Development of Prediction Model for Endocrine Disorders in the Korean Elderly Using CART Algorithm

Results from a Population-Based Study

Haewon Byeon
Department of Speech Language Pathology & Audiology
Nambu University
Gwangju, Republic of Korea

***Abstract*—The aim of the present cross-sectional study was to analyze the factors that affect endocrine disorders in the Korean elderly. The data were taken from the A Study of the Seoul Welfare Panel Study 2010. The subjects were 2111 people (879 males, 1,232 females) aged 60 and older living in the community. The dependent variable was defined as the prevalence of endocrine disorders. The explanatory variables were gender, level of education, household income, employment status, marital status, drinking, smoking, BMI, subjective health status, physical activity, experience of stress, and depression. In the Classification and Regression Tree (CART) algorithm analysis, subjective health status, BMI, education level, and household income were significantly associated with endocrine disorders in the Korean elderly. The most preferentially involved predictor was subjective health status. The development of guidelines and health education to prevent endocrine disorders is required for taking multiple risk factors into account.**

Keywords—data-mining; CART; elderly; health behavior; endocrine disorders

I. INTRODUCTION

One of the biggest difficulties in old age is health problems [1]. Among them, endocrine disorders, such as diabetes, are known as typical old-age chronic diseases [2]. According to a survey on causes of death by Statistics Korea, diabetes is the fifth major cause of death as of 2013, and fatalities from diabetes increase dramatically by over 40 times from 3.3 to 135.4 out of every 100,000 people in the population of those in their 40s to those in their 70s, respectively [3]. Given that diabetes is related to various complications, such as cerebrovascular diseases, its actual effect on death is predicted to be much greater. In addition, endocrine disorders add to the psychological and economic burden on a patient's family as well as the patient themselves in that they require consistent self-management to prevent complications after treatment.

The endocrine system is a generic term that refers to hormone secretion organs, and, based on its structure, the system consists of the pituitary, pineal, thyroid, parathyroid, and adrenal glands and the pancreas and gonads [4]. Hormones secreted from the endocrine system move to the target organs through the blood and play important roles in the growth and development of the body and the maintenance of the metabolism and homeostasis [5]. Therefore, although an imbalance of hormones does not manifest symptoms immediately, sustained hormonal problems ultimately cause metabolic disorders and fatal complications by destroying the balance of metabolic activities in the body [6, 7]. In particular, when managed poorly, diabetes, an endocrine disease, has a high possibility of causing complications, such as cerebrovascular diseases (e.g. stroke) or microvascular diseases (e.g. diabetic retinopathy) [8]; therefore, the ultimate treatment goal of endocrine disorders is controlling the disease process.

As it is necessary to manage risk factors, such as life habits, to prevent endocrine disorders and complications, it is important to elucidate related factors that affect endocrine disorders to ensure healthy aging, especially in old age. So far, a high level of education and income, participation in social activities, and positive social recognition have been reported as protective factors against endocrine disorders [2, 4, 9]. These protective factors play a role in decreasing the risk of endocrine disorders by promoting positive health-related behaviors.

On the other hand, depression, stress, irregular eating patterns, smoking, drinking, obesity, and lack of exercise have been reported as risk factors that increase the risk of endocrine disorders [10–13]. The majority of preceding studies, however, have researched risk factors on an individual basis by using regression analysis, and there is still a lack of studies that have explored various related factors in an integrated manner. In particular, a regression model requires assumptions, such as linearity, normality, and homoscedasticity, and as the distributions of some disease data are non-linear, the normality and homoscedasticity of this regression model are not suitable.

Recently, as data-mining has been used as a method of predicting diseases, complex exploration is being used for risk factors [14]. In particular, among the data-mining methods, the Classification And Regression Tree (CART) has several advantages; first, it enables nonparametric analysis; second, it enables easy understanding of disorders, as its analysis process is expressed in the tree structure; third, it enables understanding of the most closely-related factors.

Endocrine disorders frequently elude complete recoveries. In addition, since practical management must be performed by the patient themselves even when the treatment is in progress, complications should be prevented or postponed through

elimination of risk factors and sustained management [1]. Therefore, a vital theme in maintaining health in old age is determining complex factors related to old-age endocrine disorders and predicting high-risk groups.

