Original Article

Healthc Inform Res. 2014 April;20(2):125-134.
http://dx.doi.org/10.4258/hir.2014.20.2.125
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Prediction Model for Health-Related Quality of Life of Elderly with Chronic Diseases using Machine Learning Techniques

Soo-Kyoung Lee, PhD[1], Youn-Jung Son, PhD[2], Jeongeun Kim, PhD[3], Hong-Gee Kim, PhD[4], Jae-Il Lee, PhD[5], Bo-Yeong Kang, PhD[6], Hyeon-Sung Cho, MS[7], Sungin Lee, MS[4]

[1]College of Nursing, Keimyung University, Daegu; [2]Department of Nursing, Soonchunhyang University, Cheonan; [3]College of Nursing, [4]Biomedical Knowledge Engineering Lab., [5]School of Dentistry, Seoul National University, Seoul; [6]School of Mechanical Engineering, Kyungpook National University, Daegu; [7]Electronics and Telecommunications Research Institute, Daejeon, Korea

Objectives: The purposes of this study were to identify the factors that affect the health-related quality of life (HRQoL) of the elderly with chronic diseases and to subsequently develop from such factors a prediction model to help identify HRQoL risk groups that require intervention. **Methods:** We analyzed a set of secondary data regarding 716 individuals extracted from the Korea National Health and Nutrition Examination Survey from 2008 to 2010. The statistical package of SPSS and MATLAB were used for data analysis and development of the prediction model. The algorithms used in the study were the following: stepwise logistic regression (SLR) analysis and machine learning (ML) techniques, such as decision tree, random forest, and support vector machine methods. **Results:** Five factors with statistical significance were identified for HRQoL in the elderly with chronic diseases: 'monthly income', 'diagnosis of chronic disease', 'depression', 'discomfort', and 'perceived health status.' The SLR analysis showed the best performance with accuracy = 0.93 and F-score = 0.49. The results of this study provide essential materials that will help formulate personalized health management strategies and develop interventions programs towards the improvement of the HRQoL for elderly people with chronic diseases. **Conclusions:** Our study is, to our best knowledge, the first attempt to identify the influencing factors and to apply prediction models for the HRQoL of the elderly with chronic diseases by using ML techniques as an alternative and complement to the traditional statistical approaches.

Keywords: Quality of Life, Aged, Chronic Disease, Artificial Intelligence

Submitted: December 19, 2013
Revised: January 17, 2014
Accepted: January 22, 2014

Corresponding Author
Youn-Jung Son, PhD
Department of Nursing, Soonchunhyang Univeristy, 31 Soonchunhyang 6-gil, Dongnam-gu, Cheonan 330-930, Korea. Tel: +82-41-570-2487, Fax: +82-41-575-9347, E-mail: yjson@sch.ac.kr

I. Introduction

Aging degrades the health-related quality of life (HRQoL) for the elderly with chronic diseases. The HRQoL is a broad, multidimensional concept covering significant domains of daily functioning and subjective experience, such as physical functioning, social role functioning, somatic sensation, and subjective wellbeing [1]. Multidimensional analysis of the HRQoL requires considerable effort and expertise, demanding the development of more sophisticated ways to facilitate such complex, preferably automatic analysis.

Previous studies on the HRQoL have limited their focus to specific demographic groups, such as adults [2,3], women [4], and vulnerable aged men [5] or to specific diseases, such as diabetes mellitus [6], stroke [7], and cerebral palsy [8]. Previous studies have mainly used regression analysis and structural equation modeling [9] methods, with decision tree [10] as the only data mining method.

In this study, we used screening materials from a large-scale, comprehensive national survey with a prolonged analysis period, including many chronic diseases of diverse types, and influencing factors for the elderly with chronic diseases. This study used these resources to develop a HRQoL prediction model to identify HRQoL risk groups who require intervention.

Health and medical data are exponentially increasing, necessitating various means to take advantage of huge amounts of data. Big data technologies enable the fast processing of massive amounts of data [11]. Among these technologies, artificial intelligence has regained prominence as an important tool to provide intelligent services for big data, and machine learning (ML) techniques have also been used extensively for such purposes [12].

