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
0165-0327/ © 2018 Elsevier B.V. All rights reserved.

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

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات