Observational Study

Medicine®

OPEN

Sarcopenia feature selection and risk prediction using machine learning

A cross-sectional study

Yang-Jae Kang, PhD[a,b], Jun-Il Yoo, MD[c,*], Yong-chan Ha, MD[d]

Abstract

The purpose of this study was to verify the usefulness of machine learning (ML) for selection of risk factors and development of predictive models for patients with sarcopenia.

We collected medical records from Korean postmenopausal women based on Korea National Health and Nutrition Examination Surveys. A training data set compiled from simple survey data was used to construct models based on popular ML algorithms (e.g., support vector machine, random forest [RF], and logistic regression).

A total of 4020 patients ≥65 years of age were enrolled in this study. The study population consisted of 1698 (42.2%) male and 2322 (57.8%) female patients. The 10 most important risk factors in men were body mass index (BMI), red blood cell (RBC) count, blood urea nitrogen (BUN), vitamin D, ferritin, fiber intake (g/d), primary diastolic blood pressure, white blood cell (WBC) count, fat intake (g/d), age, glutamic-pyruvic transaminase, niacin intake (mg/d), protein intake (g/d), fasting blood sugar, and water intake (g/d). The 10 most important risk factors in women were BMI, water intake (g/d), WBC, RBC count, iron intake (mg/d), BUN, high-density lipoprotein, protein intake (g/d), fiber consumption (g/d), vitamin C intake (mg/d), parathyroid hormone, niacin intake (mg/d), carotene intake (μg/d), potassium intake (mg/d), calcium intake (mg/d), sodium intake (mg/d), retinol intake (μg/d), and age. A receiver operating characteristic (ROC) curve analysis found that the area under the ROC curve for each ML model was not significantly different within a gender.

The most cost-effective method in clinical practice is to make feature selection using RF models and expert knowledge and to make disease prediction using verification by several ML models. However, the developed prediction model should be validated using additional studies.

Abbreviations: ALP = alkaline phosphatase, AUC = area under the ROC curve, BMI = body mass index, BUN = blood urea nitrogen, GPT = glutamic-pyruvic transaminase, HDL = high-density lipoprotein, KNHANES = Korea National Health and Nutrition Examination Survey, ML = machine learning, P = phosphorus, PTH = parathyroid hormone, RBC = red blood cell, RF = random forest, ROC = receiver operating characteristic, WBC = white blood cell.

Keywords: feature selection, machine learning, risk prediction, sarcopenia

1. Introduction

In elderly individuals, sarcopenia is associated with increased risks of falls and fractures, which can result in increases in mortality rates of elderly patient populations.[1–4] Sarcopenia reduces the quality of life of the elderly and causes socioeconomic problems.[5–7] Recently, there has been a growing interest in sarcopenia.[8] Various risk factors for this condition have been reported.[9–12] However, effective prevention methods have not yet been proposed. Unlike osteoporosis, which concerns the relatively static organ, bone, sarcopenia concerns the dynamic organ, muscle.[13] Because various risk factors are associated with sarcopenia, it is very difficult to diagnose and treat.

Socioeconomic factors, physical activity, chronic diseases, and nutritional factors are risk factors for sarcopenia.[4,9–12] Among these factors, chronic diseases and nutritional factors have been found to be main causes of this condition.[14] Therefore, it is very important to identify and supplement the various nutrients related to sarcopenia.[15,16] However, because the various risk factors are interrelated, it is very difficult to examine them using conventional statistical methods. It is also difficult to develop cost-effective risk prediction models in clinical settings.

The purpose of this study was to verify the accuracy and validity of the use of machine learning (ML) to select risk factors and develop predictive models for patients with sarcopenia.

2. Materials and methods

2.1. Ethics statement

Data from the 2008 to 2011 Korea National Health and Nutrition Examination Surveys (KNHANESs) were reviewed

Editor: Daryle Wane.

This study was funded by the Ministry of SMEs and Startups, Republic of Korea (Project No. P0002726).

The authors have no conflicts of interest to disclose.

[a] Division of Applied Life Science Department, PMBBRC, [b] Division of Life Science Department, [c] Department of Orthopaedic Surgery, Gyeongsang National University Hospital, Jinju, [d] Department of Orthopaedic Surgery, Chung-Ang University College of Medicine, Seoul, Republic of Korea.

[*] Correspondence: Jun-Il Yoo, Department of Orthopaedic Surgery, Gyeongsang National University Hospital, 90 Chilamdong, Jinju, Gyeongnamdo 660-702, Republic of Korea (e-mail: furim@daum.net).

How to cite this article: Kang YJ, Yoo JI, Ha Yc. Sarcopenia feature selection and risk prediction using machine learning. Medicine 2019;98:43(e17699).

Received: 16 April 2019 / Received in final form: 21 September 2019 / Accepted: 27 September 2019

http://dx.doi.org/10.1097/MD.0000000000017699

and approved by the Institutional Review Board of the Korea Centers for Disease Control and Prevention (Approval No. 2008-04EXP-01-C, 2009-01CON-03-C, 2010-02CON-21-C, and 2011-02CON-06-C). Informed consent was obtained from each participant when the 2008, 2009, 2010, and 2011 KNHANESs were conducted.

2.2. Participants

This study was based on data obtained from the 2008 to 2011 KNHANESs conducted by the Korea Ministry of Health and Welfare. KNHANES is a nationwide representative cross-sectional survey of the Korean population; it uses a clustered, multistage, stratified, and rolling sampling design. It consists of 3 sections: a health interview, a health examination, and a dietary survey. More than 500 variables are examined and the survey is conducted each year. These variables are included in a health questionnaire and in laboratory findings; data on nutritional factors are also collected.[17] Survey data are collected via household interviews and direct standardized physical examinations performed in specially equipped mobile examination centers.

