Journal of Affective Disorders 231 (2018) 8–14

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea

Soo Beom Choi[a,b,1], Wanhyung Lee[c,d,e,1], Jin-Ha Yoon[c,d,e], Jong-Uk Won[c,d,e], Deok Won Kim[a,b,*]

[a] *Department of Medical Engineering, Yonsei University College of Medicine, Seoul, Republic of Korea*
[b] *Graduate Program in Biomedical Engineering, Yonsei University, Seoul, Republic of Korea*
[c] *The Institute for Occupational Health, Yonsei University College of Medicine, Seoul, Republic of Korea*
[d] *Graduate School of Public Health, Yonsei University College of Medicine, Seoul, Republic of Korea*
[e] *Incheon Worker's Health Center, Incheon, Republic of Korea*

ARTICLE INFO

Keywords:
Suicide prediction
Cox regression
Support vector machine
Deep learning

ABSTRACT

Background: Death by suicide is a preventable public health concern worldwide. The aim of this study is to investigate the probability of suicide death using baseline characteristics and simple medical facility visit history data using Cox regression, support vector machines (SVMs), and deep neural networks (DNNs).

Method: This study included 819,951 subjects in the National Health Insurance Service (NHIS)–Cohort Sample Database from 2004 to 2013. The dataset was divided randomly into two independent training and validation groups. To improve the performance of predicting suicide death, we applied SVM and DNN to the same training set as the Cox regression model.

Results: Among the study population, 2546 people died by intentional self-harm during the follow-up time. Sex, age, type of insurance, household income, disability, and medical records of eight ICD-10 codes (including mental and behavioural disorders) were selected by a Cox regression model with backward stepwise elimination. The area of under the curve (AUC) of Cox regression (0.688), SVM (0.687), and DNN (0.683) were approximately the same. The group with top .5% of predicted probability had hazard ratio of 26.21 compared to that with the lowest 10% of predicted probability.

Limitations: This study is limited by the lack of information on suicidal ideation and attempts, other potential covariates such as information of medication and subcategory ICD-10 codes. Moreover, predictors from the prior 12–24 months of the date of death could be expected to show better performances than predictors from up to 10 years ago.

Conclusions: We suggest a 10-year probability prediction model for suicide death using general characteristics and simple insurance data, which are annually conducted by the Korean government. Suicide death prevention might be enhanced by our prediction model.

1. Introduction

Suicide deaths are a preventable public health concern worldwide. According to a World Health Organization (WHO) report, nearly one million people died by suicide around the world in 2010 (World Health Organization, 2012). In the Republic of Korea, suicide deaths have increased since 1985, and suicide is the fourth leading cause of death (Korea, 2010; Jeon et al., 2016). Furthermore, the age-standardized rate of suicide in Korea is 31.2 per 100,000 people, which is the highest among Organization for Economic Cooperation and Development (OECD) countries (11.3 per 100,000) (Park et al., 2013).

Detection and assessment of vulnerable populations are considered a fundamental cornerstone for preventing suicide death. Multifocal approaches are necessary to predict risk for death by suicide, because suicide risk is related to family structure, socioeconomics, demographics, and family history of suicide and mental illness, as well as gender differences (Qin et al., 2003). However, studies of suicide death focus largely on mental health and psychiatric patients who were suffering from affective disorders, depressive symptoms, or psychological problems (Kovacs and Garrison, 1985; Goldstein et al., 1991; O'Connor and Nock, 2014). Although psychiatric problems are significantly associated with suicide death, the application of these studies to the general population may be limited; many individuals do not visit psychiatric clinicians or professionals. Moreover, surveys for mental health

[*] Correspondence to: Department of Medical Engineering Yonsei University College of Medicine, CPO Box 8044, Seoul, Republic of Korea.
E-mail address: kdw@yuhs.ac (D.W. Kim).
[1] Co-first author: both authors distributed equally.

https://doi.org/10.1016/j.jad.2018.01.019
Received 29 June 2017; Received in revised form 16 January 2018; Accepted 29 January 2018
Available online 31 January 2018

or suicidal ideations may be unreliable, because they involve sensitive issues. Some research has shown that socio-economic or demographic status could be a central issue of suicide death (Qin et al., 2003).

Therefore, the aim of the present study is to investigate the probability of suicide death using baseline characteristics and simple medical facility visit history data using multi-statistical analytic tools. Moreover, we expect that the suicide risk is concentrated in specific strata of patients, because the goal of modeling is to provide information regarding the feasibility of selective, risk-stratified preventive interventions (McCarthy et al., 2015). This model could be helpful in developing a strategy for preventing suicide death in the general population.

2. Methods

2.1. Study population

This study was conducted using data from the National Health Insurance Service–Cohort Sample Database (NHIS-CSD) from 2004 to 2013. The Korea National Health Insurance (KNHI) program provides mandatory public health insurance, offering coverage of medical care services to almost 100% of Koreans: 97% of Koreans are covered by Medicare and 3% are covered by Medicaid. Medicaid is provided to people whose income is insufficient to meet their needs and those of their families, and they are exempted from health insurance fees. Medicare patients paying health insurance fees pay approximately 10–30% of their total medical expenses when using medical facilities, and medical providers are required to submit claims for the remaining 70–90% of the medical expenses. Healthcare records of patients were not duplicated or omitted because all Korean residents receive a unique identification number at birth (Rim et al., 2015). Records of medical services and prescribed medication covered by KNHI are collected in the Korean National Health Insurance Claims Database (Kwon et al., 2015).

