TEST Engineering & Management

March - April 2020
ISSN: 0193-4120 Page No. 4249 - 4256

A Study on Prediction Model of the Hypertension Risk by Deep Learning

Sung-jun Kim[1], Dea-woo Park[*2]
[1]Doctorial course, Include Department of Convergence Engineering, Hoseo Gaeduate School of Venture, 06724, South Korea
[*2]Professor, Include Department of Convergence Engineering, Hoseo Gaeduate School of Venture, 06724, South Korea
mvstar@hanmail.net[1], prof_pdw@naver.com[*2]

Article Info
Volume 83
Page Number: 4249 - 4256
Publication Issue:
March - April 2020

Abstract

Background/Objectives: Recently, as the research on the neural network deep learning based on machine learning technique is active, there are increasing attempts to apply this technology in various industries. In particular, studies on the prediction of hypertension risk in the healthcare sector had a limitation that the questionnaire and cohort DB were not used to accurately reflect the hypertension risk prediction.

Methods/Statistical analysis: In this study, we will use Python to collect public health data for 2018, and develop a model to predict the risk of hypertension and what are the main factors affecting the development of hypertension using deep learning analysis.

For the study, Four deep learning analyzes of decision tree, random forest, gradient boosting and logistic regression were used, and among these, logistic regression showed the best prediction rate. So, in this paper, we analyzed the risk factors for hypertension using logistic regression methodology.

Findings: As a result, age, women, BMI, height, weight, waistline, systolic blood pressure, diastolic blood pressure, urinary protein, serum creatinine, and gamma-tiffy variables were associated with high risk of hypertension. And the logistic regression showed the highest predicted value of 79.41%, and showed the concordance values for each disease variable affecting the prediction. The regression equation for the risk of hypertension is Log (p / (1-p)) =-54.9372 + 0.5228 (60s) -0.1752 (Famale) -0.1652 (BMI) -0.0549 (height) +0.0502 (weight) +0.0204 (waist) Circumference) +0.2601 (constrictor blood pressure) +0.2634 (diastolic blood pressure) +0.2272 (urea protein) +0.2271 (cholesterol) -0.0029 (gamma tipi).

Improvements/Applications: As a result, if the regression formula is used to produce a program for predicting hypertension in the future, it is possible to provide a service for self-diagnosis of hypertension risk.

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 26 March 2020*

Keywords: *Deep learning, The hypertension risk prediction, Machine learning, Random Forest, Logistic regression analysis*

1. Introduction

Circulatory diseases (cardiovascular and cerebrovascular diseases) account for 24.96% of all deaths in Korea [1]. The most prevalent disease among circulatory diseases is hypertension, It is known as an important risk factor for cerebrovascular disease and coronary artery disease [2]. In addition, hypertension accounts for 12% of the 56

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 4249 - 4256

million deaths per year worldwide [3]. Hypertension is a disease at the forefront of modern civilization that, when prolonged, causes fatal complications. In order to prevent high blood pressure causing complications, proper management of patients and primary cause pevention of hypertension are important [4]. However, 50% of hypertensive patients are found in many countries, The 50% of the patients are treated, and The 50% of the treated patients have adequate blood pressure control [5]. As a result, only one eighth of all patients are properly treated. Foreign studies have reported genetic factors [6], alcohol consumption [7], and obesity [8] as risk factors associated with the hypertension. In addition, the results of epidemiologic studies on the incidence of hypertension show that the incidence of hypertension increases gradually with age, and the incidence of hypertension in men before 60 age is higher than in women, but similar or even higher in women after 60 age [9]. Meanwhile, according to a study on the risk factors for hypertension in Korea, obesity, dietary habits, smoking, drinking, and lack of exercise were reported as risk factors related to the hypertension [10]. In recent studies, high normal blood pressure and obesity have been reported as risk factors for the development of hypertension[11]. However, there have been very few studies on deep learning-based hypertension, and there has been no study on a predictive model of hypertension using data mining in Korea.

This study analyzed the analysis of the risk factors for hypertension according to gender in approximately 1,139,900 health checkup data of 20 years old or older who received health checkup for 1 year from January 1 to December 31, 2018 by Python. Four deep learning methods, decision tree, random forest, gradient boosting and logistic regression, were used to select the methodology that showed the best predictive rate and to select factors influencing hypertension.

After using the deep learning analysis method, high blood pressure risk factors for the development of hypertension were identified, and a predictive model was developed to help early detection of hypertension and prevention and management of hypertension.

