Prediction of Suicidal Ideation among Korean Adults Using Machine Learning: A Cross-Sectional Study

Bumjo Oh[1]*, Je-Yeon Yun[2,3]*, Eun Chong Yeo[4], Dong-Hoi Kim[4], Jin Kim[4], and Bum-Joo Cho[5,6,7] ✉

[1]Department of Family Medicine, SMG-SNU Boramae Medical Center, Seoul, Republic of Korea
[2]Seoul National University Hospital, Seoul, Republic of Korea
[3]Yeongeon Student Support Center, Seoul National University College of Medicine, Seoul, Republic of Korea
[4]School of Software, Hallym University, Chuncheon, Republic of Korea
[5]Department of Ophthalmology, Hallym University Sacred Heart Hospital, Hallym University College of Medicine, Anyang, Republic of Korea
[6]Medical Artificial Intelligence Center, Hallym University Medical Center, Anyang, Republic of Korea
[7]Institute of New Frontier Research, Hallym University College of Medicine, Chuncheon, Republic of Korea

Objective Suicidal ideation (SI) precedes actual suicidal event. Thus, it is important for the prevention of suicide to screen the individuals with SI. This study aimed to identify the factors associated with SI and to build prediction models in Korean adults using machine learning methods.

Methods The 2010–2013 dataset of the Korea National Health and Nutritional Examination Survey was used as the training dataset (n=16,437), and the subset collected in 2015 was used as the testing dataset (n=3,788). Various machine learning algorithms were applied and compared to the conventional logistic regression (LR)-based model.

Results Common risk factors for SI included stress awareness, experience of continuous depressive mood, EQ-5D score, depressive disorder, household income, educational status, alcohol abuse, and unmet medical service needs. The prediction performances of the machine learning models, as measured by the area under receiver-operating curve, ranged from 0.794 to 0.877, some of which were better than that of the conventional LR model (0.867). The Bayesian network, LogitBoost with LR, and ANN models outperformed the conventional LR model.

Conclusion A machine learning-based approach could provide better SI prediction performance compared to a conventional LR-based model. These may help primary care physicians to identify patients at risk of SI and will facilitate the early prevention of suicide.

Psychiatry Investig 2020;17(4):331-340

Key Words Suicidal ideation, Risk factor, Machine learning, Artificial intelligence.

INTRODUCTION

Suicide has been a major public health problem in Korea, which took second place in the suicide rate (25.8 per 100,000 population) among the OECD countries in 2016. Suicide is a complex sequential process including ideation, planning, and attempt leading to the final completion;[1] approximately one-fifth of individuals who think about suicide might eventually attempt suicide.[2] Specifically, progression from SI to suicidal attempt could be facilitated when an individual suffers from poor impulse control or poor mental health.[3,4] Of note, possible aggravating factors that contribute to suicidal ideation (SI) may include depression, loneliness in human relationships, economic difficulties, or physical pain.[5] Accordingly, attentive questioning for suffer from the suicidal ideation in the primary outpatient clinic, followed by timely consultation to the psychiatrists, could be crucial for effective reduction of suicidal risks for community population.[6] However, studies to elucidate the primal risk factors of SI for individuals visiting the primary outpatient clinic with issues other than the SI have not been sufficiently conducted yet.[7-11]

As one of the efforts to overcome the issue of generalizability for the study results per sites and to enable more valid ap-

plication of study results at individual level, big data analytics combined with the machine learning methods also have been adapted in the field of medical science. Predictive big data analytics refers to the use of algorithms, systems, and tools to extract information, generate maps, prognosticate trends, and identify patterns in a variety of past, present, or future settings.[12] With the recent advancements in technology and data science,[13] artificial intelligence using machine learning methods is utilized widely in medicine to predict the prevalence of various diseases or therapeutic outcomes.[14] These machine learning methods have shown better performance in predicting unmeasured outcomes as compared to conventional statistical methods.[15,16] For suicidal risks in association with psychological variables, previous history of suicidal attempts,[17] current suicidal ideation,[18] upcoming suicidal attempts[19] as well as the subtypes of longitudinal trajectories in changes of depressive symptoms and suicidal ideation[20] have been classified by way of the machine learning methods that applied several psychological symptoms including depressive mood, family- and pharmacotherapy-related features as explanatory features. On the other hand, few previous machine learning method-based big data studies have explored the risk factors of suicidality from the non-psychological variables to be prioritized in the primary care clinic.[21]

Therefore, the present study aimed to establish models to predict the risk of SI among Korean adults using a machine learning approach. We analyzed a large dataset of a representative Korean population to identify the factors associated with SI and to validate the performance of different machine-learning models to predict SI. To our knowledge, this is one of the largest studies to apply machine learning algorithms to identify risk factors and to build prediction models for SI ever performed.

METHODS

Data source: the Korea National Health and Nutrition Examination Survey

The dataset used in this study was acquired from the fifth and sixth Korea National Health and Nutrition Examination Survey (KNHANES) which is a nationally representative annual health survey conducted in Korea on the health and nutritional status of the general population.[22] The KNHANES is conducted by the Ministry of Health and Welfare of Korea under the leadership of the Korea Center for Disease Control. The fifth and sixth KNHANES were performed from 2010 to 2012, and 2013 through 2015, respectively. In 2014, SI was investigated only for adolescents and not for adults, and thus the 2014 KNHANES dataset was excluded from this study.