This study developed a prediction model of endocrine disorders for the Korean elderly by using data-mining and provides basic material for the prevention of old-age endocrine disorders. This paper is organized as follows: the study population and measurements are described in section 2. CART Algorithm is described in section 3. I conducted a series of experiments to verify proper performance of CART Algorithm in section 4 and the conclusions are presented in section 5.

II. Methods

A. Study population

This study analyzed raw data from Seoul Welfare Panel Study (SWPS) conducted by Seoul Welfare Foundation on citizens of Seoul from June 1, 2010 through August 31, 2010. Having acquired authorization of Statistics Korea (no. 20113) in 2009, SWPS was conducted for the purpose of investigating the level of welfare of the households and reality of vulnerable classes in Seoul and estimating the demand of welfare services [15]. The population of the study was households in Seoul at the time of 2005 Population and Housing Census and sampling was conducted on 25 districts of Seoul using stratified cluster sampling. Major survey items were income, economic level, health, living condition and demand for welfare services and survey method was Computer Assisted Personal Interviewing in which interviewers visited target households and entered responses answered according to the structured questions in notebook computers.

This study analyzed 2,111 senior citizens (879 males, 1,232 females) over the age of 60 among 7,761 people who completed SWPS.

B. Measurements

Outcome was defined as prevalence of endocrine disorders (diabetes, hypothyroidism, thyroid hyperactivity). Explanatory variables were included as sex, final education (elementary school and lower, middle, high school, over college), whether or not being engaged in economic activities (yes, no), the average monthly income of households (less than 2 million won, 2-4 million won, more than 4 million won), marital status (living with spouse, living without spouse, unmarried person), binge drinking (yes, no), smoking (non-smoker, past smoker, current smoker), BMI (underweight, normal, overweight), subjective health status (good, fair, poor), regular exercise (no, yes), experience of stress in the last 1 months, depressive symptoms in the last 1 months (yes, no). Binge drinking was defined as five or more drinks ($\geqslant$ 61 g of alcohol) per episode for men and as four or more drinks per episode ($\geqslant$ 41 g of alcohol) for women, with reference to the International Center for Alcohol Policies [16].

III. Statistical Analysis

A. Exploration on factors related to the endocrine disorders

For general characteristics, mean and percentage were presented and difference between groups based on endocrine disorders was analyzed by Chi-square test.

B. CART Algorithm

When the related factors of endocrine disorders were identified in the chi-square test, the related factors of endocrine disorders were statistically classified and a prediction model was established, using CART (Classification And Regression Tree) Algorithm.

Classification And Regression Tree (CART) is a data-mining algorithm suggested by Breiman in 1984 which measures impurities by using Gini Index and performs binary split that only forms 2 children nodes from the parent's node [17].

CART has the advantage that it enables easy interpretation of created rules and it can use both continuous variables and categorical ones. Continuous variables creates separation rules in the form of "X≤C?" or "X≥C?" while categorical binary creates separation rules in the form of "X∈{A, B}".

Gini's coefficient is a probability of two extracted elements' belonging to two different groups when the two elements are extracted from n number of elements [18]. First, misclassification probability is calculated in each node and the statistic equation runs as formula 1.

$$Gini\ Index(t) = 1 - \sum_{j}[P(j/t)]^2 \quad (1)$$

Next, all misclassification probabilities are added and the estimate of misclassification probability is calculated with formula 2.

$$G = \sum_{j=1}^{c}\sum_{i \neq j} P(i)P(j) \quad (2)$$

After the decrement of Gini's coefficient is calculated, the predictor which reduces Gini's coefficient the most is chosen as child node as the last step of the algorithm formula 3.

$$G = \sum_{j=1}^{c} P(i)(1-P(j)) = 1 - \sum_{j=1}^{c} P(j)^2 = 1 - \sum_{j=1}^{c} (n_j/n) \quad (3)$$

In the CART, the Alpha value for the criteria of splitting and merging was set at 0.05. The number of parent nodes was 200 and that of child nodes was 100, and the number of branches was limited to 5. The validity of the model was tested using the 10-fold cross-validation.

IV. Results

A. General characteristics of subjects and factors related to endocrine disorders

General characteristics of subjects and factors related to endocrine disorders are presented in Table 1.