2. Related works

Deep learning techniques are used to analyze high blood pressure risk factors. After the preprocessed data were separated into learning data and prediction data, the model showing the best prediction rate was selected by applying decision tree, random forest, gradient boosting, and logistic regression analysis methods.

2.1. Business Model Design

In this study, the flow of data movement was defined as a DCAI problem-solving methodology[Figure 1], from the step of collecting data to the step of making improvement after processing. In DCAI, D is Define Problem, C is Collect Data, A is Analysis Data I is Improve stage, and data goes through each stage to process, process, analyze, and derive insight through result confirmation. First, in the Define Problem step, the problem is defined. It defines what needs to be identified in the research you are facing and what you have to uncover through research.

At the end of this step, you need to collect data. In the data collection stage, internal data and external data can be collected separately, and internal data means data that can be collected by a company or an organization. External data refers to data collected by public institutions, data released by certain companies to achieve social contribution and research purposes, and behavioral data by consumers or programs exposed to the Internet. These data are processed and processed at the same time as collection, and analysis is performed. In the analysis phase, analysis is performed, and insights by analysis are extracted. In the Improve stage, improvement plans or application items are derived based on insights extracted by analysis. In this study, data analysis was conducted using the DCAI described above as shown in Figure 1 below.

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 4249 - 4256

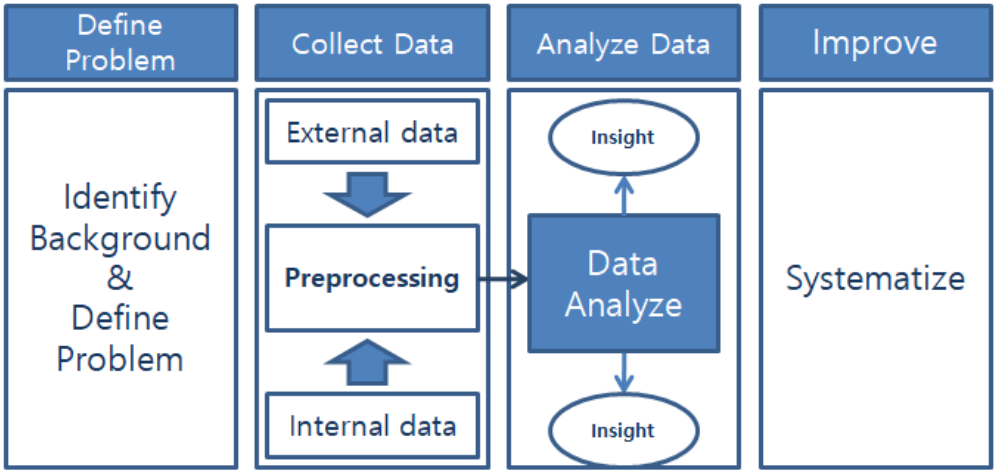

Figure 1. DCAI Problem-solving Methodology

2.2. Decision Tree Analysis

Decision trees are a technique of data mining and have two characteristics: data exploration and modeling. The basic concept of data mining is derived from the term KDD (knowledge discovery in database) and refers to the entire process of extracting knowledge from data. This technique involves segmentation, which classifies measurement data into several types, classification by classifying result variables into several classes, dimension reduction, and variable screening, which select variables that have high impact on outcome variables among several predictors. And it is very useful to know the condition of increase and decrease of explanatory amount by combination.

2.3. Random Forest

Random Forest[Figure 2] is an ensemble learning model that generates several decision trees and then makes predictions with the label that has been selected the most among the predictions of each individual tree. Ensemble learning refers to learning multiple models in machine learning and using the prediction results of those models to predict better values than a single model. Random forests as an alternative to decision trees have the advantage of higher predictive power in multifactor data and can be applied to data that can not be processed by existing statistical analysis methods.

