International Journal of Advanced Smart Convergence Vol.7 No.4 1-9 (2018)
http://dx.doi.org/10.7236/IJASC.2018.7.4.1

IJASC 18-4-1

Disease risk prediction system using correlated health indexes

Yoonjung Kim[1], Hyeon Seok Son[2] , Hayeon Kim[3*]

[1]Laboratory of Computational Biology & Bioinformatics, Institute of Public Health and Environment, Graduate School of Public Health, Seoul National University, Seoul, Korea
[2]SNU Bioinformatics Institute, Interdisciplinary Graduate Program in Bioinformatics, College of Natural Science, Seoul National University, Seoul, Korea
[3]Department of Biomedical Laboratory Science, Kyungdong University, Wonju, Gangwondo, Korea*
e-mail: {[1]okilinko, [2]hss2003}@snu.ac.kr , [3]hykim1984@kduniv.ac.kr*

Abstract

With developments in science and technology and improvement in living standards, human life expectancy is steadily increasing worldwide. For effective healthcare, it is necessary to check health conditions according to individuals' behavior and acquire prior knowledge on possible diseases. In this study, we classified the diseases that are major causes of death in Korea by referring to data provided by the Korea National Health and Nutrition Examination Survey. We selected indexes that could be used as indicators of major diseases and created the LCBB-SC. In the LCBB-SC, the data are systematically subdivided into related fields to provide integrated data related to each disease and to provide an infrastructure that can be used by researchers. In addition, by developing a web interface allowing for self-symptom assessments, this resource will be beneficial to people who want to check their own health condition using a list of diseases that might be caused by their behaviors.

Keywords: *Database, Chronic disease, Machine learning, Self-symptom checker, Bioinformatics*

1. Introduction

The recent increase in the prevalence of chronic diseases, which can be attributed to dietary changes, decreased physical activity, and an increasing elderly population, has become an important social issue. Despite the increase in the number of multi-morbidity patients who suffer two or more diseases due to an increase in chronic disease prevalence, diagnostic tests and medical care are usually concentrated on a single state of disease [1]. Multi-morbidity patients are more likely to die early than patients with a single condition, and this trend is especially common in the elderly population and low-income families, who are vulnerable to reduced immunity due to lengthy treatments and hospital stays [2]. Over the past few years, many medical

Manuscript Received: Aug. 31, 2018 / Revised: Sep. 5, 2018 / Accepted: Sep. 11, 2018
Corresponding Author: hykim1984@kduniv.ac.kr
Tel: +82-33-738-1374, Fax: +82-33-738-0243
Department of Biomedical Laboratory Science, Kyungdong University, Korea

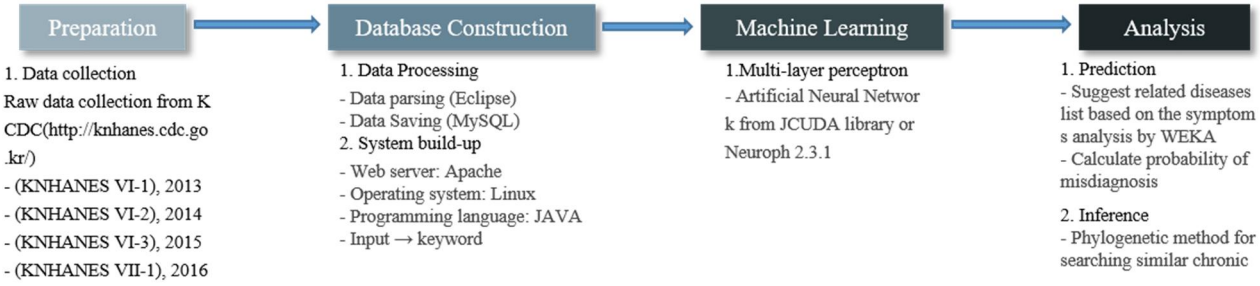


Figure 1. Workflow of the research

planners and governments worldwide have become aware of these problems and have started to study age-related diseases. In particular, the rapid development of information and communication technology, which is a major technology of the fourth industrial revolution, has attracted much attention from simple systems, such as online self-diagnosis. Therefore, to effectively manage diseases, it is important to extend the existing therapeutic disease-centered approach and establish a preventive system using machine learning methods. Additionally, it is necessary to check health conditions according to individuals' behavior and acquire prior knowledge on possible diseases.

