Detection of Suicide Attempters among Suicide Ideators Using Machine Learning

Seunghyong Ryu[1] ✉, Hyeongrae Lee[1], Dong-Kyun Lee[1], Sung-Wan Kim[2], and Chul-Eung Kim[3]

[1]Department of Mental Health Research, National Center for Mental Health, Seoul, Republic of Korea
[2]Department of Psychiatry, Chonnam National University Medical School, Gwangju, Republic of Korea
[3]Mental Health Research Institute, National Center for Mental Health, Seoul, Republic of Korea

Objective   We aimed to develop predictive models to identify suicide attempters among individuals with suicide ideation using a machine learning algorithm.

Methods   Among 35,116 individuals aged over 19 years from the Korea National Health & Nutrition Examination Survey, we selected 5,773 subjects who reported experiencing suicide ideation and had answered a survey question about suicide attempts. Then, we performed resampling with the Synthetic Minority Over-sampling TEchnique (SMOTE) to obtain data corresponding to 1,324 suicide attempters and 1,330 non-suicide attempters. We randomly assigned the samples to a training set (n=1,858) and a test set (n=796). In the training set, random forest models were trained with features selected through recursive feature elimination with 10-fold cross validation. Subsequently, the fitted model was used to predict suicide attempters in the test set.

Results   In the test set, the prediction model achieved very good performance [area under receiver operating characteristic curve (AUC)=0.947] with an accuracy of 88.9%.

Conclusion   Our results suggest that a machine learning approach can enable the prediction of individuals at high risk of suicide through the integrated analysis of various suicide risk factors.   **Psychiatry Investig 2019;16(8):588-593**

Key Words   Suicide attempt, Suicide ideation, Machine learning, Public health data.

INTRODUCTION

Suicide is considered one of the most serious and urgent public health and social issues in Korea.[1] The suicide rate in Korea is the highest among the Organization for Economic Cooperation and Development (OECD) countries, with about 36 people taking their own lives every day.[2] Korea also has a higher rate of suicide behaviors compared with other countries.[3] According to a 2016 survey of mental disorders in Korea, 15.4% of Koreans have thought about suicide and 2.4% have attempted suicide at some point in their lives.[4] Moreover, it is estimated that 2.9% of Koreans have experienced suicide ideation and 0.1% have attempted suicide at least once in the

previous year.[4] In recent years, other national and community surveys have estimated the 1-year prevalence of suicide ideation to be 5.1−9.0%, and that of suicide attempt to be 0.4−1.4%.[5,6]

Several known socio-demographic, physical, and psychological factors influence suicide mortality.[3] Suicide behaviors including suicide ideation and attempts are regarded as a major predictor of death by suicide. Even though individuals who think about suicide do not all subsequently commit suicide, people experiencing persistent and severe suicide ideation are at increased risk of attempting suicide.[7,8] Moreover, a history of suicide attempts is a strong predictor of future attempts and completed suicides.[9,10] Therefore, predicting the individuals who have engaged in suicide ideation or who have attempted suicide by screening the risk factors would be effective in preventing suicide.

Several risk factors have thus been identified for suicide behaviors, which need to be interpreted in an integrated way. Machine learning, a branch of artificial intelligence in which a computer generates predictive rules based on raw data, may provide a powerful tool to efficiently predict suicide risk and

implement suicide prevention strategies.[11,12] Some studies have attempted to predict suicide risk in a clinical setting using machine learning approaches. Passos et al.[13] distinguished suicide attempters from non-suicide attempters among patients with mood disorders with an accuracy of 64.7−72.0%, using machine learning algorithms based on demographic and clinical data. Oh et al.[14] classified individuals with a history of suicide attempts among patients with depression or anxiety disorders by applying artificial neural networks to multiple psychiatric scales and sociodemographic data, achieving an accuracy of 87.4−93.7%.

We originally attempted to develop machine learning models to predict suicide behaviors in the general population. However, the low prevalence of such behaviors in the general population could be an obstacle to the building of predictive models. Therefore, we planned to develop our models through a stepwise approach, progressing from low to high risk. In our previous study, we applied a machine learning algorithm to public health data and identified individuals experiencing suicide ideation among the general population with an accuracy of 78.1−82.1%.[15] Following our previous study, here we aimed to develop models to predict which individuals have a history of recent suicide attempts, and thus are at increased suicide risk, among those who have experienced suicide ideation. This phased approach to the prediction of suicide risk using machine learning models may efficiently screen individuals at high risk for suicide in the general population.