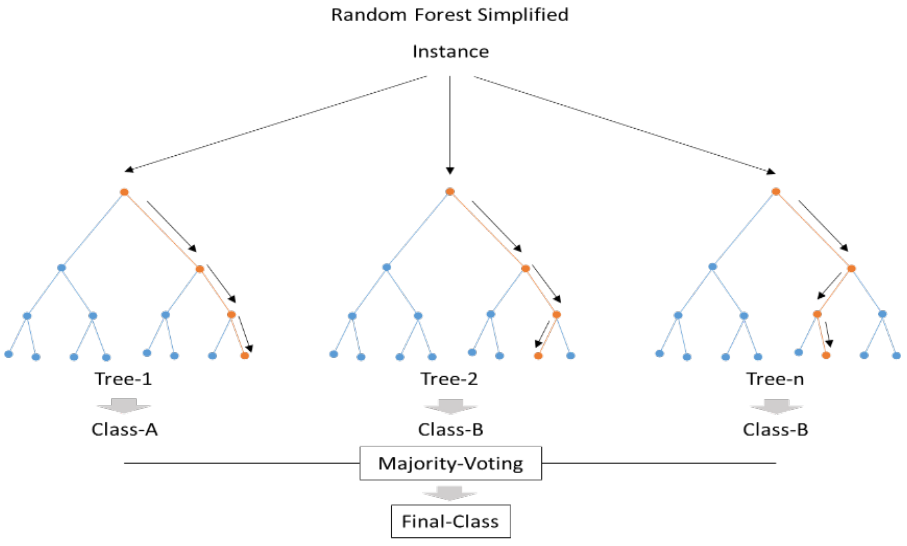

Figure 2. Random Forest Simplified

March - April 2020
ISSN: 0193-4120 Page No. 4249 - 4256

2.4. Gradient Boosting

The Boosting in machine learning refers to a way of tying relatively inaccurate Weak Learners to make them more accurate and stronger learners. Gradient boosting differentiates into tree model functions that have been learned so far by the following equation (1).

$$f_{i+1} = f_i - \sigma \frac{\delta J}{\delta f_i} \qquad (1)$$

In other words, the derivative value of the tree model in the gradient boosting model represents the weakness of the model trained to date. When fitting the next tree model, the derivative value is used to compensate for the weakness to boost performance.

2.5. Logistic Regression Analysis

Logistic regression analysis is a statistical analysis method used to classify two or more groups by using relational expression between dependent and independent variables. Logistic regression has the advantage that discrete independent variables as well as continuous independent variables can be used, and no assumptions are required. Therefore, it is preferred to discriminant analysis that performs statistical analysis with the assumption that the variance and covariance between groups are the same, following the normal distribution among independent variables. A logistic regression model with one independent variable can be easily extended to multiple logistic regression models representing multiple independent variables, and can be expressed by the following formula, where X represents each independent variable and β represents regression coefficient of an independent variable.

$$P_n = \frac{exp\,(\beta_0+\beta_1 X_1+ \beta_2 X_2+ \cdots+\beta_n X_n}{1+exp\,(\beta_0+\beta_1 X_1+ \beta_2 X_2+ \cdots+\beta_n X_n} \qquad (2)$$

3. Results Method

In this study, data from 1,139,900 people over 20 years of age who had undergone public health examinations for 1 year from January 1, 2018 to December 31, 2018, were analyzed. Exploratory data analysis was performed on general information on the patient's age, height, weight, and systolic blood pressure, diastolic blood pressure, fasting blood glucose, urine protein, hemoglobin, total cholesterol, BMI index, HDL cholesterol, and LDL cholesterol. And 19 independent variables were selected through the preprocessing process and analyzed by Python.

3.1. Analysis Object

The subjects of this study were 1,139,900 health checkup data from 20 years of age or older who received health checkup in 2018. For model estimation, the data compiled through data exploration analysis was used as a diagnosis of hypertension as a dependent variable. The independent variable was used as a common independent variable consisting of age group, gender, BMI value, height, weight, waist circumference, systolic blood pressure, diastolic blood pressure, urine protein, cholesterol, and gammatipi.

3.2. Analysis Tool

The analysis tool used Python, which is an open source that has recently been spotlighted for data analysis. This programming language gives a lot of flexibility to developers or analysts. Python provides several libraries to make data analysis easier: the Pandas library for preprocessing data, the Matplot for visualization, the Seaborn library, and the Sklearn library for data mining. Sklearn makes it easy for users to handle many of the functions related to data mining, including supervised and unsupervised learning.

3.3. Analysis Methods

The analysis is based on Python's Skearn

March - April 2020
ISSN: 0193-4120 Page No. 4249 - 4256

library. First, before the cluster analysis, silhouette analysis was performed based on health examination data to determine the appropriate number of clusters. Afterwards, clustering was performed using the K-means Clustering algorithm. After clustering, the purchasing characteristics of each cluster were identified by using a visualization technique, and the cluster with the highest prediction rate was selected and defined as an excellent cluster. The superior cluster was set as the positive of the target variable for classification analysis.

