Original Article

Healthc Inform Res. 2014 January;20(1):30-38.
http://dx.doi.org/10.4258/hir.2014.20.1.30
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Knowledge Discovery in a Community Data Set: Malnutrition among the Elderly

Myonghwa Park, PhD[1], Hyeyoung Kim, PhD[2], Sun Kyung Kim, MSN[1]

[1]College of Nursing, Chungnam National University, Daejeon; [2]Department of Nursing, Catholic Sangji College, Andong, Korea

Objectives: The purpose of this study was to design a prediction model that explains the characteristics of elderly adults at risk of malnutrition. **Methods:** Data were obtained from a large data set, 2008 Korean Elderly Survey, in which the data of 15,146 subjects were entered. With nutritional status a target variable, the input variables included the demographic and socioeconomic status of participants. The data were analyzed by using the SPSS Clementine 12.0 program's feature selection node to select meaningful variables. **Results:** Among the C5.0, C&R Tree, QUEST, and CHAID models, the highest predictability was reported by C&R Tree with the accuracy rate of 77.1%. The presence of more than two comorbidities, living alone status, having severe difficulty in daily activities, and lower perceived economic status were identified as risk factors of malnutrition in elderly. **Conclusions:** A reliable decision support model was designed to provide accurate information regarding the characteristics of elderly individuals with malnutrition. The findings demonstrated the good feasibility of data mining when used for a large community data set and its value in assisting health professionals and local decision makers to come up with effective strategies for achieving public health goals.

Keywords: Decision Trees, Data Mining, Malnutrition, Aged, Community

I. Introduction

Data mining is a recently developed technological methodology that has been used intensively and extensively in many fields. It is defined as the process of discovering previously unknown patterns or trends from stored data, building predictive models based on that information. Compared with the past healthcare system, there has been an ever increasing amount of data generated by current healthcare settings. The amount of data generated in current healthcare settings is becoming too large and complex to be analyzed by traditional statistical methods. Due to its good applicability to any field of study, data mining has gradually gained popularity in almost all areas of healthcare [1]. However, data mining has been used mostly to analyze disease-focused clinical data rather than community data gathered for primary care providers in Korea. The importance of accessing community health status has been widely recognized throughout the nation and data mining can empower local decision makers with a clear methodology for organizing and interpreting healthcare data [2].

Submitted: October 31, 2013
Revised: January 16, 2014
Accepted: January 20, 2014

Corresponding Author
Myonghwa Park, PhD
College of Nursing, Chungnam National University, 266 Munhwa-ro, Jung-gu, Daejeon 301-747, Korea. Tel: +82-42-580-8328, Fax: +82-42-580-8309, E-mail: mhpark@cnu.ac.kr

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ⓒ 2014 The Korean Society of Medical Informatics

Data mining has been recognized as a solution to handle these large amounts of complex data, transforming these data into meaningful information for the decision making process of health professionals. A comprehensive community health profile plays an important role in the development, refining and monitoring of long-term strategies designed to improve the overall health status of the population across the nation. Such information obtained at the community level can be used to ensure that the best resource allocation choices are made to improve community health [3]. Along with Korean Elderly Survey, a number of national representative studies, such as the Korean Longitudinal Study of Ageing and the Korean Retirement and Income Study, have recently been conducted in Korea [4,5]. Currently, more healthcare data are available at the community level, enabling health professionals to make a transition from the previous disease- and treatment-focused strategy to a health- and prevention-focused strategy. This also clarifies the public health role, which should be supported by unbiased data that accurately reflect a community's health status, needs, and resource allocation [6].

The proportion of elderly population in Korea reached 11.8% in 2012 which is expected to increase further accounting for 24.3% in 2030 and 37.4% by 2050 [7]. The growing elderly population has a huge impact on the healthcare system, and malnutrition is a common health problems found frequently among the elderly [8]. The leading causes of malnutrition include several individual factors, such as poor oral health, loss of vision and hearing, dementia, impaired mobility, and pain. Malnutrition is both a cause and consequence of many geriatric diseases that accounts for a significant proportion of national medical spending [9]. Nutritional status has a great influence on the immune system of the human body; the lower immune response induced by malnutrition eventually puts elderly individuals at a high risk of infection, increasing the risk of mortality [10]. A rapid and precise diagnostic method that supports clinical decision enables the earlier identification of the elderly at risk of malnutrition. This will eventually prevent negative outcomes caused by poor nutritional status, resulting in a substantial reduction in healthcare cost spent on the elderly population [11]. Most studies to date have conducted research on small samples obtained from local hospitals or specific communities using traditional statistical analysis, which may not be sufficient for discovering comprehensive knowledge [12]. In addition, the limited number of variables analyzed in those studies could not provide integrative understanding of elderly malnutrition.