TABLE I. GENERAL CHARACTERISTICS OF THE SUBJECTS BASED ON ENDOCRINE DISORDERS, N (%)

Variables	Endocrine disorders		p
	No (n=1,738)	*Yes (n=373)*	
Sex			0.098
Male	738 (84.0)	141 (16.0)	
Female	1,000 (81.2)	232 (18.8)	
Final education			0.170
≤ Elementary school	735 (80.4)	179 (19.6)	
Middle school	309 (82.8)	64 (17.2)	
High school	422 (85.1)	74 (14.9)	
≥ College	272 (82.9)	56 (17.1)	
Economic activities			0.007
Yes	309 (87.3)	45 (12.7)	
No	1,429 (81.3)	328 (18.7)	
Average monthly income of households			0.388
< 2 million won	1,113 (81.5)	253 (18.5)	
2-4 million won	409 (84.2)	77 (15.8)	
> 4 million won	75 (80.6)	18 (19.4)	
Marital status			0.978
Living with spouse	1,167 (82.2)	252 (17.8)	
Living without spouse	35 (83.3)	7 (16.7)	
Unmarried person	536 (82.5)	114 (17.5)	
Binge drinking			<0.001
Yes	438 (87.8)	61 (12.2)	
No	1,300 (80.6)	312 (19.4)	
Smoking			0.565
Current smoker	180 (84.1)	34 (15.9)	
Past smoker	373 (80.9)	88 (19.1)	
Non smoker	1,185 (82.5)	251 (17.5)	
BMI			0.059
Underweight	410 (82.5)	87 (17.5)	
Normal	1,034 (83.5)	204 (16.5)	
Overweight	294 (78.2)	82 (21.8)	
Subjective health status			<0.001
Good	533 (91.3)	51 (8.7)	
Fair	584 (83.7)	114 (16.3)	
Poor	621 (74.9)	208 (25.1)	
Regular exercise			0.842
Yes	773 (82.1)	168 (17.9)	
No	965 (82.5)	205 (17.5)	
Experience of stress in the last 1 months			0.022
Yes	472 (78.9)	126 (21.1)	
No	868 (84.4)	161 (15.6)	
Depressive symptoms in the last 1 months			0.001
Yes	421 (77.8)	120 (22.2)	
No	1,317 (83.9)	253 (16.1)	

Among the total of 2,111 subjects, number of those who have endocrine disorders was 373 (17.7%).

As the result of chi-square test, prevalence of endocrine disorders has statistically significant difference in conomic activities, alcohol drinking, BMI, subjective health status, experience of Stress, depressive symptom ($p<0.05$).

The prevalence of endocrine disorders was higher in unemployed(18.7%), non-drinkers(19.4%), obesity(21.8%) those with bad subjective health(25.1%), those who are under stress (21.1%), those with depression symptoms (22.2%), unemployed (18.7%), non-drinkers (19.4%), obesity (21.8%) and subjective poor health (25.1%), stress received (21.1%), that depressive symptoms (22.2%).

B. Prediction model for endocrine disorders using CART algorithm

Prediction model for endocrine disorders using CART algorithm is presented in Figure 1.

