International Journal of Gerontology 13 (2019) 314–319
https://doi.org/10.6890/IJGE.201912_13(4).0010

International Journal of Gerontology

journal homepage: http://www.sgecm.org.tw/ijge/

Original Article

Development of Depression Prediction Models for Caregivers of Patients with Dementia Using Decision Tree Learning Algorithm

Haewon Byeon

Department of Speech Language Pathology, College of Health Science, Honam University, Gwangju, Republic of Korea

ARTICLE INFO

Accepted 6 July 2019

Keywords:
C4.5 algorithm,
decision tree learning,
dementia,
caregiver,
depression

SUMMARY

Background: The objective of this study was to develop a depression prediction model customized for South Korean dementia caregivers by combining the RBF artificial neural network and the C4.5 algorithm.
Method: The study subjects were selected from 228,558 people who completed the community health investigation. A total of 2,421 caregivers (1109 males and 1312 females) aged 60 years or older who completed mental health questionnaires were analyzed. Depression, a result variable, was defined using the Center for Epidemiological Studies Depression (CES-D: a Korean version).
Results: This study constructed a statistical classification model using the C4.5 algorithm and the model showed that subjective health condition was the most important predictive factor associated with the depression of the dementia caregivers.
Conclusion: The results of this study suggest that it will be necessary to develop a customized program for predicting the depression of the caregivers, who take care of the demented elderly, efficiently and managing it.

1. Introduction

Dementia requires constant care and greatly burdens a family financially and physically because a patient with dementia cannot recover his or her full functions. According to Statistics Korea,[1] as of 2011, caregivers spent 6 to 9 hours per day to take care of patients with dementia on average and they annually spent 17,300 million USD as support expenses. Therefore, the increase in the number of patients with dementia raises the caregiver burden of caregivers.

As more family members need to take care of the demented elderly, the need for evidence-based health service for the health of the dementia caregiver has been increased.[2] Therefore, the caregiver burden and the quality of life have been studied continuously.[3–6] The results of these studies showed that the burden of supporting the demented elderly for an extended period negatively affected caregivers in the physical, psychological, social, and economic aspects.[3] Particularly, it is known that the caregiver burden of the demented elderly causes psychological problems such as the depression and anxiety of caregivers.[7,8] Previous studies reported that the caregivers of the demented elderly experienced depression 2–3 times more frequently[9] and they had lower life satisfaction.[10]

Only a few studies have evaluated the variables for predicting the depression of the dementia caregivers with reflecting the characteristics of Koreans,[11] despite the fact that dementia is a common disease of the elderly in South Korea and the dementia caregivers are highly stressed due to the caregiver burden. Although many studies have been conducted for understanding the caregiver burden and depression of the dementia caregivers,[12] it is difficult to generalize and apply these results to South Korea because it may vary among countries due to cultural differences.

Data mining techniques such as decision tree are widely used as a method to search for the main variables associated with health.[13,14] The decision tree is one of machine learning methods. It is easy to interpret the results of it and the model construction does not take a long time, which are advantages of this method. Particularly, the C4.5 algorithm has shown high accuracy and fast data processing speed. Therefore, it has been widely used.[15,16]

The objective of this study was to develop a depression prediction model customized for South Korean dementia caregivers by combining the RBF artificial neural network and the C4.5 algorithm. It will provide the baseline data that can be used for studying the cognitive impairment of the elderly and establishing health policies in the future.

2. Methods

2.1. Data source

This study is a secondary data analysis study using the results of the 2015 Korean Community Health Survey (KCHS). The research was approved by the Ethics Committee at Honam university (No-1041223-201801-HR-40) and was conducted in accordance with the