Data mining is known to be the best method in terms of the cost estimation, disease diagnosis, prediction of prognosis, and the discovery of hidden patterns in the healthcare system with the use of large health database [13]. A decision tree model presented in this study was developed from a combination of decision tree approaches and statistical analysis. Data mining provides information that can be the used in the analysis of risk factors for certain types of diseases. Contrasting the characteristics of ill patients with those of healthy people to find the patterns related to the occurrence of disease has been a common practice in evidence-based medicine [14]. The purposes of the study were to identify the best modeling method among data mining tools by testing the discriminatory power of individual models and to identify the patterns related to the occurrence of malnutrition in elderly. Finding patterns for the related characteristics of risky conditions that threaten the health of the community may contribute to the achievement of public health goals with effective disease prevention strategies. By identifying the associated characteristics of older adults who are more likely to be malnourished in community settings, the decision tree model allows health professionals and decision makers to intervene earlier to allocate adequate resource for nutritional support, improving the overall health of the community.

II. Methods

1. Subjects

A secondary analysis of a large community data set from the 2008 national study of Korean elderly was conducted to provide accurate information regarding the characteristics of the elderly at risk of malnutrition in Korea. The Institutional Review Board gave prior approval to the study. The sample consisted of 15,146 adults over 60 years old and community dwelling. The ten items of the 'Determine Your Nutritional Health Checklist' were used to access dietary intake and meal patterns. The tool was developed by the Nutritional Screening Initiative to screen nutritional risk in the elderly population. A summed score of 0–2 indicates good nutritional status, 3–5 moderate risks, and 6 or more high risk [15]. In this study, elderly people with summed scores of 3 or more were classified as being at risk of malnutrition as they pose problems for malnutrition.

2. Decision Tree Models

Figure 1 presents the data mining process that started with the selection of variables. The target variable was malnutrition, and the 52 input variables covered all aspects of elderly malnutrition, including general characteristics, family & social relationship, economic status, health status, health behavior,

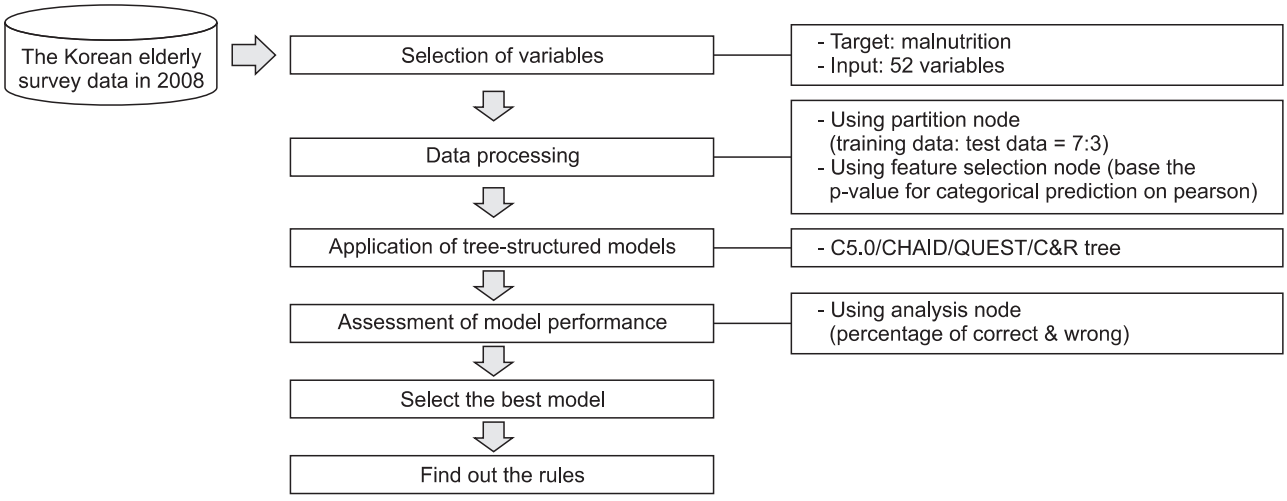

Figure 1. Process of data mining.