METHODS

Study population

This study was performed with data from the Korea National Health and Nutrition Examination Survey (KNHANES), which was conducted between 2007 and 2012 (total n=50,405). The KNHANES is a nationwide survey of the health and nutritional status of non-institutionalized civilians in Korea, and is conducted every year by the Korea Center for Disease Control and Prevention.[16] Each year, the survey uses a stratified and multistage probability sampling design to include a new sample of about 8,000 individuals. All KNHANES participants provide written consent to participate in the survey and for their personal data to be used.

Among the 38,005 individuals aged over 19 years, 35,116 subjects answered the following survey question about suicide ideation: "During the past year, have you ever felt that you were willing to die?" Among the 35,116 respondents, 5,814 (16.6%) reported experiencing suicide ideation (suicide ideators). Among them, 5,773 responded to the following survey question about suicide attempts: "Have you ever attempted suicide in the past year?" Only 331 (5.7%) of the 5,773 suicide ideators reported at-

tempting suicide (suicide attempters), while the remaining 5,442 (94.3%) denied any suicide attempt (non-suicide attempters).

The institutional review board of the National Center for Mental Health approved the protocol of this study (IRB approval number: 116271-2018-36).

Data preprocessing and set assignment

We manually selected 47 variables likely to be related to suicide risk. Subsequently, we imputed missing data with the Multiple Imputation by Chained Equations (MICE) method, and numeric data were normalized by z-scoring. The MICE is an imputation algorithm that works by running multiple regression models, and conditionally modeling each missing value depending on the observed values.[17,18]

Inputting all data into a classifier to build a learning model will usually lead to a learning bias towards the majority class, in this case non-suicide attempters (a phenomenon known as the "class imbalance problem").[19] Therefore, to obtain a more balanced dataset, we undersampled the non-suicide attempters by randomly selecting 1,330 of them, and oversampled the suicide attempters, increasing their number from 331 to 1,324, using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is one of the most popular methods for addressing class imbalance, and the general idea is to generate synthetic cases from the minority class using the information available in the data.[20,21] Thus, a total of 2,654 samples (1,324 suicide attempters and 1,330 non-suicide attempters) were finally included in this study. Then, we randomly assigned the 2,654 samples to a training set (n=1,858, 70%) and a test set (n=796, 30%), while preserving the ratio between the two classes (Figure 1).

Machine learning analysis

We used the random forest algorithm, which builds numerous classification trees in bootstrapped samples and generates an aggregate tree by averaging across them.[22] To select the smallest subset of features that most accurately classifies suicide attempters, we performed recursive feature elimination. For model development, 10-fold cross validation was used to avoid overfitting and to increase the generalization of the model. In 10-fold cross validation, data in the training set are partitioned into 10 equally sized folds, and each fold is used once as a validation set, while the other 9 folds are used for training (Figure 1). Hyperparameter optimization was performed using the grid search method.

The fitted model was then used to predict the classes in the test sets, and the predicted class was compared with the actual class. The model performance in predicting the classes was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC). We also calculated accuracy, sensitivity, specificity, positive predictive value, and negative pre-
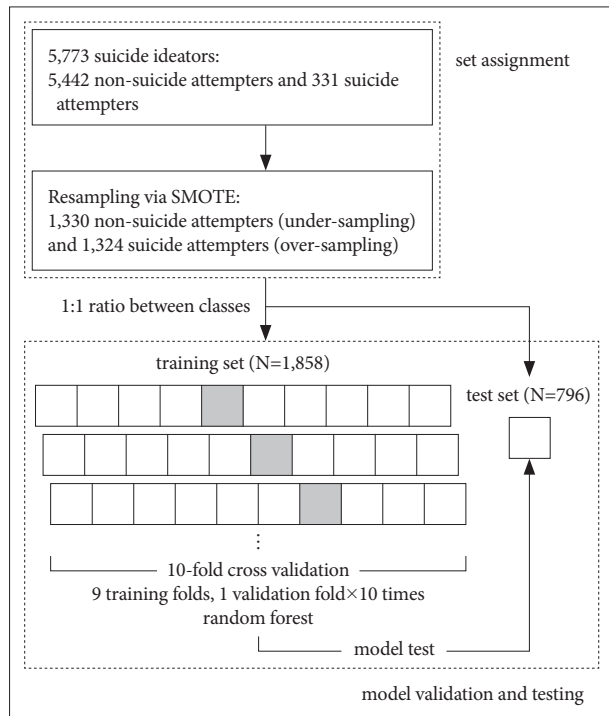
Figure 1. Schematic representation of the development of the prediction model. SMOTE: Synthetic Minority Over-sampling Technique.

