Suicide risk prediction model using machine learning algorithms for colorectal cancer patients: analyses in national health insurance data

Youngrong Lee

Department of Medicine

The Graduate School, Yonsei University

Suicide risk prediction model using machine learning algorithms for colorectal cancer patients: analyses in national health insurance data

Directed by Professor Sun Jae Jung

Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Science

Youngrong Lee

December 2022

This certifies that the Doctoral Dissertation of Youngrong Lee is approved.

Thesis Supervisor: Sun Jae Jung

Thesis Committee Member#1: Hyeon Chang Kim

Thesis Committee Member#2: Eun Lee

Thesis Committee Member#3: Aesun Shin

Thesis Committee Member#4: So-yeon Ryu

The Graduate School
Yonsei University

December 2022

<TABLE OF CONTENTS>

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Suicide risk prediction model using machine learning algorithms for colorectal cancer patients: analyses in national health insurance data

Youngrong Lee

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Sun Jae Jung)

Background: Previous studies on suicide prediction models using machine learning have consistently demonstrated high predictive performance in the general population. Patients with colorectal cancer (CRC) are known to have a higher risk of suicide than the general population; however, no study has yet investigated the risk factors and predictive performance of machine-learning models for this high-risk group. This cohort study used machine learning to examine age-, sex-, and cancer type-specific risk profiles and the prediction performance of the trained model for suicide in Korean health insurance claims data.

Method: Among the 380,569 individuals diagnosed with CRC (C18–20) between 2002 and 2018, those who died by suicide were included in the case group. The number of deaths due to suicide was 1,839 (0.48%), and to solve the problem of class imbalance, the control group was under-sampled with the same number of samples as the case group (total, n = 3,678). The performance and risk profile of each model stratified by age, sex, and cancer type were identified. Each model was trained using more than 1,600 predictors, including demographic factors, mental and physical health examinations, cancer stage, colon cancer-related surgery, prescribed medications, number of outpatient visits, emergency departments, and hospitalizations. The machine-learning models developed were classification trees and random forests. The predictors that were important in the models were evaluated using conditional logistic regression in a nested case-control study design.

Results: Prescription of psychotherapy, psychiatric medications, including sleeping pills and mood stabilizers, and the number of psychiatric outpatient visits were important predictors of suicide in all subgroups categorized by age, sex, and cancer type. Suicide risk factor profiles showed subtle differences according to age, sex, and cancer type. Recent CRC diagnoses and hospitalization-related variables, such as enema, urinary catheterization, and enteral nutrition, are prominent suicide risk factors in CRC patients. At the optimal threshold, the sensitivity of the random forest model for all CRC patients was 0.84 (84%), the specificity was 0.68 (68%), and the area under the receiver operating curve (AUC) was 0.84. The AUC of the model for the group divided by age, sex, and CRC type was approximately 0.8. CRC patients in the top 1%, 5%, 10%, and 20% of predicted risk accounted for 9.37%, 36.6%, 53.38%, and 70.81% of all suicide deaths, respectively. As a result of the nested case-control study, the associations between the found predictors and suicide were in line with the variable importance results identified in the machine-learning model.

Conclusion: This study identified the risk factors that can predict suicide in CRC patients through machine-learning techniques and suggested the possibility of clinical usage of the prediction model in a step-by-step process for cost-effective suicide prevention intervention.

Key words: machine learning, suicide, colorectal cancer, psycho-oncology

Suicide risk prediction model using machine learning algorithms for colorectal cancer patients: analyses in national health insurance data

Youngrong Lee

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Sun Jae Jung)

I. INTRODUCTION

1.Colorectal cancer as high-risk group for suicide

Patients with cancer are known to have a higher risk of suicide than the general population, and in the United States, it has been reported that it is nearly twice as high.[1] Cancer is the number one cause of death in Korea and the second-leading cause of death worldwide.[2] The main goal of cancer treatment is to survive at the expense of physical, emotional, and financial burdens. However, cancer patients and their families often overlook the long trajectory of cancer treatment. Suicide may be the pinnacle of unmanaged suffering. Although the risk factors for suicide in cancer patients are generally like those in the general population, colorectal cancer (CRC) patients undergo treatment involving colostomy surgery and chemotherapy and must adapt to changes in appearance and lifestyle, such as adapting to a stoma with a fecal bag.[3] Some studies have reported treatment-dependent suicide rates in patients with CRC. Surgical reconstruction and adjuvant treatment necessary for the management of CRC negatively affect self-image, psychological well-being, sexuality, and quality of life. CRC patients undergoing colostomies also must adapt to changes in their excretion and appearance, and these dramatic changes can affect physical and psychological functioning, increasing the likelihood of suicidal outcomes.[4]

Additionally, providing physical comfort, reducing emotional stress, and treating mental disorders are key goals in palliative care for patients with cancer. Strong evidence supports interventions to improve these important aspects of treatment to improve quality of life, including psychotherapy and pharmacotherapy.[5] Depression, sleep disturbance, anxiety, and delirium are prevalent neuropsychiatric complications in patients with cancer, which are associated with significant morbidity and mortality.[6-8] There are many reports of the association, which could either be protective or hazardous, between psychotropic medications, which are mainly prescribed for the above complications, and suicidal behavior.[9-11] In addition to the above factors, gender, age, race, the distant spread of the tumor, and an intact primary tumor were also some of the key predictors of suicide found in CRC patients.[12] As the survival rate of CRC patients continues to increase (e.g., the 5-year relative survival rate in Korea for the years 1993–2010 increased from 62.4 to 70.6% for colon cancer and from 53.4 to 73.6% for rectal cancer),[13] the prediction of high-risk suicide groups among CRC patients will become more important in terms of resolving unmet needs in the mental health of CRC patients.

2. Review of previous suicide prediction studies

Several existing suicide prediction studies have modeled risk scales by defining high-risk groups (e.g., suicide-related emergency room admissions, psychiatric hospitalization, and psychiatric hospital discharge),[14,15,16] and the general population.[17,18,19] Most of these early suicide prediction studies reported the performance of the developed model using self-reported single scales such as hopelessness, depression, overall psychopathological severity, suicidal intention, and attitude toward suicide as predictors.[20,21] Critics have argued that the predictive performance of these studies is not suitable for use in clinical settings.[22] Most of the existing studies over the past 50 years have performed suicide prediction studies using a single scale (Beck Hopeless Scale, Suicide Intent Scale, etc.) for patients defined as high-risk.[23,24] A recent meta-analysis has revealed that the prediction performance of the conventional statistical models has been weak, and no single risk factor or risk scale approach has demonstrated clear superiority, even with the aid of risk factors commonly known as "strong predictors," such as prior suicidal behavior, depression, hopelessness, or male sex.[23] Other meta-analyses about the predictive performance of single-scale-based suicide studies reported that the pooled sensitivity and specificity were 0.77 and 0.41, respectively.[22]

Recently, the trend of suicide prediction research has begun to proceed with machine-learning studies based on real-world data collected from daily administrations, such as claims and electrical health records. A study of the general population using data from a Danish registry[25] revealed data-driven risk factors for psychiatric medications (e.g., antidepressants, antipsychotics, hypnotics, and sedatives). It was also reported that different sexes may have different sets of risk profiles. Recent studies of these machine-learning algorithms have been able to accurately predict future suicide attempts in patients, mentally ill soldiers, and outpatient mental health visits from electronic health record databases.[26,27,28] Several previous studies have applied machine learning to determine the risk of suicide attempts in the general population sample and have concluded that it also needs to be applied to high-risk subgroups such as cancer patients.

However, some critics have argued that the clinical utility of these predictive studies needs further evaluation.[29] Suicide is a rare health outcome, and its low prevalence usually results in a low positive predictive value (PPV). According to a systematic review of suicide prediction research,[30] a low PPV may be the most significant impediment to the implementation of the suicide prediction model in actual clinical settings. In this study, to evaluate the clinical utility of the developed suicide prediction model, in-depth clinical feasibility was evaluated using various evaluation indicators, including the precision-recall curve[29] and the number needed to screen values.[31]

3. Machine learning algorithms in suicide prediction studies

Many existing systematic reviews have reported that a relatively small number of carefully selected sets of essential risk factors (e.g., previous suicide attempts, gender, or single risk scales) combined with conventional statistical methods are insufficient to accurately predict suicidal behavior.[22,23] Instead, a more complex conceptualization with a large number of risk sets may be necessary.[32] Conventional statistical approaches used in the field of mental health are not well suited to model complexity; in contrast, supervised machine-learning methods can model useful algorithms from complex patterns of data for predicting suicidal behavior.[26,27,33]

Machine learning has three distinct advantages over traditional approaches in each of these domains of conventional statistical approaches.[34] First, machine-learning methods determine the most accurate algorithm that maps a target outcome to the factors of interest. Traditional approaches require the researcher to determine an algorithm a priori, leading to a fairly simple model using a small set of predictors. Given the complexity of suicidal behaviors, this has repeatedly failed to yield accurate predictions.[23] Second, machine-learning algorithms can accommodate a large number of factors and simultaneously consider highly complex combinations of these factors. Recent advances in computing power have enabled the simultaneous consideration of thousands of different factors and the complex relationships among factors within a single machine-learning model. Third, machine-learning algorithms are well equipped to process overfitting, which occurs when a model utilizes the noise of a dataset. An over-fitted model would demonstrate strong performance within the dataset it was developed on, but it may perform poorly on novel datasets. The most effective machine-learning model can prevent overfitting, thereby increasing the likelihood of generalizability.

4. Objectives

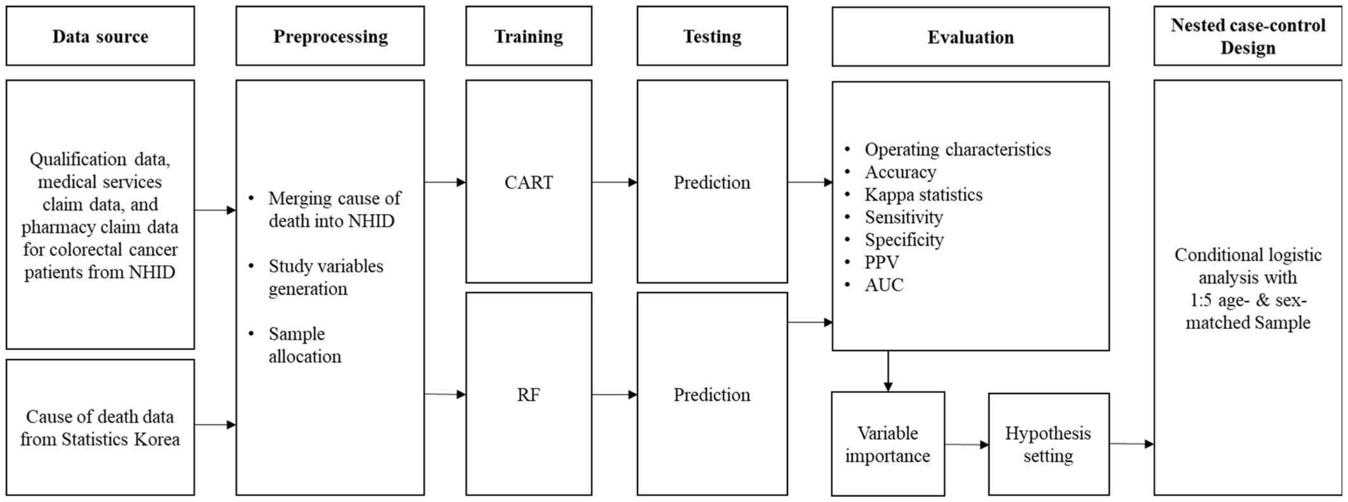
Many machine-learning studies have succeeded in discovering key predictors in large general populations.[25-28] Many researchers have concluded that machine-learning methods need to be applied to high-risk groups as well.[25] Although many individual studies on the suicide risk of colorectal cancer (CRC) have identified various risk factors for suicide, no study has yet identified a data-driven comprehensive set of risk factors using a large sample of cancer patients. The goal of the present study was to identify key predictors and develop machine-learning algorithms and models for suicide in a large nationwide CRC patient sample using data from the National Health Insurance System (NHIS). As many studies have used machine-learning algorithms to predict suicidal behaviors in the general population and the risk of suicide is higher in CRC patients, where many predictive factors have been discovered, we need to apply the machine-learning method to these high-risk subgroups.[12,35] Therefore, our study aimed to 1) explore the predictors of suicide by using machine-learning techniques in CRC patients, 2) identify the magnitude of associations among the predictors uncovered above through a conventional nested case-control study design, and 3) discuss the applicability of this model in actual clinical settings.

II. MATERIALS AND METHODS

1. Data source

The claims data from the Korean National Health Insurance Database (NHID) were analyzed. The NHID is a public database on healthcare services maintained by the Korean NHIS and contains qualification, medical service claims, and pharmacy claim data. The claims data include patient information such as age, sex, insurance premium percentile, residential regions, diagnosis information (according to the International Classification of Diseases, 10th Revision; ICD-10), and specific information on diagnostic tests, procedures, and prescriptions. The NHIS is the only insurer that provides mandatory universal health insurance that virtually covers the entire Korean population (about 97% of Korean citizens) and provides medical benefits to those in the lowest income bracket who are covered by government funding (approximately 3% of Korean citizens).