The data considered for use in this study were collected from a total of 37,753 KNHANES participants (2008 [n = 9744], 2009 [n = 10,533], 2010 [n = 8958], and 2011 [n = 8518]). However, participants were excluded if they were <65 years of age or if the data required to evaluate skeletal muscle mass and dietary intake were unavailable. After these exclusions, data from a total of 4020 participants (male: 1698; female: 2322) were included in the analysis (Fig. 1).

2.3. Measurement of appendicular skeletal muscle mass and definition of sarcopenia

Body composition was measured using whole-body dual X-ray absorptiometry (DXA) (QDR 4500A; Hologic, Inc, Waltham, MA). When the DXA was performed, each subject was asked to remove all jewelry and other personal items that could interfere with the examination and to wear paper gowns. To obtain accurate and reliable results, all body composition data were gathered by trained and quality-controlled sarcopenia examination surveyors. Bone mineral content, fat mass, and lean soft-tissue mass were measured separately for each part of the body, including the arms and legs. The total lean soft-tissue mass for the arms and legs was nearly equal to the total skeletal muscle mass. Because absolute muscle mass correlates with height, the skeletal muscle mass index (SMI) was calculated using the formula, lean mass (kg)/height (m^2); the resulting value was directly analogous to the body mass index (BMI = body weight [kg]/height [m^2]). The arm SMI was defined as arm lean mass (kg)/height (m^2). The leg SMI was defined as leg lean mass (kg)/height (m^2). The appendicular SMI was defined as the sum of the arm and leg SMIs. Sarcopenia was defined according to Asia Working Group for Sarcopenia criteria (SMI < 5.4 kg/m^2 in women and <7.0 kg/m^2 in men).[18]

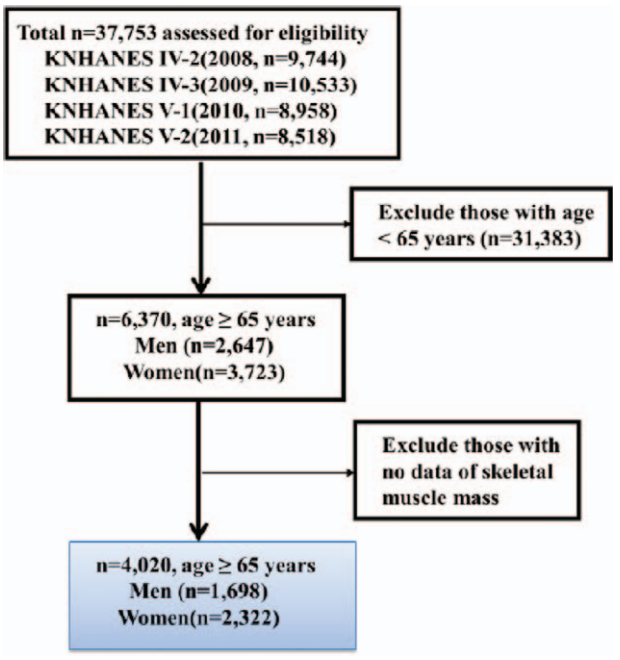

Figure 1. Study subject selection process, Korea National Health and Nutrition Examination Survey (KNHANES) IV and V (2008–2011).

2.4. Machine learning (random forest model) and knowledge-based feature selection

We collected data on 968 observed features of 1698 male and 2322 female participants from the KNHANES results. During data curation, we manually excluded columns for unrelated features and features with missing values for more than 300 individuals. We then recollected the data from individuals without missing values. A "sarcopenia" column was used as the classification label for the supervised learning. To find preliminary features that were well-related to sarcopenia, random forest (RF) classification was implemented with 2000 estimators. Notable features were those that were above a selected threshold (95th percentile of the calculated feature importance); these features selected as preliminary sarcopenia risk factors. Manual curation of these preliminary RF-based risk factors was performed by 5 orthopedic surgeons, 2 nutritionists, and 1 physician. Through consensus, they selected 17 risk variables that did not have clinically overlapping features (Figs. 2 and 3).

We used these 17 selected risk factors to build predictive models for sarcopenia based on 4 classification algorithms (i.e., logistic regression, support vector machine, gradient boosting, and RF). To determine the hyperparameter, we implemented 5-fold cross-validation of the training set with 1000 times train/test set shuffling and 50 different hyperparameters. Using each hyperparameter, we drew learning curves changing the train/test ratios to observe the convergence of cross-validation scores. We iteratively split our data set into a 75% training set and a 25% test set based on the results after repeating 1000 times. Each of the 4 classification models were built on the training set with the selected hyperparameters. The coefficient and feature importance results for each trained classification model were presented in boxplots for feature selection. Features selected at least 3 times by the 4 classification algorithms were used again to retrain with the 4 classification algorithms. Each model's predicted probability of sarcopenia classification on the test set was subjected to receiver operating characteristic (ROC) curve analysis to obtain the reliability of the final model (Figs. 4 and 5).