After separating preprocessed data into learning data and prediction data, four analysis methods such as Decision tree analysis, Random forest, Gradient Boosting, and Logistic regression analysis among deep learning techniques are applied. And the prediction model is developed through the analysis method showing the highest prediction rate among them.

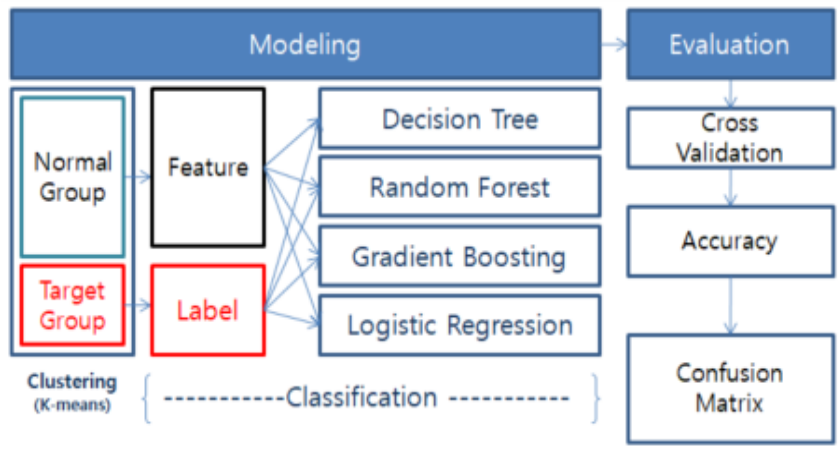

Figure 3. Analysis Method Modeling

4. Results

The four clusters of the deep learning analysis were divided and analyzed. The decision tree analysis in Figure 4 shows that the training test accuracy is 0.89 and the test set accuracy is 0.89. Error matrix results show that the probability of matching disease to disease is high only once. The random forest analysis shown in Figure 5 shows that the training test accuracy is 0.662 and the test set accuracy is 0.629, which shows somewhat lower prediction rate. According to the error matrix results, the probability of matching disease well with disease is very high in numbers 1 and 3, and other indicators show some value.

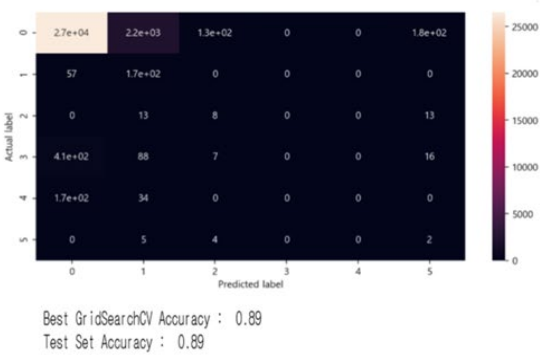

Figure 4. Decision Tree

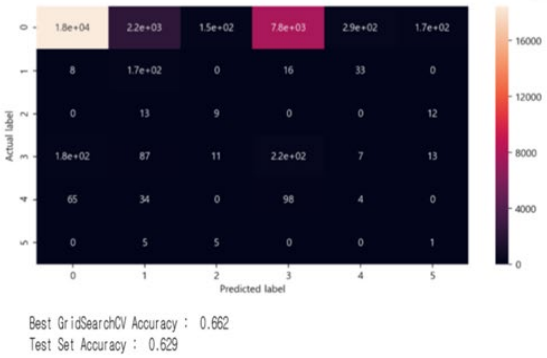

Figure 5. Random Forest

March - April 2020
ISSN: 0193-4120 Page No. 4249 - 4256

The results of the Gredient Boosting analysis in Figure 6 show that the training test accuracy is 0.955 and the test set accuracy is 0.956. The error matrix results show that the probability of matching disease well with disease generally shows a numerical value, but each disease-to- disease concordance rate is nearly zero. The logistic regression analysis of Figure 7 shows that the training test accuracy is 0.966 and the test set accuracy is 0.967, which shows a high prediction rate. According to the error matrix results, the coincidence probability that the disease fits well with the disease shows that the coincidence probability value is 0 for all diseases.