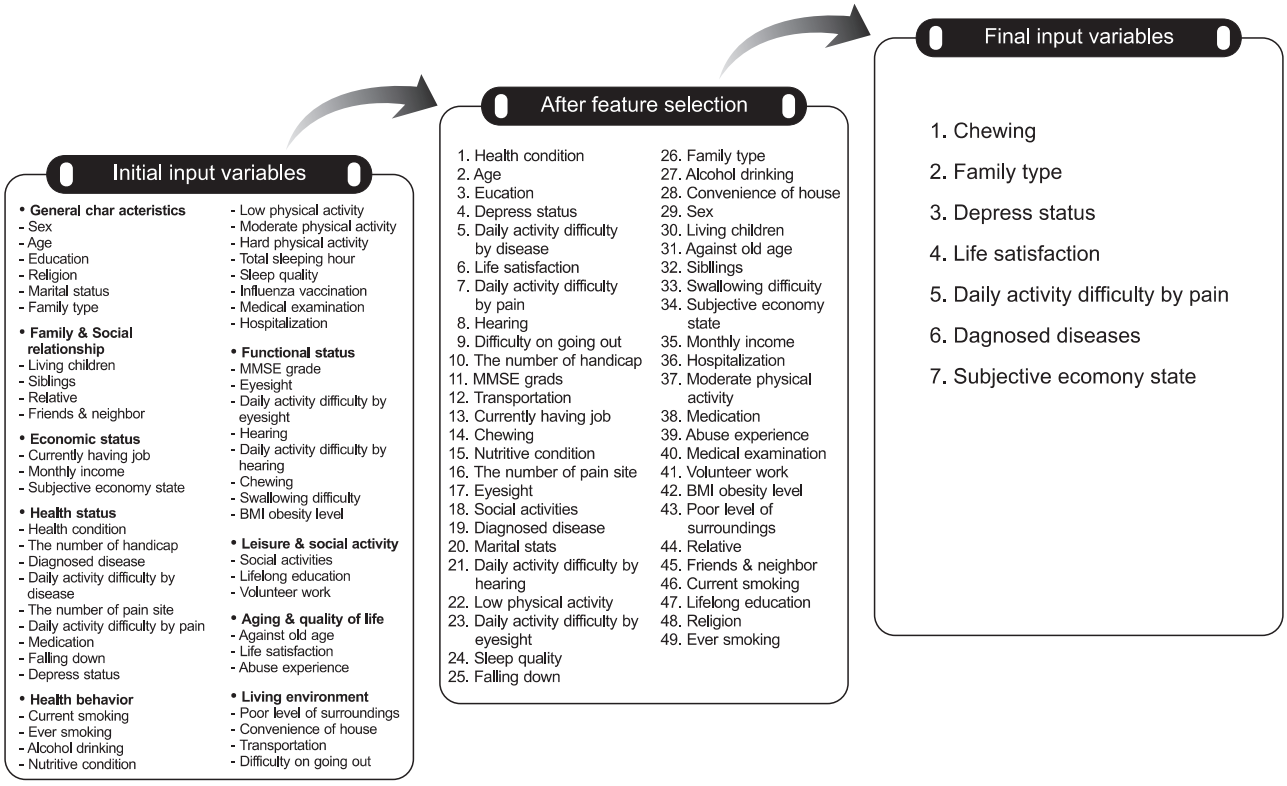

Figure 2. Input variables and feature selection.