The details of the KNHANES are described elsewhere.[23] Briefly, this annual survey extracted 4,000 households nationwide, using a stratified multi-stage clustered complex sampling

method based on age, sex, and area of residence. The selected participants were interviewed regarding their health and nutritional status. Adults aged ≥19 years were asked to answer a questionnaire regarding SI, suicide planning, and suicide attempts. The participation flowchart has been presented in Figure 1. The survey adhered to the tenets of the Declaration of Helsinki. Written informed consent was acquired from all participants. The KNHANES survey protocol was approved by the Korea Center for Disease Control Institutional Review Board (IRB no. 2010-02CON-21-C, 2011-02CON-06-C, 2012-01EXP-01-2C, 2013-07CON-03-4C, and 2013-12EXP-03-5C). Since 2014, the KNHANES has been exempted from review for research ethics, based on the Bioethics and Safety Act.

Measurement of main outcomes

Assessment of suicidal ideation

Two suicide-associated variables were investigated using self-reported health questionnaires, SI and actual suicide at-
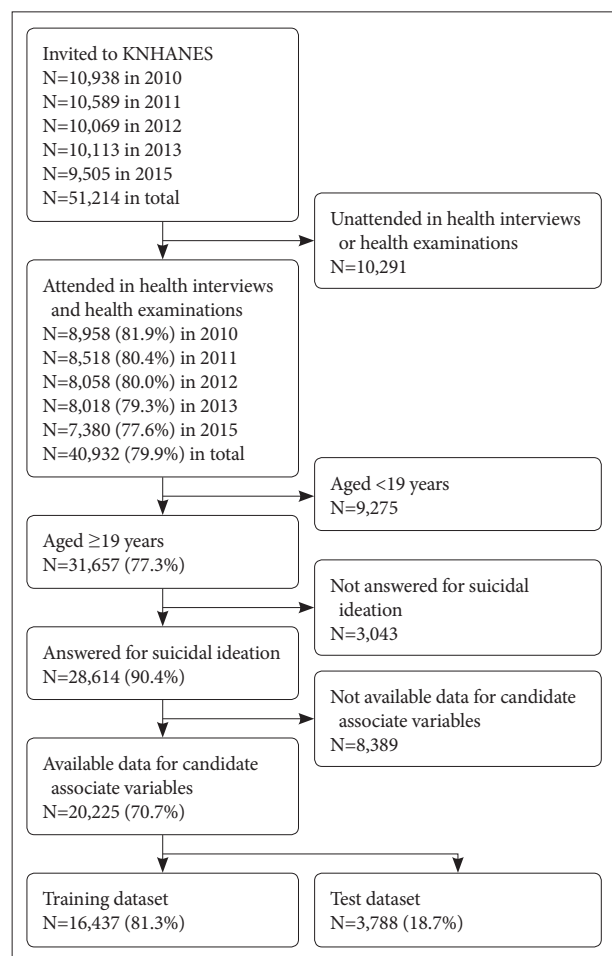

Figure 1. Flowchart for participants in the Korea National Health and Nutrition Examination Survey. KNHANES: the Korea National Health and Nutrition Examination Survey.

tempts. SI was assessed using the following question: "Have you seriously thought about wanting to die during the last one year?," to which the participants responded using "Yes" or "No" responses.

Assessment of independent variables

The KNHANES assessed over 800 variables on various aspects of the health and nutritional status of participants. Therefore, to reduce the high dimensionality of variables, before conducting the present analyses, a psychiatrist (JYY) carefully reviewed all variables included in the survey and selected candidate variables potentially associated with SI. Variables regarding suicide attempts were excluded from the study because they were directly associated with SI. Based on the consensus between a psychiatrist (JYY) and family medicine physician (BJO), 48 candidate variables were selected. Primarily, the chosen variables included participants' demographics, anthropometric features, socioeconomic status, lifestyle or health behavior variables, and systemic health conditions, including comorbidities and blood analysis results.

Demographic characteristics included age, sex, residency, marital status, and educational status. Socioeconomic variables such as occupation, personal monthly income adjusted for the number of family members, and family income were also included. Anthropometric variables included height, waist circumference, and body mass index (BMI). Health behavior factors included drinking habit, smoking, physical activity, and mental health related items such as stress and depression. High-risk alcohol consumption was defined as drinking ≥7 glasses of alcohol for men or ≥5 glasses of alcohol for women on one drinking day.[24] Alcohol dependence was evaluated using the score on the 10-item Alcohol Use Disorder Identification Test (AUDIT).[25] Physical activity was evaluated in terms of engaging in strength or flexibility exercise one or more days per week.

Stress awareness was classified into the following 4 grades: 1, very much; 2, much; 3, less; 4, rarely. Depression was determined based on whether they have felt sad or desperate continuously for ≥2 weeks during last year. Systemic comorbidities for common diseases such as hypertension, diabetes mellitus (DM), depressive disorder, or cancer were also investigated. Blood test results for complete blood count, liver function, kidney function, and coagulation panel were included.

Data analysis

Building training and testing datasets

Before constructing the training and test datasets, the whole dataset was examined for missing data. Participants who had any missing value for independent variables or SI assessment were all excluded from the study. Next, the 2010−2013 and 2015 KNHANES datasets were divided into the following mutually exclusive sub-datasets: the training dataset and the test dataset. The 2010−2013 KNHANES dataset was used as the training dataset, and the 2015 KNHANES dataset was used as the test dataset to evaluate the final performance of different machine learning models that were built using the training dataset. Preprocessing of data and development of machine learning models were performed using the Weka software (Waikato Environment for Knowledge Analysis, version 3.8.1., University of Waikato, New Zealand) and R version 3.6.2 (the R Foundation for Statistical Computing, Vienna, Austria).