Clinical nursing datasets are now being generated in an increasing number of healthcare settings, and nurse researchers have been gradually adopting large datasets for studies on nursing quality and the effectiveness of nursing intervention [13]. Though data mining has been used more widely in the business world than in nursing and healthcare, it can be an important tool for the development of healthcare knowledge and knowledge structures. It can also potentially improve the quality of decision-making by clinicians and healthcare

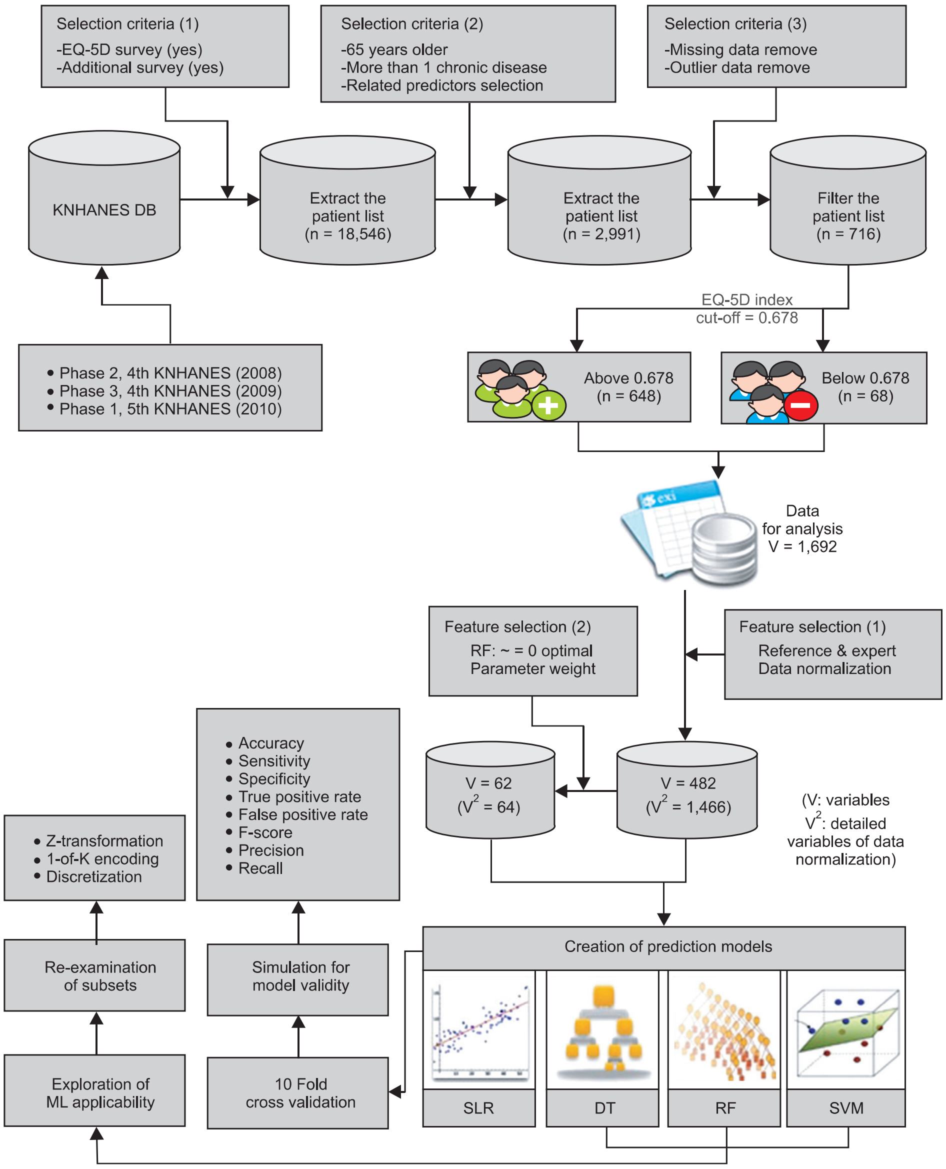

Figure 1. Procedure for extracting data and analysis. KNHANES DB: Korea National Health and Nutrition Examination Survey database, ML: machine learning, SLR: stepwise logistic regression, DT: decision tree, RF: random forest, SVM: support vector machine.

http://dx.doi.org/10.4258/hir.2014.20.2.125

administrators.

The purposes of this study were to identify the factors affecting the HRQoL of the elderly with chronic diseases and to develop a HRQoL prediction model that expressly exposes HRQoL risk groups who require intervention. ML techniques were applied and examined for their use in the analysis and prediction of HRQoL for the elderly with chronic diseases. Problems and their solutions in using ML techniques are presented. This study provides vital insights that can be incorporated into the design of personalized health management strategies and the development of interventions program, specifically for HRQoL improvement for the elderly with chronic diseases.

II. Methods

The data extraction and analysis process of this study are shown in Figure 1.

1. Data Source and Preparation

Study samples were extracted from the Korea National Health and Nutrition Examination Survey (KNHANES), which was conducted from 2008 to 2010. This survey, sponsored by the Korea Centers for Disease Control and Prevention, consisted of health interviews, screening, and nutrition examination surveys. Approximately 10,000 people were randomly selected annually. There were 2,991 elderly people (65 years of age and over) with chronic diseases out of the 18,546 cases in total. A set of 716 cases was chosen for our study after elimination of those with missing values, including outliers.