2. Methods

In this study, we used the health behavior survey data from the Korea National Health and Nutrition Examination Survey (KNHANES), which is conducted annually to calculate the statistics required for health policy establishment and evaluation. Based on the raw data provided on the home page, the relevant symptoms, indicators, major diseases, and other information were classified and sorted through references related to chronic diseases. To extract the necessary information, the raw data provided in SAS files were first converted into a text file stored in MySQL via Java programming to build the primary database. In addition, health behavior data such as drinking and smoking, which are known to cause some chronic diseases, and diagnosis information from physicians were extracted by individual ID from the survey for use as key indicators of chronic diseases to construct the secondary database. The total number of data used was 2,134,370 and the large amount of processed data extracted using Java was stored in MySQL. Four databases were built by making tables for each year so that raw data could be used for future data analysis. In addition, we constructed a database combining the fourth year of data for use in integrated analyses, as well as a processing database containing only the information used in this study. A summary of the system development environment for conducting this study is shown in Table 1.

Pre-processing of the major feature selection was performed through Weka's 'Attributeselection' object to select a subset of the relevant dataset to be used in the learning model. To evaluate the information available

Table 1. System development environment

Category	System development environment
Data	SAS
CPU	8C AMD Opteron-6128 2.0GHz x 1 (8Core)
Memory	Master node(16GB), 10compute node 8GB
HDD	500Gb SATA 7200rpm 6Gbps x 1
OS	Linux
Web server	Apache Tomcat 8.5.27
DBMS	MySQL

Algorithm	ANNs (Multilayer neural network)
Programming language	Java, JSP, Servlet, HTML, Java script

for a particular property, we added the 'evaluator' parameter and the 'ranker' parameter, which ranks the attributes according to the scores given by the 'evaluator'. To create a classification learning model, we implemented a learning model that allows faster and more precise prediction by leaving the attribute with the highest informative attributes and removing the less informative attributes. We used TreeVisualizer, a built-in function of Weka, to check the structure of the tree. From the root node, we ascertained the leaf node by descending the tree according to the attribute match. To implement multiple layer perceptron (MLP), we first created an 'MLP' package and created the 'MultiLayerPerceptrons.java' class consisting of the basic procedures of neural network algorithms and a 'HiddenLayer.java' class containing the actual back-propagation code. The 'LogisticRegression.java' class using multi-category logistic regression was used for the output layer. We added the slope and hyperbolic tangent (tanh) of the sigmoid function into the 'ActivationFunction.java' class, as well as a method that generates a random number with a normal distribution in the 'RandonGenerator.java' class. The parameters of the MLP were represented by the weight (W) and bias (b) of the 'HiddenLayer' and 'LogisticRegression' class, and W was coded to initialize randomly according to the number of units. In this case, if the initial values were not well distributed, local minimum value problems could occur frequently; therefore, random seeding was applied.

Table 2. Type and number of data used for system construction

Category	Data	Number of data	Code	Data type
Personal information	Sex	31,098	Sex	Integer
	Age	31,098	Age	Integer
	Height	29,443	HE_ht	Decimal(12)
	Weight	29,474	HE_wt	Decimal(12)
	Waist size	29,439	HE_wc	Decimal(12)
	BMI	29,441	HE_BMI	Decimal(12)
	Pulse regularity	26,060	HE_rPLS	Integer
	Pulse per minute	1,925	HE_mPLS	Integer
	Blood pressure	26,062	HE_nARM	Integer
Symptoms	Discomfort	57,738	D_2_wk	Integer
	Cough	18,306	HE_cough	Integer
	Phlegm	19,138	HE_tb2	Var(225)
	Blood phlegm	2,914	HE_tb3	Var(225)
	Chest pain	2,969	HE_tb4	Var(225)
	Dyspnea	2,920	HE_tb5	Var(225)
Symptoms	Notable weight loss	2,937	HE_tb6	Var(225)
	Tiredness	3,017	HE_tb7	Var(225)
	Fever	2,902	HE_tb8	Var(225)
	Night fever	2,906	HE_tb9	Var(225)
Health behavior	Drinking	282,490	BD1~7	Integer
	Smoking	245,129	BS1~13	Integer
Disease	Hypertension	57,740	DI1_dg, DI1_pr	Integer
	Dyslipidemia	57,740	DI2_dg, DI2_pr	Integer
	Stroke	57,740	DI3_dg, DI3_pr	Integer

