6494174	-2832708	-1789913
Maximum error	3781317	535266	3556667	2982913
Mean error	-19430.637	6199.400	-47914.090	-91948.624
Mean absolute error	554972.335	683092.894	343371.960	357254.753
Standard deviation	753496.943	1020807.415	527882.867	524473.663
Linear correlation	0.886	0.813	0.764	0.746
Occurrences	681	340	681	340

Table 3. The prediction error for CART-A and CART-B

	CART-A		CART-B	
	Training dataset	Test dataset	Training dataset	Test dataset
Minimum error	-3560630	-3995353	-2148996	-1818531
Maximum error	10036520	9021814	3557912	3299262
Mean error	0.432	38565.047	0.369	-69047.771
Mean absolute error	645226.902	754529.918	299721.426	345065.865
Standard deviation	1008764.118	1224812.987	486612.904	552560.017
Linear correlation	0.784	0.713	0.804	0.720
Occurrences	681	340	681	340

RESULTS

ANN model

Sensitivity analysis conducted while building the NN-A model reveals that the most important variables were 'hospital_stay' (relative importance 0.692) followed by 'other_Dx_group2' (0.190), 'hospital_infection' (0.109), 'other_Dx_group 5' (0.079), 'other_Dx_group7' (0.078), 'other_Dx_group17' (0.068), CTx (0.067) and 'main_diagnosis' (0.062). In the training dataset, the mean absolute error was 554,972 won and the linear correlation was 0.886 while in the test dataset, the mean absolute error was 683,093 won and the linear correlation was 0.813. These coefficients suggest that there was a strong linear correlation between the actual and the predicted hospital charges.

The important variables of the NN-B model were 'hospital _stay' (relative importance 0.299), 'other_Dx_group2' (0.168), 'other_Dx_count' (0.103), 'other_Dx_group6' (0.059), 'other _Dx_group19' (0.059), 'operation_count' (0.054), 'other_Dx_ group23' (0.053), 'patient_diff' (0.051), and 'hospital_infection' (0.048). In the training dataset, the mean absolute error was 343,372 won and the linear correlation was 0.764, while in the test dataset, the mean absolute error was 357,254 won and the linear correlation was 0.746. The linear correlation was weaker than in model NN-A, but it was still high in this model. Prediction errors among these neural net models including the mean absolute error and the linear correlation were compared in Table 2.

CART model

CART-A was created as a decision tree with 16 leaf nodes in 8 levels. Several important rules from the created decision tree are as follows: 1) IF 15.5≤'hospital_stay' THEN charge A=4,000,707 (191 cases), 2) IF 5.5≤'hospital_stay'<15.5 and 'operation_count'>=0.5 THEN charge A=2,648,716 (71 cases), 3) IF 'hospital_stay'<1.5 and 'age'>=52.5 and CTx>0.5=THEN charge A=1,437,424 (67 cases), 4) IF 'hospital_stay'<4.5 and CTx<0.5 THEN charge A=523,286 (65 cases), 5) IF 5.5≤'hospital_stay'<15.5 and 'operation_ count'<0.5 and 'other_Dx_group5'>=0.5 and 'other_Dx_ count'<1.5 THEN charge A=2,784,923 (58 cases). The numbers at the end of each rule indicate the number of patients whose charge A can be predicted by the rule. Charge A of 452 patients out of the 681 patients was able to be predicted by the above five rules. The variables and values of the above rules played an important role in predicting the charge A. Besides these variables, the 'sex' variable was also used as a split criterion. In the training dataset, the mean absolute error was 645,227 won and the linear correlation was 0.784. In the test dataset, the mean absolute error was 754,530 won and the linear correlation was 0.713.

CART-B was created using the same input variables as in CART-A. Since the total amount of charge B is relatively smaller than that of charge A, the initial tree has made many partitions even with small intervals of charge. So we pruned its branches and simplified the tree into four levels. The major variables used for the splits were 'hospital_stay', 'RDx', 'main diagnosis', 'OPx', 'other_Dx_group3', 'age', and 'sex'. In the training dataset, the mean absolute error was 299,721 won and the linear correlation was 0.804. In the test dataset, the mean absolute error was 345,066 won and the linear correlation was 0.720. Prediction errors of these neural net models were compared in Table 3.