From 2002 to 2018, the NHIS provided the data of patients who visited medical institutions in Korea and claimed medical expenses, with 40% randomly selected patients who were diagnosed with malignant neoplasms of the colon, rectum, or anus (ICD-10 code C18–21) at least once. Patients who were not diagnosed with CRC (C21) were excluded from the study. The NHID did not provide the cause of death but provided the death status and date. Therefore, deaths due to suicide attempts and other causes were extracted after merging the data on death causes provided by Statistics Korea (Figure 1). The outcome of the study, death due to a suicide attempt, included CRC patients who died with ICD-10 diagnostic code x60–84 between 2002 and 2018.


Data source
Preprocessing
Training
Testing
Evaluation
Nested case-control Design
Qualification data, medical services claim data, and pharmacy claim data for colorectal cancer patients from NHID
Cause of death data from Statistics Korea
• Merging cause of death into NHID
• Study variables generation
• Sample allocation
CART
RF
Prediction
Prediction
• Operating characteristics
• Accuracy
• Kappa statistics
• Sensitivity
• Specificity
• PPV
• AUC
Variable importance
Hypothesis setting
Conditional logistic analysis with 1:5 age- & sex-matched Sample

Abbreviation: NHID, Korean National Health Insurance Database; CART, classification and regression tree model; RF, random forest model, PPV, positive predictive value; AUC, area under receiver operating characteristic curve

Figure 1. Conceptual study analysis flow diagram

2. Study variables

Appendix 1 describes the composition of the study variables in detail. Demographic variables such as age, sex, and premium percentile were used as they were kept in their registry form in the analyses. In previous machine-learning suicide studies, a set of dummy variables reflecting temporal information based on 6, 12, 24, and 48 months before suicide was created.[35] For predictors such as diagnostic codes, medications, and procedures, time-varying dummy codes were created (i.e., diagnoses and prescriptions 0–6, 6–12, 12–24, 24–48, and 48+ month time intervals before the suicide) to examine the effect of the temporal distance of predictors for suicide onset. For patients in the non-suicidal group, random time points for each patient were selected during the follow-up period to generate time-varying dummy codes of the above predictors. These dummy variables also contain information on the number of times diagnoses and prescriptions were made during a given time period.

The diagnostic variables used in this study were largely divided into physical diseases (A00–Z99, excluding F codes) and mental diseases (F00–F99). Physical disease variables were classified by grouping diseases according to the first digit of the ICD-10 code. For example, within certain infectious and parasitic diseases (A00–B99), intestinal infectious diseases (A00–A99), and viral hepatitis (B15–B19), there were distinctions. The diagnostic codes used varied from the chapter on specific infectious and parasitic diseases (A00–B99) to the chapter on factors affecting health status and contact with medical services (Z00–Z99). However, special-purpose codes (U00–U85), external causes of morbidity (V00–Y99), and factors affecting health conditions and contact with medical services (Z00–Z99) were excluded, as they rarely appear in general claims data.

The reason for setting the classification level of the physical diagnostic code as described above is to control the number of generated variables. Diagnosis variables were created as time-considering dummy variables for the occurrence of suicide, and five variables (0–6, 6–12, 12–24, 24–48, and 48+ month intervals) were created. Therefore, using the second-digit (2,040 categories) or third-digit classification (12,121 categories) would generate an excessive number of variables. If there are too many variables, the frequency of occurrence of each variable would decrease, and the statistical and predictive power of the machine-learning model and the importance of the variables would decrease. In addition, it can cause problems when creating overfitting models that are sensitive to data noise, which can make interpretation difficult and prevent robust predictions.

Among the diagnostic codes described above, the colorectal malignant neoplasm code was generated separately as a major predictive factor. That is, variables were created by separately classifying C18–20 from malignant neoplasms of the digestive organs (C15–C26). Many studies have reported that the risk of death by suicide was highest in the first few months after diagnosis and significantly decreased with time.[36] Therefore, this variable can be related to the time of CRC diagnosis in the study samples.

For psychiatric disorders known to have a significant effect on suicide, second-digit classifications (i.e., F00–F99) were then used. For example, mood disorders (F30–F39), bipolar affective disorder (F31), and depressive episodes (F32) were distinguished. There were 72 mental disorders classified using the second-digit classification. In summary, more than 1,400 diagnostic-related dummy variables were created after excluding 19 special-purpose codes (i.e., U00–U85) and 42 V, W, Y, and Z codes rarely found in claims data from the 257 first-digit classifications and adding 72 mental disorder codes (i.e., F00–F99).

In addition, to determine whether the predictors of psychiatric disorders were an underlying disease or complication, a new set of predictors only comprised of psychiatric disorders considering time with regard to CRC diagnosis was used to develop a random forest model as a sensitivity analysis (Appendix 2). Appendix 3 presents the results.

The prescriptions used to create the time-considering dummy variable are largely divided into drugs, examinations, and procedures. The drug consists of seven psychiatric drugs (antidepressants, typical and atypical antipsychotics, mood stabilizers, sedatives, sleeping pills, and opioids) and anticancer drugs in one category (folic acid, fluorouracil, oxaliplatin, and others). The five claimed examinations used for the generation of predictors included liver metastasis ultrasound, dementia screening tests, psychiatric interviews, and neuromuscular conduction tests.

In the claimed procedure used to generate predictors, the number of visits to outpatients, emergency centers, and hospitalizations for treatment of psychiatric needs and the number of psychotherapy sessions were measured. In addition, we aimed to generate predictive variables for inpatient care, total parenteral nutrition, any type of supportive enteral nutrition, enema, urinary catheterization, rectal care, and post-colostomy care after the surgical procedure. CRC-related procedures included colostomy; surgical treatment, including colonoscopy, total colon, and total rectal resection; colectomy, including rectal and sigmoid colectomy; and rectal tumor resection; and radiology treatment, such as in vitro radiotherapy.

In a study conducted by the National Health Insurance Institute, the cancer stage was identified retrospectively by tracking the claimed examination and treatment process.[37] In the case of colon cancer (i.e., C18–19), patients without chemotherapy claims after surgical treatment were categorized as stage 1 and stage 2 if fluorouracil, capecitabine, and oral fluoropyrimidine were used during chemotherapy after surgery. Patients who received oxaliplatin during chemotherapy after surgery were classified as stage 3. Patients with rectal cancer (i.e., C20) who did not have claim records of concurrent chemoradiotherapy after surgery were classified as stage 1, and if present, stage 2 or 3. Patients who received neoadjuvant chemotherapy were classified as stage 3. After the diagnosis of CRC, if there were records of liver resection or symptomatic treatment for liver metastasis, it was classified as stage 4.

The variables explained above consist of demographic, somatic, and psychiatric diagnoses; prescription of psychiatric medications; number of psychiatric visits; and inpatient-related variables and are directly and indirectly related to suicide deaths. Appendix 4 presents the theoretical associations of the variables in this study using a directed acyclic graph.

3. Machine learning analyses

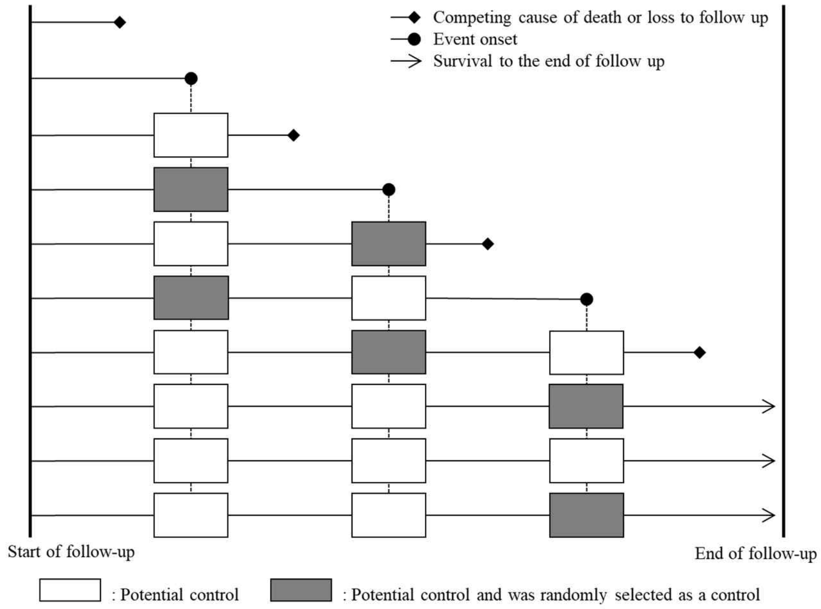
The study variables were compared and discovered by following machine-learning techniques (Figure 1). The classification and regression tree models were implemented for an initial visual evaluation of the data structure using the R (R Foundation for Statistical Computing, Vienna, Austria) package “rpart,” which uses a 10-fold cross-validation procedure. To minimize the risk of overfitting, the maximum tree depth was restricted by setting the optimized complexity parameter based on hyperparameter tuning through 10-fold cross-validation. The risk of attempted suicide was calculated for each branch of the predictor.

A random forest classifier was also implemented using the R package “randomForest.” Each random forest was built with 1,000 trees, and split candidates (features) at each node were obtained by taking twice the number of the square root of the total number of predictors, which is the default of the “randomForest” package. Each tree (among 1,000 trees) was built using all suicide observations and an equal number of randomly selected non-case cohort observations to address the class imbalance.[38,39] The mean decrease in accuracy was applied to evaluate the importance of each variable across all trees. The mean decrease in accuracy indicates the extent of outcome misclassification if a variable is excluded, either due to main effects or interactions.[40]

Prediction accuracy was calculated using accuracy, kappa statistics, sensitivity, specificity, receiver operating characteristic (ROC) curve analysis conducted in 1,000 bootstrap replicates, and the calculated area under the curve (AUC). The analysis was stratified according to age, sex, and cancer type.

4. Nested case-control analysis

A nested case-control study was performed using predictors found as a result of machine-learning algorithms using the NHID (Figure 1). In the nested case-control study, cases of a disease that occurred in a defined cohort were identified, and, for each, a specified number of matched controls were selected from among those in the cohort who did not develop the disease by the time of the disease's occurrence in the case (Figure 2). Since cases and controls have the same follow-up time, bias related to multiple follow-up times that may occur in survival analysis, such as immortal time bias, can be eliminated.[41,42] In addition, death due to a suicide attempt is highly likely to be affected by cumulative risk factors until just before the onset, so it has the advantage of including the influence of time-varying factors compared to the existing time-invariant survival analysis that assumes the proportional risk.


Competing cause of death or loss to follow up
Event onset
Survival to the end of follow up
Start of follow-up
End of follow-up
: Potential control
: Potential control and was randomly selected as a control

In incidence density or risk set sampling, a control is randomly selected from all persons at risk, excluding the index case, at the time of the index case occurrence. This is repeated for each risk set. A selected control is still eligible to be selected again as a control for another case occurring at later time, if that person still has not had the outcome of interest and is still alive and under follow-up, and may become a case at a later time in follow-up.

Figure 2. Control sampling method: incidence density sampling

According to several studies on epidemiological methodologies, there is a time-varying Cox model, the landmark method,[42] and a nested case-control study design[43] for resolving immortal time bias and considering time-dependent exposure. The simplest way is to consider exposure at the time that it occurs; that is, if a patient died by suicide 27 months after the diagnosis of cancer and was exposed to a medication or psychotherapy of interest, the patient should be compared with others who were either exposed or unexposed up to month 27 and at risk of a suicidal event going forward in time.[44]

The date of the first CRC (i.e., C18–20) diagnosis in the cohort was set as the start of the follow-up period. The control group was chosen in a one-to-five ratio (total, n

= 10,996) at the time of suicide (n = 1,839), with age- and sex-matched individuals. Covariates were adjusted for the health insurance premium percentile at the time of suicide and comorbid mental disorders such as sleep disorder, major depressive and manic episodes, bipolar disorder, schizophrenia, schizotypal disorder, brief psychotic disorder, and schizoaffective disorder.

In the machine-learning analysis, predictors with the highest variable importance were used. Psychiatric drugs (hypnotics, sedatives, antidepressants, antipsychotics, opioids, etc.), psychotherapy, colostomy surgery, and Foley catheterization, which had the highest variable importance, were analyzed. They were coded as a dummy variable by adding up the number of diagnoses and prescriptions over 0–6, 6–12, 12–24, 24–48, and 48+ month intervals, as in machine-learning analysis, to check the effect of temporal distance from suicide. The variables above were also coded as the number of diagnoses and prescriptions 6 months before the onset of suicide, identifying the effect of predictive factors that did not consider temporal distance.

Conditional logistic regression analysis was performed to investigate the magnitude of association for predictors of suicide estimated using classification, regression tree, and random forest models. Conditional logistic regression analysis yielded odds ratios and 95% confidence intervals. Analyses were performed using SAS 9.4 (SAS Institute, Inc., Cary, NC, USA).

5. Operating characteristics of high-risk thresholds

Cross-validated RF predictive probabilities were ranked, and operating characteristics were calculated for individuals within the top 50% of the predicted risk distribution. The random forest model trained with undersampled data (n = 3,678) was used to calculate the suicide risk probability for the entire CRC cohort sample (n = 380,569). Sensitivity, specificity, and PPV were calculated based on 1%, 5%, 10%, 20%, 25%, 30%, and 50%, respectively, based on the calculated risk probability and a precision-recall curve presented (Table 7 and Figure 8). Suicide cases in the training set (n = 1,380) were excluded because their inclusion would likely result in optimal performance results. A data set matching the prevalence of suicide deaths (0.48%) in the original data to the number of cases in the test set (n = 459) was randomly selected (n = 95,625) and analyzed. (Table 7 and Figure 8).