The data comprise 1016,583 nationally representative random subjects, which were produced by the KNHI using a systematic sampling method to generate a representative sample from all 48,388,112 Korean residents in 2004. The KNHI program provides coverage for all residents in the form of compulsory social insurance, which ensured the complete follow-up of study participants. If a member was censored due to death or emigration, a new member was recruited among new-borns of the same calendar year. The NHIS-CSD was proven to be a representative sample of Korean population. Detailed methods for establishing and ensuring the representativeness of the NHIS-CSD cohort were published on the KNHI website (Lee et al., 2014).

Among the 1016,583 subjects, we included 819,951 subjects after excluding 196,632 subjects who were 14 years old or younger. The institutional review board of the Yonsei University Health System approved the protocol of this study (No. 4–2017-0122).

2.2. Variables

The NHIS-CSD includes qualification and medical services claim data. Qualification data include patients' KNHI identification number, sex, age, type of insurance, household income level, disability, and mortality information (patients' cause, year and month of death). Medical facility visit history data contain information about inpatient or outpatient services an individual receives, such as diagnosis information recoded by physicians classified by the International Classification of Diseases (ICD) 10 codes (Kim et al., 2016).

Suicide deaths were identified using death causes with ICD-10 (X60~X84; intentional self-harm) in death confirmation records from 2004 to 2013. Independent variables for suicide death in this study were sex, age (15–19, 20–39, 40–59, 60–74, and above 75), type of insurance (national health insurance employee subscriber and dependent, national health insurance district subscriber and dependent, and medical aid), quartile of household income, disability, records of ICD-10 medical services claim data in 2004 (22 variables), and visits to a dental or oriental medical clinic in 2004.

2.3. Statistical analyses

The characteristics of the study participants are reported as mean ± standard deviation (SD) for continuous variables and as number (%) for categorical variables. Cox regression was performed to investigate hazard ratios (HRs) for suicide death and construct a prediction model for 10-year probability of suicide death. Among the Cox regression analysis results, HRs and 95% confidence intervals were reported. Survival time was defined as the interval between the first examinations (2004) and the date of suicide death. For feature selection, a Cox regression model with backward stepwise elimination was also performed to find risk factors for suicide death with a threshold of $p = .05$.

The dataset was divided randomly into two independent training and validation groups to test for internal validation. The training group, comprising 70% of the dataset (573,965 subjects with 1782 suicide deaths), was used to construct the Cox regression model. The validation group, comprising 30% of the dataset (245,986 subjects with 764 suicidal death), was used to assess the performance of the model for suicide death prediction. Receiver operating characteristic (ROC) curves and area under the curve (AUC) analyses were executed to verify the performance of the Cox regression model for suicide death. The predicted probability of suicide was, further, calculated for each individual and subjects were delineated into tiers according to their probabilities (the top .5%, 1%, 2%, 5%, 10%, 50%, and 90%). All statistical analyses were performed using SAS 9.4 (SAS Institute, Inc, Cary, NC). A $p < .05$ was considered statistically significant.

2.4. Machine learning

To improve the prediction of suicide death, we applied two machine learning methods to the same data set as the Cox regression model. Machine learning is an area of artificial intelligence research which uses statistical methods for data classification (Choi et al., 2014). Machine learning techniques generally have shown higher accuracy in diagnosis than classical methods in clinical settings. Support vector machines (SVMs) and artificial neural networks (ANNs) are widely used in machine learning and are the most frequently used supervised learning methods for analysing complex medical data (Choi et al., 2014).

Binary SVMs are learning and pattern-recognition algorithms that aim to distinguish classes according to a function computed from available examples. The goal is to find a hyper-plane that maximizes the separation or margin between two classes (Kim et al., 2013). The same 13 risk factors as those in the Cox regression model with backward stepwise elimination were employed for the SVM. To obtain the optimal model, we adopted a grid search, in which a range of parameter values (penalty parameter [C] of 0.01, 0.1, 1, 10, and 100 and scaling factor [σ] of 0.001, 0.01, 0.1, 1, 10, and 100) were tested using 10-fold cross-validation. The SVM models were constructed with balanced weight using MATLAB Version 2012a (Mathworks Inc., Natick, MA).

ANNs are mathematical systems which mimic biological neural networks (Yoo et al., 2016). The networks can be trained to recognize underlying patterns of diseases. Among several neural network methods, we selected the deep neural network (a 2-layer network with 10, 20, and 10 hidden units) using the Tensorflow package in Python version 3.5.2 (Python Software Foundation, Wilmington, DE).

This dataset is highly imbalanced because of the very low incidence rate of suicide death. Datasets with imbalanced classes tend to be difficult for machine learning algorithms to handle (Oronoz et al., 2015). We chose an over-sampling technique to overcome this problem, which involved matching the ratio of the major and minor groups by duplicating samples for the minor group. The training group was balanced by

Table 1
Demographic and clinical characteristics of the participants, and the association between these characteristics and death by suicide (n = 819,951).