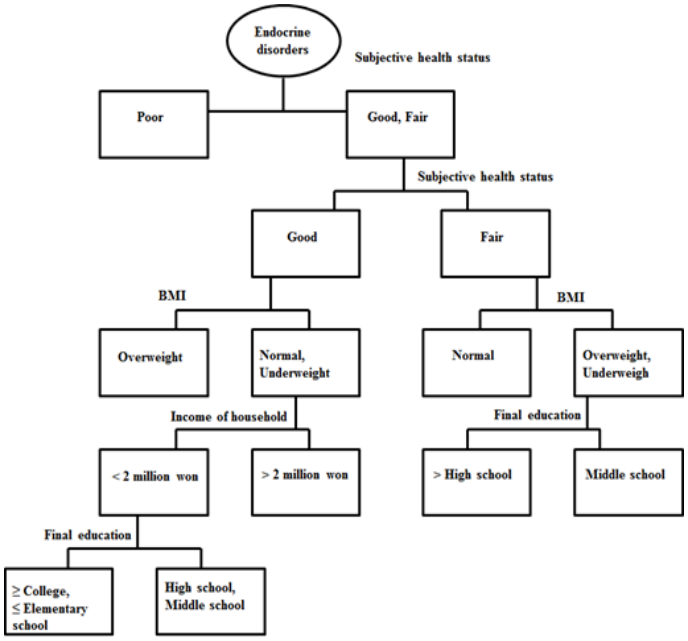

Fig. 1. Prediction model for endocrine disorders

As the result of constructing statistical classification model using CART algorithm after including variables set as factors related to endocrine disorders through chi-squared test, factors having significant effect were subjective health status, BMI, education level, and household income. The most preferentially involved predictor was subjective health status.

Table 2 is a profit chart of prediction model for endocrine disorders by CART algorithm suggested in the higher order of path for subjects' improved gain. In CART algorithm, the paths with improved gain of less than 100% are regarded as insignificant.

When this study drew out profit indicator for each node to seek out prediction paths for endocrine disorders, 2 nodes were confirmed as significant paths which effectively predict the endocrine disorders.

TABLE II. PROFIT CHART OF PREDICTION MODEL FOR ENDOCRINE DISORDERS BY CART ALGORITHM

Node no	Subjects (%)[a]	Gain n (%)[b]	Response %[c]	Gain Index %[d]	Description
12	88 (4.2)	24 (6.4)	27.3	154.4	The elderly who are middle school graduates with average subjective health and BMI which is either obese or underweight
1	829 (39.3)	208 (55.8)	25.1	142.0	The elderly who perceive their health status as poor

a. Node n(%); node number, % to 2,111

b. Gain n(%); gain number, % to 373

c. Response (%): The fraction of the endocrine disorders in the elderly

d. Gain index (%):=154.4 in total 8 node

The first path with the biggest profit indicator for the prediction of the endocrine disorders was the elderly who are middle school graduates with average subjective health and BMI which is either obese or underweight and its profit indicator was 154.4%.

The second path was the elderly who perceive their health status as poor and its profit indicator was 142.0%.

When the analysis on the prediction model by CART algorithm was completed, this study conducted 10-fold cross-validation test to assess developed prediction model. As the result of the 10-fold cross-validation test to compare stability of drawn-out model, drawn-out risk index was 0.287 and misclassification rate was 29% for cross classification model, showing the same risk index 0.288 and misclassification rate 29% of prediction model.

Surface area of Receiver Operation Characteristic (ROC) Curve (AUROC) which presents explanatory power of prediction model was 0.71, demonstrating that explanatory power of prediction model is average or medium level (Figure 2, Figure 3).

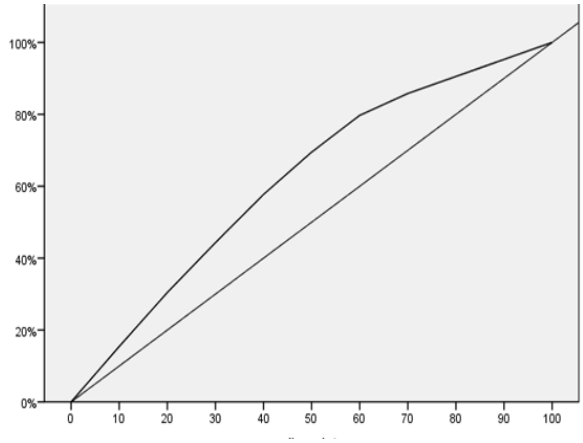

Fig. 1. Gains percentile of final model

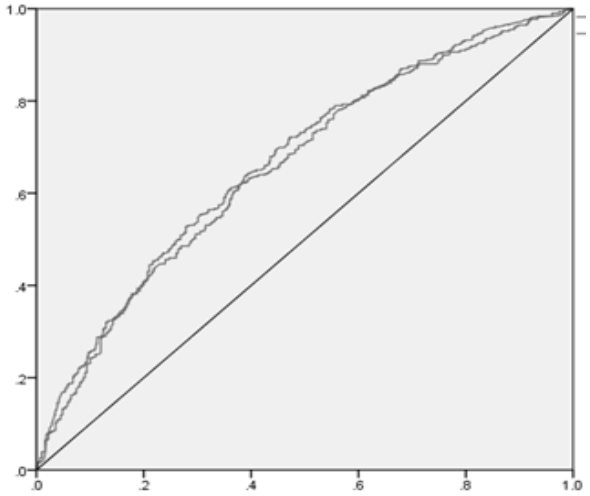

Fig. 2. ROC curve of final model

V. CONCLUSION

In order to investigate the potential factors related to endocrine disorders in the elderly over the age of 60 in local communities, this study developed a prediction model based on the CART algorithm by using epidemiological data, which represent the general Korean population.