The training and test datasets were unbalanced for the target outcome SI because the proportion of participants reporting SI was around 10% in each year. As a classification bias arising from data imbalance was expected,[26] the training dataset was re-balanced in terms of SI endowing calculated weights that made the SI and non-SI groups equal. The ClassBalancer function in the Weka software, a supervised instance filter, was used for this purpose. As a balancing method, down-sampling for the majority class may be an option, but it has the disadvantage of losing data of the majority class. Over-sampling replicates the copies of the minority class, or adds more weight to the minority class. The ClassBlancer method might be similar to over-sampling in the context of adjusting learning weight and not losing the data of the majority class. In contrast, the testing dataset was not re-balanced and was used in its original form. Data normalization was done for numerical data of the training and test datasets, rescaling the features to the rage of 0 to 1.

Feature selection

To build machine learning models, feature selection was performed using the wrapper-based feature subset evaluation method in Weka. The best-first search method with forward selection, the area under the receiver-operating characteristic (ROC) curve as a performance measure, 5-fold cross-validation, and the seed value of 1 were selected. We adopted the Bayesian network (BN), LogitBoost with logistic regression (LB), support vector machine (SVM), decision tree (DT) of J48, and artificial neural network (ANN) using multilayer perceptron algorithms in Weka. SVM was trained using the sequential minimal optimization algorithm with a linear polynomial kernel. ANN was used as a multilayer perceptron with 3 hidden layers and a learning rate of 0.01, which is a feedforward ANN. For each algorithm, a feature subset maximizing the area under the ROC curve (AUC) of the algorithm was obtained. Additionally, we used a logistic regression (LR)-based model as a reference conventional statistical method.[15,27] For this model, the backward stepwise elimination method with Akaike Information Criterion was adopted to select associated vari-

ables using the R software. The optimal cutoff point in the ROC curve was determined using the Epi package in the R software.

Construction of machine learning models

After obtaining an optimal feature subset for each algorithm, sub-datasets were created for each machine learning algorithm from the training dataset, including only the optimal feature subsets. With each training sub-dataset BN, LB, SVM, DT, and ANN machine learning models were constructed. For each algorithm, the model maximizing the AUC was adopted.

Performance evaluation

The prediction models were evaluated using the testing set. The performance of each models was compared with the others using AUC. The prediction accuracy for SI, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were also calculated.

RESULTS

Characteristics of the participants

Among the eligible 51,214 participants, 40,932 participated in the health examination and answered the questionnaire in the 2010–2013 and 2015 KNHANES (79.9% response rate). Ultimately, 16,437 participants aged ≥19 years were included in the training dataset. The test dataset comprised 3,788 cases. The demographic characteristics of the participants have been presented in Table 1. Participants who reported the presence of SI tended to be older, female, and in worse economic and educational status (p<0.001 for all) as compared to those in the non-SI group. Additionally, those with SI had higher AUDIT scores; lesser frequency of physical exercise; poor subjective health status; and higher prevalence of systemic hypertension, stroke, diabetes mellitus, renal failure, liver cirrhosis, depressive disorder, and injury within the past 1 year.

The prevalence of SI was 11.9% (1,957 of 16,437) in the training set consisted of the 2010–2013 KNHANES datasets. On the other hand, the prevalence of SI was 5.5% (210 of 3,788) in the 2015 KNHANES dataset used as the test dataset in this study.

Selected features from the machine learning models

The features selected to build a prediction model by the wrapper-based evaluation for each algorithm have been presented in Table 2. Each algorithm adopted 6 to 21 features to construct the predictive model. Decision tree algorithm selected the least number (n=6) of features and LB extracted the greatest number (n=20) of associated features which is about twice as the number of features selected in the conventional LR model (n=11).

The common variables selected unanimously in all the models were the presence of depressive disorder, stress awareness, continuous depressive mood lasting ≥2 weeks, and Euro-QoL-5D (EQ-5D) score. Household income, sex, unmet medical service needs, and AUDIT score were adopted by five algorithms except one.

Predictive performance of the machine learning models

The performance metrics of the models for the prediction of SI have been presented in Table 3. The conventional logistic regression analysis showed an AUC of 0.867. Prediction sensitivity and specificity were 79.0% and 78.5%, respectively. Of the machine learning models, the LB and ANN models showed the best performance on predicting the presence of SI (AUC, 0.877). The sensitivity of the LB model was 81.0% and its specificity was 78.7%. Otherwise, the BN model showed a similar AUC compared to the LR model (AUC, 0.867).

Figure 2 shows the decision tree suggested by the DT algorithm. The determinant at the highest branch was continuous depressive mood lasting ≥2 weeks; 88.4% of the subjects with this feature had SI. The determinants in the following steps included EQ-5D score, stress awareness, and the presence of depressive disorder, respectively.

DISCUSSION

The present study investigated the health-related risk factors for SI in Korean adults applying machine learning algorithms to the nationally representative KNHANES data. Common risk factors useful for recognizing the SI in all the machine learning models were stress awareness, sustained depressive mood more than 2 weeks, and quality of life associated with general health status (EQ-5D score). In addition, the prevalence of depressive disorder, status of household income, sex, unmet medical service needs, and alcohol use problem (AUDIT score) were adopted by five machine learning models developed in the current study. The current study showed a better performance compared to previous machine-learning based studies, reaching an AUC of 0.877.