6. Number needed to screen

To evaluate the effectiveness of the predictive models discovered in this study in the clinical environment, the number needed to screen (NNS) the model was calculated (Table 8). A PPV is useful for assessing the performance of screening tools, but its value also depends heavily on the prevalence of the disease. Most tool validity is presented as a relative risk reduction, ignoring the role of event rate in the overall clinical benefits. For example, when presented as a relative risk reduction, a highly effective screening tool for diseases with a low mortality rate would appear better than a less effective tool for diseases with higher rates. Absolute metrics, rather than relative metrics, are required for rigorous predictive model evaluation. The NNS is a statistic defined as the number of individuals who need to undergo screening to prevent one death or one side effect.[31] NNS is the reciprocal of the absolute risk reduction due to screening. Absolute risk reduction is the absolute difference between the unscreened mortality rate and the reduced mortality rate attributable to intervention after screening. NNS is an indicator that can help determine which tests should be performed first in situations where medical resources are limited. The smaller the number, the greater the benefits of screening tests. The effectiveness of the interventions due to the screening provided by the machine-learning model developed in this study is currently unknown. Therefore, the NNS was presented, assuming that the reduction rate of suicide deaths due to the intervention varied from 25 to 100%. The detailed results are summarized in Table 8 and Figure 9.

7. Suicide risk score-card

The predictive model in this study is an automatic predictive model based on passively collected claims data without evaluation by clinicians. However, a scorecard can also be created that allows clinicians to directly assess the level of suicide risk in CRC patients, referring to the existing research.[45] allows clinicians to assess a patient's risk score using the top 10 predictors found prominent in the random forest model of the study and has the potential to provide measures to prevent suicide death using a predefined cutoff score. To construct an easily used clinical scorecard (Table 9a), a logistic regression model with the top 10 prominent predictors found in the random forest model was developed (Table 9b), and the regression coefficients of the predictors from the final model were standardized by dividing all regression coefficients by the smallest coefficient and rounding off the results. To enhance the clinical utility, the final regression model was converted into a score table, which can be used as a clinical prediction model. The risk score was calculated for each participant, and the operating characteristics were generated (Table 9c).

III. RESULTS

1. Study sample selection

Figure 3 depicts the workflow for the selection of study samples. Only CRC patients (n=380,569, ICD-10 code C18-20) were extracted from the NHID consisting of patients with malignant neoplasm of colon, rectum, or anus (n=549,939). Of all CRC patients, 1,839 died from suicide attempts during the study period (2002-2018), which was 0.48% of the total CRC patients. Among non-suicidal CRC patients (n=378,730), 1,839 non-suicidal groups were selected through random sample allocation, excluding 376,891, and were used to build a machine learning model (n=3,678). 75% of this group (n=2,759) was used as a training set, and the remaining 25% (n=919) was used as a test set to measure performance. In order to analyze the magnitude of association within the cohort of major predictors obtained as a result of machine learning analysis, a nested case-control sample that matched age and gender in one-to-five ratio was extracted. The matched control group was 9,157, and 10,996 were selected for the nested case-control analysis sample.

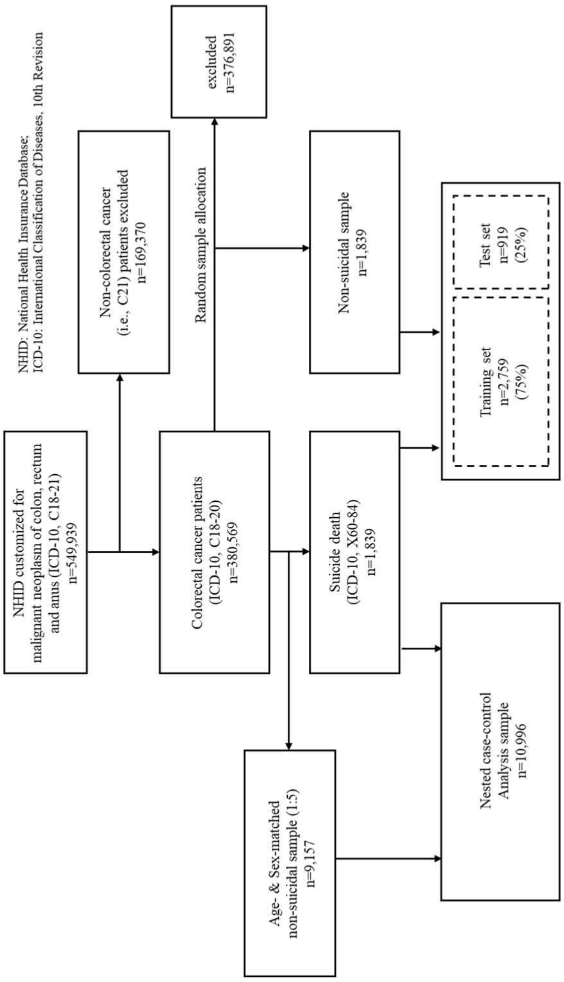

NHID: National Health Insurance Database;
ICD-10: International Classification of Diseases, 10th Revision
NHID customized for malignant neoplasm of colon, rectum and anus (ICD-10, C18-21) n=549,939
Non-colorectal cancer (i.e., C21) patients excluded n=169,370
Colorectal cancer patients (ICD-10, C18-20) n=380,569
Random sample allocation
excluded n=376,891
Non-suicidal sample n=1,839
Suicide death (ICD-10, X60-84) n=1,839
Training set n=2,759 (75%)
Test set n=919 (25%)
Age- & Sex-matched non-suicidal sample (1:5) n=9,157
Nested case-control Analysis sample n=10,996

Figure 3. Study Samples selection workflow for developing prediction model and nested case-control analysis

2. General characteristics of total sample

Table 1 summarizes the general characteristics of patients (n=380,569) who were diagnosed with CRC (ICD-10, C18-20) at least once during the study period from 2002 to 2018 among the NHID used in this study. The suicide group accounted for 0.48% (n=1,839), of the total, and deaths from other causes accounted for 27.78% (n=105,725) of the total. Subjects who died from suicide were older than subjects who neither committed suicide nor died from other causes (Mean [SD]; suicide death 63.54 [11.61]). The insurance premium decile of suicide group was lower than that of other groups (6.3 [2.85]). Male (0.36%, n=1,377) had a higher proportion of suicide than females (0.12%, n=462). The suicide was higher in the rectal cancer group (0.64%) than in the colon cancer group (0.43%). Patients with a diagnosis of any type of psychiatric disorder (0.58%) had a higher proportion of suicide than the patients without any disorder (0.41%). Specifically, suicide was higher in psychoactive disorder (F10-19) using psychoactive substances, schizophrenia, schizophrenia, and delusional disorder (F20-F29), Mood [affective] disorders (F30-F39), neurotic, stress-related and somatic disorders (F40-F48), and physiological disorders and behavioral syndromes with physical factors (F50-F59).

Table 1. Demographic characteristics of suicidal death in colorectal cancer (ICD-10 code: C18-20) patients from Korean National Health Insurance Service (2002-2018) (n=380,569)

	Suicide death, N(%)				
	Total	Survivors	Suicide death	Non suicide death	
	380,569 (100)	273,005 (71.74)	1,839 (0.48)	105,725 (27.78)	
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	p-value *
Age (in years)	61.05 (14.06)	58.05 (13.47)	63.54 (12.77)	68.78 (12.55)	<.0001
Insurance premium decile (1-10)	6.44 (2.81)	6.42 (2.78)	6.3 (2.85)	6.48 (2.86)	0.02
	N (%)	N (%)	N (%)	N (%)	
Colon cancer type					
Proximal colon	276771 (72.73)	206720 (74.69)	1199 (0.43)	68852 (24.88)	<.0001
Distal colon with rectum	23476 (6.17)	15367 (65.46)	128 (0.55)	7981 (34)	
Rectum	80322 (21.11)	50918 (63.39)	512 (0.64)	28892 (35.97)	
Sex					<.0001
Male	204202 (53.66)	140689 (68.90)	1377 (0.67)	62136 (30.43)	
Female	176367 (46.34)	132316 (75.02)	462 (0.26)	43589 (24.71)	
Colorectal cancer stage					<.0001
Stage unknown	263126 (69.14)	190947 (72.57)	1233 (0.47)	70946 (26.96)	
Stage 1	75752 (19.9)	57234 (75.55)	413 (0.55)	18105 (23.9)	
Stage 2	17836 (4.69)	11040 (61.9)	95 (0.53)	6701 (37.57)	
Stage 3	705 (0.19)	324 (45.96)	7 (0.99)	374 (53.05)	
Stage 4	23150 (6.08)	13460 (58.14)	91 (0.39)	9599 (41.46)	
Comorbidity					
All mental disorder (F10-98)					<.0001
not diagnosed	223520 (58.73)	159799 (71.49)	927 (0.41)	62794 (28.09)	
diagnosed	157049 (41.27)	113206 (72.08)	912 (0.58)	42931 (27.34)	
Mental and behavioral disorders due to the use of psychoactive substances (F10-F19)					<.0001
not diagnosed	376168 (98.94)	270188 (71.83)	1797 (0.48)	104183 (27.7)	
diagnosed	4401 (1.16)	2817 (64.01)	42 (0.95)	1542 (35.04)	
Schizophrenia, schizotypal and delusional disorders (F20-F29)					<.0001
not diagnosed	378553 (99.47)	271671 (71.77)	1818 (0.48)	105064 (27.75)	
diagnosed	2016 (0.53)	1334 (66.17)	21 (1.04)	661 (32.79)	
Mood [affect] disorders (F30-F39)					<.0001
not diagnosed	350142 (92)	250239 (71.47)	1616 (0.46)	98287 (28.07)	
diagnosed	30427 (8)	22766 (74.82)	223 (0.73)	7438 (24.45)	
Neurotic, stress-related and somatic disorders (F40-F48)					<.0001
not diagnosed	286184 (75.2)	202561 (70.78)	1345 (0.47)	82278 (28.75)	
diagnosed	94385 (24.8)	70444 (74.63)	494 (0.52)	23447 (24.84)	
Behavioral Syndrome with Physiological Disorders and Physical Factors (F50-F59)					<.0001
not diagnosed	366622 (96.34)	263380 (71.84)	1744 (0.48)	101498 (27.68)	
diagnosed	13947 (3.66)	9625 (69.01)	95 (0.68)	4227 (30.31)	
Disorders of Adult Personality and Behavior (F60-F69)					0.08
not diagnosed	380381 (99.95)	272869 (71.74)	1836 (0.48)	105676 (27.78)	

diagnosed	188 (0.05)	136 (72.34)	3 (1.6)	58 (27.49)	
Mental retardation (F70-F79)					0.59
not diagnosed	380358 (99.94)	272852 (71.74)	1839 (0.48)	105667 (27.78)	
diagnosed	211 (0.06)	153 (72.51)	0 (0)	58 (27.49)	
Mental Developmental Disorder (F80-F89)					0.80
not diagnosed	380494 (99.98)	272952 (71.74)	1839 (0.48)	105703 (27.78)	
diagnosed	75 (0.02)	53 (70.67)	0 (0)	22 (29.33)	
Behavioral and emotional disorders with a primary onset in childhood and adolescence (F90-F98)					0.06
not diagnosed	380185 (99.9)	272709 (71.73)	1838 (0.48)	105638 (27.79)	
diagnosed	384 (0.1)	296 (77.08)	1 (0.26)	87 (22.66)	

* Significant test is evaluated with t-test, ANOVA and chi-square test.

3. General characteristics of machine learning sample

Table 2 summarizes the general characteristics of sub-cohorts for machine learning analysis. Age was significantly higher in the suicide group, and there was no significant difference in insurance premium. The proportion of male was significantly higher in the suicide group (74.88%). The proportion of tumor involving rectum was significantly higher in the suicide group than in the non-suicidal group (27.84%). There was no significant difference between the two groups in the stage of CRC. The diagnosis frequency of sleep disorder (F51), schizophrenia, schizotypal and delusional disorders (F20-F25), and major depressive disorder (F32) was significantly higher in the suicide group.

Table 2. Demographic characteristics of machine learning sample from suicidal death in colorectal cancer (ICD-10 code: C18-20) patients from Korean National Health Insurance Service (2002-2018)

	Suicide death, N(%)			
	Total, N(%) 3678 (100)	No suicide 1839 (50)	Suicide 1839 (50)	
	Mean (STD)	Mean (STD)	Mean (STD)	p-value *
Age (in years)	62.63 (13.43)	61.71 (14)	63.54 (12.77)	<.0001
Insurance premium decile (0-10)	6.36 (2.83)	6.42 (2.81)	6.3 (2.85)	0.19
Variables	N (%)	N (%)	N (%)	
Colon cancer type				
Proximal colon	2530 (68.79)	1331 (72.38)	1199 (65.2)	<.0001
Distal colon with rectum	241 (6.55)	113 (6.14)	128 (6.96)	
Rectum	907 (24.66)	395 (21.48)	512 (27.84)	
Sex				0.99
Male	2368 (64.38)	991 (53.89)	1377 (74.88)	
Female	1310 (35.62)	848 (46.11)	462 (25.12)	
Colorectal cancer stage				0.21
Stage unknown	2471 (67.18)	1238 (67.32)	1233 (67.05)	
Stage 1	796 (21.64)	383 (20.83)	413 (22.46)	
Stage 2	189 (5.14)	94 (5.11)	95 (5.17)	
Stage 3	11 (0.3)	4 (0.22)	7 (0.38)	
Stage 4	211 (5.74)	120 (6.53)	91 (4.95)	
Comorbidity				
Sleep disorder (F51)				<.0001
not diagnosed	3291 (89.48)	1755 (95.43)	1536 (83.52)	
diagnosed	387 (10.52)	84 (4.57)	303 (16.48)	
Schizophrenia, schizotypal and delusional disorders (F20-F25)				<.0001
not diagnosed	3602 (97.93)	1821 (99.02)	1781 (96.85)	
diagnosed	76 (2.07)	18 (0.98)	58 (3.15)	
Major depressive disorder (F32)				<.0001
not diagnosed	3034 (82.49)	1693 (92.06)	1341 (72.92)	
diagnosed	644 (17.51)	146 (7.94)	498 (27.08)	

* Significant test is evaluated with t-test, ANOVA and chi-square test.