The survey used the EQ-5D as a HRQoL measurement tool, developed by the EuroQol Group. The EQ-5D comprises five health state dimensions (mobility, self-care, usual activity, pain/discomfort, and anxiety/depression) regarding which the respondent is asked to indicate a health state from one of three levels: no problem, some problems, and major problems [14]. A higher EQ-5D index score (range, −0.171 to 1) indicates a better HRQoL.

The HRQoL levels of the elderly with chronic diseases should inform both the content and regimen of healthcare programs. From the vantage point of cost and efficiency in healthcare policy, it is advisable to categorize HRQoL scales in ways that the most vulnerable group receives top priority in intervention. To identify the risk group, the cut-off value was set to the EQ-5D index score of 0.678, which is equivalent to the value when all of the five EQ-5D questions receive the response of 'some problems'. Based on the cut-off value, our study cases were divided into two groups.

The study cases were limited to the chronic diseases included in KHANES: hypertension, hyperlipidemia, stroke, angina pectoris, myocardial infarction, diabetes, thyroid disease, kidney failure, osteoarthritis, rheumatoid arthritis, osteoporosis, tuberculosis, asthma, gastric cancer, hepatoma, colorectal cancer, breast cancer, cervical cancer, lung cancer, other cancers, hepatitis B, hepatitis C, and liver cirrhosis. They were also limited to those who, in the morbidity question of the survey, answered 'yes' to past illnesses, and 'yes' to having been diagnosed by doctors, and who were diagnosed with diseases more than one year ago. In the data preparation stage, both the number of chronic diseases and the total duration of the diseases for each case were calculated. The duration of a disease was calculated by subtracting the current age of the case from the year when he/she was diagnosed with the disease.

Data normalization is a process in which data are transformed in ways that ensure consistency, minimal redundancy, and maximal stability of data, without data loss or unnecessary information added to the original data. For normalization, Z-transformation [15], 1-of-K encoding scheme coding [16], discretization [17,18], and new variable generation methods were used.

2. Data Analysis

The Statistical Package for the Social Sciences (SPSS) ver. 20.0 (IBM, Armonk, NY, USA) and Matrix Laboratory (MATLAB) ver. 7.14, R2012a (MathWorks, Natick, MA, USA) were used for data analysis. MATLAB is a high-level technical computing language and interactive environment for algorithm development, data analysis, visualization, and numerical computation.

We carried out descriptive statistics, χ^2-test, and stepwise logistic regression (SLR) analysis, with the significance level of $p < 0.05$. Decision tree (DT), random forest (RF), and support vector machine (SVM) algorithms were used in MATLAB.

The HRQoL was measured by EQ-5D. Using an EQ-5D score of 0.678 as a threshold value, our cases were grouped into two groups: one (n = 648) with an EQ-5D score equal to or greater than 0.678 and the other (n = 68) with a score lower than 0.678.

3. Development of Prediction Models

We built a DT model by running the 'ClassificationTree.fit' function in MATLAB on 482 variables as input data, and the 'COST' function was applied for feature weights.

RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The 'TreeBagger' function in MATLAB was used to build our RF models using 482 variables with the 'COST' function applied for feature weights.

SVM tries to model input variables by finding the separating boundary called the 'hyperplane' to achieve classification of the input variables. The 'svmparam' function in MATLAB was used to build our SVM model with the radial basis function kernel applied as its classification method. Against the 64 variables identified as important from the RF model, we assigned various values to c (cost) and γ (gamma) to measure F-scores.

SPSS was used to build our SLR model, with the *p*-value <0.05. At the 8th step of variable selection, a statistically significant result was obtained ($p = 0.000$). In terms of a goodness-of-fit measure, the regression model had a Cox & Snell R^2 value of 0.200, and a Nagelkerke R^2 value of 0.429.

III. Results

1. General Characteristics of Participants

The mean age of the participants was 70.6 years. There were 379 males (52.9%) and 337 females (47.1%). There were 136 cases (19.0%) who belonged to the top 25% in income levels and 221 cases (30.9%) who belonged to the lowest 25%. There were 420 cases (58.7%) who had elementary education or lower. There were 536 cases (74.9%) who lived with their married spouses and 157 cases (21.9%) who were widowed. There were 451 cases (63.0%) who were unemployed and 146 cases (20.4%) who were engaged in agriculture or fishing (Table 1).

2. Differences in the Level of HRQoL according to General Characteristics

Taking an average EQ-5D score for each variable, the following variables exhibited higher average EQ-5D scores when the values of the variables were male, higher education (education), higher income (monthly income), living with spouses (spouse), lower number (chronic disease number), no depression (depression), fewer days (discomfort days), fewer days (sick days), or better (perceived health status). There was no significant difference between the risk group and the non-risk group for age, job, family, chronic diseases total duration, body mass index (BMI), and drinking variables. Table 2 shows differences in levels of the HRQoL in light of the general characteristics of the subjects.