	Myocardial infarction/ angina pectoris	55,178	DI4_dg, DI4_pr	Integer
	Myocardial infarction	57,740	DI5_dg, DI5_pr	Integer
	Angina pectoris	57,740	DI6_dg, DI6_pr	Integer
	Tuberculosis	57,740	DJ2_dg, DJ2_pr	Integer
	Asthma	57,740	DJ4_dg, DJ4_pr	Integer
	Diabetes	57,740	DE1_dg, DE1_pr	Integer
	Thyroid	57,740	DE2_dg, DE2_pr	Integer
	Stomach cancer	57,740	DC1_dg, DC1_pr	Integer
	Liver cancer	57,740	DC2_dg, DC2_pr	Integer
	Lung cancer	57,740	DC6_dg, DC6_pr	Integer
	Thyroid cancer	57,740	DC7_dg, DC7_pr	Integer
	Renal failure	57,740	DN1_dg, DN1_pr	Integer
	Liver cirrhosis	57,740	DK4_dg, DK4_pr	Integer
Others	Other related information	136,553	BP8, BP1, BP5, BO3_14, BO3_05	Integer
	Weight	11,840	wt_pft, wt_ex1, wt_ex1pf	Double, Var(225)

3. Results

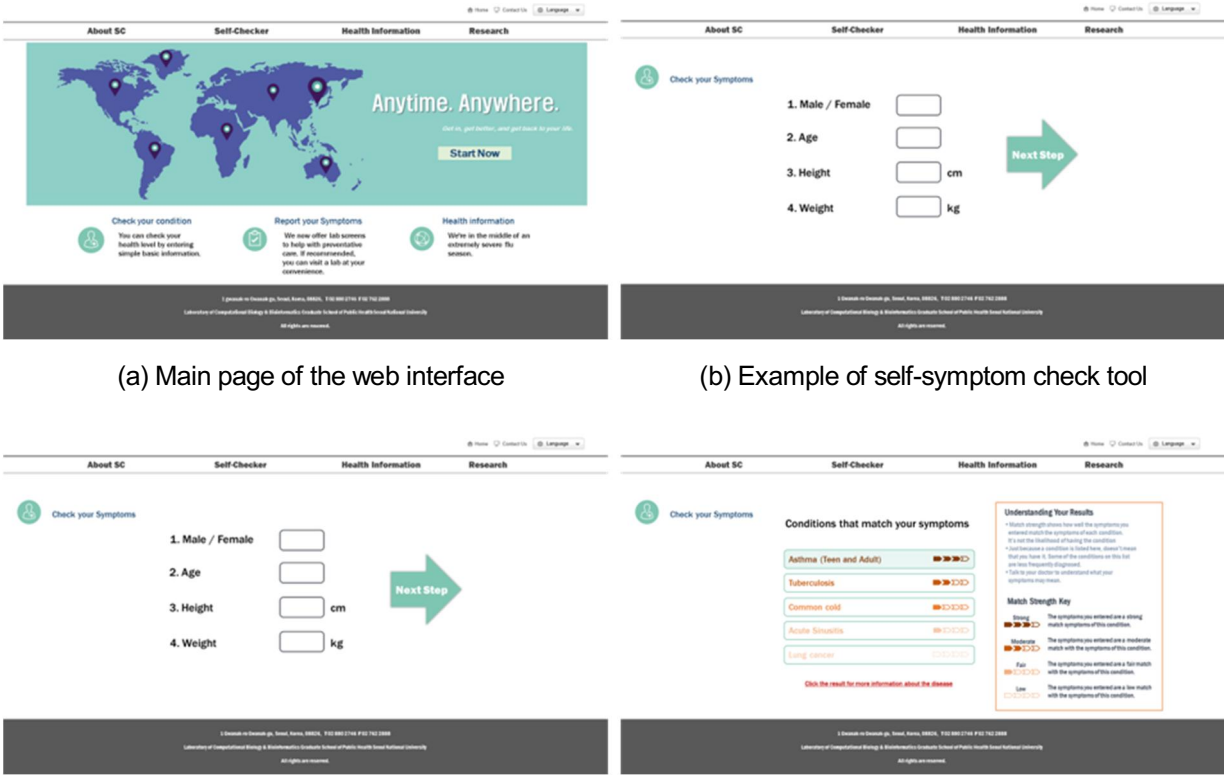
3.1 Web-interface

A total of five databases were used for the implementation of the Lab of Computer Biology and Bioinformatics - Symptom Checker (LCBB-SC), a web interface of the health index-based disease prediction system. Four databases processed each year were used for searching, and a combined database was used to provide a systematic list of relevant symptoms for the machine learning component. LCBB-SC is currently available through http://lcbb3.snu.ac.kr/LCBB-SC. The main menu tabs are 'About SC', 'Self-Checker', 'Health Information', and 'Research' (Figure 2). The 'Data & Analysis' page provides information on the source and basic information of the data used and on the machine learning algorithms used in the analysis process. There is no sub-menu within the 'Self-Checker' tab; instead, the self-diagnosis check program is accessible by clicking the tab. Basic user information, such as gender, age, height, and weight, is input at the beginning of the search. Then, the user inputs blood pressure, pulse, body mass index (BMI) and waist circumference. Variables can be left blank if the user is uncertain about the information, and figures such as BMI can be automatically calculated from height and weight information. The subsequent screen allows the user to select whether he/she is currently suffering from a disease or if a disease that has been confirmed by his/her doctor. Then, the user can select the symptoms he/she currently has. The 'Health Information' tab contains bibliographic information on various diseases. Finally, in the 'Research' tab, the user can find information about the ten databases that were used for this study, and the user can download the processed material. In addition, the related research infrastructure is provided by making the machine learning algorithm and applicable Java code used in this study available. The user can search for disease-related information through the search function in the upper right-hand corner of the website and access related links for additional information.