Depressive mood: a risk factor and possible mediator between demographic/physical health-related factors and SI

Our results suggest that the experience of continuous depressive mood is a common risk factor for SI in all models and that depressive disorder appears as a significant risk factor in most of the models. These results are in accordance with previous studies that suggested depression is significantly associated with SI, showing an increased suicidal risk as high as 3.73

Table 1. Characteristics of the participants in the training dataset

	Total subjects (N=16,437)	Without suicidal ideation (N=14,480)	With suicidal ideation (N=1,957)	p*
Age, year	49.9±15.8	49.4±15.6	53.4±16.7	<0.001
Sex, male	7,293 (44.4)	6,646 (45.9)	647 (33.1)	<0.001
Residency, urban	13,113 (79.8)	11,639 (80.4)	1,474 (75.3)	<0.001
Household income, quartile				<0.001
Lowest	2,955 (18.0)	2,346 (16.2)	609 (31.1)	
2nd	4,275 (26.0)	3,738 (25.8)	537 (27.4)	
3rd	4,554 (27.7)	4,114 (28.4)	440 (22.5)	
Highest	4,653 (28.3)	4,282 (29.6)	371 (19.0)	
Educational status				<0.001
≤Elementary school	3,716 (22.6)	2,965 (20.5)	751 (38.4)	
Middle school	1,807 (11.0)	1,571 (10.8)	236 (12.1)	
High school	5,725 (34.8)	5,171 (35.7)	554 (28.3)	
≥University graduate	5,189 (31.6)	4,773 (33.0)	416 (21.3)	
Family members	3.2±1.3	3.2±1.2	3.0±1.3	<0.001
Governmental life support	958 (5.8)	757 (5.2)	201 (10.3)	<0.001
Marital status				<0.001
Not married	1,851 (11.3)	1,655 (11.4)	196 (10.0)	
Married, live together	12,773 (77.7)	11,389 (78.7)	1,384 (70.7)	
Married, separated	105 (0.6)	84 (0.6)	21 (1.1)	
Bereavement	1,208 (7.3)	959 (6.6)	249 (12.7)	
Divorced	500 (3.0)	393 (2.7)	107 (5.5)	
High risk drinking	1,796 (10.9)	1,577 (10.9)	219 (11.2)	0.719
AUDIT score	5.7±6.4	5.7±6.3	6.1±7.5	0.015
Smoking				0.004
Never smoker	9,353 (56.9)	8,216 (56.7)	1,137 (58.1)	
Ex-smoker	3,718 (22.6)	3,330 (23.0)	388 (19.8)	
Current smoker	3,366 (20.5)	2,934 (20.3)	432 (22.1)	
Physical exercise	9,535 (58.0)	8,595 (59.4)	940 (48.0)	<0.001
Subjective health status				<0.001
Very good	746 (4.5)	698 (4.8)	48 (2.5)	
Good	4,768 (29.0)	4,452 (30.7)	316 (16.1)	
Normal	7,874 (47.9)	7,074 (48.9)	800 (40.9)	
Bad	2,503 (15.2)	1,931 (13.3)	572 (29.2)	
Very bad	546 (3.3)	325 (2.2)	221 (11.3)	
Hypertension	3,514 (21.4)	2,963 (20.5)	551 (28.2)	<0.001
Stroke	325 (2.0)	259 (1.8)	66 (3.4)	<0.001
MI or angina	428 (2.6)	347 (2.4)	81 (4.1)	<0.001
OA or RA	1,956 (11.9)	1,532 (10.6)	424 (21.7)	<0.001
DM	1,301 (7.9)	1,098 (7.6)	203 (10.4)	<0.001
Retinal failure	79 (0.5)	58 (0.4)	21 (1.1)	0.001
Liver cirrhosis	56 (0.3)	44 (0.3)	12 (0.6)	0.046
Thyroid disease	610 (3.7)	514 (3.5)	96 (4.9)	0.004
Asthma	496 (3.0)	401 (2.8)	95 (4.9)	<0.001

Table 1. Characteristics of the participants in the training dataset (continued)

	Total subjects (N=16,437)	Without suicidal ideation (N=14,480)	With suicidal ideation (N=1,957)	p*
Atopic dermatitis	365 (2.2)	311 (2.1)	54 (2.8)	0.101
Depressive disorder	663 (4.0)	403 (2.8)	260 (13.3)	<0.001
Cancer	441 (2.7)	385 (2.7)	56 (2.9)	0.655
Injury within 1 year	1,132 (6.9)	946 (6.5)	186 (9.5)	<0.001
Unmet need for medical service	2,670 (16.2)	2,035 (14.1)	635 (32.4)	<0.001
Stress awareness				<0.001
Very much	663 (4.0)	380 (2.6)	283 (14.5)	
Much	3,428 (20.9)	2,592 (17.9)	836 (42.7)	
Little	9,763 (59.4)	9,025 (62.3)	738 (37.7)	
Very little	2,583 (15.7)	2,483 (17.1)	100 (5.1)	
Continuous depressive mood ≥2 weeks	2,035 (12.4)	1,075 (7.4)	960 (49.1)	<0.001
Psychiatric consult within 1 year	322 (2.0)	167 (1.2)	155 (7.9)	<0.001
Absence from work within 1 month	535 (3.3)	401 (2.8)	134 (6.8)	<0.001
EQ-5D score	0.9±0.1	1.0±0.1	0.9±0.2	<0.001
Height	162.6±8.9	162.9±8.8	160.0±8.7	<0.001
Weight	62.8±11.5	63.0±11.5	60.9±11.4	<0.001
BMI	23.7±3.4	23.6±3.3	23.7±3.7	0.303
Waist circumference	81.0±9.9	81.0±9.8	81.4±10.6	0.064