3. Influencing Factors of HRQoL

SLR analysis and three ML techniques, DT, RF, and SVM, were used to identify the factors affecting the HRQoL of the elderly with chronic diseases. The DT analysis identified 26 variables of importance with reason for activity restriction (obesity) being the root node. By the RF analysis, important variables could be identified by variable weights of importance, and optimal parameter weights (OPW) were used to find 64 variables with OPW values being not equal to zero. There were 32 variables, identified by SVM. In the end, the number of variables to be used as input data for our study amounted to 74, and they were gathered through literature review combined with the results of the three ML techniques. Out of the 74 variables, the SLR analysis identified 8 variables as significant. Table 3 shows some of the 74 variables organized by factors that affected the HRQoL of the elderly with chronic diseases.

Table 1. Descriptive statistics for general characteristics (n = 716)

Variable	Value
Age (yr)	70.64 ± 4.69
Gender	
Male	379 (52.9)
Female	337 (47.1)
Income	
Low	136 (19.0)
Medium-low	174 (24.3)
Medium-high	185 (25.8)
High	221 (30.9)
Education level	
Below elementary	420 (58.7)
Middle school graduate	118 (16.5)
High school graduate	109 (15.2)
Above college graduate	69 (9.6)
Marital status	
Spouse (attached)	536 (74.9)
Spouse (separated)	5 (0.7)
Widowed	157 (21.9)
Divorce	13 (1.8)
Other (non-applicable)	5 (0.7)
Job	
Managers, professionals	6 (0.8)
White-collar	5 (0.7)
Service, sales	31 (4.3)
Agriculture, fishing	146 (20.4)
Functions, device	20 (2.8)
Simple labor	57 (8.0)
Unemployed	451 (63.0)

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

Table 2. Differences of health-related quality of life by general characteristics

Variable	Mean ± SD	EQ-5D group		χ^2	*p*-value[a]
		Risk	Non-risk		
Age (yr)				1.70	0.427
Former elderly (60s)	0.90 ± 0.10	34	322		
Middle elderly (70s)	0.90 ± 0.10	27	285		
Latter elderly (80s)	0.87 ± 0.11	7	41		
Gender				4.17	0.041
Male	0.93 ± 0.10	28	351		
Female	0.87 ± 1.00	40	297		
Education level				20.17	<0.001
Below elementary	0.88 ± 0.10	57	363		
Middle school graduate	0.91 ± 0.10	5	113		
High school graduate	0.93 ± 0.10	5	104		
Above college graduate	0.95 ± 0.09	1	68		
Monthly income ($)				23.31	<0.001
0–483	0.88 ± 0.11	35	180		
493–822	0.89 ± 0.10	16	129		
831–1,933	0.92 ± 0.09	13	174		
≥1,943	0.91 ± 0.10	4	165		
Spouse				4.12	0.042
Attached	0.88 ± 0.15	44	492		
Separated	0.82 ± 0.16	24	156		
Number of chronic diseases				11.98	0.017
1	0.92 ± 0.10	17	273		
2	0.90 ± 1.00	29	213		
3	0.87 ± 0.10	16	102		
4	0.85 ± 0.10	6	36		
5–8	0.86 ± 0.11	0	23		
Depression				32.96	<0.001
Yes	0.87 ± 0.10	30	102		
No	0.91 ± 0.10	38	546		
Discomfort day				61.82	<0.001
0	0.93 ± 0.09	11	387		
1–7	0.88 ± 0.10	7	82		
8–14	0.84 ± 0.10	50	179		
Sick day				74.17	<0.001
0	0.88 ± 0.13	43	586		
1–7	0.79 ± 0.20	13	53		
8–14	0.71 ± 0.19	2	4		
15–30	0.54 ± 0.21	10	5		
Perceived health status				105.7	<0.001
Very good	0.96 ± 0.08	1	30		
Good	0.93 ± 0.09	5	208		
Fair	0.92 ± 0.09	3	176		
Poor	0.85 ± 0.99	30	188		
Very poor	0.84 ± 0.10	29	46		

[a] Fisher exact test result.

