Industrial Engineering
& Management Systems
Vol 13, No 4, December 2014, pp.449-453
ISSN 1598-7248 | EISSN 2234-6473 |
http://dx.doi.org/10.7232/iems.2014.13.4.449

Prediction of Hypertension Complications Risk Using Classification Techniques

Wonji Lee, Junghye Lee, Hyeseon Lee, Chi-Hyuck Jun
Department of Industrial and Management Engineering, POSTECH, Pohang, Korea

Il-su Park
Department of Health, Uiduk University, Gyeongju, Korea

Sung-Hong Kang*
Department of Health Policy and Management, Inje University, Gimhae, Korea

(Received: November 7, 2014 / Revised: November 25, 2014 / Accepted: November 25, 2014)

ABSTRACT

Chronic diseases including hypertension and its complications are major sources causing the national medical expenditures to increase. We aim to predict the risk of hypertension complications for hypertension patients, using the sample national healthcare database established by Korean National Health Insurance Corporation. We apply classification techniques, such as logistic regression, linear discriminant analysis, and classification and regression tree to predict the hypertension complication onset event for each patient. The performance of these three methods is compared in terms of accuracy, sensitivity and specificity. The result shows that these methods seem to perform similarly although the logistic regression performs marginally better than the others.

Keywords: Hypertension Complications, Risk Prediction, Logistic Regression, LDA, CART

* Corresponding Author, E-mail: hcmkang@inje.ac.kr

1. INTRODUCTION

Despite the remarkable development of medical technology, chronic diseases such as hypertension and diabetes are still the main factors that reduce life span (Korea Ministry of Health and Welfare, 2011). Moreover, the socio-economic burden due to these chronic diseases is anticipated to increase continuously in the future. The proportion of national medical expenses in GDP is expected to increase from 6.1% in 2001 to 16.8% in 2030 (Korea National Health Insurance Corporation, 2012). Hypertension is the disease that requires the highest annual budget for its management in Korea. Especially, hypertension complications are one of the main causes to lead to fatalities (Hozawa *et al*., 2009).

Prognosis of hypertension patients depends largely on how exactly they recognize their disease condition and whether they take a suitable healthcare action (Park and Jun, 2000). However, Korean hypertension patients who exactly recognized their hypertension status were only 67.9% in 2010 (Korea Ministry of Health and Welfare, 2011). Still many hypertension patients are careless about their health conditions, although they are required to manage their diseases to prevent corresponding complications. Therefore, it would be a significant contribution to predict the risk of hypertension complications for hypertension patients using available healthcare data.

The Korea government is opening a variety of the government data to public. As a part of the policy, Korean National Health Insurance Corporation has built the sample national healthcare database, which includes medical expenses, medical treatments and health check-up records.

In this paper, we focus on the onset of hypertension complications and construct an onset risk prediction model using the above sample national healthcare data-

base. Hypertension is limited to essential (or primary) hypertension diseases and hypertension complications indicate cardiovascular diseases. Logistic regression (LR) is the most commonly used in previous research to predict the risk of diseases (Echouffo-Tcheugui *et al*., 2013). We will also use the linear discriminant analysis (LDA) and classification and regression tree (CART) in addition to LR. We will compare the prediction performance of these methods in terms of various criteria and report the hypertension complications onset rate according to major factors.

2. METHODS

The three classification methods, such as LR, LDA, and CART will be used for the risk prediction. Each method is briefly described here.

2.1 Logistic Regression

LR is the most popular technique to predict the risk of disease onset (Echouffo-Tcheugui *et al*., 2013). Unlike the ordinary regression method, LR treats binary, nominal or ordinal variables as its dependent variable (Hosmer Jr and Lemeshow, 2004). Thus it allows us to predict the probability of disease onset. Let P denote the probability of onset for a binary dependent variable. Then, LR model using k independent variables (X_j 's) is as follows when using the logit link function.

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

where β_j 's are regression coefficients. The onset probability of can be derived as follows.

$$P = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}$$

2.2 Linear Discriminant Analysis

LDA aims to find the best linear combination of features to separate objects from different classes. The linear classifier can be derived by comparing posterior probabilities of classes (Izenman, 2008).

LDA is characterized by its assumption that class-conditional probability density function of independent variables is multivariate Gaussian using the same variance-covariance matrix (denoted by Σ) for all classes (Press and Wilson, 1978). The discriminant function of class i (i = 1, 2) is given by

$$U_i = \mu_i^T \Sigma^{-1} x - \frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i + \ln \pi_i$$

where μ_i is the mean vector of independent variables for class i and π_i is the prior probability for class i. Then we can classify an object into class 1 if $U_1 > U_2$ and vice versa. Note that the above discriminant function is linear in x.

Based on the Bayes theorem, we can compute the posterior probability for class i as follows.

$$P(\text{class } i \mid x) = \frac{\pi_i f_i(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}, \text{ i} = 1, 2$$

where $f_i(x)$ is the multivariate normal density function for class i.