* Corresponding author. Department of Speech Language Pathology, School of Public Health, Honam University, 417, Eodeung-daero, Gwangsan-gu, Gwangju, Republic of Korea.
E-mail address: bhwpuma@naver.com (H. Byeon)

ethical standards of the Declaration of Helsinki. The KCHS is a representative national epidemiological survey of South Korea in order to establish and evaluate the local health care plans and produce comparable local health statistics according to the Regional Public Health Act Paragraph 4. It has been conducted at 251 public health centers nationwide upon the approval of Statistics Korea (Approval No-11775) since 2008.[17] The complex sample design was used as the experiment design of the study. The subjects of the 2015 KCHS were adults (≥ 19 years old; who were born before Jul 31, 1995) living in the sample households at the time of the survey and the survey units were households (household survey) and members of the households (individual survey). KCHS is composed of 177 items in 18 domains encompassing sociodemographic factors (e.g., occupation and education), health behavior, oral health, mental health, health examination, and medical care service utilization. The survey was conducted from Aug 16 to Oct 31, 2015. Trained surveyors visited the sample households in person and conducted 1:1 interviews using a laptop loaded with the survey program. This study analyzed 2,421 adults (1,109 males and 1,312 females), who were equal to or older than 19 years, were main caregivers of the demented elderly (≥ 60 years), and completed the health behavior investigation and the mental health survey. The study subjects were selected from 228,558 people who completed the community health investigation. A flow chart of this study is shown in Figure 1.

2.2. Measurement of variables

Depression, a result variable, was defined using the Center for Epidemiological Studies Depression (CES-D: a Korean version).[18] CES-D is a self-reporting type depression index, which was developed by the National Institute of Mental Health. It is a primary screening tool for depression. The total score was 60 points and the score of 25 points or more was defined as a depression.

The explanatory variables of this study were age (19–39, 40–59, or 60 ≤), gender, the highest level of education (≤ elementary school, junior high school graduation, high school graduation, or college graduation and above), social activity in the past one month (yes or no), economic activity (yes or no), mean household income (less than 2 million KRW, 2–4 million KRW, or more than 4 million KRW), marital status (living together, bereavement/separation, or single), drinking (less than once a week or twice or more per week), smoking (smoker or nonsmoker), subjective health condition (good, average, or poor), number of walking (less than one day per week or two days or more per week), disease/accident/addiction experience within the past two weeks (yes or no), subjective stress (yes or no), the frequency of meeting a neighbor (less once a month or two times or more per month), and the frequency of meeting a relative (less once a month or two times or more per month).

2.3. Analysis method

Factors related to the depression of dementia caregivers were explored using the radial basis function (RBF) artificial neural network, which uses RBF as the coupling function of the hidden layer. The artificial neural network is a nonlinear model that finds a hidden pattern in data by learning the real data repetitively.[19,20] This study included the variables with a relative importance of inputs equal to or greater than 0.1 were considered as main explanatory variables influencing the decision of result variables and they were included in the C4.5 based decision tree model. The predictive power of the artificial neural network was verified using Arear under receiver operating characteristic (AUROC).

The prediction model related to the depression of the dementia caregivers were developed using the C4.5 algorithm. C4.5 can handle continuous data and incomplete data. Moreover, it can resolve the over-fitting issue using pruning.[21] There are the advantages of this algorithm.[20] The C4.5 algorithm makes decision variables for classification by using each data attribute. In this case, the decision variable is determined by using the attribute with the maximum information gain after examining the information gain of each dataset, which is the result of selecting an attribute of classified data. The information gain ratio between the labeled dataset *S* and the class attribute *X* is defined as Equation (1).

$$IGR\left(S|X\right)=\frac{IG\left(S|X\right)}{-\sum_{i}\frac{|S_i|}{S}\log_2\frac{|S_i|}{S}} \qquad (1)$$

Si is a subset of *S* and it means the value of attribute *X*, while |*S*| is the number of instances. The C4.5 algorithm predicts the class of new instance after constructing a set of rules in the learning stage.[20] In the model of this study, the splitting and merging standard (α) of the decision rule was set to 0.05. The number of parent nodes, that of child nodes, and that of branches were limited to 100, 60, and 5, respectively.

When the incident rate is low such as the prevalence of a disease, the screening of defective products, or the decision variables of this study, the unbalanced data can be generated. The unbalanced data can cause issues to the classification algorithm, such as classifying with multiple outcome categories. This study adjusted the unbalanced data by setting the weights of misclassification costs asymmetrically with considering the prevalence of depression in the South Korean elderly in order to compensate for the unbalanced issue of the data. The feasibility of C4.5 was tested using 10-fold cross-validation. In order to evaluate the performance of the developed C4.5, this study calculated the ratio of correct classification after dividing the KCHS data into training data (70%) and test data (30%). It was compared with the ratio of correct classicization estimated from logistic regression analysis.