functional status, leisure & social activity, quality of life, and living environment (Figure 2). The ratio between the training data and test data was 7:3 by random division of the full data set to which four tree-structured models, including C5.0, CHAID, QUEST, and classification and regression (C&R) Tree, were applied. In an assessment of model performance, analysis node was used to calculate the percentage of correct and wrong classification of each model, and then the correct classification rates of the models were compared to select the best model. In this study, the C&R Tree showed the best predictive performance, and it was selected to draw the final decision support tree [16]. A decision tree is a visual representation of obtained knowledge using a tree form in which each branch represents an output of the test. This decision tree enables easy understanding and interpretation of data as the nodes and branches are organized hierarchically. It generates reliable knowledge, promoting accuracy in clinical decision-making [17]. The C&R Tree node is a tree-based classification

http://dx.doi.org/10.4258/hir.2014.20.1.30

and prediction method. The C&R Tree approach provides a very simple representation that displays accumulated knowledge well. This method allows the tree to grow large before pruning, which results in smaller trees, and this ensures better cross-validation properties of decision tree modeling [18]. The recursive partitioning process splits the records into segments with similar output field values, and two subgroups are defined by the first split. The split ends when one of stopping criteria is triggered [19].

3. Statistical Analysis

The default experimental parameters of Clementine ver. 12.0 (SPSS Inc., Chicago, IL, USA) were utilized to construct the decision tree model with the usage of the C&R Tree component in this study. SPSS ver. 12.0 for Windows (SPSS Inc.) was used to perform the statistical analysis in which important variables were chosen using the feature selection method. Among Pearson, Likelihood ratio, Cramer's V, and Lambda-a, Pearson correlation analysis was selected to prioritize the meaningful variables [20].

Table 1. General characteristics

Variable	Total (n = 15,146)	Normal (n = 7,769)	Malnutrition (n = 7,377)	χ^2	*p*-value
Sex				182.49	0.001
Male	6,185 (40.8)	3,581 (57.9)	2,604 (42.1)		
Female	8,961 (59.2)	4,188 (46.7)	4,773 (53.3)		
Age (yr)				558.26	0.001
60–69	7,090 (46.8)	4,278 (60.3)	2,812 (39.7)		
70–79	6,192 (40.8)	2,790 (45.1)	3,402 (54.9)		
>80	1,864 (12.4)	701 (37.6)	1,163 (62.4)		
Education (yr)				746.08	0.001
0–6	5,908 (39.0)	3,137 (53.1)	2,771 (46.9)		
7–12	3,646 (24.1)	2,284 (62.6)	1,362 (37.4)		
>12	837 (5.5)	584 (69.8)	253 (30.2)		
Marital status				1,031.72	0.001
Single	54 (0.4)	14 (25.9)	40 (74.1)		
Married	9,380 (61.9)	5,770 (61.5)	3,610 (38.5)		
Separated	5,712 (37.7)	1,985 (34.8)	3,727 (65.2)		
Family type				1,132.29	0.001
Living alone	3,484 (23.0)	925 (26.5)	2,559 (73.5)		
Living with spouse	6,653 (43.9)	4,017 (60.4)	2,636 (39.6)		
Living with children	4,256 (28.1)	2,432 (57.1)	1,824 (42.9)		
Monthly income (1,000 US $)				1,106.11	0.001
<0.5	4,229 (27.9)	1,356 (32.1)	2,873 (67.9)		
0.5-0.9	3,877 (25.6)	1,908 (49.2)	1,969 (50.8)		
1-1.9	3,501 (23.1)	2,153 (61.5)	1,348 (38.5)		
2-2.9	1,685 (11.1)	1,100 (65.3)	585 (34.7)		
≥3	1,854 (12.2)	1,252 (67.5)	602 (32.5)		
Employment				257.36	0.001
Yes	5,204 (34.4)	3,138 (60.3)	2,066 (39.7)		
No	9,942 (65.6)	4,631 (46.6)	5,311 (53.4)		
Religion				5.19	0.392
Buddhist	5,085 (33.6)	2,648 (52.1)	2,437 (47.9)		
Protestantism	2,905 (19.2)	1,481 (51.0)	1,424 (49.0)		
Catholic	1,144 (7.6)	566 (49.5)	578 (50.5)		
No religion	5,872 (38.8)	3,010 (51.3)	2,862 (48.7)		

Values are presented as number (%).