*comparison with no suicidal ideation group, not adjusted for any covariate. MI: myocardial infarction, DM: diabetes mellitus, OA: osteoarthritis, RA: rheumatic arthritis, BMI: body mass index

times in the elderly population with comorbid depressive disorder or 2.17 times in the university students with depressive symptom, as compared to those without.[28,29] Public psychoeducational programs for depression have been found effective in reducing the incidence of suicide attempts.[30-32]

In terms of the mediating factors between these demographic or physical health-related variables, stress cognition, and depressive mood have been identified as some of the most influential factors in SI. Previous researches have reported that various demographic risk factors increase suicidality, including older age,[33] being an unmarried male,[34] and having lower economic status.[35] In a meta-analysis, alcohol abuse was also reported as a significant risk factor for SI [odds ratio (OR)= 1.86] as well as for suicide attempts and completed suicide.[36] In addition, poor health-related quality of life has also been associated with SI or suicidal behavior in many studies.[37,38] Physical illness and functional disability were other factors associated with suicide.[39] Similarly, the various machine learning algorithms adopted in the current study, including SVM, ANN, and BN, disclosed several risk factors that were reported in previous studies. Although some variables, such as smoking or physical exercise, were determined as risk factors only in part of the algorithms, the risk factors identified in this study could provide medical explanations in terms of the association of these variables with SI.

Better performance in recognizing the SI compared to other machine learning studies

The current study included the general population regardless of the mood status, but showed a better performance compared to previous machine-learning based studies, reaching an AUC of 0.877. Thus far, many studies have tried to predict the risk of SI using various methods. In the American general population with depressive mood, a prediction model using logistic regression showed an AUC of 0.809 (95% CI, 0.779−0.840) in the validation dataset.[40] Specifically, Liu et al.[41] modified the conventional logistic regression method by shrinking the coefficients with heuristics, to prevent the overfitting of the model and to improve model performance. In another study, a prediction model for SI was built using about 15 variables in Chinese patients with depression.[42] Specifically, Fang et al.[42] used a stepwise logistic regression analysis and the AUC of the model was 0.80 (95% CI, 0.78−0.81), which was lower than that observed in the present study.

The superior performance achieved in the current study for recognizing the presence of SI in general population might be attributed to larger number of input variables or different algorithms of the machine learning-based approach, and to the large sample size. Most of the machine learning algorithms tested in our study showed performances comparable to that of the LR-based model, using fewer variables. On the contrary, the

tested machine learning algorithms did not surpass the performance of the simple LR-based model. The reason for this is not clear thus far, but the association of risk factors, such as the EQ-5D score or the presence of continuous depressive mood with SI, could have a sigmoid-like pattern, thus leading to the

better performance of the LR-based model. Collectively, simpler models built by machine learning algorithms may help to understand the risk factors for SI more clearly, and they may aid the evaluation of the SI risk of an individual more easily.

Study implication: how to recognize the patients with higher risks of SI at primary medical institutions

Considering the under-diagnosis of SI in primary care settings and the low rate of consultation with psychiatrists,[43] the present study might help primary care physicians to identify individuals who are at a higher risk of SI and to initiate timely co-working with psychiatrists.[44] A previous study found that those who had attempted suicide had more than twice the rate of visiting non-mental health facilities before such an attempt.[45] This can be interpreted as the social prejudice against psychiatric illness and treatment, rather than as a cause of depression and suicidal thoughts. Therefore, the screening and intervention of suicidal risk at primary medical institutions may be very important for patients with or without depressive disorder. Active interactions with patients requiring mental health care services and removing trigger factors for suicidal behavior are also important. As shown in the present study, interventions for smoking cessation and physical exercise related health education, and reducing stress and depression could be helpful for preventing suicide among elderly individuals.

Study limitations

This study has some limitations that need to be considered. First, this study used one question to evaluate SI that required the participants to respond as "yes" or "no" to indicate its "presence" or "absence." The related paucity of more quantitative data about the severity of SI might have restricted a more detailed evaluation of the validity and reliability of the predictive models developed in this study. Second, the measurement of

Table 2. Selected features associated with suicidal ideation using machine learning algorithms

Features	Machine learning algorithm					
	BN	LB	SVM	DT	ANN	LR
Sex	O	O	O		O	O
Age		O			O	
Household income	O	O	O		O	O
Education	O	O	O			O
Marital status		O				
Occupation			O			
High risk drinking	O					
AUDIT score	O	O	O		O	O
Smoking		O	O			
Physical exercise	O	O	O		O	
Sleep duration			O			
Subjective health status		O	O		O	O
Stroke	O					
Renal failure	O	O		O		
Liver cirrhosis	O			O	O	
OA or RA		O	O			
Atopic dermatitis			O			
Depressive disorder	O	O	O	O	O	O
Cancer			O			
Injury within 1 year	O	O				
Unmet need for medical service	O	O	O		O	O
Stress awareness	O	O	O	O	O	O
Continuous depressive mood ≥2 weeks	O	O	O	O	O	O
Psychiatric consult within 1 year	O	O			O	O
Absence from work within 1 month	O	O	O		O	
EQ-5D score	O	O	O	O	O	O
Height					O	
BMI	O					
Systolic blood pressure			O			
Frequency of eating out			O			