	Suicidal death			
	No (n = 817405)	**Yes (n = 2546)**	**P-value**	**HR (95% CI)**
Duration of follow-up (month)	113.5 ± 20.5	61.7 ± 33.7		
Male	402913 (49.3)	1696 (66.6)	< 0.001*	2.590 (2.360–2.843)
Age				
15–19	66358 (8.1)	99 (3.9)	< 0.001*	0.640 (0.516–0.793)
20–39	342189 (41.9)	736 (28.9)		Reference
40–59	275860 (33.8)	888 (34.9)	< 0.001*	1.349 (1.216–1.497)
60–74	102502 (12.5)	589 (23.1)	< 0.001*	2.279 (2.010–2.584)
Above 75	30496 (3.7)	234 (9.2)	< 0.001*	4.286 (3.638–5.050)
Type of insurance				
NHI employee subscriber	185930 (22.8)	331 (13.0)		Reference
NHI employee dependent	241361 (29.5)	729 (28.6)	< 0.001*	1.862 (1.623–2.136)
NHI district subscriber	168677 (20.6)	834 (32.8)	< 0.001*	2.134 (1.873–2.430)
NHI district dependent	194176 (23.8)	471 (18.5)	< 0.001*	2.100 (1.809–2.437)
Medical aid	27261 (3.3)	181 (7.1)	< 0.001*	2.566 (2.082–3.162)
Household income (n, %)				
High	217425 (26.6)	517 (20.3)		Reference
Moderate–high	181679 (22.2)	487 (19.1)	0.002*	1.214 (1.072–1.375)
Moderate–low	222695 (27.2)	686 (26.9)	< 0.001*	1.454 (1.295–1.632)
Low	195606 (23.9)	856 (33.6)	< 0.001*	1.875 (1.669–2.108)
Disability	33184 (4.1)	243 (9.5)	< 0.001*	1.408 (1.226–1.618)
Dental or oriental medical clinic	350138 (42.8)	1074 (42.2)	0.320	0.958 (0.880–1.043)
ICD-10 Code				
A00 - B99	138069 (16.9)	412 (16.2)	0.557	0.979 (0.875–1.095)
C00 - D48	32721 (4.0)	124 (4.9)	0.247	1.121 (0.931–1.350)
D50 - D89	13016 (1.6)	42 (1.65)	0.698	1.069 (0.785–1.456)
E00 - E90	75661 (9.3)	346 (13.6)	0.025*	1.156 (1.021–1.308)
F00 - F99	46808 (5.7)	401 (15.8)	< 0.001*	2.725 (2.423–3.065)
G00 - G99	54676 (6.7)	296 (11.6)	< 0.001*	1.274 (1.115–1.454)
H00 - H59	155460 (19.0)	448 (17.6)	0.004*	0.860 (0.772–0.959)
H60 - H95	60966 (7.5)	176 (6.9)	0.004*	0.806 (0.688–0.943)
I00 - I99	116066 (14.2)	559 (22.0)	0.425	1.050 (0.940–1.173)
J00 - J99	406184 (49.7)	1145 (45.0)	< 0.001*	0.852 (0.779–0.931)
K00 - K93	287797 (35.2)	974 (38.3)	0.897	1.006 (0.916–1.105)
L00 - L99	179404 (22.0)	556 (21.8)	0.963	1.013 (0.916–1.119)
M00 - M99	234131 (28.6)	841 (33.0)	0.084	0.942 (0.854–1.039)
N00 - N99	143134 (17.5)	421 (16.5)	0.011*	1.164 (1.039–1.304)
O00 - O99	13575 (1.7)	14 (0.6)	0.087	0.618 (0.352–1.086)
P00 - P96	130 (0.0)	1 (0.0)	0.113	5.073 (0.698–36.852)
Q00 - Q99	1474 (0.2)	4 (0.2)	0.741	0.852 (0.319–2.273)
R00 - R99	137023 (16.8)	514 (20.2)	0.094	1.120 (1.007–1.246)
S00 - T98	167295 (20.5)	627 (24.6)	0.002*	1.176 (1.068–1.295)
U00 - U99	4 (0.0)	0 (0.0)	0.960	–
V01 - Y98	1375 (0.2)	7 (0.3)	0.532	1.268 (0.603–2.667)
Z00 - Z99	33508 (4.1)	100 (3.9)	0.038*	1.256 (1.015–1.554)

HR, hazards ratio; CI, confidence interval; NHI, national health insurance; ICD, International Classification of Diseases.
A00-B99, Certain infectious and parasitic diseases; C00-D48, Neoplasms; D50-D89, Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism; E00-E90, Endocrine, nutritional and metabolic diseases; F00-F99, Mental and behavioural disorders; G00-G99, Diseases of the nervous system; H00-H59, Diseases of the eye and adnexa; H60-H95, Diseases of the ear and mastoid process; I00-I99, Diseases of the circulatory system; J00-J99, Diseases of the respiratory system; K00-K93, Diseases of the digestive system; L00-L99, Diseases of the skin and subcutaneous tissue; M00-M99, Diseases of the musculoskeletal system and connective tissue; N00-N99, Diseases of the genitourinary system; O00-O99, Pregnancy, childbirth and the puerperium; P00-P96, Certain conditions originating in the perinatal period; Q00-Q99, Congenital malformations, deformations and chromosomal abnormalities; R00-R99, Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; S00-T98, Injury, poisoning and certain other consequences of external causes; U00-U99, Codes for special purposes; V01-Y98, External causes of morbidity and mortality; Z00-Z99, Factors influencing health status and contact with health services.
Cox regression controlled for each variable.

* P-value < .05.

over-sampling and the validation group remained unchanged.