III. Results

1. Demographic Characteristics of the Subjects

A total of 15,146 elderly subjects were included in this study, and the sample consisted of 6,185 males (40.8%) and 8,961 females (59.2%). Of the total participants, 7,769 (51.3%) had normal nutritional status while 7,377 (48.7%) had poor nutritional status. Compared with elderly subjects showing good nutritional status, the mean age of elderly subjects at risk of malnutrition was higher (Table 1). About three-fourths of older people living alone were categorized as having poor nutritional status, and there was a greater proportion of elderly subjects with good nutritional status in the higher monthly income group. Apart from 54 participants (0.4%), the majority of participants had been married at least once. While malnourished elderly subjects comprised about one third of the married group, the proportion of malnourished elderly subjects was much greater in the separated or living alone groups, accounting for 73.5% and 65.2%, respectively.

2. Selecting Important Attributes

As a result of the feature selection process, 48 important attributes were identified to be important variables of a total of 52 input variables shown in Figure 2. Seven variables including chewing ability, level of life satisfaction, depression status, health status, number of diagnosed diseases, difficulties in daily activities caused by pain, subjective economic state, and monthly income were identified to be most significant variables.

3. Comparing Predictability of the Models using Analysis Node

The highest percentage of correct prediction in training data was reported in C5.0; however, the C&R Tree model showed relatively higher percentages of correct prediction in both training and test data with 78.10% and 80.95%, respectively (Table 2).

4. Modeling of the Data using C&R Tree

Figure 2 presents seven finalized input variables that were selected during feature selection. The final decision support model was designed using the C&R Tree algorithm and is shown in Figure 3.

In this study, a total of six rules were identified to be associated with the risk of malnutrition in the elderly. The six decision rules were the following: 1) good and very good chewing ability, depressed, level of life satisfaction less than 2.89; 2) good and very good chewing ability, depressed, level of life satisfaction above 2.89, severe difficulty in daily activity caused by pain; 3) low chewing ability, living with spouse or children; low chewing ability, living with spouse or children, depressed; 5) low chewing ability, living with spouse or children, not depressed, number of disease above 2.5; 6) low chewing ability, living with spouse or children, not depressed, number of disease less than 2.5, subjective income status is poor.

5. Variables in the Final Modeling

The traditional statistical method and chi-square test were used to find out whether there are significant differences between the elderly at risk of malnutrition and normal subjects in the seven final variables identified by the C&R Tree algorithm. All seven variables showed statistical significance with a p-value of <0.05 (Table 3). Chewing ability, in particular, showed the biggest differences between well-nourished participants and malnourished participants. The majority of participants with low to moderate chewing ability had poor nutrition status (84.1% and 70.2%, respectively) while eight out of ten participants with either good or very good chewing ability were well-nourished. In addition higher proportions of individuals with depression, low life satisfaction, many difficulties in daily activities and poorly perceived economic status were found to be at risk of malnutrition.

IV. Discussion

This study explored the applicability of a prediction model using data mining of a large community data set. To improve the accuracy of knowledge regarding the characteristics of the elderly at risk of malnutrition, we identified significant rules during the process of data mining. Elderly subjects at risk of malnutrition were compared to normal and well-nourished elderly subjects to find the patterns associated with the occurrence of malnutrition in the elderly population. A tree-structured decision model was designed, and its potential application to large amounts of public health data was examined. Malnutrition is a commonly reported condition among the elderly, and it is considered to be both a cause and consequence of many age-related diseases. There has been growing concern over poor nutrition status in Korea as we are becoming an aging population. Therefore, this study was carried out to find patterns related to malnutrition in the elderly. Using the C&R Tree model, a well-designed prediction model was developed, which showed good performance in finding associated rules [18]. The C&R Tree uses a recursive partitioning method that provides a very simple representation that displays accumulated knowledge with a well-organized structure. To predict continuous

http://dx.doi.org/10.4258/hir.2014.20.1.30

Table 2. Predictive performance according to modelling methods

Modeling method	Training data		Test data	
	Correct (%)	Wrong (%)	Correct (%)	Wrong (%)
C5.0	85.68	14.32	80.68	19.32
C&R Tree	78.10	21.90	80.95	19.05
QUEST	77.25	22.75	78.34	21.66
CHAID	78.59	21.41	79.29	20.71