BN: Bayesian network, LB: LogitBoost with logistic regression, SVM: support vector machine, DT: decision tree, ANN: artificial neural network, LR: logistic regression, MI: myocardial infarction, OA: osteoarthritis, RA: rheumatic arthritis, BMI: body mass index, EQ-5D: Euro-QoL-5D

Table 3. Performance of the prediction model for suicidal ideation using machine learning algorithms

	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	Positive PV	Negative PV
		Machine learning algorithm				
BN	0.867	75.6	81.9	75.2	16.2	98.6
LB	0.877	78.8	81.0	78.7	18.2	98.6
SVM	0.794	81.0	77.6	81.2	19.5	98.4
DT	0.843	71.9	81.0	71.3	14.2	98.5
ANN	0.877	77.1	81.4	76.8	17.1	98.6
LR	0.867	78.5	79.0	78.5	17.8	98.5

AUC: area under the receiver operating characteristic curve, PV: predictive value, BN: Bayesian network, LB: LogitBoost with logistic regression, SVM: support vector machine, DT: decision tree, ANN: artificial neural network, LR: logistic regression
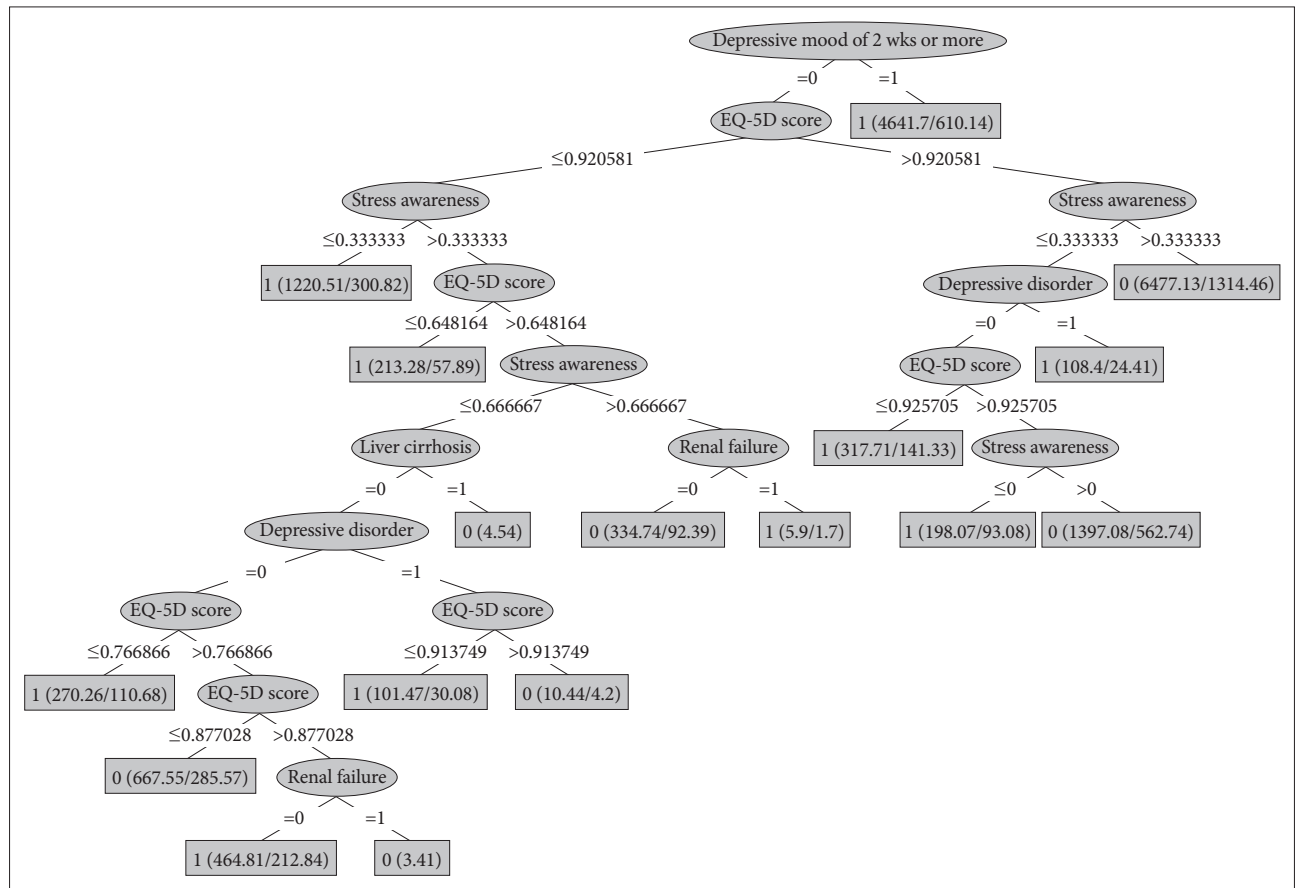
Figure 2. Decision tree to predict suicidal ideation. EQ-5D: Euro-QoL-5D.

SI through a self-report questionnaire was not accompanied by a more facilitative clinical interview with a clinician (as in the database used in this study). This may have led to higher chances of false-negative predictions of SI. Third, there was a discrepancy for the prevalence of SI in the training dataset and the test dataset. The lower prevalence of the minority class in the test dataset might have increased the performance of machine-learning algorithms. Fourth, direct comparison for the prediction performance using the AUC value between LR and other machine learning models may not be appropriate because machine learning models were developed using the Weka, while the LR model was developed using the R software. However, considering the amount of time that would be required to acquire this massive amount of data through face-to-face interviews, the strategy of using the KNHANES data—a nationwide representative survey of the adult population in South Korea—could have been a more favorable option.