3. Results

3.1. General characteristics of study subjects

Among the 2,421 subjects, 327 (13.5%) caregivers who took care of the demented elderly experienced depression during the

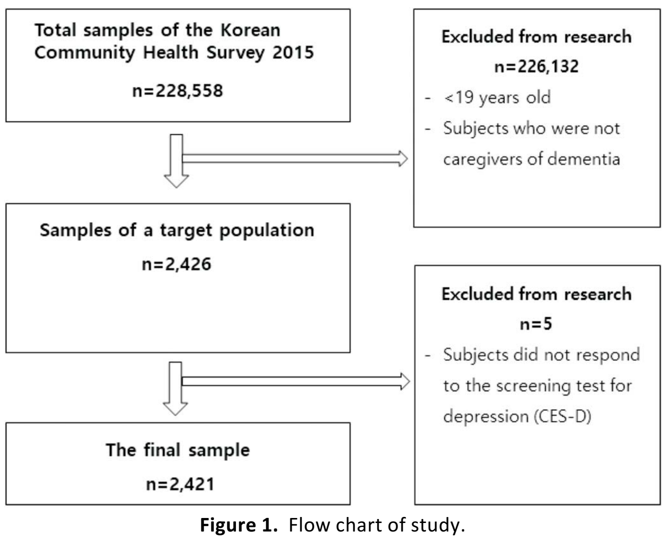

Figure 1. Flow chart of study.

past year. The results of the cross-tabulation analysis showed that caregivers who experienced depression and those who did not experience depression had significant ($p < 0.05$) differences in age, gender, the highest level of education, the social activity in the past one month, economic activity, monthly mean household income, marital status, subjective health condition, disease/accident/addition experience in the past two weeks, and subjective stress. Specifically, caregivers who were 60 years or older (16.1%), were female (16.8%), graduated from junior high graduation or below (17.4%), did not have a social activity in the past month (15.9%), had the mean monthly household income equal to or less than 2 million KRW (17.4%), were bereaved or separated of a spouse (17.3%), were in poor subjective health (21.9%), experienced a disease/accident/addiction in the past two weeks (24.4%), or experienced subjective stress (16.0%) experienced depression more frequently.

3.2. Exploring factors associated with depression using artificial neural network

An artificial neural network analysis was conducted for 1,403 training samples (59.2%), 730 test samples (30.8%) and 236 verification samples (10.0%). The results of the analysis drew ten hidden layers generating the error of test errors. The classification accuracy was 86.6, 86.8, and 86.4% for the training samples, test samples, and verification samples, respectively. Moreover, AUROC was 0.78, and the fitness and explanatory power of the classification model were found excellent.

The results of analyzing the normalized importance of the neural network model revealed that subjective health condition, subjective stress level, disease and accident experience in the past two weeks, the frequency of meeting a relative, economic activity, gender, mean monthly household income, and walking performance were important in the order of magnitude. Among them, subjective health condition had the highest normalized importance weight of the depression experience of the dementia caregivers in Figure 2.

3.3. Depression prediction model of dementia caregivers using C4.5 algorithm

This study constructed a statistical classification model using the C4.5 algorithm in Figure 3 and the model showed that subjective health condition was the most important predictive factor associ-