Table 4. Comparison of ANN and CART models

Prediction models		Pearson correlation coefficients		Mean absolute error	
		Training dataset	Test dataset	Training dataset	Test dataset
Charge A	NN-A	0.886	0.813	554,972	683,093
	CART-A	0.784	0.713	645,227	754,530
Charge B	NN-B	0.764	0.746	343,372	357,254
	CART-B	0.804	0.720	299,721	345,066

Comparison of the two models

We have built four predictive models for hospital charge, NN-A, NN-B, CART-A and CART-B. These models were compared with respect to the linear correlation coefficient and the mean absolute error (Table 4). In predicting the charge A, NN-A showed a better the linear correlation coefficient than CART-A for both datasets. And in predicting the charge B, NN-B showed a better linear correlation coefficient than CART-B only for the test dataset. In the aspect of the mean absolute error, CART-A was inferior to NN-A but CART-B was superior to NN-B in both datasets.

DISCUSSION

In the previous studies related to the hospital charge of

cancer care patients, statistical models showed limited explanatory power because of the limitations both in the data type and in the number of input variables (3-5, 8, 9). There were several researches to build prediction models with regression analysis for the health care costs (3, 4, 10). Penberthy et al. developed the prediction models for medicare cost of elderly patients who had suffered from breast, colorectal, lung or prostate cancers, resulting R-square values of 0.38 to 0.49 (4). Brooks et al. analyzed the health care cost for the patients undergoing hysterectomy for endometrial carcinoma (3). They used linear, stepwise and three-stage regression analyses to build a prediction model with resulting R-square value of 0.71. Tollestrup et al. applied Tobit regression model to breast cancer dataset and investigated whether the hospital charge for Hispanics was different from that of non-Hispanics (5). But the prediction error was not described in this article. Therefore, these regression models had significant limitations to be used for the prediction of hospital charge.

Up to the present, studies about the prediction of hospital charge of cancer patients using a data mining method are very rare. There have been some artificial neural network analyses in the medical field for other concern. Ismael et al. developed the predictive ANN models of hospital charge for acute coronary syndrome patients and compared the performances of these models with one another (11). They used 16 input variables representing the patient's status and complication and built four models for hospital charge. Among them, two models have shown the classification accuracy of 79%. Marshall et al. combined Bayesian belief network and phase-type distribution to build models for predicting the number of days of hospital stay for geriatric patients (12). Walczak and Scharf built models to predict the amount of blood which is needed in operations using ANN based on the radial basis function (13). They showed that the prediction accuracy of ANN models is better than that of MSBOS (Maximum Surgical Blood Order Schedule). ANN has also been used to compare cancer patient's prognosis. Burke et al. reported that the ANN was significantly more accurate than TNM staging system when both use the TNM prognostic factors alone (14).

Though the success of ANN in medical field is being demonstrated in many areas, the ANN is not widely used as an alternative of the statistical method to predict hospital charges until recently. In fact, there are debates whether the performance of ANN is better than statistical methods. In an article comparing ANN with statistical method, Sargent suggested that the artificial neural network should not replace standard statistical approaches as the method of choice for the classification of medical data (15). They referred the reason why ANN is not universally outperforming a regression technique to the limited amount of data and incorrectly measured variables. Regarding this aspect, Michie et al. reported that the success of the ANN is correlated directly with the success of the statistical procedures used (16). As a consequence, to have a good prediction result in ANN analysis, it is important to analyze integrated medical data to select appropriate features. Thus in our study, the feature selection was performed by medical domain experts who can identify and differentiate the diverse medical data so as to result in a good prediction performance. ANN being very useful when creating predictive models where the interrelationships and the behavior of the various problem parameters are unknown (17), it may allow users to create predictive models without explicitly specifying such information.