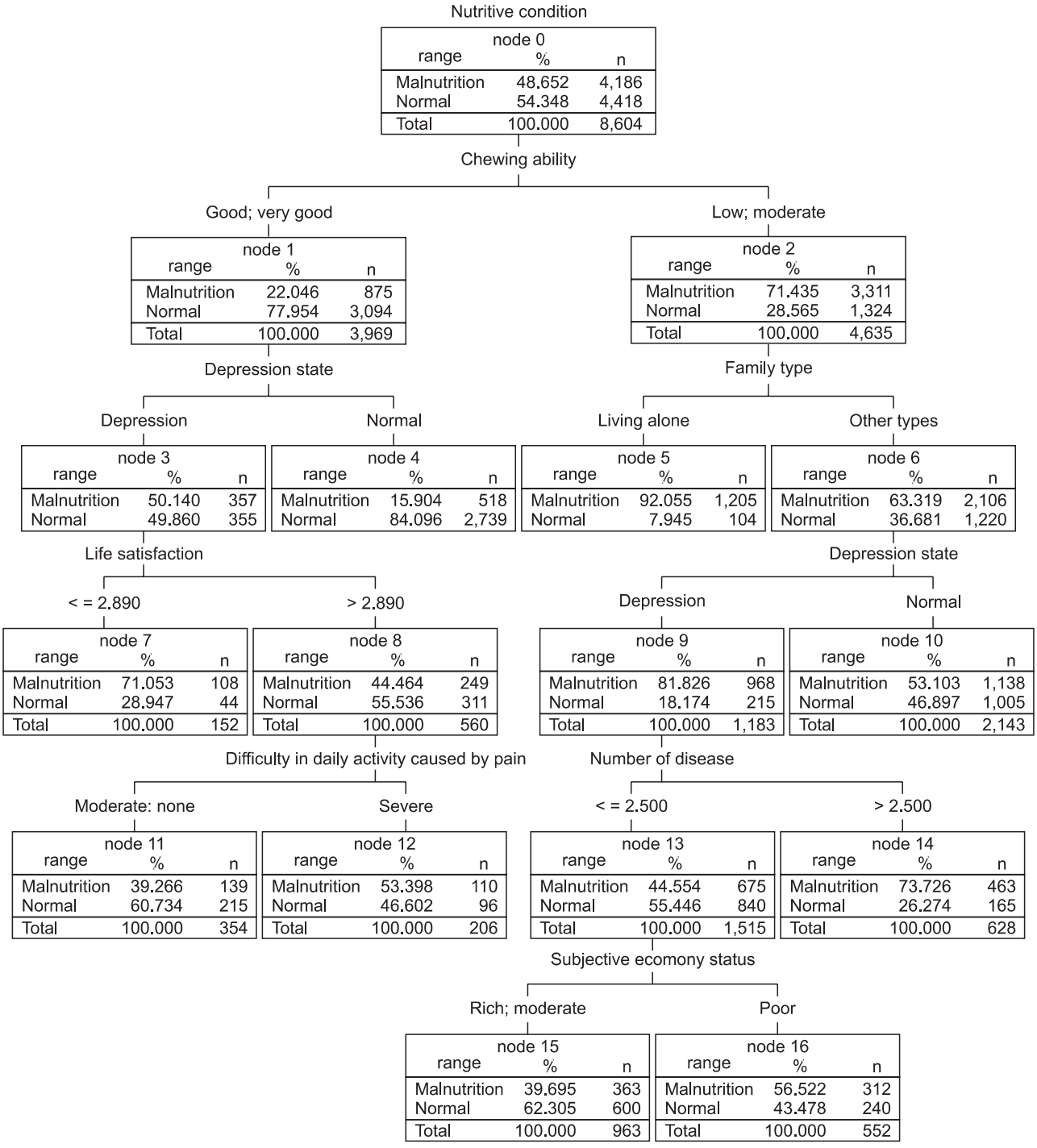

Figure 3. Decision tree model based on C&R Tree.

dependent variables (regression) and categorical predictor variables (classification), the C&R Tree builds a classification and regression tree [15]. Compared with other classification technique used for classification or regression of problems, there are many benefits that can be obtained using the C&R Tree [21]. The simplicity of the C&R Tree enables clinicians to make rapid classification of new clinical observations. Moreover, when there is a little prior knowledge, tree methods are known to be well suited for data mining tasks for data from healthcare settings [22].