4. Classification and regression trees

Figures 4a-i are classification tree diagrams showing the predictors of suicide in CRC patients claimed to NHIS from 2002 to 2018, and the predictors were analyzed by stratification by age, gender and cancer diagnosis type. Among total patients (n=3,678), the highest predictive factor was the prescribing of psychotherapy with 0-6 months prior to the onset of the suicide (Figure 4a). Gender was more predictive of suicide in men than in women. Sleeping pills prescribed 0-6 months before the onset of the suicide were also predictive factors for suicide, followed by CRC diagnosis record within 0-6 month, mood stabilizer prescription, and urinary catheterization 0-6 months prior.

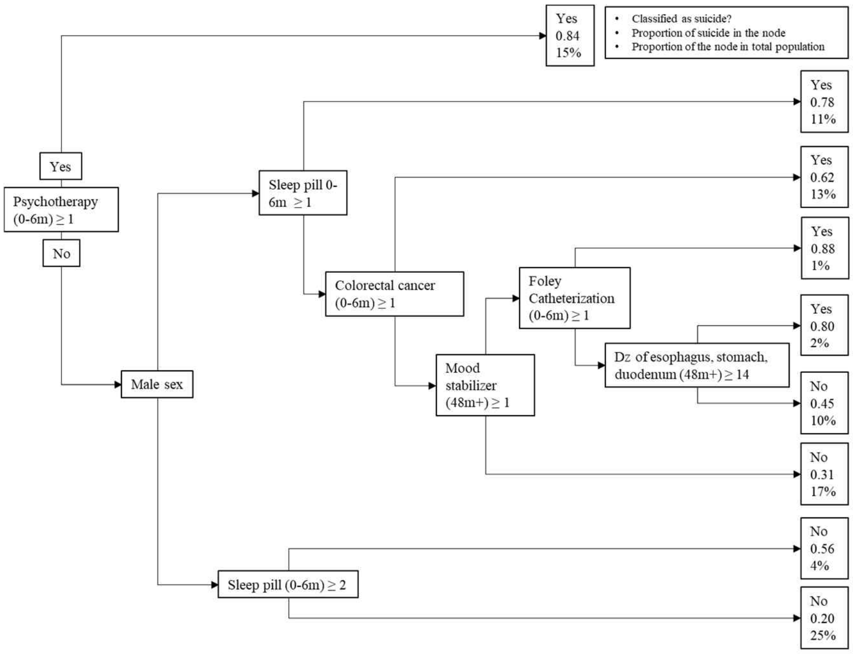
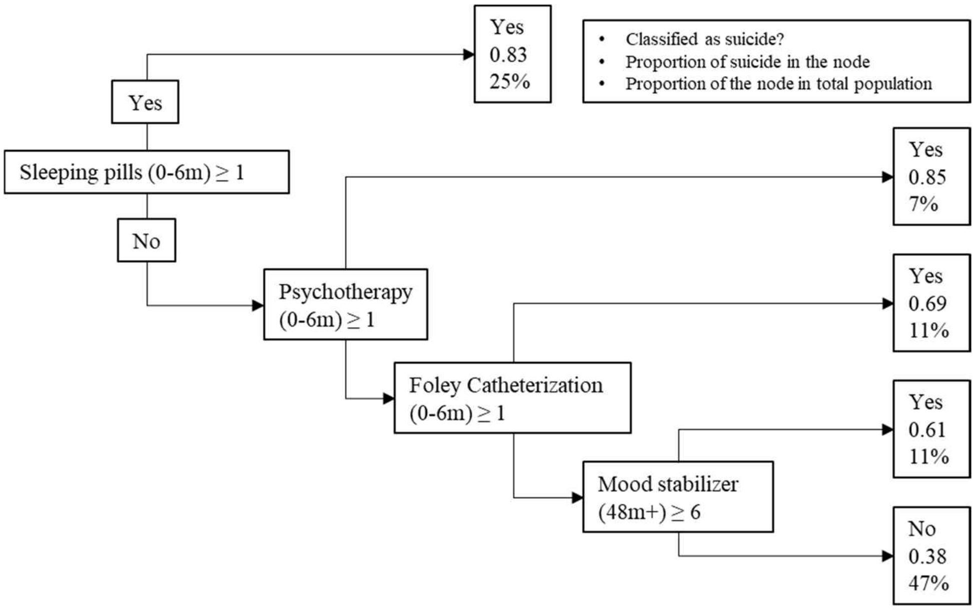

Yes
0.84
15%
• Classified as suicide?
• Proportion of suicide in the node
• Proportion of the node in total population
Yes
0.78
11%
Yes
Psychotherapy (0-6m) ≥ 1
No
Sleep pill 0-6m ≥ 1
Yes
0.62
13%
Yes
0.88
1%
Colorectal cancer (0-6m) ≥ 1
Foley Catheterization (0-6m) ≥ 1
Yes
0.80
2%
Dz of esophagus, stomach, duodenum (48m+) ≥ 14
Male sex
Mood stabilizer (48m+) ≥ 1
No
0.45
10%
No
0.31
17%
No
0.56
4%
Sleep pill (0-6m) ≥ 2
No
0.20
25%

Figure 4a. Classification tree depicting suicide attempt predictors among total colorectal cancer patient in Korea, 2002–2018. (N=3,678)

Since gender was found to be a prominent predictor, the predictor search was stratified by gender (Figure 4b&4c). Figure 4b is a classification tree for predictors in men. Prescription of sleeping pills 1 or more times 0-6 months before the onset of suicide was the most first predictor of the classification tree for men. Once or more prescription of psychotherapy 0-6 months before onset, 6 or more mood stabilizer prescriptions 48 months before the onset, and Foley catheterization prescription (0-6m) followed. For women (Figure 4c), the prescribing individual psychotherapy 0-6 months before the onset was the first predictive factor of the tree for women. It was followed by prescriptions of sleeping pills 2 or more times 0-6 months before the suicide, enteral nutrition (0-6m) before, psychotherapy (12-24m) and enema (0-6m).



Yes 0.83 25%
Classified as suicide?
Proportion of suicide in the node
Proportion of the node in total population
Yes
Sleeping pills (0-6m) ≥ 1
No
Yes 0.85 7%
Psychotherapy (0-6m) ≥ 1
Yes 0.69 11%
Foley Catheterization (0-6m) ≥ 1
Yes 0.61 11%
Mood stabilizer (48m+) ≥ 6
No 0.38 47%

Figure 4b. Classification tree depicting suicide attempt predictors among male colorectal cancer patient in Korea, 2002–2018. (N=2,368)

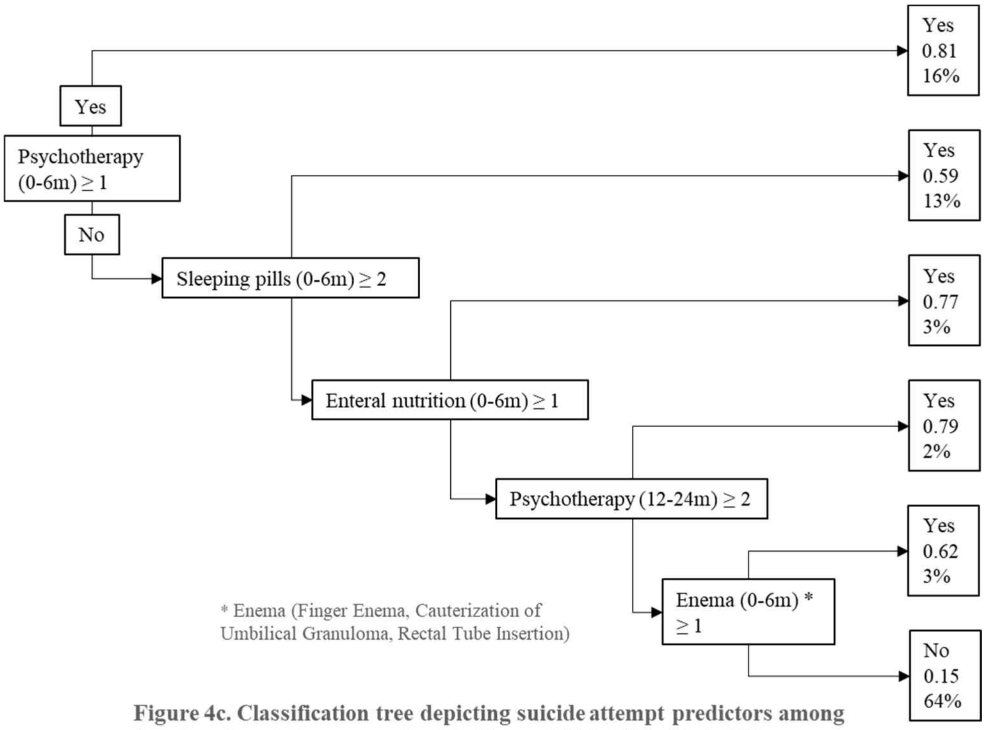

Psychotherapy (0-6m) ≥ 1
Yes
No
Yes 0.81 16%
Sleeping pills (0-6m) ≥ 2
Yes 0.59 13%
Enteral nutrition (0-6m) ≥ 1
Yes 0.77 3%
Psychotherapy (12-24m) ≥ 2
Yes 0.79 2%
Enema (0-6m) * ≥ 1
Yes 0.62 3%
No 0.15 64%
* Enema (Finger Enema, Cauterization of Umbilical Granuloma, Rectal Tube Insertion)

Figure 4c. Classification tree depicting suicide attempt predictors among female colorectal cancer patient in Korea, 2002–2018. (N=1,310)

For patients in their 10s and 20s, the predictive factor for suicide was 2 or more psychiatric outpatient visits, followed by use of sedative more than once 48 month prior. For 30s, use of mood stabilizer more than once 48 months before was the first predictor of the tree, followed by acute lower respiratory infection (48m+), disease of digestive system (24-48m), sleeping pills (0-6m), and recent diagnosis of CRC (6-12m) (Figure 4d).

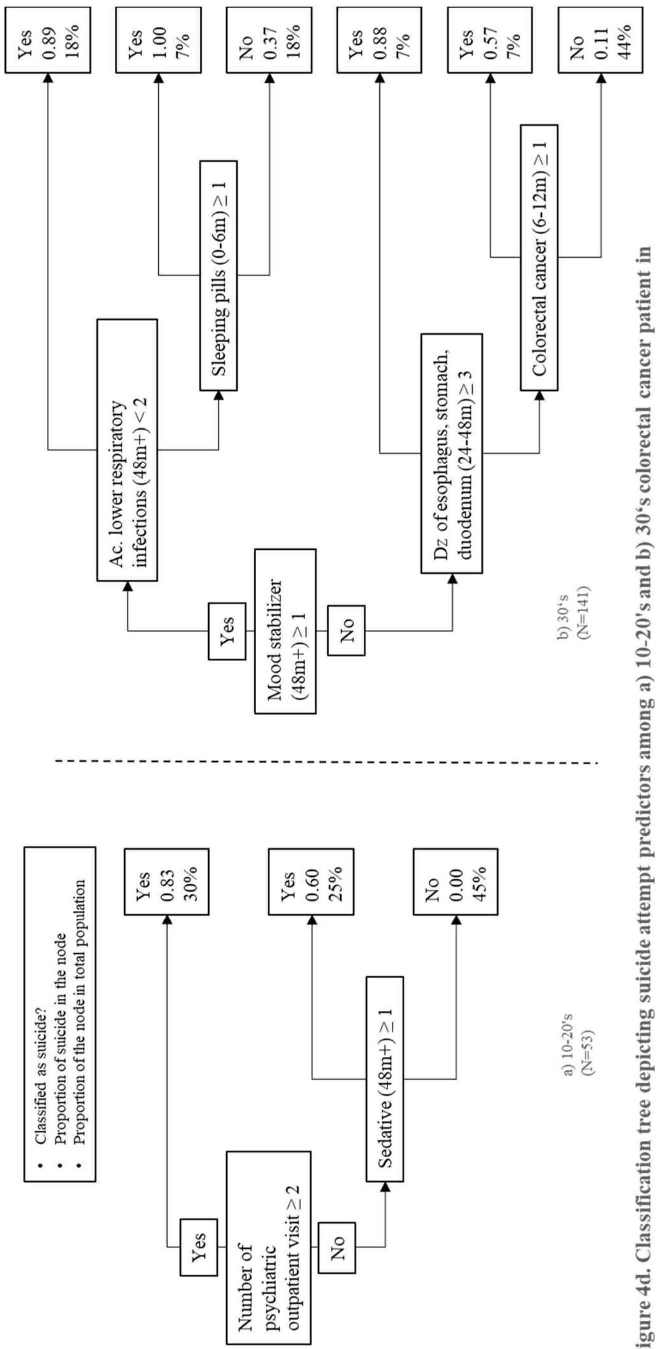

Classified as suicide?
Proportion of suicide in the node
Proportion of the node in total population
Number of psychiatric outpatient visit ≥ 2
Yes
No
Yes 0.83 30%
Sedative (48m+) ≥ 1
Yes 0.60 25%
No 0.00 45%
a) 10-20's (N=53)
Mood stabilizer (48m+) ≥ 1
Yes
No
Ac. lower respiratory infections (48m+) < 2
Yes 0.89 18%
Sleeping pills (0-6m) ≥ 1
Yes 1.00 7%
No 0.37 18%
Dz of esophagus, stomach, duodenum (24-48m) ≥ 3
Yes 0.88 7%
Colorectal cancer (6-12m) ≥ 1
Yes 0.57 7%
No 0.11 44%
b) 30's (N=141)

Figure 4d. Classification tree depicting suicide attempt predictors among a) 10-20's and b) 30's colorectal cancer patient in Korea, 2002–2018. a) (N=53), b) (N=141)

The most prominent suicide predictor of patients in their 40s was the prescription of psychotherapy 0-6 months ago. Male sex and disorder diagnosis related to skin were also prominent predictors (Figure 4e). The predictive factors for suicide in their 50s were psychotherapy 0-6 months ahead, male sex, CRC diagnosis within 6-12 months and prescription of a sleeping pills 5 or more times 0-6 months ago (see Figure 4f). The first predictors appeared in the tree for CRC patients in their 60s were the prescription of sleeping pills 0-6 months ago, followed by urinary catheterization (0-6m), male sex, disorder of lens (48m+) and recent CRC diagnosis (Figure 4g).