Conclusions

The current study demonstrated the clinical utility of the machine learning approach in providing performance for SI prediction that was comparable to that of the conventional LR-based model. Our machine learning-based predictive models for SI might help primary care physicians to identify patients at a higher risk of experiencing SI, thereby facilitating the early and active implementation of preventive interventions, including psychiatric consultation and modification of suicidal risk-related medical conditions. The present results might be helpful in designing and implementing suicide prevention programs effectively.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Bumjo Oh, Je-Yeon Yun, Jin Kim, Bum-Joo Cho. Data curation: Je-Yeon Yun, Eun Chong Yeo, Bum-Joo Cho. Formal analysis: Eun Chong Yeo, Dong-Hoi Kim, Bum-Joo Cho. Funding acquisition:

Bum-Joo Cho. Investigation: Bumjo Oh, Je-Yeon Yun, Bum-Joo Cho. Methodology: Dong-Hoi Kim, Jin Kim, Bum-Joo Cho. Project administration: Jin Kim, Bum-Joo Cho. Resources: Dong-Hoi Kim, Jin Kim, Bum-Joo Cho. Software: Eun Chong Yeo, Bum-Joo Cho. Supervision: Dong-Hoi Kim, Jin Kim, Bum-Joo Cho. Validation: Bumjo Oh, Bum-Joo Cho. Visualization: Eun Chong Yeo, Bum-Joo Cho. Writing—original draft: Bumjo Oh, Je-Yeon Yun, Bum-Joo Cho. Writing—review & editing: Bumjo Oh, Je-Yeon Yun, Bum-Joo Cho.

ORCID iDs

Bumjo Oh	https://orcid.org/0000-0002-2468-0755
Je-Yeon Yun	https://orcid.org/0000-0002-5531-2410
Eun Chong Yeo	https://orcid.org/0000-0002-3809-1388
Dong-Hoi Kim	https://orcid.org/0000-0001-9929-5408
Jin Kim	https://orcid.org/0000-0002-2015-3624
Bum-Joo Cho	https://orcid.org/0000-0002-0244-388X

REFERENCES

1. Neeleman J, de Graaf R, Vollebergh W. The suicidal process; prospective comparison between early and later stages. J Affect Disord 2004; 82:43-52.
2. Borges G, Nock MK, Haro Abad JM, Hwang I, Sampson NA, Alonso J, et al. Twelve-month prevalence of and risk factors for suicide attempts in the World Health Organization World Mental Health Surveys. J Clin Psychiatry 2010;71:1617-1628.
3. Brailovskaia J, Forkmann T, Glaesmer H, Paashaus L, Rath D, Schonfelder A, et al. Positive mental health moderates the association between suicide ideation and suicide attempts. J Affect Disord 2018;245:246-249.
4. Gvion Y, Apter A. Aggression, impulsivity, and suicide behavior: a review of the literature. Arch Suicide Res 2011;15:93-112.
5. Bergfeld IO, Mantione M, Figee M, Schuurman PR, Lok A, Denys D. Treatment-resistant depression and suicidality. J Affect Disord 2018; 235:362-367.
6. Ten Have M, van Dorsselaer S, de Graaf R. Prevalence and risk factors for first onset of suicidal behaviors in the Netherlands Mental Health Survey and Incidence Study-2. J Affect Disord 2013;147:205-211.
7. de Heer EW, Ten Have M, van Marwijk HWJ, Dekker J, de Graaf R, Beekman ATF, et al. Pain as a risk factor for suicidal ideation. A population-based longitudinal cohort study. Gen Hosp Psychiatry 2018 [Epub ahead of print].
8. Holden KB, Bradford LD, Hall SP, Belton AS. Prevalence and correlates of depressive symptoms and resiliency among African American women in a community-based primary health care center. J Health Care Poor Underserved 2013;24:79-93.
9. Landa A, Skritskaya N, Nicasio A, Humensky J, Lewis-Fernandez R. Unmet need for treatment of depression among immigrants from the former USSR in the US: a primary care study. Int J Psychiatry Med 2015;50:271-289.
10. Lehmann M, Kohlmann S, Gierk B, Murray AM, Lowe B. Suicidal ideation in patients with coronary heart disease and hypertension: Baseline results from the DEPSCREEN-INFO clinical trial. Clin Psychol Psychother 2018;25:754-764.
11. Strupp J, Ehmann C, Galushko M, Bucken R, Perrar KM, Hamacher S, et al. Risk factors for suicidal ideation in patients feeling severely affected by multiple sclerosis. J Palliat Med 2016;19:523-528.
12. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive big data analytics: a study of parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. PLoS One 2016;11:e0157077.
13. Zhang Y, Guo SL, Han LN, Li TL. Application and exploration of big data mining in clinical medicine. Chin Med J (Engl) 2016;129:731-738.
14. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2:230-