3. Results

Among the 4020 patients ≥65 years of age who were included in the study population, 1698 (42.2%) were men and 2322 (57.8%)

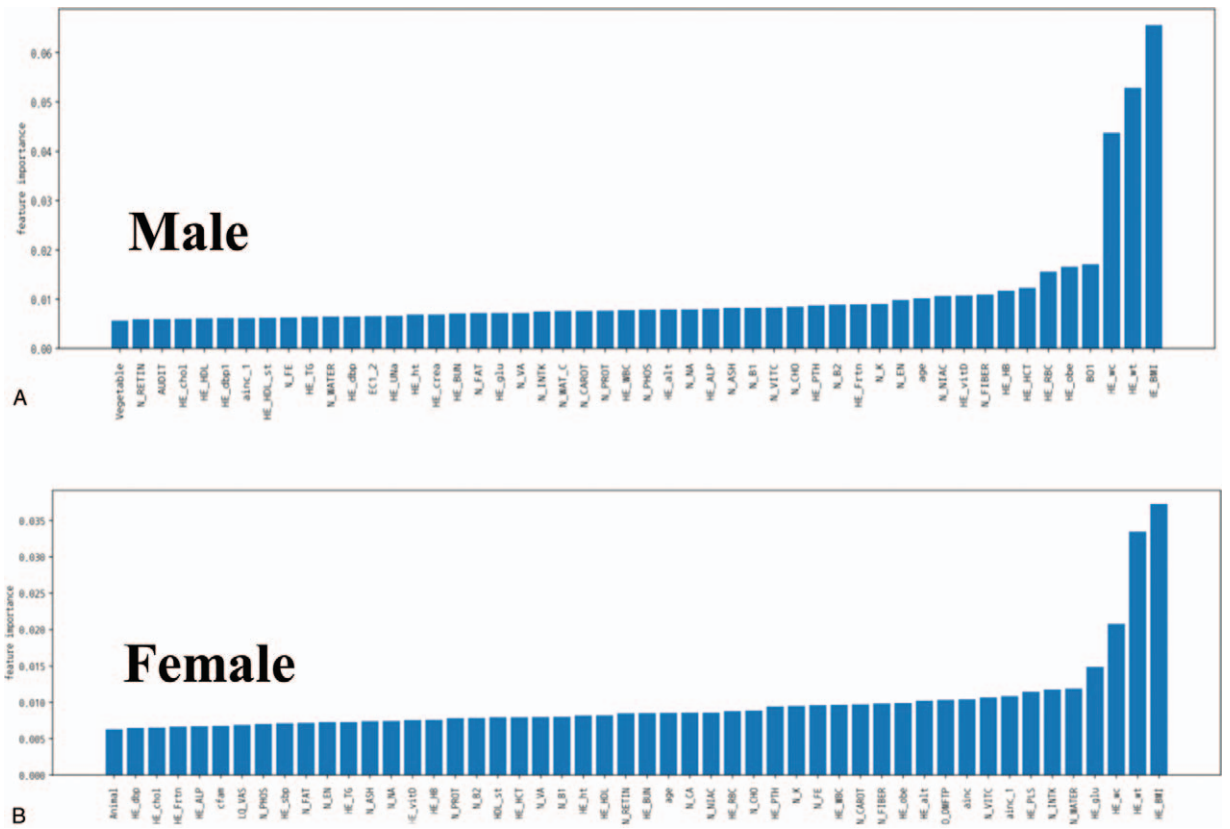

Figure 2. (A) Feature selection using random forest (RF) machine learning, data from elderly male participants. (B) Feature selection using RF machine learning, data from elderly female participants.

were women. The features selected for men using the RF method were BMI, body weight, waist circumference, subjective body recognition, obesity, red blood cell (RBC) count, hematocrit, hemoglobin, fiber intake (g/d), vitamin D, niacin intake (mg/d), thiamine intake (mg/d), ash intake (g/d), alkaline phosphatase (ALP) (mg/d), potassium intake (g/d), ferritin, riboflavin (mg/d), sodium intake (mg/d), glutamic-pyruvic transaminase (GPT), phosphorus (P) intake (mg/d), white blood cell (WBC) count, protein intake (g/d), carotene intake (μg/d), water intake (g/d), vitamin A intake (μg/d), fasting blood glucose, fat intake (g/d), blood urea nitrogen (BUN), blood creatinine, height, urinary sodium, occupation, primary diastolic blood pressure, triglyceride, iron intake (mg/d), high-density lipoprotein (HDL) cholesterol, cholesterol, AUDIT, retinol intake (μg/d), and vegetable protein intake (g/d). The 10 most important risk factors in men were BMI, RBC count, BUN, vitamin D, ferritin, fiber intake (g/d), primary diastolic blood pressure, WBC count, fat intake (g/d), age, GPT, niacin intake (mg/d), protein intake (g/d), fasting blood sugar, and water intake (g/d).

The features selected for women using the RF method were BMI, body weight, waist circumference, fasting blood glucose, water intake (g/d), food intake (g/d), pulse regularity, household income, vitamin C intake (mg/d), carotene intake (μg/d), WBC count, iron intake (mg/d), potassium intake (mg/d), parathyroid hormone (PTH), carbohydrate intake (g/d), RBC count, niacin intake (mg/d), calcium intake (mg/d), age, BUN, retinol intake (μg/d), HDL, height, thiamine intake (mg/d), riboflavin intake (mg/d), protein intake (g/d), hemoglobin, vitamin D, sodium intake (mg/d), ash intake (g/d), triglycerides, energy intake (kcal), daily intake of fat (g), primary systolic blood pressure, daily intake (mg), quality of life, number in household, ALP, ferritin, total cholesterol, primary diastolic blood pressure, and animal protein intake (g/d). The 10 most important risk factors in women were BMI, water intake (g/d), WBC, RBC count, iron intake (mg/d), BUN, HDL, protein intake (g/d), fiber consumption (g/d), vitamin C intake (mg/d), PTH, niacin intake (mg/d), carotene intake (μg/d), potassium intake (mg/d), calcium intake (mg/d), sodium intake (mg/d), retinol intake (μg/d), and age.