3.2 Machine learning analysis

We used Weka (v.3.8.2; [3]) to analyze correlations between health indicators and diseases provided in

the raw data and to compare the accuracy of predictions. Analyzing the acquired information with the evaluator parameter through the 'Attributeselection' object, the codes with the highest informativeness are as follows: (HE_tb4_D), 30 (HE_tb7_D), 34 (HE_tb9_D), 38 (BD1_11), 41 (BD2_32), 58 (DI1_dg), 71 (DJ2_pr), 74 (DE1_dg), and 95 (wt_pft). To develop a model that can make more accurate and rapid predictions, we removed the non-core attributes that were not essential for running the dataset and made a classification learning model using the selected data to re-implement Quinlan's decision tree learner [4]. Using the J48 class, the output tree had a total of 48 nodes, 22 of which were endpoints. Through the 'ActivationFunction.java' class, we applied a function that activated the output value of the neuron with a net



(a) Main page of the web interface

(b) Example of self-symptom check tool

(c) Example of self-symptom check tool

(d) Example page of result

Figure 2. The Lab of Computer Biology and Bioinformatics - Symptom Checker (LCBB-SC)

value was larger than the threshold value and deactivated the output value of the non-neuron, which was analyzed according to the 48 sigmoid nodes. In addition, a 10-fold crossover evaluation model was applied to evaluate the performance of the model and to confirm the correlation between the diagnosis and presence of a disease. In the case of hypertension, there was a high correlation between systolic blood pressure and diastolic blood pressure, BMI (>30kg/m), and current average daily smoking amount. Among chronic diseases, there was a positive correlation with stroke, myocardial infarction, angina, diabetes, thyroid, and renal failure. Dyslipidemia was also correlated with BMI and drinking, and there was a correlation between stroke, diabetes, and renal failure among chronic diseases. Stroke had a relatively high correlation with the number of days of fever and smoking, and it was associated with hypertension, dyslipidemia, and diabetes. Myocardial infarction showed a high correlation with chest pain and dyspnea, and correlated with angina and diabetes in chronic diseases. Angina pectoris was also highly correlated with the expression of chest pain and