Table 3. Important variables of four models

Factor	Variable	DT	RF	SVM	SLR
Characteristics of the individual	Age	O			
	Education level		O	O	
Characteristics of environment	Monthly income	O	O	O	O
	Job		O	O	
Physiological factor	Number of chronic diseases		O	O	
	Duration of chronic diseases		O		
	Duration of rheumatoid arthritis	O	O	O	
	Diagnostic time of hypertension		O	O	
	Duration of osteoarthritis	O	O	O	
	Prevalence of osteoporosis	O	O	O	
	Diagnostic time of stroke	O	O	O	
	Duration of diabetes		O	O	
	Diagnostic time of lung cancer	O			
	Treatment of myocardial infarction, angina		O	O	O
	Sick days (past 1 month)		O	O	O
	T-score of femoral bone		O	O	
	Insulin (blood test)	O	O		
	Waist circumference	O	O	O	
	Body mass index		O	O	
	Fasting hours	O	O	O	
	Number of permanent teeth caries experience	O	O		
	Mandibular prosthesis status		O	O	
Symptom experience	Discomfort day (past 2 weeks)	O	O	O	O
	Continuous depression (for more than 2 weeks)	O	O	O	O
	Having thought of suicide (within 1 year)	O			
	Duration of rhinitis (wk)	O	O		
	Restriction of activity (old-age)		O	O	O
	Restriction of activity (obesity)	O	O		
General health perceptions	Perceived health status	O	O	O	O
Health promoting behaviors	Walking day (per a week)		O	O	
	Duration of moderate physical activity		O	O	
	Smokers cigarettes (per a day)	O	O		
	Reason of outpatient utilization	O	O	O	O

DT: decision tree, RF: random forest, SVM: support vector machine, SLR: stepwise logistic regression.

By applying the SLR and complementary ML techniques, additional variables were identified that helped predict the HRQoL of the elderly with chronic diseases: duration of osteoarthritis, prevalence of osteoporosis, diagnostic time of stroke, mandibular prosthesis status, having thought of suicide, and others. However, these variables, though affecting the HRQoL, were excluded from the current study because they could not be validated statistically.

Univariate and multivariate analyzes were performed on the variables identified from the ML analyses. The variables used in the multivariate analysis did not show any statistical significance; however, in the univariate analysis, duration of osteoarthritis, prevalence of osteoporosis, diagnostic time of stroke, mandibular prosthesis status, and having thought of suicide were significantly different between the two groups. In other words, the elderly people with chronic diseases had lower HRQoL scores when they had longer periods of osteoarthritis, suffered from osteoporosis, were diagnosed with stroke at a later stage of their life, had thought of suicide within one year, and had extensive mandibular prosthesis.

4. Resulting Prediction Models

In the resulting DT model, the diagnostic time of stroke was the dividing factor for the HRQoL for those subjects for which the restriction of activity was not due to obesity and the duration of osteoarthritis was less than 2.5 years. Also, the prevalence of osteoporosis was the dividing factor for those subjects for which the restriction of activity was not due to obesity and the duration of rheumatoid arthritis was more than 2.5 years. On the other hand, when the restriction of activity was obesity, if the discomfort day (recent 2 weeks) was less than 5.5 days, the diagnostic time of osteoarthritis was an important variable, or else the perceived health status was a variable of significance.

In the resulting RF model, the number of trees built was increased from 1 to 200 consecutively to measure the performance of each tree. Four trees showed the highest F-score of 0.346.

In our SVM model, the best performance was achieved with an F-score of 0.507 when c = 0.819, γ = 0.3012, the number of used variables = 50, and the feature weight (w) = 10.

In our SLR model, 8 variables were found to be significant as shown in Table 4: continuous depression for more than 2 weeks (p = 0.002), treatment of myocardial infarction, angina (p = 0.038), reason of outpatient utilization (p = 0.004), monthly income (p = 0.007), perceived health status (p = 0.011), sick days (past 1 month) (p = 0.007), discomfort days (past 2 weeks) (p = 0.002), and restriction of activity (old-age) (p = 0.002).

5. Evaluation of Prediction Models

To validate each prediction model, we used a 10-fold cross validation. In 10-fold cross-validation, the data set is divided into 10 parts. Then training is carried out with 9 and testing with1; the process is repeated until all parts have been tested. A confusion matrix was built to gauge the performance of each model. Each model's performance was measured by 8 parameters, such as accuracy, sensitivity, precision, recall, and F-score. As shown in Table 5 and the corresponding box-plots provided in Figure 2, the SLR model ranked the highest (accuracy = 0.93, F-score = 0.49). After the SLR model the SVM (0.90, 051), RF (0.87, 0.33), and DT (0.82, 0.23) models rank in decreasing order of performance.