3. Results

3.1. Baseline characteristics

The mean ± SD of survival time of the study population was 113.4 ± 20.8 months. Among the study population, 2546 people died by intentional self-harm during the follow-up time. Table 1 lists the demographic and clinical characteristics of the study population. In the Cox regression results, suicide death was significantly correlated with the fourteen variables shown with asterisks, which were sex, age, type of insurance, household income, disability, medical records of ICD-10 codes of E00–E90 (Endocrine, nutritional and metabolic diseases), F00–F99 (Mental and behavioural disorders), G00–G99 (Diseases of the nervous system), H00–H59 (Diseases of the eye and adnexa), H60–H95 (Diseases of the ear and mastoid process), J00–J99 (Diseases of the respiratory system), N00–N99 (Diseases of the genitourinary system), S00–T98 (Injury, poisoning and certain other consequences of external causes), and Z00–Z99 (Factors influencing health status and contact with health services).

3.2. Risk factors for suicidal death

Sex, age, type of insurance, household income, disability, medical records of ICD-10 codes of E00–E90, F00–F99, G00–G99, H60–H95,

Table 2
Feature selection for suicide death using a Cox regression model with backward stepwise elimination in training set (n = 573,965).

	Suicidal death				
	No (n = 572183)	**Yes (n = 1782)**	**Total (n = 573965)**	**Beta**	**HR (95% CI)**
Male	282150 (49.3)	1204 (67.6)	283354 (49.4)	1.000	2.717 (2.431–3.037)
Age					
15–19	46585 (8.1)	70 (3.9)	46655 (8.1)	−0.415	0.661 (0.511–0.853)
20–39	239686 (41.9)	499 (28.0)	240185 (41.9)		Reference
40–59	192872 (33.7)	634 (35.6)	193506 (33.7)	0.362	1.437 (1.270–1.625)
60–74	71688 (12.5)	422 (23.7)	72110 (12.6)	0.904	2.469 (2.137–2.852)
Above 75	21352 (3.7)	157 (8.8)	21509 (3.8)	1.485	4.414 (3.639–5.352)
Type of insurance					
NHI employee subscriber	130256 (22.8)	229 (12.9)	130485 (22.7)		Reference
NHI employee dependent	169124 (29.6)	507 (28.5)	169631 (29.6)	0.642	1.901 (1.612–2.241)
NHI district subscriber	117948 (20.6)	594 (33.3)	118542 (20.7)	0.780	2.182 (1.868–2.550)
NHI district dependent	135769 (23.7)	323 (18.1)	136092 (23.7)	0.765	2.149 (1.797–2.571)
Medical aid	19086 (3.3)	129 (7.2)	19215 (3.4)	0.949	2.584 (2.023–3.301)
Household income (n, %)					
High	152214 (26.6)	360 (20.2)	152574 (26.6)		Reference
Moderate–high	127161 (22.2)	326 (18.3)	127487 (22.2)	0.160	1.174 (1.010–1.364)
Moderate–low	155899 (27.3)	488 (27.4)	156387 (27.3)	0.407	1.502 (1.309–1.724)
Low	136909 (23.9)	608 (34.1)	137517 (24.0)	0.661	1.936 (1.685–2.225)
Disability	23221 (4.1)	176 (9.9)	23397 (4.1)	0.368	1.445 (1.227–1.701)
ICD-10 Code					
E00 - E90	52878 (9.2)	249 (14.0)	53127 (9.3)	0.184	1.201 (1.043–1.383)
F00 - F99	32605 (5.7)	298 (16.7)	32903 (5.7)	1.116	3.052 (2.665–3.494)
G00 - G99	38317 (6.7)	207 (11.6)	38524 (6.7)	0.238	1.268 (1.083–1.484)
H60 - H95	42639 (7.5)	113 (6.3)	42752 (7.5)	−0.310	0.734 (0.604–0.891)
J00 - J99	284197 (49.7)	797 (44.7)	284994 (49.7)	−0.144	0.866 (0.782–0.959)
M00 - M99	163770 (28.6)	567 (31.8)	164337 (28.6)	−0.127	0.881 (0.784–0.989)
N00 - N99	100060 (17.5)	291 (16.3)	100351 (17.5)	0.163	1.177 (1.031–1.344)
S00 - T98	117244 (20.5)	430 (24.1)	117674 (20.5	0.141	1.152 (1.027–1.291)

HR, hazards ratio; CI, confidence interval; NHI, national health insurance; ICD, International Classification of Diseases.
E00-E90, Endocrine, nutritional and metabolic diseases; F00-F99, Mental and behavioural disorders; G00-G99, Diseases of the nervous system; H60-H95, Diseases of the ear and mastoid process; J00-J99, Diseases of the respiratory system; M00-M99, Diseases of the musculoskeletal system and connective tissue; N00-N99, Diseases of the genitourinary system; S00-T98, Injury, poisoning and certain other consequences of external causes.
Cox regression with backward stepwise elimination involved a threshold of $p = .05$.

J00–J99, M00–M99 (Diseases of the musculoskeletal system and connective tissue), N00–N99, and S00–T98 were selected by the Cox regression model with backward stepwise elimination as shown in Table 2.

In Table 2, men were more vulnerable to suicide death then women with a HR of 2.72. When the 20–39 age group was used as the reference group, the 15–19, 40–59, 60–74, and above 75 age groups had HRs of 0.66, 1.44, 2.47, and 4.41 (the largest), respectively. When national health insurance employee subscribers were used as the reference group, employee dependent, district subscriber, and medical aid groups had HRs of 1.90, 2.18, 2.15, and 2.58, respectively. When high household income was used as the reference group, moderate–high, moderate–low, and low household income had HRs of 1.17, 1.50, and 1.94, respectively. Disability had a HR of 1.45. The suicide death risk was higher in subjects with medical records of ICD-10 codes of E00–E90 (HR: 1.20), F00–F99 (HR: 3.05), G00–G99 (HR: 1.27), N00–N99 (HR: 1.18), or S00–T98 (HR: 1.15). The suicidal death risk was lower in the subjects with medical records of ICD-10 codes of H60–H95 (HR: 0.73), J00–J99 (HR: 0.87), and M00–M99 (HR: 0.88).