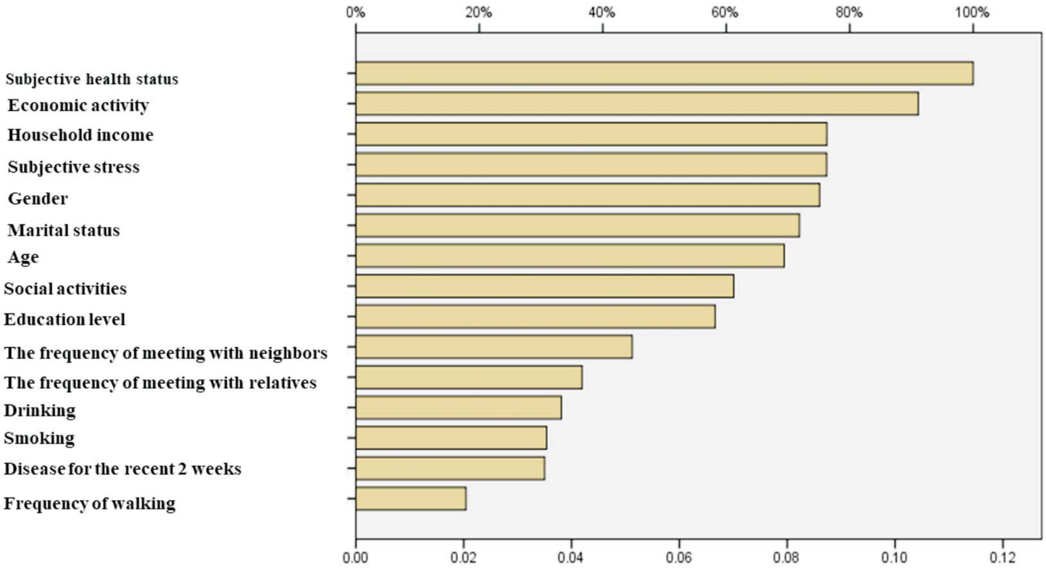

Figure 2. Normalized importance.

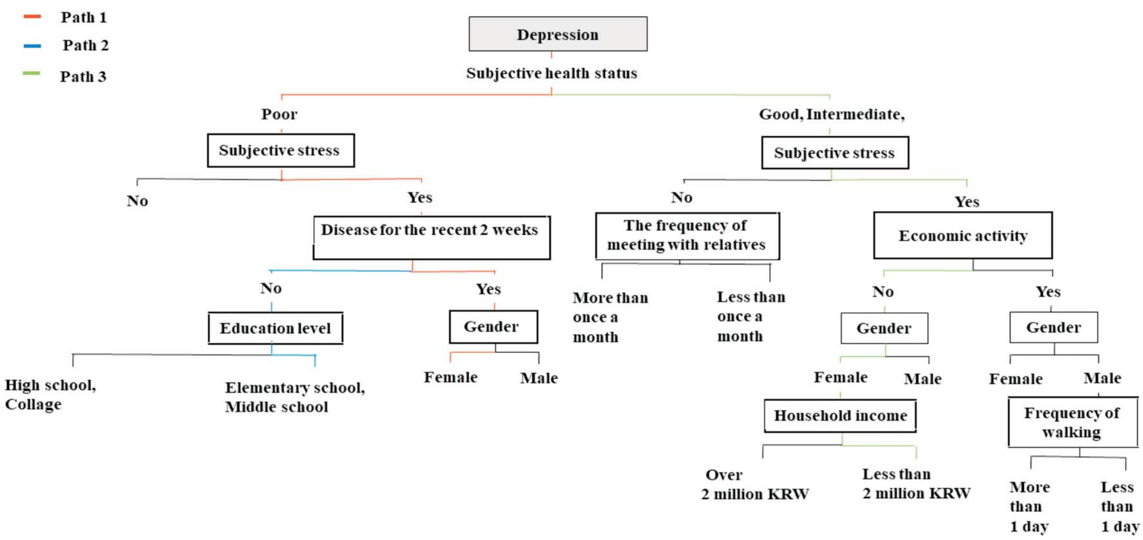

Figure 3. Depression prediction model of dementia caregivers using C4.5 algorithm.

ated with the depression of the dementia caregivers. Three nodes, out of total 11 paths, were significant paths for predicting depression (Table 1). The path 1 had the largest gain index in predicting the depression of the dementia caregivers. The path 1 was the female who experienced disease/accident/addition in the past two weeks, was aware of stress, and perceived that her subjective health was in the poor condition. The gain index of the path 1 was 275.7%. The path 2 was the person who did not experience disease/accident/addition in the past two weeks, was in a poor subjective health condition, was aware of stress, and had the highest level of education equal to or below junior high graduated. The gain index of the path 2 was 184.5%. The path 3 was the female who had an average or better subjective health condition, was aware of stress, had a mean monthly household income below 2 million KRW, and did not do an economic activity. The gain index of the path 3 was 178.4%.

The developed prediction model was evaluated using a 10-fold cross-validity test. The stability of the derived model was compared and the results showed that the risk index and misclassification rate of the cross-classification model were 0.304 and 30%, respectively.