smoking, and high correlation coefficients were found among myocardial infarction, thyroid disease, and diabetes mellitus. In the case of pulmonary tuberculosis, there was a positive correlation between coughing and dyspnea for three consecutive months or more. Asthma had a high correlation with respiratory difficulty and exposure to secondhand smoke. Diabetes mellitus had the highest correlation coefficient, with a high correlation with most chronic diseases and parameters such as age, weight, BMI, blood pressure, drinking, smoking, hypertension, and dyslipidemia. Thyroid was correlated with age, hypertension, and angina. In malignant neoplasms, negative correlations were frequently seen. In stomach cancer, there was a high correlation with the number of days of fever and a negative correlation with sex, indicating that the disease is more frequent in males than in females. For liver cancer, there was a high correlation between gender, monthly drinking rate, fatigue appearance, and cirrhosis. On the other hand, there was a negative correlation between daily mean sleep time and continuous depression for two weeks or more. Lung cancer exhibited a high correlation with the experience of discomfort during the past two weeks, drinking, the presence of secondhand smoke in the rectum, thyroid cancer, perceived stress level, and average sleep time. Thyroid carcinoma was highly correlated with age, drinking, myocardial infarction, and hypertension. Renal failure was associated with a high incidence of fatigue, myocardial infarction, angina, hypertension, and dyslipidemia, but only had a positive correlation with liver cancer.

Only results with a p-value < 0.0001, with significant differences among multiple correlation analysis results, were extracted, from which a distance matrix was created. Correlation coefficients that close to zero were not shown in the neural network algorithm. Using the results of the correlation analysis, we found that the predictions of the algorithms were 4.7% for hypertension, 9.8% for stroke, 3.2% for myocardial infarction, and 11.3% for diabetes mellitus. In addition, the predicted rate of angina pectoris was 2.1%; however, similar results were obtained for various malignant neoplasms. In the case of cancer patients, the difference in data from the conventional method is not clear, because the amount of data was not very large. Meanwhile, diabetes and hypertension showed higher prediction rates because of the large amount of data available compared with the cancer patient data. In addition, the correlation analysis can be interpreted as a result before the adjustment of parameters and weight values. As a result of model learning by applying the feed forward method, the weight value was adjusted, and a more accurate and specific weight was used to perform the analysis.

In the results of the machine learning analysis, the effect of BMI on hypertension was the highest (10.7784) and the average daily smoking amount was also high (6.3455). Diabetes (5.5124), renal failure (2.2643), and stroke (2.2548) were the most frequently associated diseases with hypertension, myocardial infarction, or angina pectoris (8.4531). In the case of dyslipidemia, BMI was the highest (7.6924), and it was also highly correlated with drinking (6.9944). The high value (7.3455) in the present study is consistent with the findings of a number of studies on smoking and stroke [5-9]. However, there was a negative correlation (-7.3686) with the number of days of fever, which could be interpreted as an increase in the incidence of stroke with little or no fever days, but there was no evidence or study to support it. Therefore, it is necessary to investigate additional literature or modify the relevant components of the algorithms. Stroke-related diseases were hypertension (7.2667), dyslipidemia (5.9564), and diabetes (4.5675). Myocardial infarction and angina pectoris were highly correlated, with values of 22.4495 and 20.9892, respectively, and the risk factors associated with these two diseases were similar to those of dyspnea (10.8972/8.8985), chest pain (7.9632/9.5541), smoking (17.7637), and diabetes mellitus, as a common chronic disease (6.4541/4.1142). In the case of pulmonary tuberculosis, cough fever days (9.1154) and dyspnea (9.3122) were the most influential factors. In the case of diabetes, there was a correlation with many of the factors tested. BMI (9.7784), weight (8.2416), and age (6.4286) were the main factors influencing diabetes, in addition to the

monthly amount of drinking (4.5417) and smoking amount (2.4875). In the case of thyroid disease, age (7.9614) and weight loss (4.5495) were the highest positively correlated variables. Hypertension had the highest value of 6.3412, and that of angina pectoris was 2.7319, which was relatively low compared with other chronic diseases. In the case of stomach cancer, a negative correlation coefficient of -4.9784 was found for gender, which can be interpreted as the disease being mostly seen in males, given that the male gender was set to 1 and female was set to 2. In the case of lung cancer, the correlation between actual drinking and gastric cancer has been confirmed in numerous studies and recent meta-analyses [10-12]. In the case of renal failure, the number of days of fatigue was found to have the greatest effect (2.955), while high blood pressure and diabetes were relatively high (2.0641 and 2.0013, respectively) (Table 3).