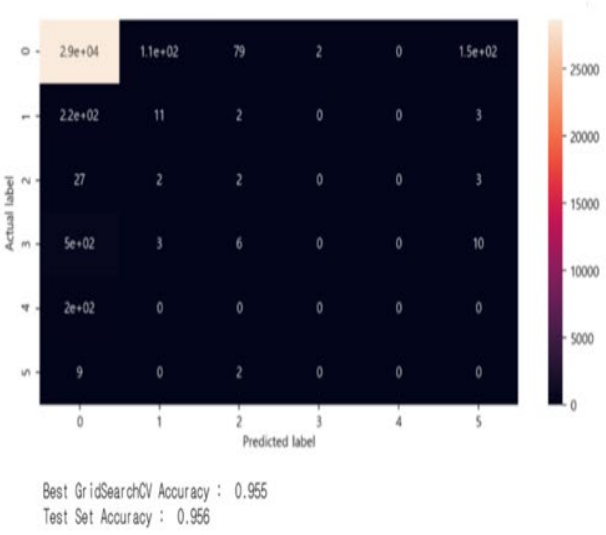

Figure 6. Gradient Boosting

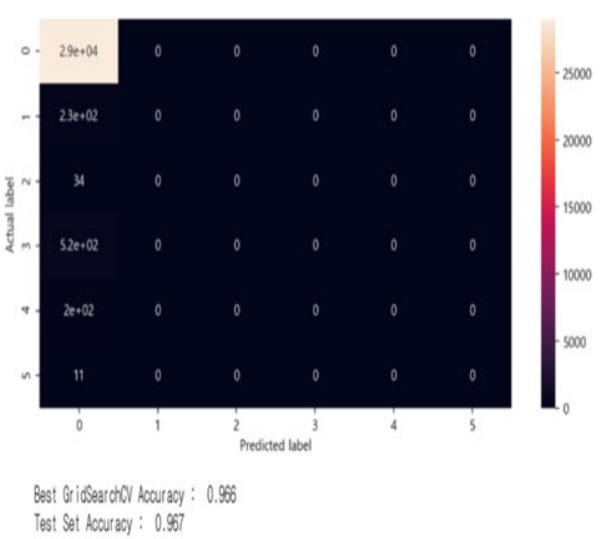

Figure 7. Logistic regression analysis

As a result, it can be seen that Logistic regression analysis is superior to other methods for identifying high blood pressure risk factors. Therefore, I examined the factors affecting hypertension using Logistic regression analysis. [Table 1] and [Table 2] show the results obtained through logistic regression analysis.

Table 1. Logit Regression Results

Dep. Variable	: Hypertension Symptom	No. Observations	: 391,724
Model	: Logit	Df Residuals	: 391,704
Method	: MLE	Df Model	: 19
Date	: Mon, 13 Jan 2020	Pseudo R-squ	: 0.7941
Time	: 15:40:38	Log-Likelihood	: -11585.
Converged	: True	LL-Null	: -56253.
		LLR p-value	: 0.000

Table 2. Logit Regression Results

	Coef	Std err	Z	P > | z |	[0.025	0.975]
Intercept	-54.9372	1.930	-28.462	0.000	-58.720	-51.154
C(65years old or older)	0.5228	0.043	12.132	0.000	0.438	0.607
C(female)	-0.1752	0.064	-2.728	0.006	-0.301	-0.049

BMI	-0.1652	0.035	-4.673	0.000	-0.234	-0.096
Height	-0.0549	0.011	-4.772	0.000	-0.077	-0.032
Weight	0.0502	0.014	3.686	0.000	0.024	0.077
Waistline	0.0204	0.003	5.830	0.000	0.014	0.027
Systolic blood pressure	0.2601	0.003	98.519	0.000	0.255	0.265
Diastolic blood pressure	0.2634	0.003	83.984	0.000	0.257	0.270
Fasting blood sugar	-0.0009	0.001	-0.667	0.505	-0.004	0.002
Urine protein	0.2272	0.097	2.352	0.019	0.038	0.417
Serum creatinine	0.2271	0.114	1.992	0.046	0.004	0.451
Total cholesterol	-0.0025	0.001	-1.915	0.056	-0.005	5.86e-05
Hemoglobin	-0.0287	0.016	-1.795	0.073	-0.060	0.003
Triglycerides	0.0004	0.000	1.608	0.108	-9.78e-05	0.001
HDL cholesterol	0.0015	0.002	0.977	0.329	-0.001	0.004
LDL cholesterol	0.0017	0.001	1.441	0.150	-0.001	0.004
AST	0.0037	0.003	1.228	0.219	-0.002	0.010
ALT	-0.0021	0.002	-1.124	0.261	-0.006	0.002
Gamma Tippie	-0.0029	0.001	-2.748	0.006	-0.005	-0.001

The analysis results in Table 2 show that the model has a good residual model of 391704, 19 degrees of freedom, 79.41% of explanatory diagram of the model and 5% of P-value. The coefficients of factors affecting hypertension were derived for each independent variable. And age, female, BMI, height, weight, waist circumference, systolic blood pressure, diastolic blood pressure, urinary protein, serum creatinine, and gamma-tipipi variables affected high blood pressure turn out to be a factor giving.