Use of the features selected according to each ML model resulted in different findings. The features used in more than 2 of the 4 ML models were BMI, daily vitamin C intake (mg), GPT, daily fiber intake (g), PTH, RBC count, daily niacin intake (mg), and BUN in men. They were BMI, daily water intake (g), 15-second pulse rate, daily vitamin C intake (mg), daily fiber intake (g), WBC count, daily iron intake (mg), PTH, age, and BUN in women. Daily protein intake (g) was: In men, BMI, vitamin C intake (mg/d), GPT, fiber intake (g/d), PTH, RBC, niacin intake (mg/d), and BUN. In women, BMI, water intake (g/d), 15-second pulse rate, vitamin C intake (mg/d), fiber intake (g/d), WBC count, iron intake (mg/d), PTH, age, BUN, protein intake (g/d) (Tables 1 and 2).

The ROC analysis revealed that the area under the ROC curve (AUC) for each ML model was not significantly different for either gender (RF: 0.82 in men, 0.78 in women; support vector: 0.80 in men, 0.81 in women; gradient boosting: 0.81 in men, 0.81 in women; logistic regression: 0.82 in men, 0.80 in women) (Table 3).

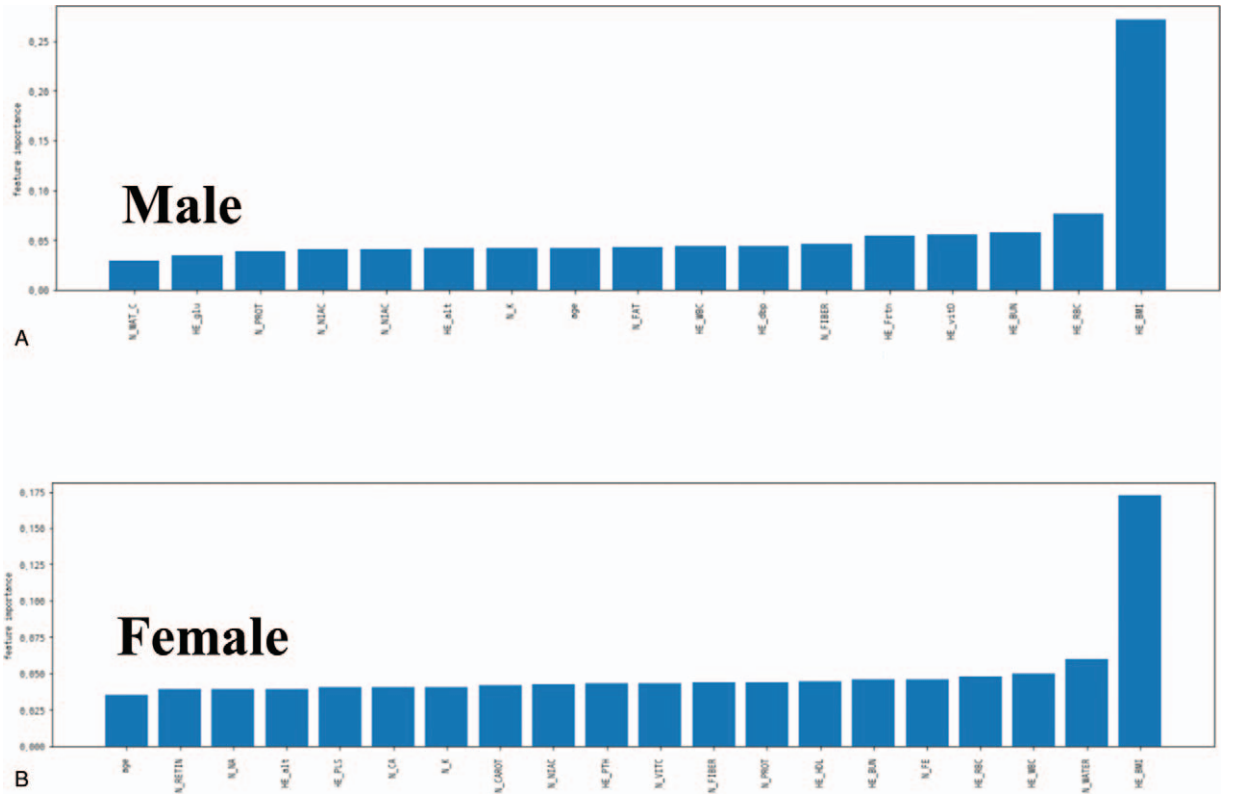

Figure 3. (A) Feature selection using random forest (RF) machine learning after application of selection category, data from elderly male participants. (B) Feature selection using RF machine learning after application of selection category, data from elderly female participants.

4. Discussion

Many studies have identified possible disease prediction models using ML. Various models that include human anatomical measurements have been developed. However, the development of a disease prediction model using ML is still along the boundary between artificial and human intelligence. It is also impossible to cost-effectively measure many risk factors in the clinical field. In this study, we found that different features were selected when different algorithms were applied to different ML models. However, for any feature, the accuracy of the diagnostic prediction was found to be similar after ML. When the number of features increases, the accuracy of the prediction model might increase as diagnostic accuracy increases. However, tests and survey tools that can be used in the clinical field are still limited. Therefore, a disease prediction model that uses ML is one of the most cost-effective artificial intelligence methods available at this time.