dictive value from the confusion matrix. In addition, in order to confirm the generalized performance of the fitted model, we randomly divided the test set samples into 10 subsets, preserving the class ratio, and calculated the respective prediction scores. After 10 iterations of this process, the prediction scores of the 100 subsets were averaged.

All analyses were conducted in R (version 3.4.3, https://www.r-project.org/) and its packages, including "DMwR," "mice," and "caret."

RESULTS

Model training and validation

The feature selection process showed that a model trained with 41 features (Table 1) achieved the highest value of Kappa (0.705) (Figure 2). The top 10 features in order of importance were the following: "days of feeling sick or in discomfort," "Alcohol Use Disorders Identification Test (AUDIT) score," "amount of daily smoking," "average work week," "household composition," "EuroQoL-Visual Analogue Scale (VAS)," "age," "frequency of drinking," "number of household members," "depressed mood over two weeks". After model training and validation, the fitted model achieved an accuracy of 0.865 in the training set.

Model testing

In the whole test set, the fitted model showed very good

Table 1. Selected features (in order of decreasing importance)

1	Days of feeling sick or in discomfort
2	AUDIT score
3	Amount of daily smoking
4	Average work week
5	Household composition
6	EQ-VAS
7	Age
8	Frequency of drinking
9	Number of household members
10	Depressed mood over two weeks
11	Days of walking per week
12	Average sleep time
13	Level of education
14	Reasons for unemployment
15	Father's level of education
16	Amount of drinking
17	Days of moderate physical activity per week
18	Marriage stability
19	Stress level in daily life
20	Subjective body perception
21	Subjective health status
22	Mother's level of education
23	EQ-5D: anxiety/depression
24	National basic livelihood security
25	Type of health insurance
26	EQ-5D: usual activities
27	EQ-5D: pain/discomfort
28	Household income
29	Job position
30	Smoking preference
31	Weight change
32	Home ownership
33	EQ-5D: self-care
34	EQ-5D: mobility
35	Limitation of daily life and social activities
36	Feeling sick or in discomfort
37	Economic activity status
38	Being in bed sick in the last month
39	Arthritis
40	Sex
41	Hypertension

AUDIT: Alcohol Use Disorders Identification Test, EQ-5D: Euro-Qol-5D standardized instrument for use as a measure of health outcome, VAS: Visual Analogue Scale

performance (AUC=0.947) in predicting suicide attempters (Figure 3), with accuracy 0.889 {95% CI: 0.866−0.910, no information rate=0.501, p value [accuracy>no information rate (ACC>NIR)]<0.001}, sensitivity 0.859, specificity 0.920, positive predictive value 0.914, and negative predictive value 0.868 (Table 2). The model also showed very good performance in the 100 subsets used to assess its generalized performance, with mean AUC=0.947±0.001, mean accuracy 0.892±0.001, mean sensitivity 0.862±0.002, mean specificity 0.921±0.001, mean positive predictive value 0.918±0.002, and mean negative predictive value 0.872±0.002.