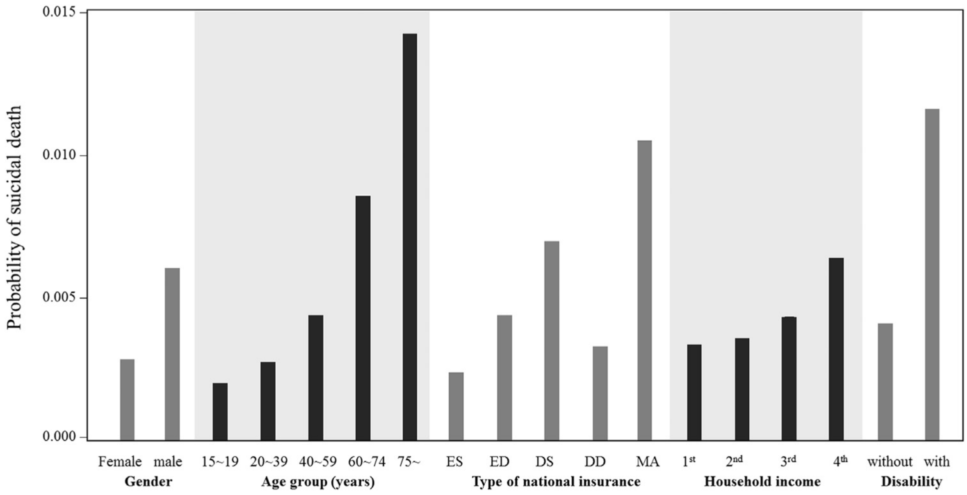

Fig. 1. Mean probabilities of suicide death calculated by Cox regression model for each baseline characteristic group. ES, national health insurance employee subscriber; ED, national health insurance employee dependent; DS, national health insurance district subscriber; DD, national health insurance district dependent; MA, medical aid.

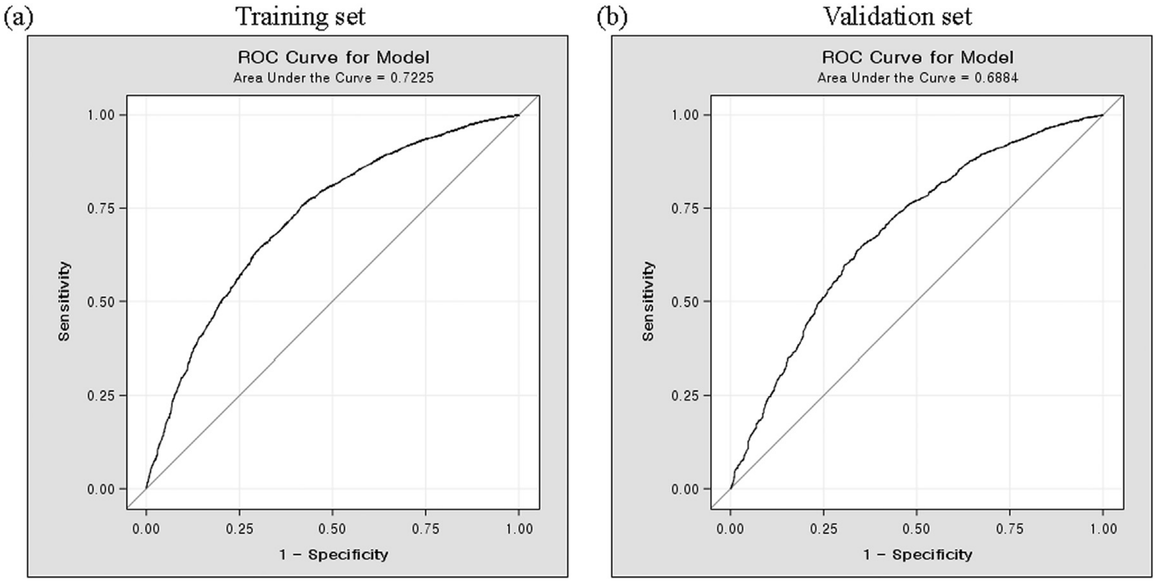

Fig. 2. Receiver operating characteristic curves and areas under the curve (AUCs) of Cox regression model for suicide death in training (a) and validation (b) sets.

3.3. Performance of prediction models of suicide death

Fig. 1 shows the mean probabilities of suicide death calculated by the Cox regression model in each baseline characteristic groups. In Fig. 1, elderly men with low socio-economic status and a disability had a high probability of suicide death. Fig. 2(a) and (b) demonstrate the performance of our Cox regression models for prediction of suicide death in the training and validation sets, respectively. The AUCs of Cox regression were 0.723 and 0.688, respectively, in the training and validation sets.

We obtained the AUCs for the validation set to evaluate the performance of Cox regression, SVM, and DNN with non-changed and over-sampling training sets (Table 3). Cox regression models performed almost the same for non-changed and over-sampling training sets, with an AUC of 0.688. The SVM and DNN models for non-changed training sets had poor AUCs, of 0.576 and 0.632, respectively. The over-sampling technique improved the AUCs of the SVM and DNN models to 0.687 and 0.683, respectively, but the Cox regression model had the highest AUC, 0.688.