2.3 Classification and Regression Tree

CART is a recursive partitioning method for classification and regression (Breiman *et al*., 1983; Loh, 2011). It splits the data space by choosing one variable at a time repeatedly like a decision tree. Splitting criterion at each node is to reduce the impurity of the children nodes via Gini index. After a fully grown tree is obtained by the procedure of binary splitting, pruning can be applied to generate several simpler trees using the cost-complexity measure. Then, the best pruned tree is chosen by the cross-validation.

Each leaf node of the best pruned tree has two classes—one for the onset event and another for the non-onset event. The onset risk is predicted by the proportion of onset events at the leaf node corresponding to an individual observation.

2.4 Prediction Evaluation Method

The above three methods are applied to the sample national healthcare data for the prediction model learning and the performance evaluation. The performance evaluation is based on the cross validation with the training set and the test set. The onset events are blinded for the test set. The onset event is predicted as being occurred when the onset probability is greater than the cutoff value. As performance measures we consider accuracy, sensitivity and specificity.

Due to the issue of determining the cutoff value, we use area under ROC (AUC) as an additional performance measure. ROC curve is a plot of the false positive rate (1-specificity) versus the true positive rate (sensitivity) (Zhu *et al*., 2010).

3. DATA AND PRE-PROCESSING

3.1 Data

The sample national healthcare database is established by Korean National Health Insurance Corporation and includes one million subscribers, which were sam-

pled as 2% of the population with regard to the strata of age, sex and income level. It includes socio-demographic variables, records of health check-up, and other information related to medical treatments and medical expenses for nine years from 2002 to 2010.

Table 1. Independent variables

Names of variables
Socio-demographic variables
Age (yr)
Sex (male/female)
Income (twenty classes)
Medical treatment records
The number of health check-ups (2002-2007)
The days of hospital treatment (2002-2008)
The days of outpatient treatment (2002-2008)
Prescription compliance (%)
Health check-up indices
Body mass index (%)
Waistline (cm)
Maximum blood pressure (mmHg)
Minimum blood pressure (mmHg)
Fasting plasma glucose level (mg/dL)
Total cholesterol level (mg/dL)
Hemoglobin level (g/dL)
SGOT (U/L)
SGPT (U/L)
γ-GPT (U/L)
Behavior variables
Smoking status (smoking, non-smoking, past smoking)
Smoking period (yr)
Daily smoking amount
Weekly number of drinking
Drinking quantity per one time
Family history (presence/absence)
Hypertension
Stroke
Cardiac diseases
Diabetes
Cancers
Own medical history
Diabetes (presence/absence)

We set the year 2008 as the reference year and establish the cohort for this study. First, among people who took health check-up, we select patients who were diagnosed as hypertension by 2008 but exclude people who have received medical treatment for complications. Then we identify their hypertension complications onset for those people during the follow-up period from the health check-up in 2008 to the end of 2010. In the case of people without medical records during the follow-up period, we define that the complications do not occur. Also we exclude people dying during the follow-up period. Based on previous studies on hypertension and its complications, we selected independent variables as in Table 1. We also generate the dependent variable, a binary variable indicating whether complications has occurred or not during the follow-up period.

3.2 Data Pre-processing

We conducted data pre-processing for treating missing values and outliers. Some variables for behavior and family history have missing values. In fact, these variables were recorded from questionnaires. It is assumed here that unanswered questions are regarded as negative responses. For example, there are three questions (smoking status, smoking period, and daily smoking amount) related to smoking. When these are unanswered, these variables are imputed as 'non-smoking,' 0 and 0. Except these variables, we exclude observation with any missing value from data set. We delete observations whose health check-up indices are out of the normal ranges according to the national standards (Ministry of Health and Welfare, 2013).

The pre-processed data set consists of 10,814 hypertension patients, which includes 1,739 hypertension complications patients or 16.08% of the whole observations. The class of hypertension complications is minor as compared to the non-onset (major) class, so under-sampling is performed to randomly select the major class to be equal to the minor class. We repeat this step five times to prepare five such data sets. For each under-sampled data set, we conduct five-fold cross-validation for each data set to obtain the performance measures and take the average of 25 values from 5 data sets.

4. RESULT

Table 2 shows the onset rates of hypertension complications according to several independent variables. Because age is not discrete, we classify it into three groups. The group of younger than 40 is not included here because no hypertension patients in the group were found to have complications. The income level is classified as three groups (low, medium and high).

It is seen that the onset rate for female is a little higher than male. The onset rate becomes higher as age increases. Hypertension patients with the family history of stroke and cardiac disease are likely to have high risk of hypertension complications.

The performance of three prediction models is compared in Table 3. Four performance measures of each model are listed for the training and the test sets. The bold face indicates the best performance among three models. Because we have already adjusted imbalance between classes by under-sampling, it is reasonable to set cutoff value to 0.5 for classifying into onset or non-onset event when computing the above measures.

LR and LDA show almost same performance in terms of four evaluation measures although LR is known to be generally preferable to LDA (Kurt *et al*., 2008; Press and Wilson, 1978). Compared to LR and LDA, CART is better in training sets, but worse in test sets. CART is known to be sensitive to data because the structure of

tree varies depending on data. This can be also interpreted that CART does not reflect features individually well because features are used for only constructing splitting rules, but not directly for computing individual's risk. As a result, it is natural that the performance in the case of test set is worse than one in the case of training set.