The accuracy measure was given the highest importance because the main goal of the study is to find a prediction model that best divides the variables. The SLR model not only exhibited the highest accuracy, but also used the fewest variables (8 variables). The DT model used 1,482 variables, the RF 64, and the SVM (c = 0.819, γ = 0.3012, w = 10) 32. Hence, the SLR was found to be the optimal model that efficiently predicted the HRQoL groups for the elderly with

Table 5. Comparison of performance in prediction models

Name	DT	RF	SVM	SLR
True positive rate	0.22	0.29	0.53	0.37
False positive rate	0.12	0.07	0.06	0.01
Specificity	0.88	0.93	0.94	0.99
Sensitivity	0.22	0.29	0.53	0.37
Precision	0.17	0.31	0.53	0.74
Recall	0.22	0.29	0.53	0.37
Accuracy	0.82	0.87	0.90	0.93
F-score	0.23	0.33	0.51	0.49

DT: decision tree, RF: random forest, SVM: support vector machine, SLR: stepwise logistic regression.

Table 4. Result of stepwise logistic regression

Variable	β	SE	p-value	Exp (β)	95% CI for Exp (β)	
					Lower	Upper
Continuous depression for more than 2 weeks (1 = yes, 2 = no)	1.018	0.330	0.002	2.767	1.449	5.285
Treatment of myocardial infarction, angina (1 = yes, 0 = no)	−0.312	0.150	0.038	0.732	0.545	0.983
Reason of outpatient utilization (1 = disease, 2 = accident, 3 = others)	0.165	0.058	0.004	1.179	1.053	1.321
Monthly income	0.004	0.002	0.007	1.004	1.001	1.007
Perceived health status (1 = very good, …, 5 = very poor)	−0.490	0.193	0.011	0.612	0.419	0.894
Sick day (past 1 month)	−0.068	0.025	0.007	0.934	0.889	0.982
Discomfort day (past 2 weeks)	−0.086	0.028	0.002	0.918	0.870	0.969
Restriction of activity (old-age) (1 = yes, 0 = no)	0.167	0.055	0.002	1.181	1.061	1.316
Constant	3.413	1.617	0.035	30.367		

SE: standard error, CI: confidence interval.