3.4. The prediction power of the model

In order to identify the prediction power of the developed C4.5 model, the logistic regression model and the ratio of correct classification were compared (Table 2). The analysis of training data showed that C4.5 showed high classification accuracy (71.8%) and the accuracy of logistic regression model was lower (69.6%) than that. For the test data, the accuracy of C4.5 was 70.8%, while that of logistic regression analysis was 69.5%. These results revealed that C4.5 was more accurate than logistic regression model in both training and test datasets.

4. Discussion

In this study, 13.5% of community caregivers supporting patients with dementia experienced depression. The results of this study agreed with Epstein-Lubow (2009),[22] who reported that 1/3 of dementia caregivers experienced major depression symptoms and more than half of them suffered from chronic depression. Demented Elderly Factual Survey[23] conducted by the Ministry of Health and Welfare showed that 62% of caregivers for demented elderly experienced depressive disorder and 20% of them received psychiatric treatment. The probability of a family with a demented member to be affected by depression is two or three times higher than that without a demented member.[22] Pinquart and Sorensen (2003)'s meta-study also revealed that dementia caregivers had significantly higher psychological distress and stress levels and significantly lower self-efficacy and physical health level than other caregivers. Moreover, they showed that these differences were even greater than noncaregivers.[24] Therefore, social intervention is necessary to maintain the dementia caregivers mentally healthy. The developed model for predicting the depression of the dementia caregivers using C4.5 showed that major risk factors for predicting the depression of the dementia caregivers were subjective health condition, subjective stress level, disease and accident experience in the past two weeks, the frequency of meeting a relative, an economic activity, the highest level of education, monthly mean household income, and walking. Previous studies,[3,9] which evaluated the factors affecting the caregiver burden, indicated that the perceived health condition of the family caregiver was the most important factor and reported that the caregiver burden of a family caregiver increased as the health level decreased. These results agreed with the results of this study. The dementia caregivers tend to perceive their health negatively due to their psychological burden. It may adversely influence on the physical aspect as well as the emotional aspect (e.g., depression).[25] Therefore, it is necessary to develop a program for preventing depression with a priority given to the dementia caregiver who perceives a health problem subjectively.

The burden of caregivers depends on the health condition of caregivers. When spouses are caregivers, they have to care for the demented elderly while their physical functions are deteriorated due to aging. Consequently, the caregiving is physically more demanding to the spouses than others including children.[26] It has been reported that caregivers who support the demented elderly have greater physical and psychological burdens than the elderly with other chronic diseases such as hypertension or stroke.[27] Moreover, Schulz & Beach (1999)[28] reported that the mortality rate increased by 63% when caregivers were senior citizens. Moreover, Pinquart & Sörensen (2007),[26] a meta-analysis study, confirmed that depression was the most important factor affecting the physical health of caregivers. Therefore, the health status of caregivers would be as important as the burden due to the disability of the demented elderly.

The results of this study also revealed that the gender of a caregiver was a factor affecting the caregiver burden. Many previous studies showed that females felt the caregiver burden more, which could be because the female mainly took care of the demented

Table 2
The prediction power of the model.

Data	Model	Accuracy (%)
Training data	Logistic regression model	69.6
	C4.5	71.8
Test data	Logistic regression model	69.5
	C4.5	70.8

Table 1
Gain index for predictive model.

Node no.	Node n (%)[1]	Gain n (%)[2]	Response %[3]	Gain index %[4]	Characteristic
15	174 (7.2)	65 (19.8)	37.4	275.7	Subjective health condition = poor; subjective stress = yes; disease/accident/addiction experience within the past two weeks = yes; gender = female
14	328 (13.5)	82 (25.0)	25.0	184.5	Subjective health condition = poor; subjective stress = yes; disease/accident/addiction experience within the past two weeks = no; level of education = elementary school or junior high school graduation
22	120 (5.0)	29 (8.8)	24.2	178.4	Subjective health condition = good or average; subjective stress = yes; economic activity = no; gender = female; mean household income = less than 2 million KRW

[1] Node n (%); node number, % to 2,421.
[2] Gain n (%); gain number, % to 37.4.
[3] Response (%): The fraction of the depression.
[4] Gain index (%): = 275.7 in total 11 node.

elderly and the female received the informal support less than the male.[14]