Table 2. Hypertension complications onset rate

Characteristic	Onset rate (%)
Sex (male/female)	
Male	15.45
Female	16.79
Age (yr)	
40-49	8.51
50-59	11.80
60-69	17.89
Income level	
Low	16.36
Mid	15.56
High	16.34
Family medical history [a)]	
Hypertension	16.31 / 16.00
Stroke	19.14 / 15.80
Cardiac diseases	20.73 / 15.91
Diabetes	13.83 / 16.30
Cancers	16.81 / 15.97
Own diabetes history [a)]	16.05 / 18.40

[a)] Ratio of the presence/absence.

Besides, the predictive performance values are relatively lower than those predicted previously in the case of hypertension onset (Echouffo-Tcheugui *et al*., 2013). It seems because hypertension complications has more complex mechanism and it is more difficult to predict. Besides Srinivas *et al*. (2010) showed better performance to predict the risk of cardiovascular disease with more detailed medical data, but our research has the limitation of screening test data drawn from biyearly health check-ups.

5. CONCLUSION

This paper analyzed the risk of hypertension complications in Korea for the first time using the national healthcare database. For predicting the risk of hypertension complications, we used three methods: LR, LCA, and CART. These methods show relatively low performances due to the limitation of screening test data and complex mechanism of hypertension complications. The result shows that all three methods perform equally well, but the LR performs marginally better than the other two.

Whereas past studies on Korean's hypertension are usually limited to some specific groups, our study considers broader variables and observations. Furthermore, compared to the studies in other countries, our data set provides definitely a large number of observations, thus offering more reliable results which may be extended to other chronic diseases and their complications.

ACKNOWLEDGMENTS

This research was supported by a grant of the Korea Health technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (No. HI13C0790).

REFERENCES

Breiman, L., Friedman, J., Olshen, R., Stone, C., Steinberg, D., and Colla, P. (1983), *CART: Classification and Regression Trees*, Wadsworth, Belmont, CA.

Dreiseitl, S. and Ohno-Machado, L. (2002), Logistic regression and artificial neural network classification models: a methodology review, *Journal of Biomedical Informatics*, **35**(5), 352-359.

Echouffo-Tcheugui, J. B., Batty, G. D., Kivimäki, M., and Kengne, A. P. (2013), Risk models to predict hypertension: a systematic review, *PloS One*, **8**(7), e67370.

Hosmer Jr, D. W. and Lemeshow, S. (2004), *Applied Logistic Regression*, John Wiley and Sons, Hoboken, NJ.

Hozawa, A., Kuriyama, S., Kakizaki, M., Ohmori-Matsuda, K., Ohkubo, T., and Tsuji, I. (2009), Attributable risk fraction of prehypertension on cardiovas-

Table 3. Comparison of risk prediction models for training set and test set

	AUC		Accuracy		Sensitivity		Specificity	
	Training	Test	Training	Test	Training	Test	Training	Test
LR	0.6307	**0.6072**	0.5957	**0.5782**	0.5894	0.5738	0.6017	**0.5841**
LDA	0.6304	0.6068	0.5955	0.5779	0.5897	**0.5740**	0.6010	0.5832
CART	**0.7061**	0.5704	**0.6497**	0.5533	**0.6537**	0.5659	**0.6456**	0.5406

LR: logistic regression, LDA: linear discriminant analysis, CART: classification and regression tree, AUC: area under ROC.

cular disease mortality in the Japanese population: the Ohsaki study, *American Journal of Hypertension*, **22**(3), 267-272.

Izenman, A. J. (2008), Linear discriminant analysis. In: *Modern Multivariate Statistical Techniques*, Springer, Heidelberg, 237-280.

Korea Ministry of Health and Welfare (2011), *2010 Korean Health Statistics Report*, Korea Ministry of Health and Welfare, Seoul.

Korea National Health Insurance Corporation (2012), *2011 Health Insurance Statistics Annual Report*, Korea National Health Insurance Corporation, Seoul.

Kurt, I., Ture, M., and Kurum, A. T. (2008), Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease, *Expert Systems with Applications*, **34**(1), 366-374.

Loh, W. Y. (2011), Classification and regression trees, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **1**(1), 14-23.

Park, Y. I. and Jun, M. H. (2000), The effect of a self-regulation program for hypertensives in rural areas, *Journal of Korean Academy of Nursing*, **30**(5), 1303-1317.

Press, S. J. and Wilson, S. (1978), Choosing between logistic regression and discriminant analysis, *Journal of the American Statistical Association*, **73**(364), 699-705.

Srinivas, K., Rao, G. R., and Govardhan, A. (2010), Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques, *Proceedings of the 5th International Conference on Computer Science and Education (ICCSE2010)*, Hefei, China, 2010, 1344-1349.

Zhu, W., Zeng, N., and Wang, N. (2010), Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations, *Proceedings of the 23rd Annual Conference on Northeast SAS Users Group* (*NESUG): Health Care and Life Sciences*, Baltimore, MD.