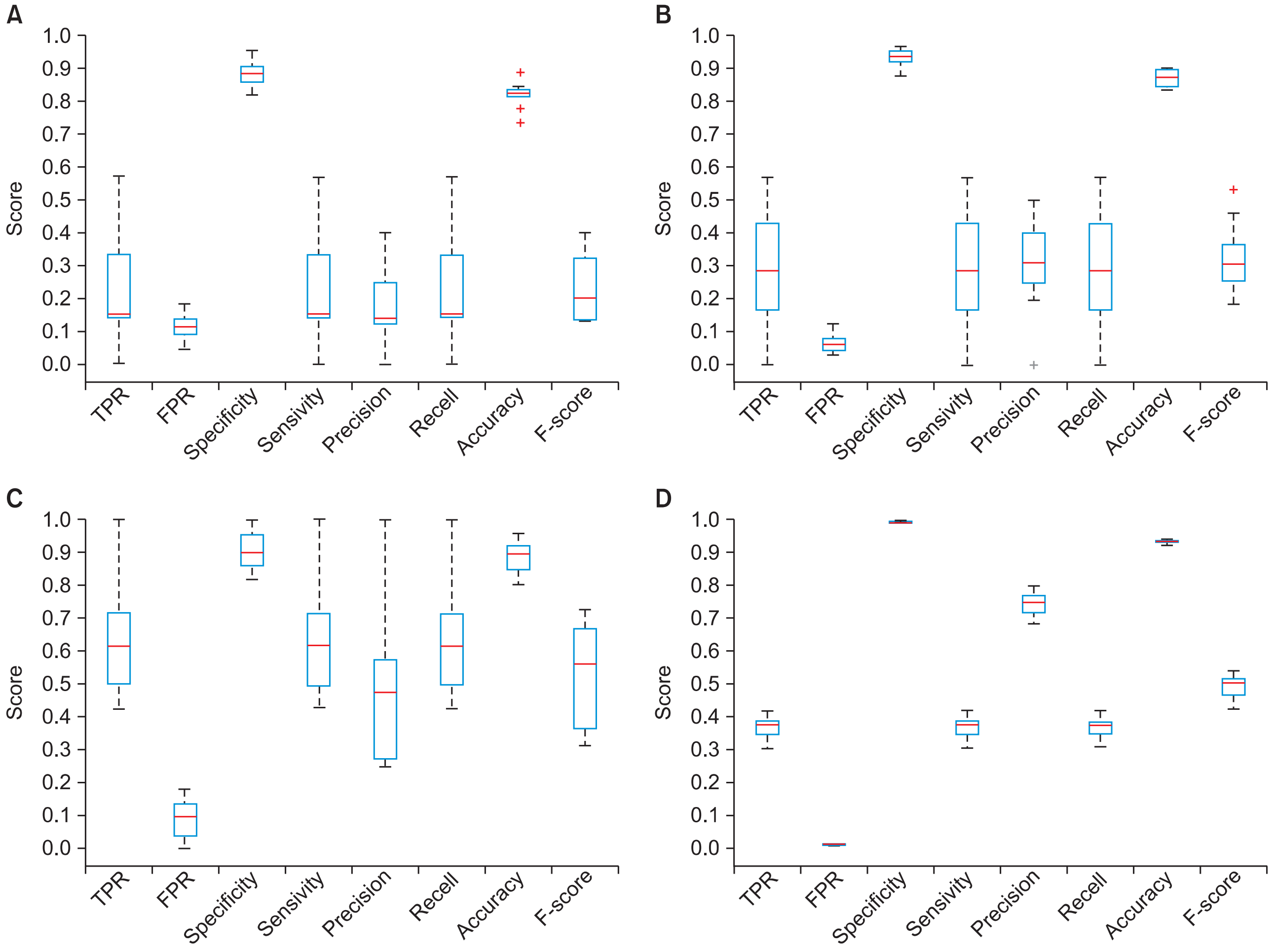

Figure 2. Boxplot of performance in four prediction models (A) decision tree, (B) random forest, (C) support vector machine, and (D) stepwise logistic regression. TPR: true positive rate, FPR: false positive rate.

chronic diseases.

A series of SVM models was created by applying 1-of-K encoding and discretization normalization schemes. For example, the 11 variables—number of chronic diseases, total duration of chronic diseases, age, femoral bone T-score, discomfort days (recent 2 weeks), sick days (past 1 month), duration of osteoarthritis, duration of rheumatoid arthritis, duration of diabetes, BMI, copayments for outpatient–from the hierarchical clustering were normalized by the application of discretization and 1-of-K encoding. They were then used to create an SVM model, which was in turn compared against the SVM model that was not normalized. The accuracy of the model with data normalization was 0.87, whereas that of the model without normalization was 0.72. In our tests, though limited in scope, we found that the performance of an SVM model increased as variables were normalized by the application of discretization and 1-of-K encoding.

IV. Discussion

Out of the 8 variables or factors that the SLR model demonstrated to be significant, 5 factors were found: monthly income, diagnosis of chronic disease, depression, discomfort, and perceived health status. The values of these factors dictate the HRQoL of the elderly with chronic diseases.

On the other hand, variables, such as age, gender, education level, job, family, number of chronic diseases, duration of chronic diseases, BMI, physical activity, smoking, drinking, did not seem to significantly affect the HRQoL. It is, therefore, necessary to steer more focus on, and to build health management strategies and intervention in accordance with, those variables that affect the HRQoL. In other words, considering the fact that factors affecting each chronic disease vary, intervention methods should be developed to address such differences. Also, previous studies [19,20] on adults or the elderly showed that employment and the types of employment act as significant factors, but when restricted to the elderly with chronic diseases, employment exerted no influence on the HRQoL. Hence, for the elderly with chronic diseases, intervention focus should be directed towards ways to help them adapt to diseases, rather than towards employment.

As one study reported [21], factors, such as cultivation of self-confidence, levels of self-management, and acceptance of chronic diseases, decide health status and eventually quality of life. According to the univariate analysis in the present study, the number of chronic diseases and the combined period of chronic diseases do not always correspond to the HRQoL levels. That is, for example, a higher number of chronic diseases do not necessarily translate into lower levels of HRQoL; rather, as a patient adapts more to his/her illnesses, the HRQoL level bounces back to higher levels at certain point in time.

Of course, an elderly person with chronic diseases will show a lower level of HRQoL than the elderly without them do. Intervention focus, however, should not be so much on diseases themselves, but on comprehensive intervention methods that alleviate depression and/or discomfort, and promote higher perceived health status.

The ML models in the study showed relatively less promising accuracy and efficiency levels than the SLR model did. To ascertain the problems, we tested SVM by applying different data normalization schemes as follows. A sample re-test set was created, which consisted of the 8 variables found to be significant in the SLR model. Eleven additional variables were selected by experts from 74 variables that were clustered by way of hierarchical clustering [22].

This study had the following limitations. 1) It was limited to examining the impacts of individual variables; we did not examine how each variable affects others; nor did we study the nature of direct or indirect influencing factors. 2) Since the survey used EQ-5D, a general instrument without any specific population groups being targeted, it may be necessary to develop HRQoL measurement tools for the elderly with chronic diseases that address disease-specific requirements and factors related to both individual and complex chronic diseases.

Our study is, to the best of our knowledge, the first attempt to use ML techniques to identify the influencing factors and to apply prediction models for the HRQoL of the elderly with chronic diseases as an alternative and complement to the traditional statistical approaches. Our source data were designed to be used for statistical analysis, which may explain the reason for relatively poor performance of the ML models [23]. Nevertheless, ML models may open new possibilities to find health-related factors that otherwise would be hidden in traditional analysis methods.