DISCUSSION

In this study, application of a machine learning algorithm to public health data achieved good performance and high accuracy in distinguishing individuals with a history of suicide attempt from those with suicide ideation without suicide attempt. This result suggests that machine learning approaches may be useful to detect suicide risk in the general population. In particular, using machine learning techniques, we selected several variables related to physical health, substance use, and socioeconomic status as important features for detecting suicide attempters. In the previous studies using classical statis-
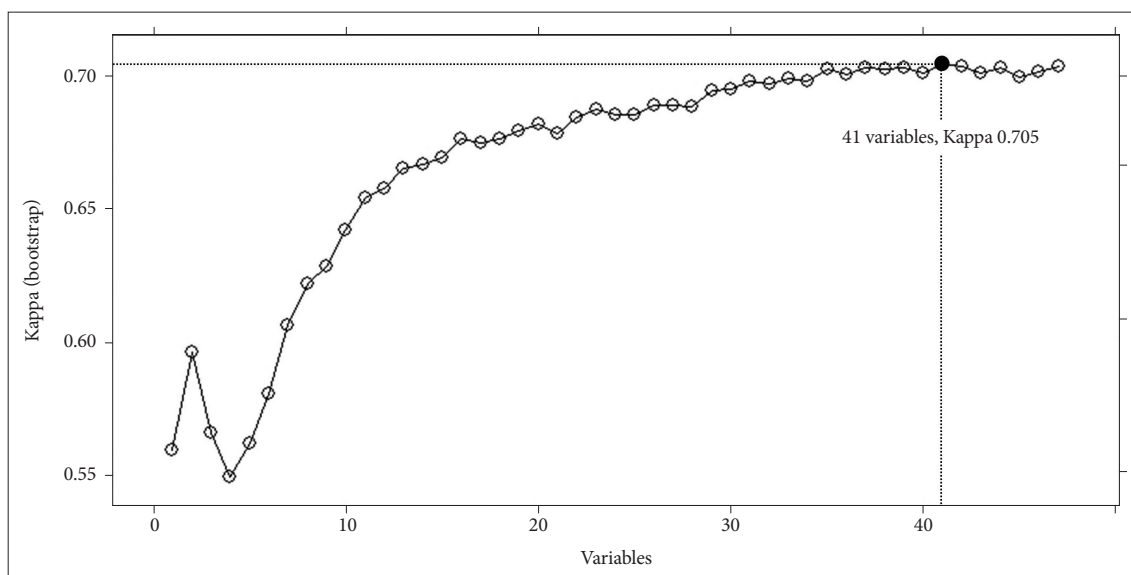

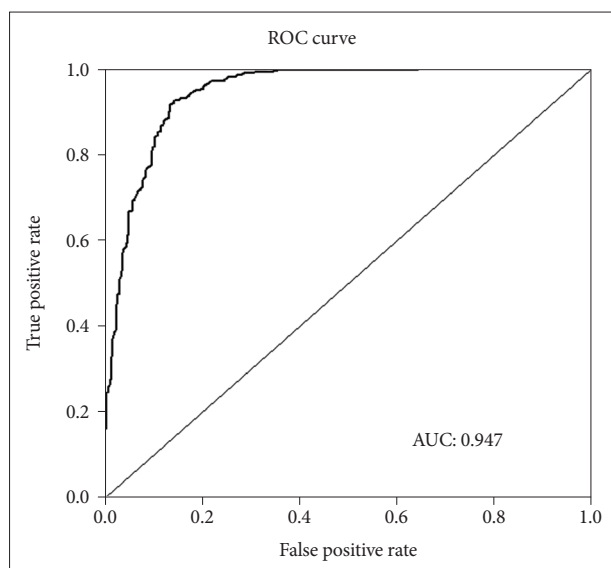**Figure 2.** A plot of feature selection by recursive feature elimination.

Figure 3. Receiver operating characteristic (ROC) curve. AUC: Area under ROC curve.

Table 2. Performance of model predicting suicide attempters in the test set (N=796)

	Confusion matrix		
	Actual class		Total
Predicted class	Suicide attempt +	Suicide attempt -	
Suicide attempt +	341	32	373
Suicide attempt -	56	367	423
Total	397	399	796
Prediction scores			
Accuracy*	0.889		
Sensitivity	0.859		
Specificity	0.920		
Positive predictive value	0.914		
Negative predictive value	0.868		

*95% CI: 0.866−0.910, p value (ACC>NIR)<0.001. CI: confidence interval, ACC: accuracy, NIR: no information rate

tical methods, these suicide risk factors have been explored individually. However, the present study showed that machine learning approaches enable to predict individuals at high risk for suicide by analyzing various suicide risk factors in an integrated way.