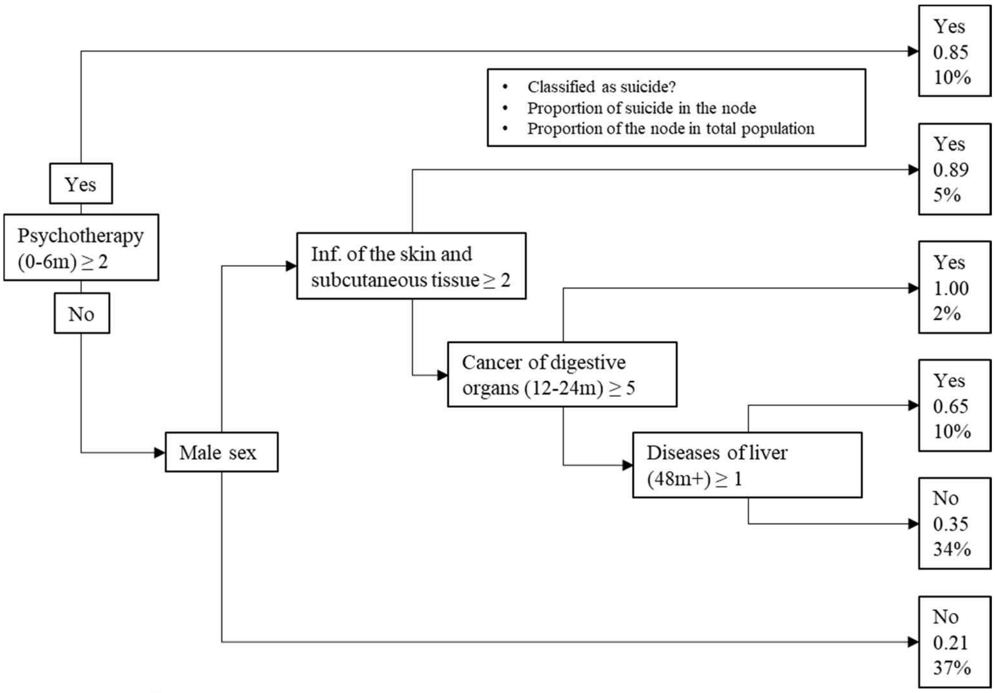

Yes
0.85
10%
• Classified as suicide?
• Proportion of suicide in the node
• Proportion of the node in total population
Yes
0.89
5%
Yes
Psychotherapy (0-6m) ≥ 2
No
Inf. of the skin and subcutaneous tissue ≥ 2
Yes
1.00
2%
Cancer of digestive organs (12-24m) ≥ 5
Yes
0.65
10%
Male sex
Diseases of liver (48m+) ≥ 1
No
0.35
34%
No
0.21
37%

Figure 4e. Classification tree depicting suicide attempt predictors among 40's colorectal cancer patient in Korea, 2002–2018. (N=440)

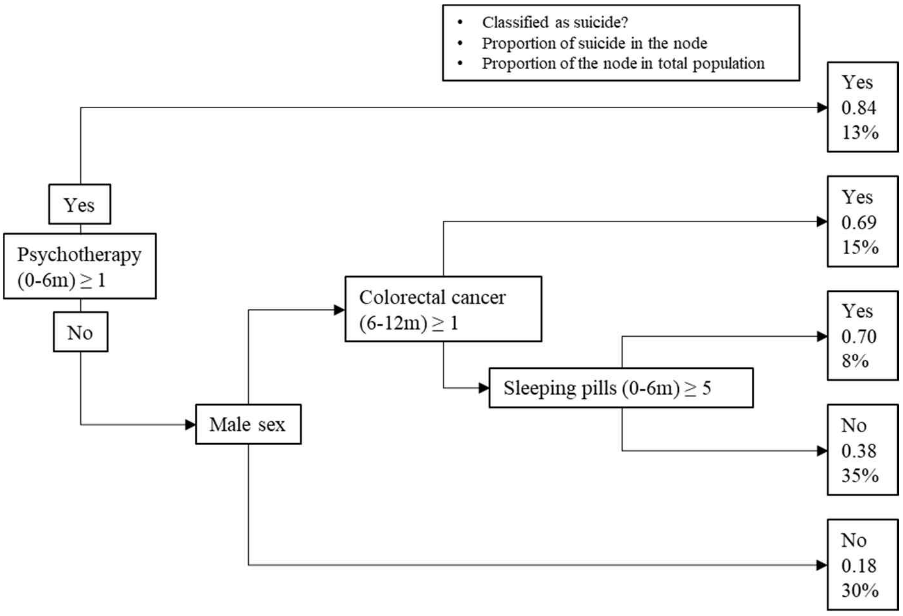

Classified as suicide?
Proportion of suicide in the node
Proportion of the node in total population
Yes
Psychotherapy (0-6m) ≥ 1
No
Male sex
Colorectal cancer (6-12m) ≥ 1
Sleeping pills (0-6m) ≥ 5
Yes 0.84 13%
Yes 0.69 15%
Yes 0.70 8%
No 0.38 35%
No 0.18 30%

Figure 4f. Classification tree depicting suicide attempt predictors among 50's colorectal cancer patient in Korea, 2002–2018. (N=758)

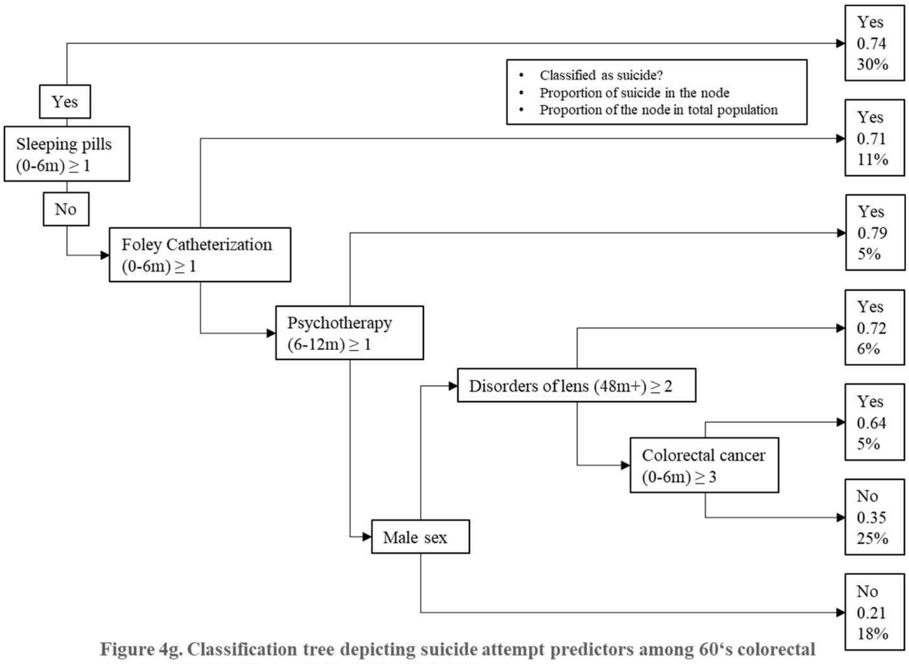

Classified as suicide?
Proportion of suicide in the node
Proportion of the node in total population
Yes
Sleeping pills (0-6m) ≥ 1
No
Foley Catheterization (0-6m) ≥ 1
Psychotherapy (6-12m) ≥ 1
Disorders of lens (48m+) ≥ 2
Colorectal cancer (0-6m) ≥ 3
Male sex
Yes 0.74 30%
Yes 0.71 11%
Yes 0.79 5%
Yes 0.72 6%
Yes 0.64 5%
No 0.35 25%
No 0.21 18%

Figure 4g. Classification tree depicting suicide attempt predictors among 60's colorectal cancer patient in Korea, 2002–2018. (N=2,286)

Among CRC patients, in the group diagnosed with ICD-10 codes C18 and C19 (N=2,771), the first appeared predictor was prescribing psychotherapy 0-6 months prior. Male sex and one or more prescriptions of sleeping pills 0-6 months ago, recent CRC diagnosis within 6-12 month, one or more diagnosis of spondylopathy (12-24m), recent urinary catheterization (0-6m) were the followed (Figure 4h). In the group with rectal cancer (N=907), the first appeared predictor of suicide was recent urinary catheterization within 0-6-month prior, followed by male sex and recent prescription of psychotherapy within 0-6-month (Figure 4i).

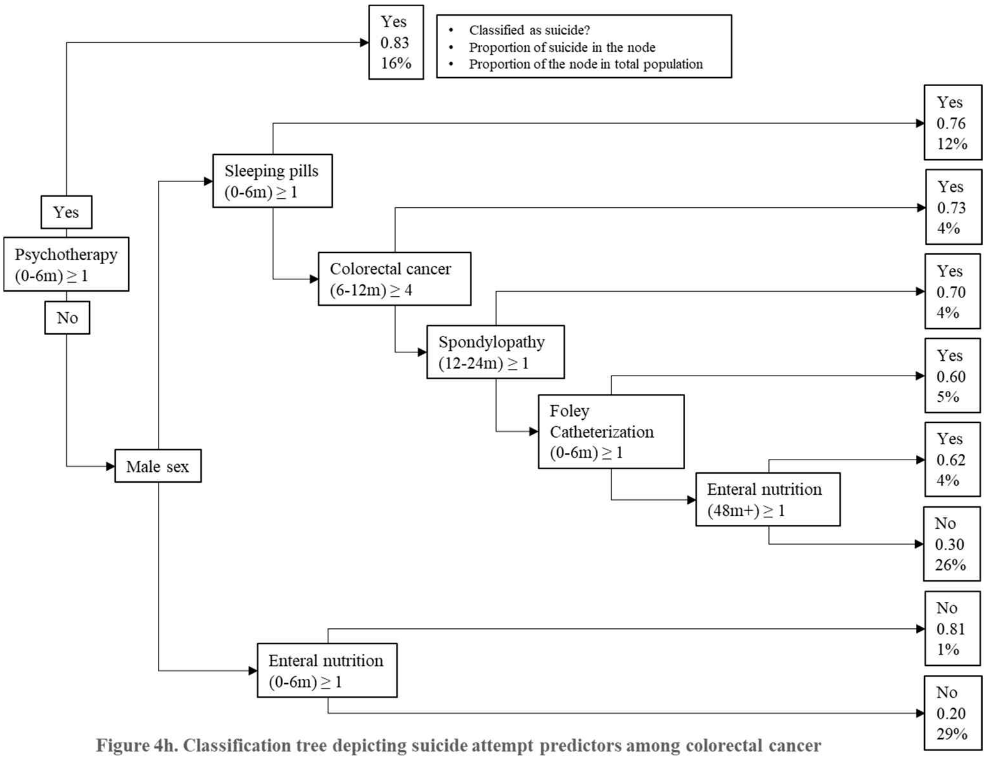

Yes
0.83
16%
Classified as suicide?
Proportion of suicide in the node
Proportion of the node in total population
Yes
Psychotherapy (0-6m) ≥ 1
No
Male sex
Sleeping pills (0-6m) ≥ 1
Colorectal cancer (6-12m) ≥ 4
Spondylopathy (12-24m) ≥ 1
Foley Catheterization (0-6m) ≥ 1
Enteral nutrition (48m+) ≥ 1
Enteral nutrition (0-6m) ≥ 1
Yes 0.76 12%
Yes 0.73 4%
Yes 0.70 4%
Yes 0.60 5%
Yes 0.62 4%
No 0.30 26%
No 0.81 1%
No 0.20 29%