This study used data from the KNHANES to analyze factors that affected the HRQoL of the elderly with chronic diseases. We used ML techniques as a supplement to the SLR to develop prediction models for HRQoL risk groups. New influencing factors were identified, with incidental insights that, for ML techniques, data normalization of mixed-type data and careful selection of variables had significant impact on the performance and efficiency of the techniques.

Our study can be used as data in healthcare for the development of new clinical assessment and interventions for the elderly with chronic diseases. In other words, it would be possible to develop, specifically for the elderly with chronic illnesses, an HRQoL measurement tool that helps prioritize intervention for HRQoL risk groups. Based on the identified influencing factors, this study could also provide guidelines for healthcare staff in caring for the elderly and could help fine-tune and improve healthcare intervention in practice.

Conflict of Interest

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

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2010-0025461). This work was supported by the Soonchunhyang University Research Fund (2014).

References

1. Kempen GI, Ormel J, Brilman EI, Relyveld J. Adaptive responses among Dutch elderly: the impact of eight chronic medical conditions on health-related quality of life. Am J Public Health 1997;87(1):38-44.
2. Kil SR, Lee SI, Yun SC, An HM, Jo MW. The decline of health-related quality of life associated with some diseases in Korean adults. J Prev Med Public Health 2008;41(6):434-41.
3. An HM. Factors of health related quality of life of Korea male and female adults according to life cycle: by using 4th National Health and Nutrition Examination Survey [thesis]. Seoul: Yonsei University; 2010.
4. Kim M. The study of comparing the factors of affecting on the quality of life for young-old women and old-old women. Korean J Soc Welf 2006;58(2):197-222.
5. Jeon EY, Choi YH. Factors affecting the health-related

quality of life according to age in vulnerable aged men. J Korean Acad Nurs 2010;40(3):400-10.

6. Porojan M, Poanta L, Dumitrascu DL. Assessing health related quality of life in diabetic patients. Rom J Intern Med 2012;50(1):27-31.
7. Alguren B, Fridlund B, Cieza A, Sunnerhagen KS, Christensson L. Factors associated with health-related quality of life after stroke: a 1-year prospective cohort study. Neurorehabil Neural Repair 2012;26(3):266-74.
8. Lee BH, Ko JY. Influential factors for the health related quality of life in children with mental retardation and cerebral palsy. J Spec Educ Rehabil Sci 2010;49(2):105-26.
9. Heckman TG. The chronic illness quality of life (CIQOL) model: explaining life satisfaction in people living with HIV disease. Health Psychol 2003;22(2):140-7.
10. Choe SY. Associated factors of health-related quality of life using data mining: data from Korean National Health and Nutrition Examination Survey in 2005 [thesis]. Seoul: Korea University; 2009.
11. Son MS, Moon PS. Big data era of Korea, if you do not want to Galapagos [Internet]. Seoul: LG Business Insight: c2012 [cited at 2014 Mar 15]. Available from: http://www.lgeri.com/uploadFiles/ko/pdf/ind/LGBI1188-02_20120313130223.pdf.
12. Lim SJ, Min OK. Machine learning technology trends for big data processing. Electron Telecommun Trends 2012;27(5):55-63.
13. Maas ML, Delaney C. Nursing process outcome linkage research: issues, current status, and health policy implications. Med Care 2004;42(2 Suppl):II40-8.
14. EuroQol Group. EuroQol: a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.
15. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. J Mol Diagn 2003;5(2):73-81.
16. Bishop CM. Pattern recognition and machine learning. New York (NY): Springer; 2006.
17. Wettergren L, Bjorkholm M, Axdorph U, Langius-Eklof A. Determinants of health-related quality of life in long-term survivors of Hodgkin's lymphoma. Qual Life Res 2004;13(8):1369-79.
18. Garcia S, Luengo J, Saez JA, Lopez V, Herrera F. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Trans Knowl Data Eng 2013;25(4):734-50.
19. Jeong YH. A report on the health related quality of life in Korea. Health Welf Policy Forum 2011;(182):6-14.
20. Lee DH, Bin SO. Structure relationships for diseased and health-related quality of life in the elderly. J Korea Content Assoc 2011;11(1):216-24.
21. Song MS. Self-management education model based on concept of health promotion for older adults with chronic illness. J Korean Gerontol Nurs 2004;6(2):228-42.
22. Embrechts MJ, Gatti CJ, Linton J, Roysam B. Hierarchical clustering for large data sets. In: Georgieva P, Mihaylova L, Jain LC. Advances in intelligent signal processing and data mining: theory and applications. Berlin: Springer; 2013. p. 197-233.
23. Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res 2011;17(4):232-43.

http://dx.doi.org/10.4258/hir.2014.20.2.125