Table 3. Multiple layer perceptron result for chronic disease (10-folds)

Code	Data (No. of data)	Code	Data (No. of data)	Code	Data (No. of data)
DI1_dg	6.3455 * BS3_2 + 2.2548 * DN1_dg + 2.2643 * DN1_pr + 11.2264 * HE_sbp1 + 10.9765 * HE_dbp1 + 10.7784 * HE_BMI + 8.4531 * DI4_pr + 5.5124 * DE1_dg 4.4894 * DE1_pr	DI6_dg	20.9892 * DI4_dg + 8.8985 * HE_tb4 + 17.7637 * DI5_dg + 9.5541 * HE_tb4_d + 4.3039 * BS3_1 + 6.4484 * BS3_2 + 6.8841 * DE2_dg 4.1142 * DE1_dg	DE1_dg	6.4286 * age + 8.2416 * HE_wt + 9.7784 * HE_BMI + 5.2264 * HE_sbp1 + 3.3714 * HE_dbp1 + 2.4197 * BD13 + 4.5417 * dr_month + 8.8743 * BS3_2 + 2.4875 * BS6_2 + 8.4495 * DI1_dg + 6.7314 * DI2_dg + 3.7112 * DI3_dg + 2.4495 * DI5_dg + 1.9411 * DI3_dg +
DI2_dg	2.2131 * DI4_pr + -2.2208 * DC6_pr+ 7.6924 * HE_BMI + 6.9944 * dr_month + 4.2927 * DI3_pr + 6.2867 * DE1_dg 5.4841 * DE1_pr	DJ2_dg	-2.2545 * DI4_pr 2.2961 * DI5_pr + 9.1154 * HE_cough1 9.3122 * HE_tb5 + 7.4487 * HE_tb5_d	DE2_dg	4.5495 * HE_tb6 + 7.9614 * age + 6.3412 * DI1_dg + 2.7319 * DI6_dg +
DI3_dg	-7.3683 * HE_tb8 4.3205 * HE_tb9 + 7.3455 * BS3_2 + 7.2667 * DI1_dg + 5.9564 * DI2_dg + 4.5675 * DE1_dg	DJ4_dg	6.4167 * HE_tb5 + 4.8871 * BS8_2 + 3.1451 * BS9_2 + 4.5771 * BS13 +	DC1_dg	6.4167 * HE_tb8 + -4.9784 * sex 3.0547 * BD1_11 +
DI4_dg	0.8016 * DI5_dg + 0.9171 * DI6_dg +	DI5_dg	22.4495 * DI4_dg + 16.2391 * DI6_dg + 10.8972 * HE_tb4 + 7.9632 * HE_tb4_d + 9.6544 * HE_tb5 + 6.3554 * HE_tb5_d +	DC7_dg	8.3645 * age + 2.9465 * dr_month + 3.4998 * DI5_dg + 2.0107 * DI1_dg_2 + 2.0013 * DE1_dg +

DC2_dg	4.7984 * dr_month + 3.1194 * HE_tb7_d + 2.1974 * DK4_dg + -6.1173 * sex+ -2.4855 * BP8 -1.5487 * BP8	DN1_dg	6.4541 * DE1_dg 2.9545 * HE_tb7 + 1.9451 * DI5_dg + 2.5498 * DI6_dg + 2.0641 * DI1_dg + 1.7498 * DI2_dg +	DC6_dg	2.7984 * D_2_1 + 6.7064 * BS8_2 + 3.9322 * dr_month + 2.0494 * DC7_dg +