Figure 4h. Classification tree depicting suicide attempt predictors among colorectal cancer patient diagnosed with C18 and C19, 2002–2018. (N=2,771)

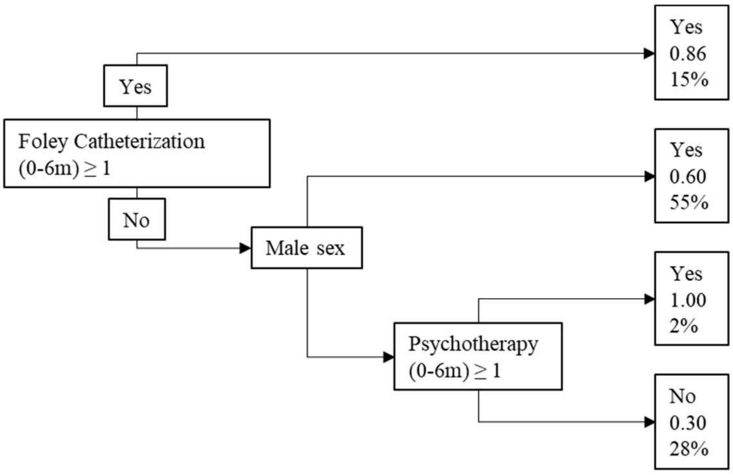


Figure 4i. Classification tree depicting suicide attempt predictors among colorectal cancer patient diagnosed with C20, 2002–2018. (N=907)

5. Predictor evaluations and variable importance

Tables 3a-c summarizes the top 10 predictors of suicide found in the classification tree and random forest in order of variable importance. These tables present only the top 10 items in the order of importance of each variable depicted in Figures 5a-j. The top 10 predictors of the full-fledged tree without hyper-parameter tuning and the those of the pruned tree with hyper-parameter tuning are very similar (Table 3a, b). The important predictors for the total sample were gender, prescription of sleeping pills, psychotherapy, and mood stabilizer. The importance of these variables showed a tendency to be greater the closer to the time of suicide. For example, sleeping pill prescribed 0-6 months prior appeared to be a more important predictor than that prescribed 12-24 months prior.

It is noteworthy to mention that there is a marked difference in the ranking of predictors between males and females. In both sexes, psychotherapy, prescription of sleeping pills, and recent CRC diagnosis were found to be common high-ranking predictors. However, in women, treatments related to hospitalization and severity of the cancer, such as enema and enteral nutrition seem to be important predictors, while in men, it was not included in the top 10 predictors.

Another notable difference was found among different age groups. For those in their 10′s, 20′s, and 30′s, outpatient visits to psychiatry, psychotherapy, digestive and cardiopulmonary complications, Acute upper respiratory infections, and insurance premium were the major predictors. As age increased, sex become prominent predictive factors while sleeping pills, psychotherapy, recent diagnosis of CRC, and number of psychiatric inpatients and outpatient visit were also prominent.

Table 3a. Top 10 predictor in variable importance results for classification and regression trees (full-fledged tree)

	Total population top 10 predictors	Sex-specific top 10 predictors		Age-specific top 10 predictors					Type-specific top 10 predictors	
Rank	**Total full-fledged tree**	**Male full-fledged tree**	**Female full-fledged tree**	**10-20's full-fledged tree**	**30's full-fledged tree**	**40's full-fledged tree**	**50's full-fledged tree**	**60's full-fledged tree**	**ICD C18, 19**	**ICD C20**
1	sleeping pills (0-6m)	Psychotherapy (12-24m)	sleeping pills (0-6m)	Symptoms involving the digestive system and abdomen_48m	Diseases of esophagus, stomach and duodenum (24-48m)	Sleeping pills (0-6m)	Sex	Cancer of colon, rectosigmoid junction, and rectum (48m+)	Sleeping pills (0-6m)	Psychotherapy (0-6m)
2	Sex	Psychotherapy (0-6m)	Psychotherapy (12-24m)	Number of psychiatric outpatient visit	Psychotherapy (48m+)	Cancer of digestive organs (12-24m)	Sleeping pills (0-6m)	Psychotherapy (6-12m)	Sleeping pills (24-48m)	Sleeping pills (0-6m)
3	Foley Catheterization (0-6m)	Sleepin pills (0-6m)	Psychotherapy (0-6m)	Psychotherapy 12-24m	Mood stabilizer (48m+)	Psychotherapy (48m+)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Sleepin pills (0-6m)	Foley Catheterization (0-6m)	Other diseases of intestines (0-6m)
4	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Mood stabilizer (48m+)	Number of psychiatric outpatient visit	insurance premium	sleeping pills (12-24m)	Foley Catheterization (0-6m)	Cancer of colon, rectosignmoid junction, and rectum (6-12m)	Sex	Psychotherapy (0-6m)	Sex
5	Psychotherapy (0-6m)	Cancer of colon, rectosignmoid junction, and rectum (24-48m)	Enema (0-6m)	Disorders of conjunctiva_48m	Psychotherapy (0-6m)	Psychotherapy (0-6m)	Psychotherapy (0-6m)	Psychotherapy (0-6m)	Psychotherapy (6-12m)	Cancer of colon, rectosigmoid junction, and rectum (6-12m)
6	Psychotherapy (12-24m)	Sleepin pills (6-12m)	Psychotherapy (6-12m)	Influenza and pneumonia_48m	Cancer of colon, rectosigmoid junction, and rectum (6-12m)	Other soft tissue disorders (48m+)	Psychotherapy (12-24m)	Cancer of colon, rectosigmoid junction, and rectum (24-48m)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Psychotherapy (12-24m)
7	Psychotherapy (6-12m)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Other acute lower respiratory infections_48m	Number of psychiatric outpatient visit	Diseases of liver (48m+)	insurance premium	Cancer of colon, recto rectosigmoid signmoid junction, and rectum (0-6m)	Psychotherapy (12-24m)	Other diseases of upper respiratory tract (48m+)
8	sleeping pills (24-48m)	Cancer of colon, rectosigmoid junction, and rectum (6-12m)	Foley Catheterization (0-6m)	Sedatives 48m Prior	Sleeping pills (24-48m)	Sex	Cancer of colon, rectosigmoid junction, and rectum (12-24m)	sleeping pills (6-12m)	Sex	Foley Catheterization (0-6m)
9	Mood stabilizer (48m+)	Sleepin pills (24-48m)	Tubal nutrition (0-6m)	Opioid 24-48m	sleeping pills (0-6m)	Psychotherapy (6-12m)	Other diseases of intestines (48m+)	Psychotherapy (24-48m)	Mood stabilizer (48m+)	Enema (0-6m)
10	sleeping pills (48m+)	Enema (0-6m)	Psychotherapy (48m+)		insurance premium	Psychotherapy (24-48m)	Psychotherapy (6-12m)	Sleeping pills (24-48m)	Opioids (48m+)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)

Table 3b. Top 10 predictor in variable importance results for classification and regression trees (pruned tree)

	Total population top 10 predictors	Sex-specific top 10 predictors		Age-specific top 10 predictors					Type-specific top 10 predictors	
Rank	**Total pruned tree**	**Male pruned tree**	**Female pruned tree**	**10-20's pruned tree**	**30's pruned tree**	**40's pruned tree**	**50's pruned tree**	**60's pruned tree**	**ICD C18, 19**	**ICD C20**
1	Sleeping pills (0-6m)	Psychotherapy (0-6m)	Sleeping pills (0-6m)	Number of psychiatric outpatient visit	Diseases of esophagus, stomach and duodenum (24-48m)	Sleepin pills (0-6m)	Sex	Psychotherapy (6-12m)	Sleeping pills (0-6m)	Psychotherapy (0-6m)
2	Sex	Psychotherapy (12-24m)	Psychotherapy (12-24m)	Sedative (48m+)	Psychotherapy (48m+)	Malignant neoplasms of digestive organs (12-24m)	Sleepin pills (0-6m)	Cancer of colon, rectosigmoid junction, and rectum (48m+)	Sleeping pills (24-48m)	Sex
3	Foley Catheterization (0-6m)	Sleeping pills (0-6m)	Psychotherapy (0-6m)	Psychotherapy (12-24m)	Mood stabilizer (48m+)	Psychotherapy (0-6m)	Psychotherapy (0-6m)	Sex	Foley Catheterization (0-6m)	Sleeping pills (0-6m)
4	Psychotherapy (0-6m)	Mood stabilizer (48m+)	Enema (0-6m)	Acute upper respiratory infections (6-12m)	Sleepin pills (12-24m)	Psychotherapy (48m+)	Cancer of colon, rectosigmoid junction, and rectum (6-12m)	Sleepin pills (0-6m)	Psychotherapy (0-6m)	Other diseases of intestines (0-6m)
5	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Sleeping pills (6-12m)	Number of psychiatric outpatient visit	Other diseases of intestines (48m+)	Psychotherapy (0-6m)	Sex	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Psychotherapy (0-6m)	Psychotherapy (6-12m)	Foley Catheterization (0-6m)
6	Psychotherapy (12-24m)	Sleeping pills (24-48m)	Psychotherapy (6-12m)	Opioids (48m+)	Cancer of colon, rectosigmoid junction, and rectum (6-12m)	Psychotherapy (6-12m)	Cancer of colon, rectosigmoid junction, and rectum (12-24m)	Cancer of colon, rectosigmoid junction, and rectum (24-48m)	Sex	Enema (0-6m)
7	Psychotherapy (6-12m)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Foley Catheterization (0-6m)	Other acute lower respiratory infections (48m+)	Number of psychiatric outpatient visit	Foley Catheterization (0-6m)	Psychotherapy (12-24m)	Sleeping pills (6-12m)	Psychotherapy (12-24m)	Psychotherapy (12-24m)
8	Sleeping pills (24-48m)	Psychotherapy (6-12m)	Tubal nutrition (0-6m)	Disorders of conjunctiva (48m+)	Sleeping pills (24-48m)	Psychotherapy (24-48m)	Psychotherapy (6-12m)	Sleeping pills (24-48m)	Mood stabilizer (48m+)	Major depressive disorder, single episode (24-48m)
9	Mood stabilizer (48m+)	Psychotherapy (24-48m)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	age (in year)	Sleeping pills (0-6m)	Number of psychiatric inpatient visit	Other diseases of intestines (48m+)	Diseases of male genital organs (48m+)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Cancer of colon, rectosigmoid junction, and rectum (6-12m)
10	Sleeping pills (48m+)	Foley Catheterization (0-6m)	Psychotherapy (48m+)		insurance premium	Diseases of liver (48m+)	Foley Catheterization (0-6m)	Psychotherapy (24-48m)	Opioids (48m+)	Sleeping pills (6-12m)