Disease prediction models that use genome-wide sequencing and wearable devices are still being developed. They can be used as examples of enhancement of diagnostic prediction using multiple features. However, it is ethically and morally problematic to determine a patient's treatment plan using unverified data and the currently available algorithms. Therefore, our approach represents one of the best examples of using artificial intelligence that uses the best features and models as a tool to assist with diagnosis.

Goodman et al[19] performed a study using National Health and Nutrition Examination Survey (NHANES) data to identify predictors of low skeletal muscle mass in older adults. Their goal is to develop a practical clinical assessment tool for use by clinicians to identify patients requiring DXA screening for muscle mass composition. They found that BMI is strongly associated with a low SMI and that BMI might be an informative predictor in the primary care setting. However, because sarcopenia has many risk factors, depending on the condition of the patient, predicting sarcopenia by simply using BMI might be problematic. Using NHANES data in Korea in this study, we found that a prediction model could be developed that accurately predicts simple risk factors that can be used in an outpatient setting.

4.1. Clinical relevance

It is very important to construct a simple disease prediction model that can be used in the clinical field. The appropriate conditions needed for a disease prediction model used in an outpatient setting include: a minimum number of appropriate risk factors; the ability to measure each variable at the outpatient clinic; variables that are less likely to fluctuate depending on daily patient health conditions; and standardization of test items should be done. This study revealed that ML could be used to develop a prediction model to select risk factors for sarcopenia. When multiple risk factors are analyzed at the same time, selection of risk factors is constrained by statistical limitations such as multi-collinearity. However, this ML technique also has the advantage of finding new risk factors that previously could not be predicted. The most important prerequisite for this disease prediction model is labeling the disease. Therefore, it would be a prerequisite to develop a model with high accuracy to clarify characteristics of the patient and control groups and substitute

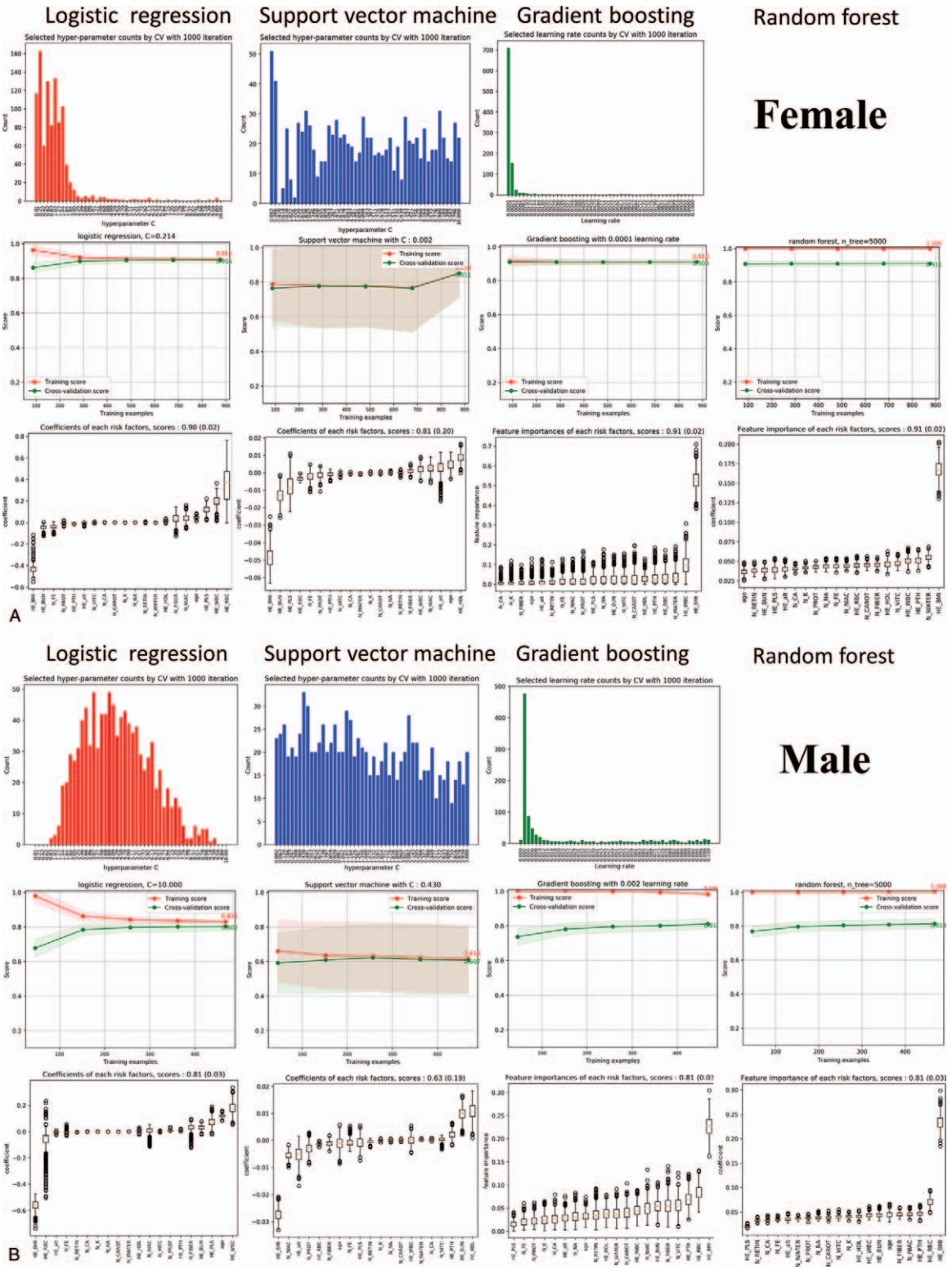

Figure 4. (A) Process of risk prediction model construction using machine learning (ML), data from elderly female participants. (B) Process of risk prediction model construction using ML, data from elderly male participants.

for the disease the features of the disease that are clearly identified (Fig. 6).