In this study, seven variables and six rules for the prediction of malnutrition risk in elderly subjects were identified. Those associated rules provide information regarding co-occurrence and relationships between risk factors that may increase the probability of elderly suffering from malnutrition. According to the rules identified by the decision tree model, elderly people living alone with low chewing ability, depression, low life satisfaction, comorbidity, functional limitation and subjective economic issues are more likely to be malnourished. Findings are consistent with previous studies conducted to identify patient factors associated with the risk of malnutrition that was strongly associated with age-related changes of older adults. Elderly people with swallowing difficulties due to oral problems, such as absence of teeth, poor oral hygiene, and gum infections, have a decreased food intake and are vulnerable to malnutrition [23]. In addition, several studies found that along with functional changes, social changes in elderly people's lives (loss, changed family structure, and decreased income) may result in depression and reduced life satisfaction, increasing the risk of being malnourished [24].

The final results of the tree method demonstrated the usability of large public health data sets with good feasibility of decision model in the classification of elderly with malnutrition. Through repeated testing and refinement of data mining and the C&R Tree in particular, it is anticipated that new knowledge will be discovered by more sophisticated analysis of healthcare data at the community level [13]. As a result of

Table 3. Statistical analysis to variables used final modeling

Variable	Total (n = 15,146)	Normal (n = 7,769)	Malnutrition (n = 7,377)	χ^2 or t	p-value
Chewing ability				3,507.49	0.001
Low	390 (2.6)	62 (15.9)	328 (84.1)		
Moderate	7,770 (51.3)	2,315 (29.8)	5,455 (70.2)		
Good	6,586 (43.5)	5,080 (77.1)	1,506 (22.9)		
Very good	400 (2.6)	312 (78.0)	88 (22.0)		
Family type				1,132.29	0.001
Living alone	3,484 (23.0)	925 (26.5)	2,559 (73.5)		
Couple	6,653 (43.9)	4,017 (60.4)	2,636 (39.6)		
With children	4,256 (28.1)	2,432 (57.1)	1,824 (42.9)		
Others	753 (5.0)	395 (52.5)	358 (47.5)		
Depression status				1,927.58	0.001
Depression	4,423 (29.5)	1,051 (23.8)	3,372 (76.2)		
Normal	10,547 (70.5)	6,652 (63.1)	3,895 (36.9)		
Life satisfaction	3.42 ± 0.45	3.58 ± 0.37	3.26 ± 0.46	-38.08	0.001
Daily activity difficulty				1,054.79	0.001
None	839 (6.8)	561 (66.9)	278 (33.1)		
Little	7,483 (60.4)	4,120 (55.1)	3,363 (44.9)		
A lot	4,064 (32.8)	1,052 (25.9)	3,012 (74.1)		
Diagnosed disease	2.00 ± 1.54	1.55 ± 1.30	2.47 ± 1.62	46.38	0.001
Subjective economy state				1,320.78	0.001
Poor	6,949 (45.9)	2,464 (35.5)	4,485 (64.5)		
Moderate	7,401 (48.9)	4,748 (64.2)	2,653 (35.8)		
Rich	755 (5.0)	543 (71.9)	212 (28.1)		
No answer	41 (0.3)	14 (34.1)	27 (65.9)		

Values are presented as mean ± standard deviation or number (%).

http://dx.doi.org/10.4258/hir.2014.20.1.30

this study, a reliable decision support model was designed that provides accurate information regarding the characteristics of the elderly with malnutrition. The algorithm used to construct the decision tree showed high accuracy, and it is expected to facilitate the discovery of discriminatory knowledge for the targeted problem. The C&R Tree, which was based on the C&R Tree method, provided excellent discrimination of the characteristics associated with malnutrition in the elderly. This decision tree can be utilized to identify community residing elderly individuals who are at high risk of malnutrition; this will eventually contribute to significantly reducing healthcare costs spent on treating malnutrition and its complications in the elderly.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea grant funded by the Korean government (No. 2010-0024922).