Table 4 demonstrates the suicide risk concentration by tier of predicted probability, and the HR between suicidal death and tier of calculated probabilities. Considering the lowest 10% of predicted probability as the reference group, the group with top .5% of predicted probability had a HR of 26.21. Moreover, the percentages of suicidal death gradually increase to the top of each group. This result indicates the feasibility of risk-stratified preventive interventions using tiers of predicted probability of suicidal death calculated by our model.

Table 4
Suicide risk concentration by tier of predicted probability calculated by the Cox regression model (n = 819,951).

Tier of predicted probability, %	Suicidal death, n (%)	Subjects, n (%)	HR (95% CI)
top .5%	69 (2.02)	3412 (0.4)	26.206 (19.105–35.946)
0.5–1	79 (1.65)	4787 (0.6)	21.045 (15.518–28.541)
1–2	105 (1.42)	7390 (0.90)	16.616 (12.506–22.078)
2–5	224 (0.99)	22677 (2.8)	11.012 (8.597–14.107)
5–10	355 (0.81)	43718 (5.3)	8.646 (6.839–10.930)
10–50	1170 (0.36)	327965 (40.0)	3.473 (2.794–4.319)
50–90	457 (0.14)	326600 (39.8)	1.330 (1.058–1.673)
90–100	87 (0.10)	83402 (10.2)	Reference
Total	2546 (0.31)	819951 (100.0)	

HR, hazards ratio; CI, confidence interval.
HRs were calculated in Cox regression by suicidal death and tier of calculated probabilities.

4. Discussion

The aim of the current study was to investigate predictive models of 10-year probability of suicide death using baseline characteristics and medical facility visits history with ICD-10 codes from the population-based NHIS-CSD. Male gender, older age, lower income, medical aid, and disability were linked to increased risk for suicide death at 10-year follow-up. Furthermore, suicide deaths at the follow-up period were significantly associated with participants who visited a medical facility during one year from baseline due to endocrine, nutritional and metabolic diseases; mental and behavioural disorders; diseases of the nervous system, the ear and mastoid process, the respiratory system, the musculoskeletal system and connective tissue, and the genitourinary system; or injury, poisoning and certain other consequences of external causes. The Cox regression model effectively predicted suicide death from the analysis, and had a slightly higher AUC than the SVM and DNN.

Table 3
Area under the receiver operating characteristic curve of prediction models for suicide death using original and over-sampling training sets in validation set (n = 245,986).

	Original set	Over-sampling set
Cox regression	0.688	0.688
Support vector machine	0.576	0.687
Deep neural network	0.632	0.683

Previous studies reported gender differences of suicide, that women have higher risk for suicidal ideation and attempt but lower suicidal death rate than men (Canetto and Sakinofsky, 1998; Möller-Leimkühler, 2003; Choi et al., 2017). In this study, men were also found to be more vulnerable to suicide death than women. Men use more violent and aggressive methods for suicide, such as drugs, hanging, and poisoning, than women (Denning et al., 2000). Therefore, it is assumed that men have higher rates of suicidal success than women.

Age group was selected as a risk factor of suicide death. In Korea, an increasing trend in suicide deaths had been observed since the economic crisis in 1997, which is attributed to the increase in suicide deaths in older age groups (Kwon et al., 2009). Previous reports indicated that increased risks of completed suicide were found in elderly people compared with any other age group (O'connell et al., 2004). The highest HR for suicide death was shown in the population over 75 years old in the current analysis. This result could be explained by the fact that older population suffers from various physical and/or mental illnesses, increased functional disability, and social isolation (Waern et al., 2002; Kiosses et al., 2014; Liu et al., 2014; Fässberg et al., 2016).

Low socio-economic status is strongly and closely associated with suicide death (Cubbin et al., 2000; Shah et al., 2008; Crump et al., 2014). The current study also demonstrated that people receiving medical aid and with low household incomes had a higher probability of death by suicide than other groups (Fig. 1). Physical or intellectual limitations due to disability were also closely linked to suicide death. Previous study indicated a significant relationship between suicide and disability when controlling various potential confounders, including both age and income level (Meltzer et al., 2012). Hopeless and depressive symptoms of individuals with disability could be associated with suicide (Hamzaoglu et al., 2010; Mezuk et al., 2012).

Individuals who visited a medical facility with F00–F99 (mental and behavioural disorders) and S00–T98 (injury, poisoning, and certain other consequences of external causes) were found to have a significantly increased risk for suicide death. Psychiatric and emergency medical professionals might have a key role in preventing suicide death, because vulnerable populations were more likely to visit a hospital's psychiatric or emergency department before successful suicide (Appleby et al., 1999; Gairin et al., 2003). This information was recorded by the government to manage and maintain the national health insurance service; further research with detailed medical facility visit history is warranted to set a strategy for preventing suicide deaths.

Although we constructed prediction models using machine learning to improve prediction performance, the AUCs of the SVM and DNN were slightly lower than that of the Cox regression model. The characteristics of the dataset were highly imbalanced (due to the extremely low prevalence of suicide death). Given a highly imbalanced dataset, cost-insensitive classifiers tend to predict everything as the majority class (Oronoz et al., 2015). To overcome this problem, we used over-sampling, which improved the AUCs of both the SVM and DNN, though not sufficiently. Moreover, we tried to balance the weight parameters for the SVM models, but in vain. The performance of the Cox regression models did not differ for either the original or over-sampling training sets. Cox regression analysis is easier to implement in a clinical setting and less demanding of computational resources than the machine learning method. Whereas Cox regression analysis provides risk factors with hazard ratio, machine learning does not because of its black-box characteristic. Therefore, if the performances of Cox regression analysis and machine learning method are approximately the same, the regression model may be more appropriate in a clinical environment.