This study had some few limitations. First, risk factors were selected differently depending on the ML model used. Thus, other risk factors might have been selected (i.e., a set different from the 17 selected) if models other than the RF model were used. However, the RF model used in this study is the most intuitive and convenient way to examine strong associations such as biological causality because it uses an algorithm to determine feature significance using a decision tree.[20] Second, the number

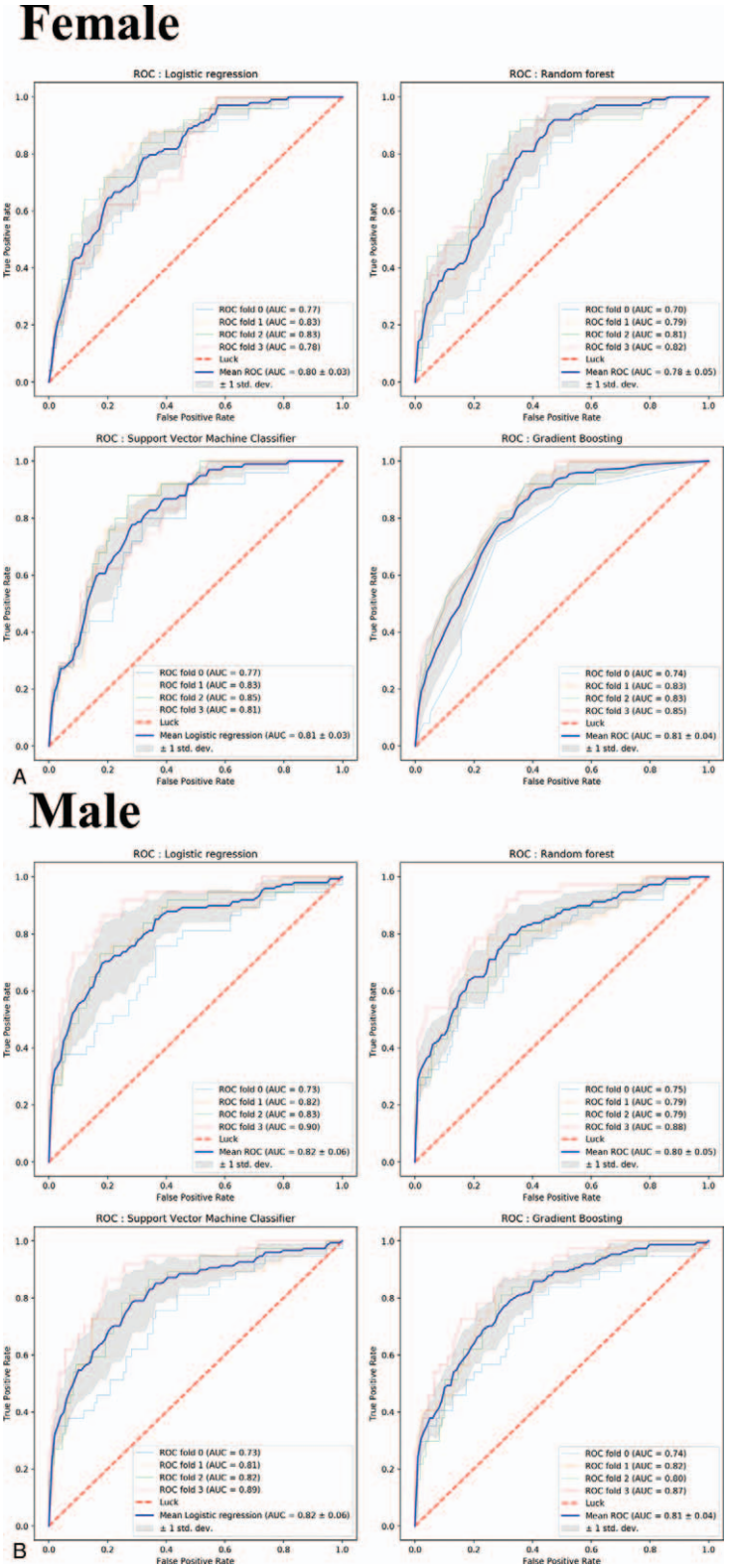

Figure 5. (A) Receiver operator characteristic (ROC) curve analysis for various machine learning (ML) models, data from elderly female participants. (B) ROC curve analysis for various ML models, data from elderly male participants.

of risk factors was limited by the investigators. If the number of risk factors was increased, the accuracy of the model would have increased. However, from a cost-benefit perspective, only a limited number of risk factors can be used in a real clinical setting. Therefore, choosing the correct number of risk factors is very important. Further research is needed to determine the appropriate number of risk factors for the methodology used in this study to determine model fit. Third, the choice of risk

Table 1

Scoring for common risk factors according to different machine learning models, data from elderly male participants.