References

1. Abbott PA. Knowledge discovery in large data sets: a primer for data mining applications in health care. In: Ball MJ, Hannah KJ, Newbold SK, Douglas JV, editors. Nursing informatics: where caring and technology meet. New York: Springer; 2000. p. 139-48.
2. Studnicki J, Hevner AR, Berndt DJ, Luther SL. Comparing alternative methods for composing community peer groups: a data warehouse application. J Public Health Manag Pract 2001;7(6):87-95.
3. Berndt DJ, Hevner AR, Studnicki J. Data warehouse dissemination strategies for community health assessments. Upgrade 2001;2(1):48-54.
4. Jang SN, Cho SI, Chang J, Boo K, Shin HG, Lee H, et al. Employment status and depressive symptoms in Koreans: results from a baseline survey of the Korean Longitudinal Study of Aging. J Gerontol B Psychol Sci Soc Sci 2009;64(5):677-83.
5. Kim HS. Income in old ages and role of children. Proceedings of the 1st Korean Retirement and Income Study (KReIS) Conference; 2008 Jun 24; Seoul, Korea.
6. Cropper S. Collaborative working and the issue of sustainability. In: Huxham C, editor. Creating collaborative advantage. London: SAGE Publications; 1996. p.80-100.
7. Korea Ministry for Health and Welfare. 2008 Korean Elderly Survey. Seoul, Korea: Ministry for Health and Welfare; 2009.
8. Alzheimer's Society. Food for thought. London: Alzheimer's Society; 2012.
9. Fitzpatrick J. Oral health care needs of dependent older people: responsibilities of nurses and care staff. J Adv Nurs 2000;32(6):1325-32.
10. National Collaborating Centre for Acute Care. Nutrition support for adults: oral nutrition support, enteral tube feeding and parenteral nutrition: methods, evidence & guidance. London: National Collaborating Centre for Acute Care; 2006.
11. Jones JM. The methodology of nutritional screening and assessment tools. J Hum Nutr Diet 2002;15(1):59-71.
12. Park M, Choi S, Shin AM, Koo CH. Analysis of the characteristics of the older adults with depression using data mining decision tree analysis. J Korean Acad Nurs 2013;43(1):1-10.
13. Ahn SY. ADL, IADL and cognition of elders living alone. J Korean Gerontol Nurs Soc 2007;9(1):68-75.
14. Li J, Fu AW, Fahey P. Efficient discovery of risk patterns in medical data. Artif Intell Med 2009;45(1):77-89.
15. Barrocas A, White JV, Gomez C, Smithwick L. Assessing health status in the elderly: the nutrition screening initiative. J Health Care Poor Underserved 1996;7(3):210-8.
16. Fayyad UM, Piatetsky-Shapiro G, Smyth P. Advances in knowledge discovery and data mining. Menlo Park (CA): AAAI Press; 1996.
17. Austin PC, Tu JV, Lee DS. Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. J Clin Epidemiol 2010;63(10):1145-55.
18. Garcia S, Fernandez A, Herrera F. Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. Appl Soft Comput 2009;9(4):1304-14.
19. Hallick JN. Analytics and the data warehouse. Health Manag Technol 2001;22(6):24-5.
20. Huh J, Jeong KS, Huh SH, Choi HK. Clementine 7 manual. Seoul, Korea: Data Solution; 2003.
21. Huh MH, Lee YG. Data mining modeling and case. 2nd ed. Seoul: Hannarae; 2008.
22. Koh HC, Leong SK. Data mining applications in the context of casemix. Ann Acad Med Singapore 2001;30(4 Suppl):41-9.
23. Sheiham A, Steele J. Does the condition of the mouth

and teeth affect the ability to eat certain foods, nutrient and dietary intake and nutritional status amongst older people? Public Health Nutr 2001;4(3):797-803.
24. Vanderwee K, Clays E, Bocquaert I, Gobert M, Folens B, Defloor T. Malnutrition and associated factors in elderly hospital patients: a Belgian cross-sectional, multi-centre study. Clin Nutr 2010;29(4):469-76.

http://dx.doi.org/10.4258/hir.2014.20.1.30