243.
15. Kim SK, Yoo TK, Oh E, Kim DW. Osteoporosis risk prediction using machine learning and conventional methods. Conf Proc IEEE Eng Med Biol Soc 2013;2013:188-191.
16. Lee SK, Son YJ, Kim J, Kim HG, Lee JI, Kang BY, et al. Prediction model for health-related quality of life of elderly with chronic diseases using machine learning techniques. Healthc Inform Res 2014;20:125-134.
17. Jung JS, Park SJ, Kim EY, Na KS, Kim YJ, Kim KG. Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. PLoS One 2019;14:e0217639.
18. Colic S, J Richardson D, James Reilly P, Gary Hasey M. Using machine learning algorithms to enhance the management of suicide ideation. Conf Proc IEEE Eng Med Biol Soc 2018;2018:4936-4939.
19. Carson NJ, Mullin B, Sanchez MJ, Lu F, Yang K, Menezes M, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PLoS One 2019;14:e0211116.
20. Gong J, Simon GE, Liu S. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. PLoS One 2019;14:e0222665.
21. Burke TA, Ammerman BA, Jacobucci R. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. J Affect Disord 2019;245:869-884.
22. Chun MY, Cho BJ, Yoo SH, Oh B, Kang JS, Yeon C. Association between sleep duration and musculoskeletal pain: The Korea National Health and Nutrition Examination Survey 2010-2015. Medicine (Baltimore) 2018;97:e13656.
23. Yoon KC, Choi W, Lee HS, Kim SD, Kim SH, Kim CY, et al. An overview of ophthalmologic survey methodology in the 2008-2015 Korean National Health and Nutrition Examination Surveys. Korean J Ophthalmol 2015;29:359-367.
24. Witkiewitz K, Hallgren KA, Kranzler HR, Mann KF, Hasin DS, Falk DE, et al. Clinical validation of reduced alcohol consumption after treatment for alcohol dependence using the World Health Organization risk drinking levels. Alcohol Clin Exp Res 2017;41:179-186.
25. Reinert DF, Allen JP. The Alcohol Use Disorders Identification Test (AUDIT): a review of recent research. Alcohol Clin Exp Res 2002;26:272-279.
26. Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: a review. GESTS Int T Comput Sci Eng 2006;30:25-36.
27. Hsieh CH, Lu RH, Lee NH, Chiu WT, Hsu MH, Li YC. Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. Surgery 2011; 149:87-93.
28. Awata S. Prevention of suicide in the elderly. Seishin Shinkeigaku Zasshi 2005;107:1099-1109.
29. Wang YH, Shi ZT, Luo QY. Association of depressive symptoms and suicidal ideation among university students in China: A systematic review and meta-analysis. Medicine (Baltimore) 2017;96:e6476.
30. Hegerl U, Althaus D, Stefanek J. Public attitudes towards treatment of depression: effects of an information campaign. Pharmacopsychiatry 2003;36:288-291.
31. Jorm AF, Christensen H, Griffiths KM. Public beliefs about causes and risk factors for mental disorders: changes in Australia over 8 years. Soc Psychiatry Psychiatr Epidemiol 2005;40:764-767.
32. Paykel ES, Hart D, Priest RG. Changes in public attitudes to depression during the Defeat Depression Campaign. Br J Psychiatry 1998;173:519-522.
33. Lee H, Seol KH, Kim JW. Age and sex-related differences in risk factors for elderly suicide: Differentiating between suicide ideation and attempts. Int J Geriatr Psychiatry 2018;33:e300-e306.
34. Kyung-Sook W, SangSoo S, Sangjin S, Young-Jeon S. Marital status integration and suicide: A meta-analysis and meta-regression. Soc Sci Med 2018;197:116-126.
35. Ki M, Seong Sohn E, An B, Lim J. Differentiation of direct and indirect

socioeconomic effects on suicide attempts in South Korea. Medicine (Baltimore) 2017;96:e9331.

36. Darvishi N, Farhadi M, Haghtalab T, Poorolajal J. Alcohol-related risk of suicidal ideation, suicide attempt, and completed suicide: a meta-analysis. PLoS One 2015;10:e0126870.

37. Goldney RD, Fisher LJ, Wilson DH, Cheok F. Suicidal ideation and health-related quality of life in the community. Med J Aust 2001;175:546-549.

38. Kim JH, Kwon JW. The impact of health-related quality of life on suicidal ideation and suicide attempts among Korean older adults. J Gerontol Nurs 2012;38:48-59.

39. Fassberg MM, Cheung G, Canetto SS, Erlangsen A, Lapierre S, Lindner R, et al. A systematic review of physical illness, functional disability, and suicidal behaviour among older adults. Aging Ment Health 2016;20:166-194.

40. Liu Y, Sareen J, Bolton JM, Wang JL. Development and validation of a risk prediction algorithm for the recurrence of suicidal ideation among general population with low mood. J Affect Disord 2016;193:11-17.

41. Liu X, Liu X, Sun J, Yu NX, Sun B, Li Q, et al. Proactive Suicide Prevention Online (PSPO): machine identification and crisis management for Chinese social media users with suicidal thoughts and behaviors. J Med Internet Res 2019;21:e11705.

42. Fang X, Zhang C, Wu Z, Peng D, Xia W, Xu J, et al. Prevalence, risk factors and clinical characteristics of suicidal ideation in Chinese patients with depression. J Affect Disord 2018;235:135-141.

43. Choi YJ, Lee WY. The prevalence of suicidal ideation and depression among primary care patients and current management in South Korea. Int J Ment Health Syst 2017;11:18.

44. Park SW, Lee JH, Lee EK, Song JJ, Park HS, Hwang SY, et al. Development of the Suicide Risk Scale for Medical Inpatients. J Korean Med Sci 2018;33:e18.

45. McDowell AK, Lineberry TW, Bostwick JM. Practical suicide-risk management for the busy primary care physician. Mayo Clin Proc 2011;86:792-800.