In terms of gender, female caregivers had a heavier caregiver burden because they experienced aggravated roles such as housework, child-rearing and child-education, vocational activities, and spouse roles.[29] Therefore, it is important to pay attention to the role stress and role conflicts of women and help them utilize time and financial resources well. Moreover, it is critical to provide counseling programs after finding out whether there is a conflict between the caregiving for the demented elderly and other duties such as housework and child-rearing. In terms of social support, it was found that the frequency of meeting a relative was a significant factor associated with the depression of the dementia caregivers. The caregivers of the demented elderly generally experience isolation and loneliness while devoting most of their time to take care of the patients. It, as a consequence, limits the time with friends and relatives, weakens the social network, and causes depression to caregivers.[27] The results of this study implied that effective external support measures should be prepared to mitigate the caregiver burden of caregivers. It is urgent to detect the depression high-risk group as soon as possible and provide proper education and promotion, service network establishment, and regular health counseling and management for them in order to support them emotionally.

Another finding of this study was that the model, made by combining RBF artificial neural network and C4.5, was somewhat more accurate than logistic regression model. It could be because the sample size of this study (2,421 people) was large enough to provide enough training data for C4.5 model to demonstrate full capacity. It is necessary to minimize the overfitting hindering appropriate learning in order to develop an efficient artificial neural network model. Adeli & Wu (1998)[30] suggested a weight regularization model as a way to prevent overfitting in neural network models. This study combined artificial neural network with C4.5 to enhance the prediction power of the developed model while minimizing overfitting. Therefore, the developed model predicts the depression of dementia caregivers by analyzing only the variables with high importance. However, it may still have an overfitting issue, causing noise in the prediction results by studying non-major variables. Future studies are needed to explore the most optimal prediction model by using ensemble model that utilizes a weight regularization[31] method for artificial neural networks and applies bootstrap sampling to each decision tree for tree-based prediction models.

The limits of this study are as follows. First, this study could not analyze the severity of dementia. Future studies are needed to analyze the depression of dementia caregivers according to the severity of dementia by using measurement tools such as Mini–Mental State Examination. Secondly, there may be potential confounding variables for depression that are not included in this model. Thirdly, since this study is a cross-sectional study, the results withdrawn from this study cannot be interpreted as a causal relationship even if a relationship is found. Therefore, the results of this study should be interpreted carefully.

5. Conclusions

The results of this study proved that the model for predicting the depression of the dementia caregivers, which was built by combining the RBF artificial neural network and the C4.5, was a model with high explanatory power and predictive power. The results of this study suggest that it will be necessary to develop a customized program for predicting the depression of the caregivers, who take care of the demented elderly, efficiently and managing it.

Acknowledgment

This study was supported by research fund from Honam University (2017) and National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041091; NRF-2019S1A5A8034211).

Competing interests

The author(s) declare no competing interests.

Funding

This study was supported by research fund from Honam University (2017) and National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041091; NRF-2019S1A5A8034211).

Author contributions

HB conceived and designed the experiments and analyzed the low data andwrotethe paper.