The 1-year prevalence rate of suicide attempt is rather low, less than 1%,[5,6] so it is very difficult to directly detect suicide attempters among the general population. However, those who experienced suicide ideation, especially in a mild or fleeting form, are relatively more common than those who attempted suicide. Moreover, the proportion of suicide attempters is higher in those who have experienced suicide ideation than in the general population. Therefore, we assumed that it would be more efficient to identify suicide ideators in the general population first, and then to classify them into suicide ideators and attempters. For this reason, we adopted a 2-step method to detect those who have suicide risk in the general population. In our previous work, we had already developed a machine learning model predicting suicide ideators in the general population with an accuracy of 78.1−82.1%.[15] As the next step, in this study, we developed models classifying suicide attempters and ideators by applying a machine learning algorithm to the KNHANES data. When predicting suicide attempters in the test sets, the machine learning model showed very good performance (AUC=0.947) with an accuracy of 88.9%.

This study used a resampling method different from the one we used in our previous study predicting suicide ideators. Also in this study we faced a class imbalance problem, which had to be handled prior to the development of the prediction model. As mentioned in the Methods section, the number of suicide attempters (n=331) in the data was very low compared with that of non-suicide attempters (n=5,442). Such an imbalance between two classes can lead to biased learning in favor of the majority class (non-suicide attempters). The problem can be attenuated by resampling methods, which produce class-balanced data. It is known that undersampling is generally helpful, while random oversampling is not.[20] Therefore, in order to balance the two classes, we used the SMOTE algorithm, which is an oversampling approach that creates synthetic minority class samples.[20,21] Undersampled non-suicide attempter data (n=1,330) and oversampled (partially synthetic) suicide attempter data (n=1,324) constituted the input used to train the models.

In this study, features for physical health (days of feeling sick or in discomfort, days of walking per week), substance use (AUDIT score, amount of daily smoking), and socioeconomic status (average work week, household composition) played an important role in classifying suicide attempters and suicide ideators. However, in our previous model predicting suicide ideators among the general population, features such as depressed mood, stress level, and quality of life showed greater importance.[15] This difference suggests that suicide ideation might be mainly induced by internal and psychological factors, while suicide attempt might be triggered by external and environmental factors.[23,24] In addition, the features we identified for the prediction of suicide attempters have been previously reported to be risk factors for suicide attempt in the Korean population.[25] In particular, it is noteworthy that variables related to alcohol use and smoking status, such as the AUDIT score and amount of daily smoking, were selected as highly important features in our prediction model. A recent study analyzing the KNHANES data reported that the combination of alcohol use and smoking was associated with greater suicide risk than alcohol or smoking separately.[26]

This study has some methodological limitations. First, the data from the KNHANES include information about suicide risk and psychological status obtained through very simple questions and scales, which might affect model performance. Second, we used only one machine learning algorithm, namely a random forest. Additional analyses are needed to compare the model performance with other machine learning algorithms, such as support vector machines and artificial neural networks. Third, our prediction model was built with class-balanced data including synthetic samples generated by a resampling algorithm. Further studies are needed to confirm the model's performance on actual data affected by biased class ratio.

In conclusion, our results demonstrate that machine learning models based on public health data can successfully detect individuals at high risk for suicide in the general population. Further studies are needed to apply our models to the prediction of individuals at high risk for suicide in clinical or community settings.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Seunghyong Ryu. Data curation: Hyeongrae Lee, Dong-Kyun Lee. Funding acquisition: Seunghyong Ryu. Methodology: Seunghyong Ryu, Hyeongrae Lee, Dong-Kyun Lee. Project administration: Seunghyong Ryu, Hyeongrae Lee. Supervision: Sung-Wan Kim, Chul-Eung Kim. Visualization: Dong-Kyun Lee. Writing—original draft: Seunghyong Ryu, Sung-Wan Kim. Writing—review & editing: Seunghyong Ryu.