5. Conclusion

I developed a model for predicting hypertension risk between men and women using running techniques. For fair model evaluation, the total data was divided into train data and test data, and the model made by using data mining technique was applied to the test data.

Overall, it can be seen that systolic blood pressure, diastolic blood pressure, urine protein, serum creatinine, weight and waistline are the most important risk factors for hypertension in women than men. In particular, systolic and diastolic blood pressure, urine protein can be seen that the risk is quite large.

In the hypertension prediction model, the logistic regression showed the highest predicted value of 79.41%, and showed the concordance values for each disease variable affecting the prediction. Also, significant factors (factors affecting hypertension) within 5% of P.value were constants –54.9372, 60s (0.5228), women (-0.1752), BMI (-0.1652), height (-0.0549), and weight (0.0502), Waist circumference (0.0204), systolic blood pressure (0.2601), diastolic blood pressure (0.2634), urine protein (0.2272), cholesterol (0.2271), gamma typhi (-0.0029).

The regression equation for the risk of hypertension is Log (p / (1-p)) =-54.9372 + 0.5228 (60s) -0.1752 (Famale) -0.1652 (BMI) -0.0549 (height) +0.0502 (weight) +0.0204 (waist) Circumference) +0.2601 (constrictor blood pressure) +0.2634 (diastolic blood pressure) +0.2272 (urea protein) +0.2271

March - April 2020
ISSN: 0193-4120 Page No. 4249 - 4256

(cholesterol) -0.0029 (gamma tipi). As a result, if the regression formula is used to produce a program for predicting hypertension in the future, it is possible to provide a service for self-diagnosis of hypertension risk.

References

1. Statistical Office. http://www.nso.go.kr
2. MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, et al.. Blood pressure, stroke and coronary Heart discase, Part 1. Prolonged differences in blood pressure; prospective observational studies corrected for the regression dilution bias. Lancer. 1990 Vol. 335 :765-773.
3. WHO, Reducing Risks and Promoting Life. World Health Report. 2002.
4. Cook NR, Cohen J, Hebert P,. Implications of small reduction in diastolic blood pressure for primary prevention. Arch Intern Med. 1995 Vol.155:701-709
5. Klungel OH, de Boer A, Paes AHP, Seidell JC, Nagelkerke NJD, et al. Undertreatment of hypertension in a population-based study in The Netherlands. J hypertension, 1998 Vol.16 : 1371-1378
6. Williams RR, Hunt SC, Hasstedt SJ, Hopkins PN, Wu LL, et al. Are there interaction and relations between genetic and enviromental factors predisopsing to high blood pressure? . Hypertension. 1991 Vol.18: I-29-I-37
7. Kojima S, Kawano Y, Abe H, et al. Acute effects of alcohol ingestion on blood pressure and erythrocyte sodium concertration. J hypertension; 2005. Vol.11: 185-190
8. He J, Muntner P, Chen J, Roccella EJ, Streiffer RH, Whelton PK. Factors associated with hypertension control in the general population of the United States.: Arch Intern Med; 2002. Vol.162 no.9:1051-1058
9. Dannenberg AL, Garrison RJ, Kannel WB. Incidence of hypertension in the Framingham Study.: Am J public Health; 1998. Vol.78:676-679
10. Hae Sook Sohn, Chae Un Lee, Jin Ho Chun, Jung Hak Kang, Hwi Dong Kim, Kui Oak Jung, Kyu Il Cho. Risk Factors of Hypertension and The Effect of These Factors on Blood Pressure: Korean Journal of Epidemiology, 1995 vol.17 no.2:201-213, https://www.e-epih.org/journal/view.php?number=272
11. Jong-Myon Bae, Yoon-ok Ahn. A Nested Case-Control Study on the High Normal Blood Pressure as a Risk Factor of Hypertension in Korean Middle-aged Men. J Korean Med Sci. 2002 Jun vol.17 no.3: 328-36, https://www.ncbi.nlm.nih.gov/pubmed/12068135