References

1. Statistics Korea. *Population Projections for Korea: 2010–2060*. Daejeon, South Korea: Statistics Korea; 2011. Available at http://kostat.go.kr/portal/eng/pressReleases/8/8/index.board?bmode=read&bSeq=&aSeq=273109&pageNo=1&rowNum=10&navCount=10&currPg=&searchInfo=&sTarget=title&sTxt=. Accessed October 29, 2019.
2. Epstein-Lubow GP, Beevers CG, Bishop DS, et al. Family functioning is associated with depressive symptoms in caregivers of acute stroke survivors. *Arch Phys Med Rehabil.* 2009;90:947–955.
3. Chiao CY, Wu HS, Hsiao CY. Caregiver burden for informal caregivers of patients with dementia: A systematic review. *Int Nurs Rev.* 2015;62:340–350.
4. Hsu CC, Wang YM, Huang CR, et al. Sustained benefit of a psycho-educational training program for dementia caregivers in Taiwan. *Int J Gerontol.* 2017;11:31–35.
5. Etters L, Goodall D, Harrison BE. Caregiver burden among dementia patient caregivers: A review of the literature. *J Am Acad Nurse Pract.* 2008;20:423–428.
6. Thomas P, Lalloué F, Preux PM, et al. Dementia patients caregivers quality of life: The PIXEL study. *Int J Geriatr Psychiatry.* 2006;21:50–56.
7. van der Lee J, Bakker TJ, Duivenvoorden HJ, et al. Multivariate models of subjective caregiver burden in dementia: A systematic review. *Ageing Res Rev.* 2014;15:76–93.
8. Papastavrou E, Kalokerinou A, Papacostas SS, et al. Caring for a relative with dementia: Family caregiver burden. *J Adv Nurs.* 2007;58:446–457.
9. Schoenmakers B, Buntinx F, Delepeleire J. Factors determining the impact of care-giving on caregivers of elderly patients with dementia. A systematic literature review. *Maturitas.* 2010;66:191–200.
10. Tay KCP, Seow CCD, Xiao C, et al. Structured interviews examining the burden, coping, self-efficacy, and quality of life among family caregivers of persons with dementia in Singapore. *Dementia.* 2016;15:204–220.
11. Kim MD, Hong SC, Lee CI, et al. Caregiver burden among caregivers of Koreans with dementia. *Gerontology.* 2009;55:106–113.
12. Schulz R, Savla J, Czaja SJ, et al. The role of compassion, suffering, and intrusive thoughts in dementia caregiver depression. *Aging Ment Health.* 2017;21:997–1004.
13. Byeon H. The risk factors of laryngeal pathology in Korean adults using a decision tree model. *J Voice.* 2015;29:59–64.
14. Byeon H. A prediction model for mild cognitive impairment using random forests. *IJACSA.* 2015;6:8–12.
15. Hssina B, Merbouha A, Ezzikouri H, et al. A comparative study of decision tree ID3 and C4.5. *IJACSA, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications*. 2014:13–19.
16. Byeon H, Koh HW. A feature analysis of risk factors for stroke in the mid-

dle-aged adults focused on perception of sudden speech and language problem. *IJACSA*. 2015;6:95–99.
17. Kim YT, Choi BY, Lee KO, et al. Overview of Korean Community Health Survey. *JKMA.* 2012;55:74–83. [In Korean, English abstract]
18. Cho MJ, Kim KH. Use of the center for epidemiologic studies depression (CES-D) scale in Korea. *J Nerv Ment Dis*. 1998;186:304–310.
19. Cho S, Yu S, Byeon H. A prediction model for benign laryngeal disease using supervised learning techniques. *IJBSBT*. 2016;8:105–110.
20. Quinlan JR, *C4. 5: Programs for Machine Learning*. 1th ed. Burlington, MA: Morgan Kaufmann; 2014.
21. Schulz R, Martire LM. Family caregiving of persons with dementia: Prevalence, health effects, and support strategies. *Am J Geriatr Psychiatry.* 2004;12:240–249.
22. Epstein-Lubow G. Family caregiver health: What to do when a spouse or child needs help. *R I Med J.* 2009;92:106–107.
23. Korea Institute for Health and Social Affairs. *Analysis of the Survey of Living Conditions and Welfare Needs of Korean Older Persons*. Sejong, South Korea: Korea Institute for Health and Social Affairs; 2012. Available at http://repository.kihasa.re.kr/handle/201002/9784. Accessed October 29, 2019. [In Korean, English abstract]
24. Pinquart M, Sörensen S. Differences between caregivers and noncaregivers in psychological health and physical health: A meta-analysis. *Psychol Aging.* 2003;18:250–267.
25. Penning MJ, Wu Z. Caregiver stress and mental health: Impact of caregiving relationship and gender. *Gerontologist*. 2015;56:1102–1113.
26. Pinquart M, Sörensen S. Correlates of physical health of informal caregivers: A meta-analysis. *J Gerontol B Psychol Sci Soc Sci*. 2007;62:126–137.
27. Brodaty H, Donkin M. Family caregivers of people with dementia. *Dialogues Clin Neurosci.* 2009;11:217–228.
28. Schulz R, Beach SR. Caregiving as a risk factor for mortality: The caregiver health effects study. *JAMA.* 1999;282:2215–2219.
29. Torti FM Jr, Gwyther LP, Reed SD, et al. A multinational review of recent trends and reports in dementia caregiver burden. *Alzheimer Dis Assoc Disord.* 2004;18:99–109.
30. Adeli H, Wu M. Regularization neural network for construction cost estimation. *J Constr Eng M Asce*. 1998;124:18–24.
31. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–1958.