ORCID iD

Seunghyong Ryu   https://orcid.org/0000-0001-6127-760X

REFERENCES

1. Kim SW, Yoon JS. Suicide, an urgent health issue in Korea. J Korean Med Sci 2013;28:345-347.
2. OECD. OECD Health Statistics 2017. Available from: https://stats.oecd.org/Index.aspx?DataSetCode=HEALTH_STAT. Accessed April 29, 2019.
3. Jeon HJ, Lee JY, Lee YM, Hong JP, Won SH, Cho SJ, et al. Lifetime prevalence and correlates of suicidal ideation, plan, and single and multiple attempts in a Korean nationwide study. J Nerv Ment Dis 2010;198:643-646.
4. Ministry of Health and Welfare. The Survey of Mental Disorders in Korea 2016. Sejong: Ministry of Health and Welfare; 2017.
5. Ju YJ, Jang SI. Suicide related indicators and trend of Korea in 2015. Health Policy Manag 2017;27:75-79.
6. Kim SW, Bae KY, Kim JM, Kim SY, Yoo JA, Yang SJ, et al. Psychosocial correlates of attempted suicide and attitudes toward suicide. J Korean Neuropsychiatr Assoc 2010;49:367-373.
7. May AM, Klonsky ED, Klein DN. Predicting future suicide attempts among depressed suicide ideators: a 10-year longitudinal study. J Psychiatr Res 2012;46:946-952.
8. Miranda R, Ortin A, Scott M, Shaffer D. Characteristics of suicidal ideation that predict the transition to future suicide attempts in adolescents. J Child Psychol Psychiatry 2014;55:1288-1296.
9. Bostwick JM, Pabbati C, Geske JR, McKean AJ. Suicide attempt as a risk factor for completed suicide: even more lethal than we knew. Am J Psychiatry 2016;173:1094-1100.
10. Bostwick JM, Pankratz VS. Affective disorders and suicide risk: a reexamination. Am J Psychiatry 2000;157:1925-1932.
11. Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, et al. Letter to the Editor: Suicide as a complex classification problem: machine learning and related techniques can advance suicide prediction-a reply to Roaldset (2016). Psychol Med 2016;46:2009-2010.
12. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. Am J Psychiatry 2018;175:951-960.
13. Passos IC, Mwangi B, Cao B, Hamilton JE, Wu MJ, Zhang XY, et al. Identifying a clinical signature of suicidality among patients with mood disorders: a pilot study using a machine learning approach. J Affect Disord 2016;193:109-116.
14. Oh J, Yun K, Hwang JH, Chae JH. Classification of suicide attempts through a machine learning algorithm based on multiple systemic psychiatric scales. Front Psychiatry 2017;8:192.
15. Ryu S, Lee H, Lee DK, Park K. Use of a machine learning algorithm to predict individuals with suicide deation in the general opulation. Psychiatry Investig 2018;15:1030-1036.
16. Kweon S, Kim Y, Jang MJ, Kim Y, Kim K, Choi S, et al. Data resource profile: the Korea National Health and Nutrition Examination Survey (KNHANES). Int J Epidemiol 2014;43:69-77.
17. Raghunathan TE, Solenberger PW, Van Hoewyk J. IVEware: Imputation and Variance Estimation Software User Guide. Michigan: University of Michigan; 2002.
18. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;16:219-242.
19. Li DC, Liu CW, Hu SC. A learning method for the class imbalance problem with medical data sets. Comput Biol Med 2010;40:509-518.
20. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013;14:106.
21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:341-378.
22. Breiman L. Random forests. Mach Learn 2001;45:5-32.
23. Kim SM, Baek JH, Han DH, Lee YS, Yurgelun-Todd DA. Psychosocial-environmental risk factors for suicide attempts in adolescents with suicidal ideation: findings from a sample of 73,238 adolescents. Suicide Life Threat Behav 2015;45:477-487.
24. Souetre E, Wehr TA, Douillet P, Darcourt G. Influence of environmental factors on suicidal behavior. Psychiatry Res 1990;32:253-263.
25. Choi SB, Lee W, Yoon JH, Won JU, Kim DW. Risk factors of suicide attempt among people with suicidal ideation in South Korea: a cross-sectional study. BMC Public Health 2017;17:579.
26. Jung M. The relationship between alcohol abuse and suicide risk according to smoking status: a cross-sectional study. J Affect Disord 2019;244:164-170.