Variables	LR	GB	SVM	RF	Score
BMI	o	o	O	o	4
Water intake, g/d	x	x	x	x	0
15-s pulse rate	0	x	x	x	1
Vitamin C, mg/d	x	o	o	x	2
GPT	0	x	o	x	2
Fiber intake, g/d	0	0	x	o	3
Carotene intake, μg/d	x	x	x	x	0
WBC	0	x	x	x	1
Iron intake	0	x	x	x	1
Potassium intake	x	x	x	x	0
PTH, pg/mL	x	o	o	o	3
RBC	o	o	o	o	4
Niacin intake, mg/d	x	x	0	0	2
Calcium intake, mg/d	0	0	x	0	1
Age	0	x	x	x	1
BUN, mg/dL	0	x	0	x	2
Retinol intake	0	x	x	x	1
HDL, mg/dL	0	0	x	0	1
Protein intake, g/d	0	0	x	0	1
Sodium intake, g/d	x	x	x	x	0

BMI = body mass index, BUN = blood urea nitrogen, GB = gradient boosting, GPT = glutamic-oxaloacetic transaminase, HDL = high-density lipoprotein, LR = logistic regression, PTH = parathyroid hormone, RBC = red blood cell, RF = random forest, SVM = support vector machine, WBC = white blood cell.

factors was not selected using only ML. Selection was also performed by researchers. In addition, different models used different methods to determine the importance of a feature during the selection of risk factors. Therefore, risk factors commonly selected in this study would be more important. However, there were no significant differences in the accuracies of the different ML models. Therefore, using ML to develop a prediction model is a meaningful approach. Nevertheless, it is important to note that the use of well-constructed and labeled national data (e.g., the KNHANES data) for a low-cost but accurate method to produce a predictive model of sarcopenia is an important conclusion of this study. Finally, the results of this study may not be applicable to populations of different ethnicities and races because genetic factors could have affected the results. Therefore, further research will be required to validate this predictive model of sarcopenia using various country-specific data.

In conclusion, the most cost-effective method in clinical practice is to perform feature selection using an RF model and expert knowledge. Several ML models should be used to verify disease prediction results. The developed prediction model approach should be validated using additional studies.

Table 2

Scoring for common risk factors according to different machine learning models, data from elderly female participants.

Variables	LR	GB	SVM	RF	Score
BMI	1	1	1	1	4
Water intake, g/d	0	1	0	1	2
15-s pulse rate	1	0	1	0	2
Vitamin C, mg/d	0	1	0	1	2
GPT	0	0	0	0	0
Fiber intake, g/d	0	1	1	0	2
Carotene intake, μg/d	0	0	0	0	0
WBC	1	0	1	1	3
Iron intake	1	0	1	0	2
Potassium intake	0	0	0	0	0
PTH, pg/mL	1	0	1	1	3
RBC	1	0	0	0	1
Niacin intake, mg/d	1	0	0	0	1
Calcium intake, mg/d	0	0	0	0	0
Age	1	0	1	0	2
BUN, mg/dL	1	0	1	0	2
Retinol intake	0	0	0	0	0
HDL, mg/dL	0	0	1	0	1
Protein intake, g/d	1	0	1	0	2
Sodium intake, g/d	0	1	0	0	1

BMI = body mass index, BUN = blood urea nitrogen, GB = gradient boosting, GPT = glutamic-oxaloacetic transaminase, HDL = high-density lipoprotein, LR = logistic regression, PTH = parathyroid hormone, RBC = red blood cell, RF = random forest, SVM = support vector machine, WBC = white blood cell.

Author contributions

Conceptualization: Yang-Jae Kang, Jun-Il Yoo, Yong-chan Ha.
Data curation: Yang-Jae Kang, Jun-Il Yoo, Yong-chan Ha.
Formal analysis: Yang-Jae Kang, Jun-Il Yoo.
Investigation: Yang-Jae Kang, Jun-Il Yoo.
Methodology: Yang-Jae Kang.
Project administration: Yang-Jae Kang.
Supervision: Yang-Jae Kang, Yong-chan Ha.
Visualization: Yong-chan Ha.

Table 3

Accuracy of prediction models derived using various machine learning techniques.

Prediction model	AUC
Machine learning	
Random forest	0.82 in male, 0.78 in female
Support vector	0.80 in male, 0.81 in female
Gradient boosting	0.81 in male, 0.81 in female
Logistic regression	0.82 in male, 0.80 in female

AUC = area under the curve.