Table 3c. Top 10 predictor in variable importance results for random forests

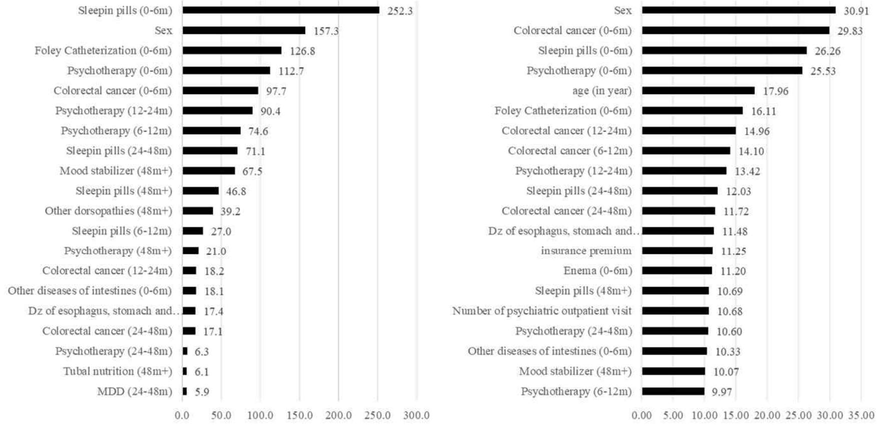
	Total population top 10 predictors	Sex-specific top 10 predictors		Age-specific top 10 predictors					Type-specific top 10 predictors	
Rank	**Total Random forest**	**Male Random forest**	**Female Random forest**	**10-20s Random forest**	**30's Random forest**	**40's random forest**	**50's random forest**	**60's Random forest**	**ICD C18, 19**	**ICD C20**
1	Sex	sleeping pills (0-6m)	Psychotherapy (0-6m)	Dermatitis and eczema (48m+)	Mood stabilizer (48m+)	sleeping pills (0-6m)	Malignant neoplasms of colon, rectosignmoid junction, and rectum (0-6m)	sleeping pills (0-6m)	sleeping pills (0-6m)	age (in year)
2	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	sleeping pills (0-6m)	Diseases of esophagus, stomach and duodenum (48m+)	Diseases of esophagus, stomach and duodenum (24-48m)	Sex	Sex	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Psychotherapy (0-6m)	Sex
3	sleeping pills (0-6m)	age (in year)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Injuries to the abdomen, lower back, lumbar spine, pelvis and external genitals (48m+)	Number of psychiatric outpatient visit	Psychotherapy (0-6m)	Psychotherapy (0-6m)	Other diseases of intestines (0-6m)	Sex	Cancer of colon, rectosigmoid junction, and rectum (0-6m)
4	Psychotherapy (0-6m)	Psychotherapy (0-6m)	Psychotherapy (12-24m)	Diseases of liver (48m+)	Sleepin pills (0-6m)	Cancer of digestive organs (12-24m)	Cancer of colon, rectosigmoid junction, and rectum (6-12m)	Psychotherapy (0-6m)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	Foley Catheterization (0-6m)
5	age (in year)	Mood stabilizer (48m+)	Psychotherapy (48m+)	age (in year)	Mood stabilizer (24-48m)	Cancer of colon, rectosigmoid junction, and rectum (12-24m)	sleeping pills (0-6m)	Foley Catheterization (0-6m)	Foley Catheterization (0-6m)	Enema (0-6m)
6	Foley Catheterization (0-6m)	sleeping pills (6-12m)	Enema (0-6m)	Other acute lower respiratory infections (24-48m)	Psychotherapy (48m+)	Foley Catheterization (0-6m)	Psychotherapy (12-24m)	Sex	age (in year)	Psychotherapy (0-6m)
7	Cancer of colon, rectosigmoid junction, and rectum (12-24m)	Cancer of colon, rectosigmoid junction, and rectum (6-12m)	Number of psychiatric outpatient visit	Acute upper respiratory infections (6-12m)	Symptoms involving the circulatory and respiratory systems (48m+)	Cancer of colon, rectosigmoid junction, and rectum (0-6m)	insurance premium	age (in year)	Psychotherapy (12-24m)	Cancer of colon, rectosigmoid junction, and rectum (24-48m)
8	Cancer of colon, rectosigmoid junction, and rectum (6-12m)	Cancer of colon, rectosigmoid junction, and rectum (12-24m)	Foley Catheterization (0-6m)	Atypical antipsychotic (48m+)	Psychotherapy (0-6m)	Infections of the skin and subcutaneous tissue (48m+)	Cancer of colon, rectosigmoid junction, and rectum (12-24m)	Cancer of colon, rectosigmoid junction, and rectum (24-48m)	Cancer of colon, rectosigmoid junction, and rectum (6-12m)	Cancer of colon, rectosigmoid junction, and rectum (6-12m)
9	Psychotherapy (12-24m)	insurance premium	Tubal nutrition (0-6m)	Psychotherapy (48m+)	Sleepin pills (12-24m)	Psychotherapy (6-12m)	Psychotherapy (6-12m)	Cancer of colon, rectosigmoid junction, and rectum (12-24m)	Mood stabilizer (48m+)	sleeping pills (0-6m)
10	Sleepin pills (24-48m)	Foley Catheterization (0-6m)	Psychotherapy (24-48m)	Other acute lower respiratory infections (48m+)	Malignant neoplasms of colon, rectosignmoid junction, and rectum (6-12m)	Number of psychiatric outpatient visit	Foley Catheterization (0-6m)	Malignant neoplasms of colon, rectosignmoid junction, and rectum (48m+)	Psychotherapy (6-12m)	Malignant neoplasms of colon, rectosignmoid junction, and rectum (12-24m)

In the group stratified by type of the cancer, sleeping pills, psychotherapy, gender and recent CRC diagnosis were shared as the most common predictive factors, but the order of importance was slightly different. In the CRC (i.e., C18-19) group, sleeping pill prescription was the main predictor in the order of temporal proximity to onset of suicide, followed by psychotherapy, sex, mood stabilizer and recent diagnosis of CRC. In the rectal cancer (i.e., C20) group, psychotherapy was a more important factors than sleeping pills, and gastrointestinal comorbidity, Foley catheterization, enema, diagnosis of major depressive disorder and recent diagnosis of CRC were followed as major predictors.

Table 3c also summarizes the variable importance found in random forests in order. When using a random forest, it is often very difficult to identify the tree structures (random forest creates 1,000 trees by default), therefore checking the importance of variables would be the most feasible way to identify major predictors. As shown in Table 3c and Figures 5a-j, age and gender are two of the most important predictors in the random forest, which could be data-driven rationale for applying the classification tree and random forest model by subdividing them into age and gender. Recent diagnosis of CRC, sleeping pills prescription, psychiatric outpatient visit, diagnosis of gastrointestinal complications, psychotherapy, and urinary catheterization were important predictors in the total sample.

In men, sleeping pill, recent diagnosis of the cancer, age, psychotherapy and mood stabilizer, insurance premium, and urinary catheterization, in order, were major predictors of suicide. In the case of women, psychotherapy, psychiatric outpatient visits and urinary catheterization, and enteral nutrition were important predictors than prescription of psychiatric drugs. In the age group in their 10s and 20s, dermatitis, gastrointestinal complication, lower and upper respiratory infection were important predictor, while in 30s, prescription of psychiatric drugs was more important. Gender, psychotherapy, recent diagnosis of CRC were the most important predictors in the age of 40 or older. In stratification by the cancer type (Figure 5i-j), age, gender, sleeping pills, urinary catheterization and recent diagnosis of CRC, psychotherapy were the major predictors shared in both. In CRC (i.e., C18-19), sleeping pill prescription were the main factors, whereas in rectal cancer (i.e., C20), recent diagnosis of CRC and urinary catheterization were the main factors. The details of variable importance are presented in detail in Figures 5a-j.

Psychiatric disorders such as major depressive disorder are included as important predictors. Appendix 1 shows the results of the random forest model in which a variable set is created and run only with psychiatric disorders. The variables with the highest predictive power among psychiatric disorders were major depressive disorder, anxiety disorder, and somatoform disorder. For these diagnostic factors, diagnosis records both before and after the diagnosis of CRC were included as major predictors.

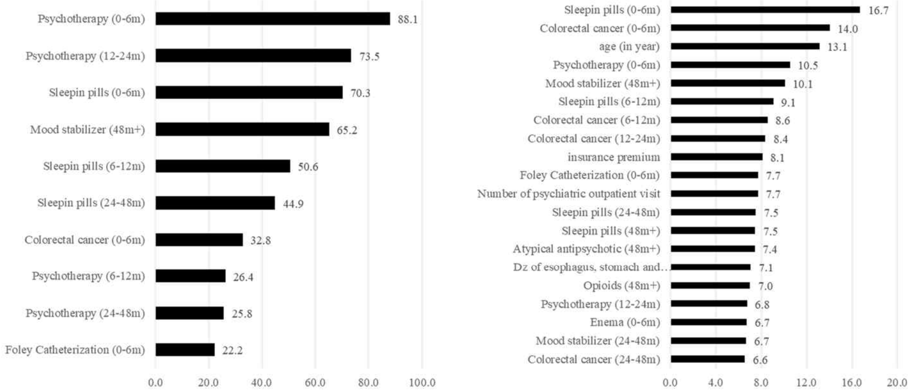


The x-axis represents variable importance (i.e., mean decrease in accuracy).

a) classification tree

b) random forest

Figure 5a. Variable importance for a) classification tree and b) random forest in total colorectal cancer patients (n=3,678)

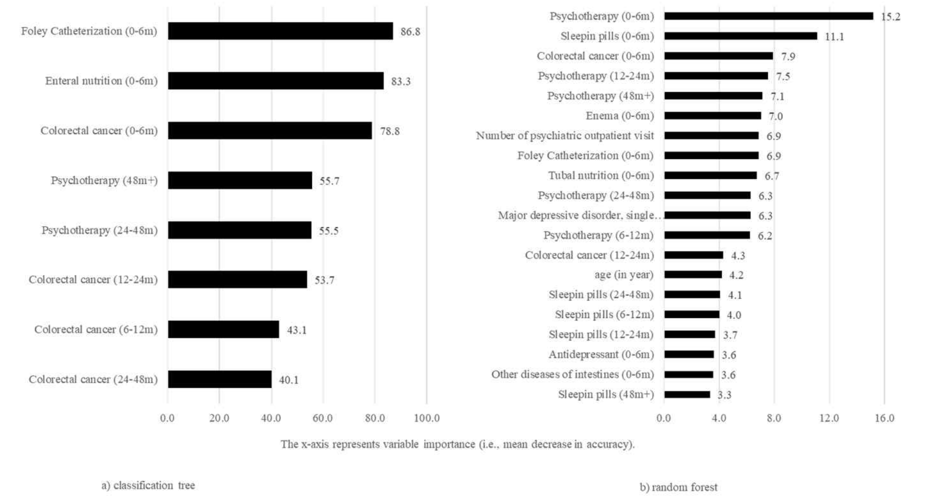


The x-axis represents variable importance (i.e., mean decrease in accuracy).

a) classification tree

b) random forest

Figure 5b. Variable importance for a) classification tree and b) random forest in male colorectal cancer patients (n=2,368)

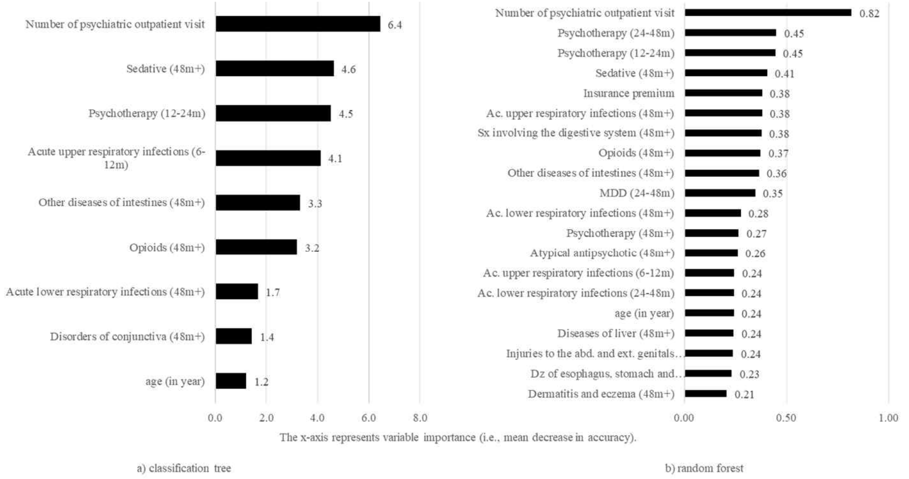


The x-axis represents variable importance (i.e., mean decrease in accuracy).

a) classification tree

b) random forest

Figure 5c. Variable importance for a) classification tree and b) random forest in female colorectal cancer patients (n=1,310)

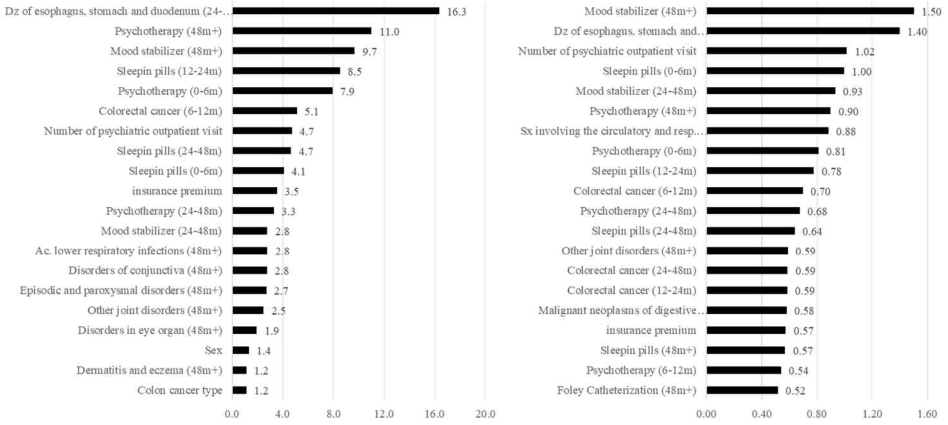


The x-axis represents variable importance (i.e., mean decrease in accuracy).

a) classification tree

b) random forest

Figure 5d. Variable importance for a) classification tree and b) random forest in 10-20's colorectal cancer patients (n=53)

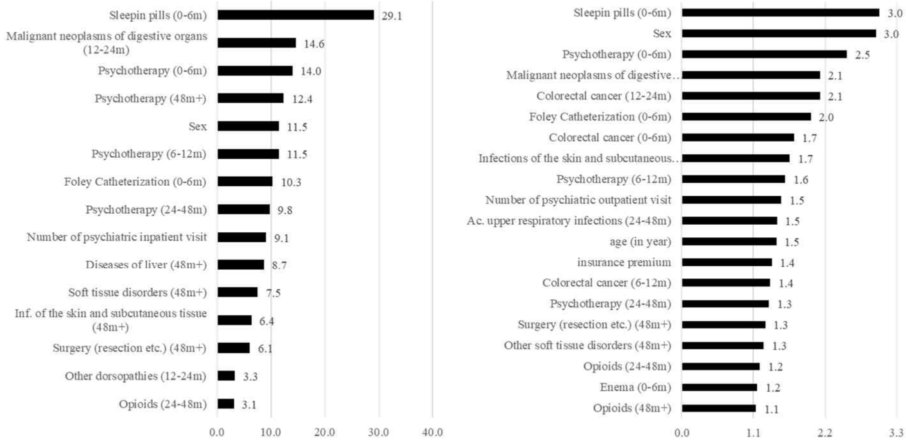


The x-axis represents variable importance (i.e., mean decrease in accuracy).

a) classification tree

b) random forest

Figure 5e. Variable importance for a) classification tree and b) random forest in 30's colorectal cancer patients (n=141)