In the prediction of endocrine disorders in old age in this study, the overriding factor was subjective health status. Subjective health is known to recognize physiological and biological changes more correctly and have a greater effect on interactions among the nerve system, endocrine system, and immune system than the objective measurement of health [19]. Additionally, the 2008 National Survey on Living Conditions and Welfare Needs of the Elderly conducted by the Korea Institute for Health and Social Affairs reported that 90.0% of the elderly who perceive their health as poor had more than one chronic disease [20]. Therefore, the reason why subjective health is highlighted as the most compelling factor for endocrine disorders in this study is because the elderly with poor subjective health may have perceived problems of the endocrine system by responding to physical problems more sensitively than the elderly with good subjective health; second, in contrast, the elderly with endocrine disorders might

have perceived their health as poor. As subjective health is a major motivating factor that influences health-promoting behaviors [21], in order to prevent endocrine disorders, it is necessary to implement regular check-ups and health education for the elderly with poor subjective health.

This study identified two high-risk groups who are vulnerable to old-age endocrine diseases. According to this prediction model, "the elderly who are middle school graduates with average subjective health and are either obese or underweight" and "the elderly who perceive their health status as poor" are at a high-risk for endocrine disorders.

According to preceding studies that investigated factors related to endocrine diseases, a high level of education was a protective factor against endocrine disorders, and, in contrast, a low level of education increased the risk of endocrine disorders [1]. Moreover, in elderly people, not only was there a relationship between health-promoting living habits and subjective health status and were socio-demographic variables, such as education, major variables of living habits [22], but the elderly with subjectively good health also practice positive living habits more [23], which demonstrates that the results of preceding studies support those of this study.

Furthermore, this study implies that the risk factors identified in preceding studies, such as socio-demographic characteristics and living habits, not only affect endocrine diseases on an individual basis, but work in synergy with other risk factors as well. Considering that subjective health status, health-promoting behaviors, and desirable living habits are correlated, it is necessary to develop programs for preventing endocrine diseases that take level of education, subjective level of health, and living habits into consideration.

The limitations of this study are as follows; first, there is a possibility that potential factors exist that may influence endocrine disorders other than the factors included in the study model. Second, as this study is based on cross-sectional research, the results of the study cannot be interpreted as causal relationships.

The elderly who are middle school graduates with average subjective health, who are either obese or underweight, and who perceive their health status as poor were high-risk groups for endocrine disorders. The development of guidelines and health education for preventing endocrine disorders is required for taking multiple risk factors into account. Furthermore, longitudinal studies to explore the causal relationship between multiple risk factors and endocrine disorders are required.

ACKNOWLEDGMENT

The author wish to thank the Seoul Welfare Foundation on citizens of Seoul that provided the raw data for analysis.

REFERENCES

[1] A. Steptoe, C. Wright, S. R. Kunz-Ebrecht, and S. Iliffe, Dispositional optimism and health behaviour in community-dwelling older people: associations with healthy ageing. British Journal of Health Psychology, vol. 11, no. 1, pp.71–84, 2006.

[2] D. Kapoor, and T. H. Jones, Smoking and hormones in health and endocrine disorders. European Journal of Endocrinology, vol. 152, no. 4, pp. 491–499, 2005.

[3] Available at: http://kostat.go.kr/portal/korea/index.action, accessed in 15/8/2015

[4] S. H. Golden, K. A. Robinson, I, Saldanha, B, Anton, and P. W. Ladenson, Prevalence and incidence of endocrine and metabolic disorders in the United States: a comprehensive review. The Journal of Clinical Endocrinology & Metabolism, vol. 94, no. 6, pp. 1853–1878, 2009.