The key strength of this study is that it demonstrates the possibility of suicide death prediction using demographic characteristics and simple insurance data, which are annually conducted and managed by the Korean government. The NHIS-CSD including death confirmation records from 2004 to 2013 do not have any missing or censored data. Moreover, approximately a million subjects in this dataset were selected among the entire Korean population because the KNHI program provides mandatory public health insurance. Therefore, our results are relatively less biased than conventional survey studies.

This study is limited by the lack of information on other potential covariates such as educational level and occupation because those were not available in the NHIS-CSD. Second, we could not investigate suicidal ideation and attempts. We included death confirmation codes with intentional self-harm, which indicated success of suicide. We could not include failed cases of suicide. Although our results may be underestimated, the selected risk factors were significantly associated with suicide death. The process of investigating attempted suicide is further complicated by the necessity of a more specific for definition than the ICD-10 code for intentional self–harm, because this code includes multiple subcategories for different self–harm methodologies (e.g., poisoning, using sharp object, jumping from high places, or crashing a motor vehicle), potentially indicative of different risk factors. Third, it should additionally be noted that the 2004 characteristics reflect aged information because of the necessary decade–long gap between when predictors were evaluated and suicides were measured. Further research analysing the difference between suicidal and non-suicidal deaths using predictors gathered within a 12–24 month time span prior to death is warranted, as this data is expected to have improved performance for predicting suicide risk than predictors up to 10 years old. Finally, no information regarding psychotropic medication was included in this study. Including prescription information for psychotropic medications or certain classes of psychotropic medications (e.g. antidepressants, antipsychotics, mood stabilizers, or anti-anxiety) could improve the suicide risk predictions. Expanding input variables with ICD-10 code subcategories, drug information, and more recent predictors could improve prediction model performance.

This study suggests a 10-year probability prediction of suicide death using Cox regression analysis with basic characteristics and simple national insurance data. Suicide death prevention might be enhanced by considering basic information from general population, collected by the NHIS. Moreover, it may be possible to utilize information on detailed medical facility visit history to predict psychiatric problems as well as suicidal death.

Acknowledgements

This study used National Health Insurance Service National Sample Cohort (NHIS-NSC) data made by National Health Insurance Service (NHIS).

References

Appleby, L., Shaw, J., Amos, T., McDonnell, R., Harris, C., McCann, K., Kiernan, K., Davies, S., Bickley, H., Parsons, R., 1999. Suicide within 12 months of contact with mental health services: national clinical survey. BMJ 318, 1235–1239.

Canetto, S.S., Sakinofsky, I., 1998. The gender paradox in suicide. Suicide Life Threat. Behav. 28, 1–23.

Choi, S.B., Kim, W.J., Yoo, T.K., Park, J.S., Chung, J.W., Lee, Y.-h., Kang, E.S., Kim, D.W., 2014. Screening for prediabetes using machine learning models. Comput. Math. Methods Med. 618976, 1–8.

Choi, S.B., Lee, W., Yoon, J., Won, J., Kim, D.W., 2017. Risk factors of suicide attempt among people with suicidal ideation in South Korea: a cross-sectional study. BMC Public Health 17, 579.

Crump, C., Sundquist, K., Sundquist, J., Winkleby, M., 2014. Sociodemographic, psychiatric and somatic risk factors for suicide: a Swedish national cohort study. Psychol. Med. 44, 279–289.

Cubbin, C., LeClere, F.B., Smith, G.S., 2000. Socioeconomic status and injury mortality: individual and neighbourhood determinants. J. Epidemiol. Community Health 54, 517–524.

Denning, D.G., Conwell, Y., King, D., Cox, C., 2000. Method choice, intent, and gender in completed suicide. Suicide Life Threat. Behav. 30, 282–288.

Fässberg, M.M., Cheung, G., Canetto, S.S., Erlangsen, A., Lapierre, S., Lindner, R., Draper, B., Gallo, J.J., Wong, C., Wu, J., 2016. A systematic review of physical illness, functional disability, and suicidal behaviour among older adults. Aging Ment. Health 20, 166–194.

Gairin, I., House, A., Owens, D., 2003. Attendance at the accident and emergency department in the year before suicide: retrospective study. Br. J. Psychiatry 183, 28–33.

Goldstein, R.B., Black, D.W., Nasrallah, A., Winokur, G., 1991. The prediction of suicide: sensitivity, specificity, and predictive value of a multivariate model applied to suicide among 1906 patients with affective disorders. Arch. Gen. Psychiatry 48, 418–422.

Hamzaoglu, O., Ozkan, O., Ulusoy, M., Gokdogan, F., 2010. The prevalence of hopelessness among adults: disability and other related factors. Int. J. Psychiatry Med. 40, 77–91.

Jeon, S.Y., Reither, E.N., Masters, R.K., 2016. A population-based analysis of increasing rates of suicide mortality in Japan and South Korea, 1985–2010. BMC Public Health 16, 1.

Kim, K.-A., Choi, J.Y., Yoo, T.K., Kim, S.K., Chung, K., Kim, D.W., 2013. Mortality prediction of rats in acute hemorrhagic shock using machine learning techniques. Med. Biol. Eng. Comput. 51, 1059–1067.