If we consider the results only from the test dataset in terms of the linear correlation, NN-A and NN-B were better than CART-A and CART-B. But for the mean absolute error, NN-A was better than CART-A while NN-B was worse than CART-B. Though we can not generalize which one is absolutely superior to the other, we suggest that ANN model is better than CART model to predict the charges of the colorectal cancer patients in Kyunghee University Hospital.

Our study has some limitations. First, when the target variable is a binary variable or a nominal variable, the models can be compared with the accuracy matrix generated by data mining system. But the target variables in this study were neither the cases so we cannot compare them with such an accuracy matrix. Second, the dataset we used did not include all the clinical information such as disease stages and pathologic cell types of the cancer. These variables, when added to our dataset, may enable us to build a more accurate prediction model in various aspects but they are not available in a digitalized form, yet. Our current study was designed to focus on the colorectal cancer patients for this reason. The colorectal cancer patients had a relatively homogenous clinical stage and pathologic cell type. That is, more than 80% of patients were advanced stage and the pathologic cell types were adenocarcinoma in more than 95%. Thus the patients shared almost the same treatment protocol including surgery, chemotherapy and radiotherapy to the extent that the effect of staging and pathologic cell type has been considerably offset. But we are planning to build a complete model including the entire patient's data to predict the charges of whole spectrum of cancer patients with the introduction of electric medial record (EMR) in the hospital.

REFERENCES

1. Korea National Cancer Center. *National Cancer Statistics. Seoul: The Institute: 2001.*
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; 17: 604-10.*
3. 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; 92: 950-8.*
4. Penberthy L, Retchin SM, McDonald MK, McClish DK, Desch CE,

Riley GF, Smith TJ, Hillner BE, Newschaffer CJ. *Predictors of medicare costs in elderly beneficiaries with breast, colorectal, lung, or prostate cancer. Health Care Manag Sci 1999; 2: 149-60.*

5. Tollestrup K, Frost FJ, Stidley CA, Bedrick E, McMillan G, Kunde T, Petersen HV. *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; 66: 25-31.*
6. Dayhoff JE, DeLeo JM. *Artificial neural networks: opening the black box. Cancer 2001; 91 (Suppl 8): 1615-35.*
7. 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 Inf 2001; 63: 41-50.*
8. Roche K, Paul N, Smuck B, Whitehead M, Zee B, Pater J, Hiatt MA, Walker H. *Factors affecting workload of cancer clinical trials: results of a multicenter study of the National Cancer Institute of Canada Clinical Trials Group. J Clin Oncol 2002; 20: 545-56.*
9. Goldman DP, Schoenbaum ML, Potosky AL, Weeks JC, Berry SH, Escarce JJ, Weidmer BA, Kilgore ML, Wagle N, Adams JL, Figlin RA, Lewis JH, Cohen J, Kaplan R, McCabe M. *Measuring the incremental cost of clinical cancer research. J Clin Oncol 2001; 19: 105-10.*
10. Fireman BH, Fehrenbacher L, Gruskin EP, Ray GT. *Cost of care for patients in cancer clinical trials. J Natl Cancer Inst 2000; 92: 136-42.*
11. 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.*
12. Marshall AH, McClean SI, Shapcott CM, Millard PH. *Modelling patient duration of stay to facilitate resource management of geriatric hospitals. Health Care Manag Sci 2002; 5: 313-9.*
13. Walczak S, Scharf JE. *Transfusion cost containment for abdominal surgery with neural networks. Neural Processing Letters 2000; 11: 229-38.*
14. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr, Marks JR, Winchester DP, Bostwick DG. *Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997; 79: 857-62.*
15. Sargent DJ. *Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer 2001; 91 (Suppl 8): 1636-42.*
16. Michie D, Spiegelhalter DJ, Taylor CC. *Machine learning, neural and statistical classification. New York: Ellis Horwood; 1994.*
17. Rodvold DM, McLeod DG, Brandt JM, Snow PB, Murphy GP. *Introduction to artificial neural networks for physicians: taking the lid off the black box. Prostate 2001; 46: 39-44.*