4. Discussion

Based on the large amount of data related to existing chronic diseases, the results of this study can be presented as a basis for the selection of a suitable algorithm for predicting related chronic diseases using a health index. Further, we can derive a reasonable prediction result of the weighted neural network by quantifying it. In addition, this study presents not only an analysis of disease risk factors but also the results of quantified associations between chronic diseases, differentiating it from previous studies. However, because the analysis was conducted based on survey data provided by the general public, a respondent could possibly misrepresent their health data, and non-responses reduced the number of valid samples, affecting the quality of the data and reducing the accuracy of the survey estimates. In the future, we will provide a more diverse and reliable infrastructure for chronic disease research by adding updated eating habits and genetic data from the survey that were not included in this study. These results can be used to infer or predict the correlations among new and unknown diseases in research, such as thyroid disease, hypertension, and cardiovascular disease. It is also expected that the machine-running algorithms used in the analysis process will supplement the analysis of results that are judged to be errors through literature research or published books, and ultimately contribute to laying a foundation for and supporting the expansion of health research studies through the establishment of a reliable and independent chronic disease database.

5. Conclusion

As the aging population increases, the number of chronic disease patients who need constant care is steadily increasing worldwide, most of whom have multiple diseases at the same time, making comprehensive approaches to common ailments necessary. Therefore, effective management of diseases is necessary to extend the existing treatment-oriented approach to a preventive focus and to check an individual's health condition in advance. It is also important to acquire preventive knowledge about possible diseases and identify and evaluate the various complications that may arise from complex multiple disease states. In this study, by establishing a chronic disease management system specialized in Korea using local information and technology, we have identified various disease-related factors that are the main causes of death in Korea to provide a basis for promoting a healthy life span and public health.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT and MOE) (No. 2016R1C1B2015511 and No. 2017R1D1A1B03033413).

References

[1] M. Fortin, G. Bravo, C. Hudon, A. Vanasse and L. Lapointe. Prevalence of multimorbidity among adults seen in family practice. The Annals of Family Medicine. Vol. 3, No. 3, pp. 223-228, May 2005.

[2] C. Vogeli, A. Shields, T.A. Lee, T.B. Gibson, W.D. Marder, K.B. Weiss and D. Blumenthal. Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. Journal of general internal medicine. Vol. 22 No. 3, pp. 391-395, Dec 2007.

[3] I.H. Witten, E. Frank, M.A. Hall and C.J Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. Oct 1, 2016.

[4] J.R. Quinlan. Induction of decision trees. Machine learning. Vol. 1, No. 1, pp. 91-106, Mar 1986.

[5] G.A Colditz, R. Bonita, M. J. Stampfer, W. C. Willett, B. Rosner, F.E. Speizer and C.H. Hennekens. Cigarette smoking and risk of stroke in middle-aged women. New England Journal of Medicine. Vol. 318, No. 15, pp. 937-941, Apr 1988.

[6] M. Higa and Z. Davanipour. Smoking and stroke. Neuroepidemiology. Vol. 10, No. 4, pp. 211-222, 1991.

[7] R. Shinton and G. Beevers. Meta-analysis of relation between cigarette smoking and stroke. Bmj. Vol. 298, No. 6676, pp. 789-794, Mar 1989.

[8] P.A. Wolf, R.B. D'Agostino, W.B. Kannel, R. Bonita and A.J. Belanger. Cigarette smoking as a risk factor for stroke: The Framingham Study. Jama. Vol. 259, No. 7, pp. 1025-1029, Feb 1988.

[9] G.J. Hankey. Smoking and risk of stroke. Journal of cardiovascular risk. Vol. 6, No. 4, pp. 207-211, Aug 1999.

[10] W.H. Chow, C.A. Swanson, J. Lissowska, F.D. Groves and L.H. Sobin, A. Nasierowska□Guttmejer, J. Radziszewski, J. Regula, A.W. Hsing, S. Jagannatha and W. Zatonski. Risk of stomach cancer in relation to consumption of cigarettes, alcohol, tea and coffee in Warsaw, Poland. International journal of cancer. Vol. 81, No. 6, pp. 871-876, Jun 1999.

[11] P. Boffetta and M. Hashibe. Alcohol and cancer. The lancet oncology. Vol. 7, No. 2, pp. 149-156, Feb 2006.

[12] K. Ma, Z. Baloch, T.T. He and X. Xia. Alcohol consumption and gastric cancer risk: A meta-analysis. Medical science monitor: international medical journal of experimental and clinical research. Vol. 28, pp. 238, 2017.