Kim, S., Shin, D.W., Yun, J.M., Hwang, Y., Park, S.K., Ko, Y.-J., Cho, B., 2016. Medication adherence and the risk of cardiovascular mortality and hospitalization among patients With newly prescribed antihypertensive medications novelty and significance. Hypertension 67, 506–512.

Kiosses, D.N., Szanto, K., Alexopoulos, G.S., 2014. Suicide in older adults: the role of emotions and cognition. Curr. Psychiatry Rep. 16, 1–8.

Korea, S., 2010. Annual Report on the Cause of Death Statistics. Statistics Korea, Daejeon.

Kovacs, M., Garrison, B., 1985. Hopelessness and eventual suicide: a 10-year prospective study of patients hospitalized with suicidal ideation. Am. J. Psychiatry 1, 559–563.

Kwon, J.-W., Chun, H., Cho, S.-i., 2009. A closer look at the increase in suicide rates in South Korea from 1986–2005. BMC Public Health 9, 72.

Kwon, J.-W., Park, E.-J., Jung, S.-Y., Sohn, H., Ryu, H., Suh, H., 2015. A large national cohort study of the association between bisphosphonates and osteonecrosis of the jaw in patients with osteoporosis: a nested case-control study. J. Dent. Res. 94, 212S–219S.

Lee, J., Kim, K., Lee, J., 2014. Establishment of a Nation Cohort Sample Database using National Health Insurance Data. National Health Insurance Service, Seoul (Korea).

Liu, L., Gou, Z., Zuo, J., 2014. Social support mediates loneliness and depression in elderly people. J. Health Psychol. 21, 750–758.

Möller-Leimkühler, A.M., 2003. The gender gap in suicide and premature death or: why are men so vulnerable? Eur. Arch. Psychiatry Clin. Neurosci. 253, 1–8.

McCarthy, J.F., Bossarte, R.M., Katz, I.R., Thompson, C., Kemp, J., Hannemann, C.M., Nielson, C., Schoenbaum, M., 2015. Predictive modeling and concentration of the risk of suicide: implications for preventive interventions in the US department of veterans affairs. Am. J. Public Health 105, 1935–1942.

Meltzer, H., Brugha, T., Dennis, M.S., Hassiotis, A., Jenkins, R., McManus, S., Rai, D., Bebbington, P., 2012. The influence of disability on suicidal behaviour. ALTER-Eur. J. Disabil. Res./Rev. Eur. De. Rech. Sur Le. Handicap 6, 1–12.

Mezuk, B., Edwards, L., Lohman, M., Choi, M., Lapane, K., 2012. Depression and frailty in later life: a synthetic review. Int. J. Geriatr. Psychiatry 27, 879–892.

O'connell, H., Chin, A.-V., Cunningham, C., Lawlor, B.A., 2004. Recent developments: suicide in older people. BMJ 329, 895–899.

O'Connor, R.C., Nock, M.K., 2014. The psychology of suicidal behaviour. Lancet Psychiatry 1, 73–85.

Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A.D., Casillas, A., 2015. On the creation of a clinical gold standard corpus in Spanish: mining adverse drug reactions. J. Biomed. Inform. 56, 318–332.

Park, S., Choi, J.W., Yi, K.K., Hong, J.P., 2013. Suicide mortality and risk factors in the 12 months after discharge from psychiatric inpatient care in Korea: 1989–2006. Psychiatry Res. 208, 145–150.

Qin, P., Agerbo, E., Mortensen, P.B., 2003. Suicide risk in relation to socioeconomic, demographic, psychiatric, and familial factors: a national register–based study of all suicides in Denmark, 1981–1997. Am. J. Psychiatry 160, 765–772.

Rim, T.H., Kim, D.W., Han, J.S., Chung, E.J., 2015. Retinal vein occlusion and the risk of stroke development: a 9-year nationwide population-based study. Ophthalmology 122, 1187–1194.

Shah, A., Bhat, R., MacKenzie, S., Koen, C., 2008. A cross-national study of the relationship between elderly suicide rates and life expectancy and markers of socioeconomic status and health care. Int. Psychogeriatr. 20, 347–360.

Waern, M., Rubenowitz, E., Runeson, B., Skoog, I., Wilhelmson, K., Allebeck, P., 2002. Burden of illness and suicide in elderly people: case-control study. BMJ 324, 1355.

World Health Organization, 2012. Public health action for the prevention of suicide: a framework.

Yoo, T.K., Kim, D.W., Choi, S.B., Park, J.S., 2016. Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. PLoS One 11, e0148724.

Dr. Deok Won Kim was born in Seoul, Korea in 1952. He received the B.S degree from Seoul National University, Seoul, Korea in 1976, the M.S. degree in electrical engineering from Northwestern University, Evanston, IL, USA, and the Ph.D. degree in biomedical engineering from the University of Texas, Austin, TX, USA in 1980 and 1986, respectively. He is currently a Professor in the Medical Engineering Department, Yonsei University College of Medicine, Seoul, Korea, where he has been since 1987. His areas of research interest are electromagnetic field hazards and multiple classification using machine learning in medicine. Publication: Simple Scoring System and Artificial Neural Network for Knee Osteoarthritis Risk Prediction: A Cross-Sectional Study, PLoS ONE, 2016. Risk factors of suicide attempt among people with suicidal ideation in South Korea: a cross-sectional study, BMC Public Health, 2017.