[5] U. Meier, and A. M. Gressner, Endocrine regulation of energy metabolism: review of pathobiochemical and clinical chemical aspects of leptin, ghrelin, adiponectin, and resistin. Clinical Chemistry, vol. 50, no. 9, pp. 1511–1525, 2004.

[6] C. G. Campbell, S. E. Borglin, F. B. Green, A. Grayson, E. Wozei, and W. T. Stringfellow, Biologically directed environmental monitoring, fate, and transport of estrogenic endocrine disrupting compounds in water: a review. Chemosphere, vol. 65, no. 8, pp. 1265–1280, 2006.

[7] L. Adorini, G. Penna, Control of autoimmune diseases by the vitamin D endocrine system. Nature Clinical Practice Rheumatology, vol. 4, no. 8, pp. 404–412, 2008.

[8] J. M. Forbes, and M. E. Cooper, Mechanisms of diabetic complications. Physiological reviews, vol. 93, no. 1, pp. 137–188, 2013.

[9] A. Monzani, F. Prodam, A. Rapa, S. Moia, V. Agarla, S. Bellone, and G. Bona, Endocrine disorders in childhood and adolescence: natural history of subclinical hypothyroidism in children and adolescents and potential effects of replacement therapy: a review. European Journal of Endocrinology, vol. 168, no. 1, pp. 1–11, 2013.

[10] D. L. Musselman, and C. B. Nemeroff, Depression and endocrine disorders: focus on the thyroid and adrenal system. British Journal of Psychiatry, vol. 168, no. 30, pp. 123–128, 1996.

[11] S. Safea, Endocrine disruptors and human health: is there a problem. Toxicology, vol. 205, no. 1, pp. 3–10, 2004.

[12] S. H. Golden, A. Brown, J. A.Cauley, M. H. Chin, T. L. Gary-Webb, C. Kim, J. A. Sosa, A. E. Sumner, and B. Anton, Health disparities in endocrine disorders: biological, clinical, and nonclinical factors—an Endocrine Society scientific statement. The Journal of Clinical Endocrinology & Metabolism, vol. 97, no. 9, pp.1579–1639, 2012.

[13] S. Melmed, K. S. Polonsky, P. R. Larsen, and H. M. Kronenberg, Williams textbook of endocrinology. PA: Elsevier & Saunders, 2011.

[14] H. Byeon, The risk factors of laryngeal pathology in Korean adults using a decision tree model. Journal of Voice, vol. 29, no. 1, pp. 59–64, 2015.

[15] Seoul Welfare Foundation, Seoul Welfare Panel Study 2010. Seoul, Seoul Welfare Foundation, 2010.

[16] International Center for Alcohol Policies. ICAP blue book: practical guides for alcohol policy and prevention approaches. Available at: http://www.icap.org/PolicyTools/ICAPBlueBook/ accessed in 15/8/2015

[17] C. El Moucary, Data Mining for Engineering Schools: Predicting Students' Performance and Enrollment in Masters Programs. International Journal of Advanced Computer Science and Applications, vol. 2, no. 10, pp. 1–9, 2011.

[18] H. Byeon, and R. Lee, Prediction model for the smoking in Korean adolescent using CART algorism. Information-An International Interdisciplinary Journal, vol. 17, no. 12A, pp. 6273–6278, 2014.

[19] N. J. Kim, The effect of regular exercise on the subjective health and social activity of the elderly. Journal of Korean Physical Education, vol. 39, no. 1, pp. 149–158, 2000.

[20] Korea Institute for Health and Social Affairs, A study of 2008 living profiles and welfare service needs of older persons in Korea Survey. Seoul, Korea Institute for Health and Social Affairs, 2010.

[21] C. N. Wei, K. Harada, K. Ueda, K. Fukumoto, K. Minamoto, and A. Ueda, Assessment of health-promoting lifestyle profile in Japanese university students. Environmental health and preventive medicine, vol. 17, no. 3, pp. 222–227, 2012.

[22] O. Simsekoglu, and N. Aydemir, Health-seeking practices and psycho-social factors related to health behaviors in a Turkish sample. Journal of Behavioral Health, vol. 3, no. 1, pp. 17–24, 2014.

[23] J. Ogden, Health psychology. Berkshire, McGraw-Hill Education, 2012.