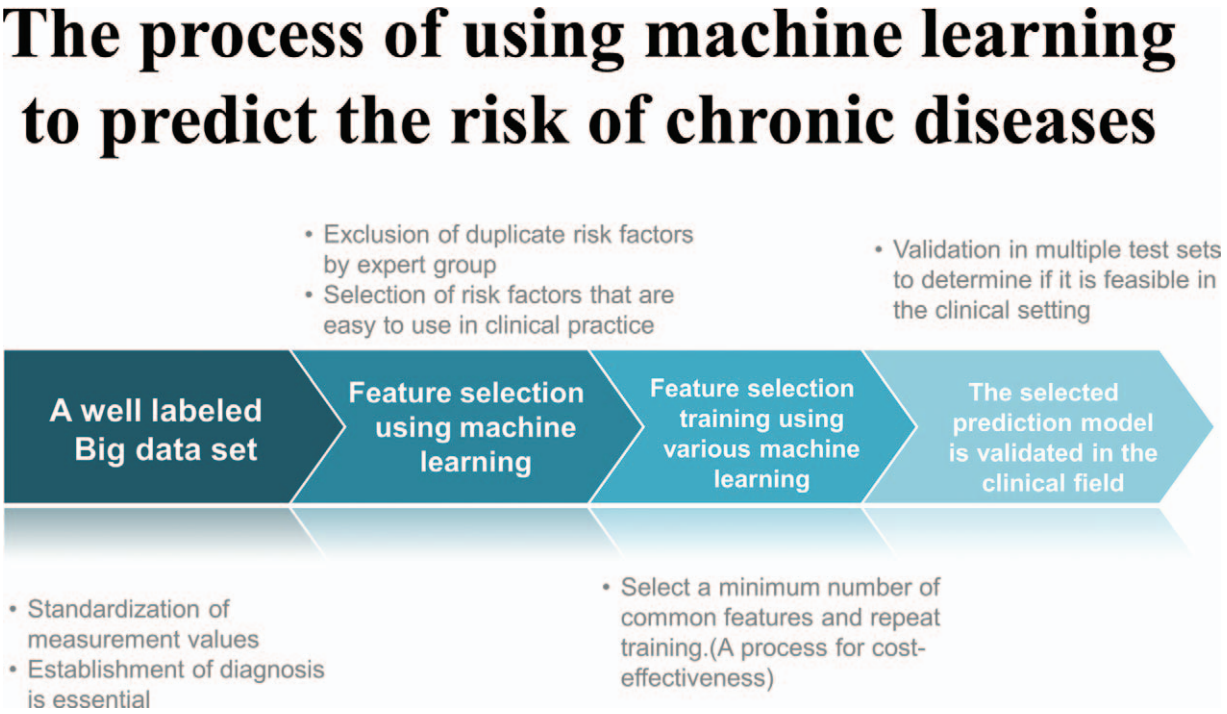

Figure 6. The machine learning process used to predict the risks of chronic diseases.

References

[1] Brown JC, Harhay MO, Harhay MN. Sarcopenia and mortality among a population-based sample of community-dwelling older adults. J Cachexia Sarcopenia Muscle 2016;7:290–8.

[2] Deren ME, Babu J, Cohen EM, et al. Increased mortality in elderly patients with sarcopenia and acetabular fractures. J Bone Joint Surg Am 2017;99:200–6.

[3] Szulc P, Feyt C, Chapurlat R. High risk of fall, poor physical function, and low grip strength in men with fracture-the STRAMBO study. J Cachexia Sarcopenia Muscle 2016;7:299–311.

[4] Tsutsumimoto K, Doi T, Makizako H, et al. Aging-related anorexia and its association with disability and frailty. J Cachexia Sarcopenia Muscle 2018;9:834–43.

[5] Beaudart C, Reginster J-Y, Geerinck A, et al. Current review of the SarQoL®: a health-related quality of life questionnaire specific to sarcopenia. Expert Rev Pharmacoecon Outcomes Res 2017;17:335–41.

[6] Tsekoura M, Kastrinis A, Katsoulaki M, et al. Sarcopenia and its impact on quality of life. Adv Exp Med Biol 2017;987:213–8.

[7] Dorosty A, Arero G, Chamar M, et al. Prevalence of sarcopenia and its association with socioeconomic status among the elderly in Tehran. Ethiop J Health Sci 2016;26:389–96.

[8] Han A, Bokshan SL, Marcaccio SE, et al. Diagnostic criteria and clinical outcomes in sarcopenia research: a literature review. J Clin Med 2018;7:

[9] Yoo J-I, Ha Y-C, Lee Y-K, et al. High levels of heavy metals increase the prevalence of sarcopenia in the elderly population. J Bone Metab 2016;23:101–9.

[10] Yoon B-H, Lee J-K, Choi D-S, et al. Prevalence and associated risk factors of sarcopenia in female patients with osteoporotic fracture. J Bone Metab 2018;25:59–62.

[11] Yoo J-I, Choi H, Song S-Y, et al. Relationship between water intake and skeletal muscle mass in elderly Koreans: a nationwide population-based study. Nutr Burbank Los Angel Cty Calif 2018;53:38–42.

[12] Yoo J-I, Ha Y-C, Lee Y-K, et al. High prevalence of sarcopenia among binge drinking elderly women: a nationwide population-based study. BMC Geriatr 2017;17:114.

[13] Argilés JM, Campos N, Lopez-Pedrosa JM, et al. Skeletal muscle regulates metabolism via interorgan crosstalk: roles in health and disease. J Am Med Dir Assoc 2016;17:789–96.

[14] Yoo J-I, Ha Y-C, Choi H, et al. Malnutrition and chronic inflammation as risk factors for sarcopenia in elderly patients with hip fracture. Asia Pac J Clin Nutr 2018;27:527–32.

[15] Hickson M. Nutritional interventions in sarcopenia: a critical review. Proc Nutr Soc 2015;74:378–86.

[16] Yanai H. Nutrition for sarcopenia. J Clin Med Res 2015;7:926–31.

[17] Kweon S, Kim Y, Jang M, et al. Data resource profile: the Korea national health and nutrition examination survey (KNHANES). Int J Epidemiol 2014;43:69–77.

[18] Chen L-K, Liu L-K, Woo J, et al. Sarcopenia in Asia: consensus report of the Asian working group for sarcopenia. J Am Med Dir Assoc 2014;15:95–101.

[19] Goodman MJ, Ghate SR, Mavros P, et al. Development of a practical screening tool to predict low muscle mass using NHANES 1999-2004. J Cachexia Sarcopenia Muscle 2013;4:187–97.

[20] Goldstein BA, Polley EC, Briggs FBS. Random forests for genetic association studies. Stat Appl Genet Mol Biol 2011;10:32.