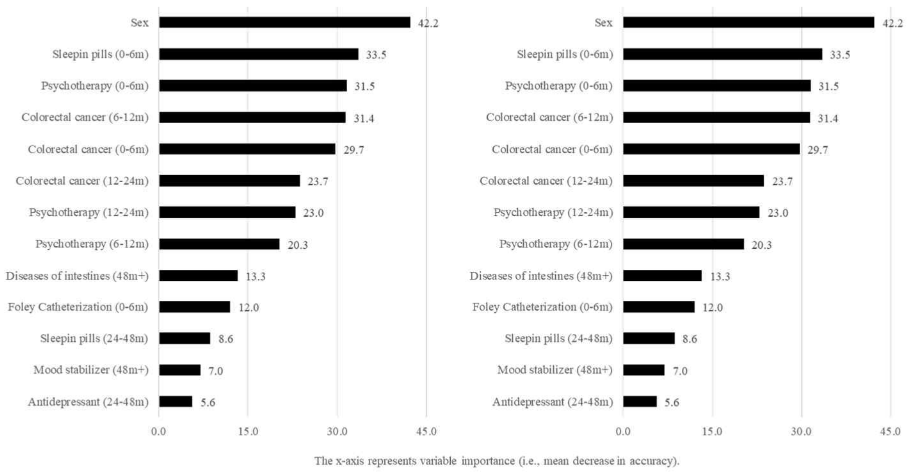


The x-axis represents variable importance (i.e., mean decrease in accuracy).

a) classification tree

b) random forest

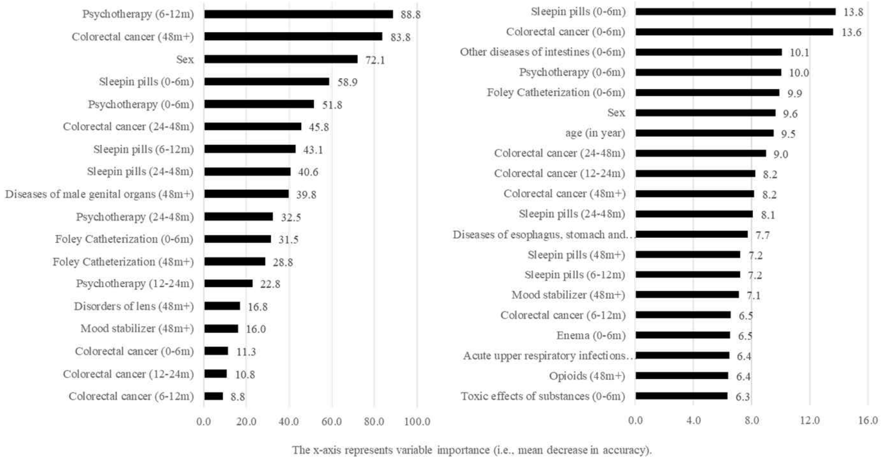
Figure 5f. Variable importance for a) classification tree and b) random forest in 40’s colorectal cancer patients (n=440)



a) classification tree

b) random forest

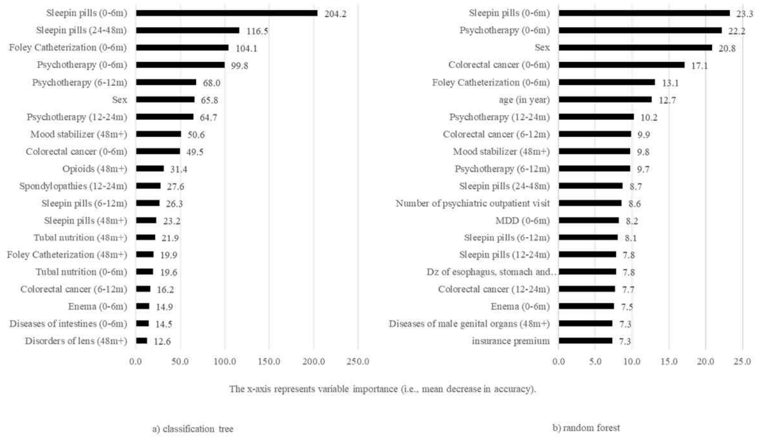
Figure 5g. Variable importance for a) classification tree and b) random forest in 50's colorectal cancer patients (n=758)



a) classification tree

b) random forest

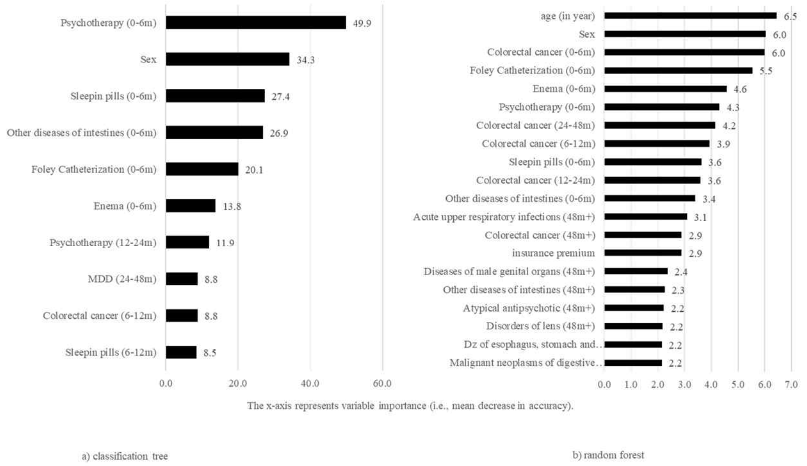
Figure 5h. Variable importance for a) classification tree and b) random forest in 60's colorectal cancer patients (n=2,286)



a) classification tree

b) random forest

Figure 5i. Variable importance for a) classification tree and b) random forest in colorectal cancer patients diagnosed with C18&19 (n=2,771)



a) classification tree

b) random forest

Figure 5j. Variable importance for a) classification tree and b) random forest in colorectal cancer patients diagnosed with C20 (n=907)

6. Model prediction performance of classification tree and random forest

Table 4 and Figures 6 summarize the prediction performance of the classification tree (i.e., both full-fledged and pruned tree) and random forest for the total sample and the sub-sample stratified by age, gender, and cancer type. Table 4 summarizes the accuracy of the prediction model calculated using the confusion table, 95% confidence interval, Kappa statistics, sensitivity, specificity, PPV, and area under the receiver operating curve (AUC). Figure 6 shows the receiver operating curve and AUC of the classification tree and random forest model. In general, in all models, the AUC of the random forest model is larger than the AUC of the general classification trees. The threshold of the optimal point (the cut-off points of the risk probability for classifying suicide) of the random forest model in the total sample (Figure 6a) was 0.489 (48.9%), and the sensitivity and specificity at this point were 84.1% and 68.9%, respectively. Sensitivity and specificity in Table 4 were different in values since those values for predictive performance were calculated at which threshold was 0.50 (50%), and were 76.25% and 63.83%, respectively. The predictive power of the random forest model was (except for the male sub-sample with AUC 0.764) all exceeded 0.80 (80%). The AUCs of the random forest model in the male and female subsamples were 0.764 (76.4%) and 0.814 (81.4%), respectively.

Table 4a. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among colorectal cancer sample (N=3,678)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.73 (0.70, 0.75)	0.70 (0.67, 0.73)	0.75 (0.72, 0.78)
Kappa statistics *	0.45	0.40	0.50
Sensitivity *	0.76	0.68	0.83
Specificity *	0.69	0.72	0.67
PPV *	0.71	0.71	0.72
AUC	0.78	0.75	0.84

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

Table 4b. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among male colorectal cancer patient (N=2,368)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.69 (0.65, 0.72)	0.68 (0.64, 0.72)	0.72 (0.68, 0.75)
Kappa statistics *	0.33	0.35	0.38
Sensitivity *	0.80	0.70	0.91
Specificity *	0.53	0.66	0.45
PPV *	0.70	0.74	0.70
AUC	0.67	0.69	0.76

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

Table 4c. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among female colorectal cancer patient (N=1,310)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.73 (0.68, 0.78)	0.72 (0.67, 0.77)	0.78 (0.73, 0.82)
Kappa statistics *	0.42	0.41	0.51
Sensitivity *	0.64	0.66	0.64
Specificity *	0.78	0.76	0.85
PPV *	0.62	0.60	0.70
AUC	0.73	0.74	0.81

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

Table 4d. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among colorectal cancer patient age between 10-20s (N=53)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.54 (0.25, 0.81)	0.54 (0.25, 0.81)	0.69 (0.39, 0.91)
Kappa statistics *	0.03	0.03	0.24
Sensitivity *	0.40	0.40	0.20
Specificity *	0.63	0.63	1.00
PPV *	0.40	0.40	1.00
AUC	0.56	0.56	0.90

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

Table 4e. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among colorectal cancer patient age in 30s (N=141)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.59 (0.41, 0.75)	0.59 (0.41, 0.75)	0.59 (0.41, 0.75)
Kappa statistics *	0.14	0.14	0.14
Sensitivity *	0.40	0.40	0.40
Specificity *	0.74	0.74	0.74
PPV *	0.55	0.55	0.55
AUC	0.58	0.58	0.80

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

Table 4f. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among colorectal cancer patient age in 40s (N=440)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.61 (0.51, 0.70)	0.66 (0.56, 0.75)	0.75 (0.66, 0.83)
Kappa statistics *	0.19	0.27	0.48
Sensitivity *	0.52	0.41	0.61
Specificity *	0.67	0.84	0.86
PPV *	0.53	0.66	0.76
AUC	0.65	0.68	0.81

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

Table 4g. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among colorectal cancer patient age in 50s (N=758)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.69 (0.62, 0.75)	0.69 (0.62, 0.76)	0.76 (0.69, 0.82)
Kappa statistics *	0.36	0.36	0.52
Sensitivity *	0.56	0.52	0.75
Specificity *	0.79	0.84	0.77
PPV *	0.69	0.72	0.73
AUC	0.74	0.74	0.82

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

Table 4h. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among colorectal cancer patient age in 60s (N=2,286)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.72 (0.68, 0.76)	0.71 (0.67, 0.75)	0.75 (0.72, 0.79)
Kappa statistics *	0.43	0.42	0.50
Sensitivity *	0.80	0.75	0.87
Specificity *	0.63	0.67	0.61
PPV *	0.71	0.72	0.72
AUC	0.74	0.72	0.84

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

Table 4i. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among colorectal cancer patient diagnosed C18 & 19 (N=2,775)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.73 (0.70, 0.77)	0.72 (0.69, 0.76)	0.76 (0.73, 0.79)
Kappa statistics *	0.47	0.44	0.52
Sensitivity *	0.72	0.67	0.77
Specificity *	0.75	0.78	0.75
PPV *	0.72	0.73	0.74
AUC	0.76	0.75	0.83

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

Table 4j. Comparison of prediction performance of Classification and regression tree and random forest models that predict suicide results among colorectal cancer patient diagnosed C20 (N=903)

	Classification and regression tree model (Full-fledged tree)	Classification and regression tree model (pruned tree)	Random Forest model
Accuracy * (95% CI)	0.67 (0.60, 0.73)	0.60 (0.53, 0.66)	0.72 (0.65, 0.77)
Kappa statistics *	0.32	0.13	0.40
Sensitivity *	0.73	0.83	0.86
Specificity *	0.58	0.30	0.53
PPV *	0.70	0.61	0.71
AUC	0.74	0.65	0.81

* The performance of the model was calculated based on a classification probability of 0.5 or greater.

Abbreviation: AUC, area under the receiver operating characteristic curve; PPV, Positive predictive value

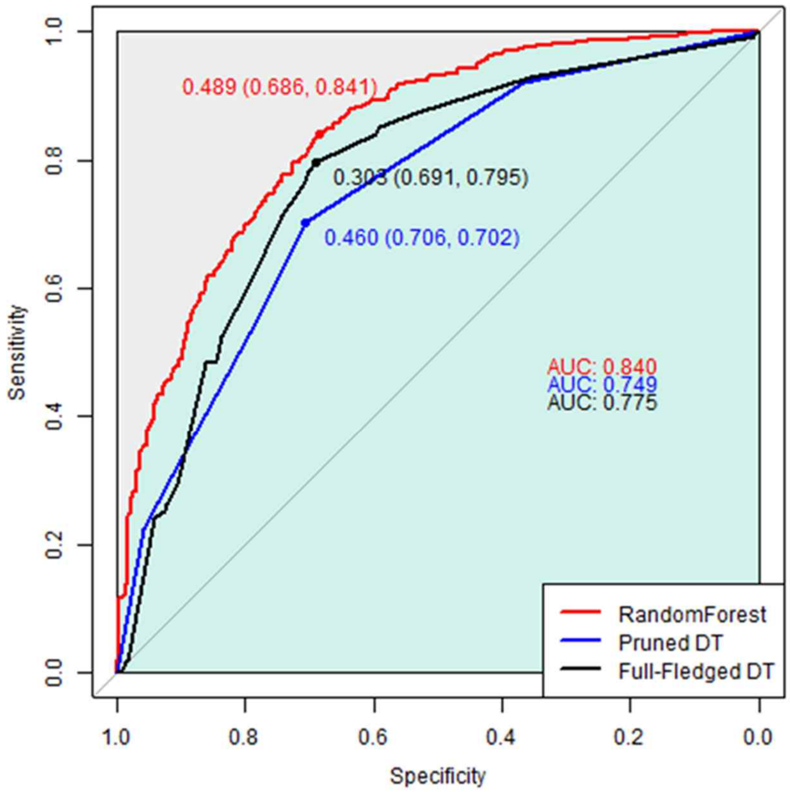
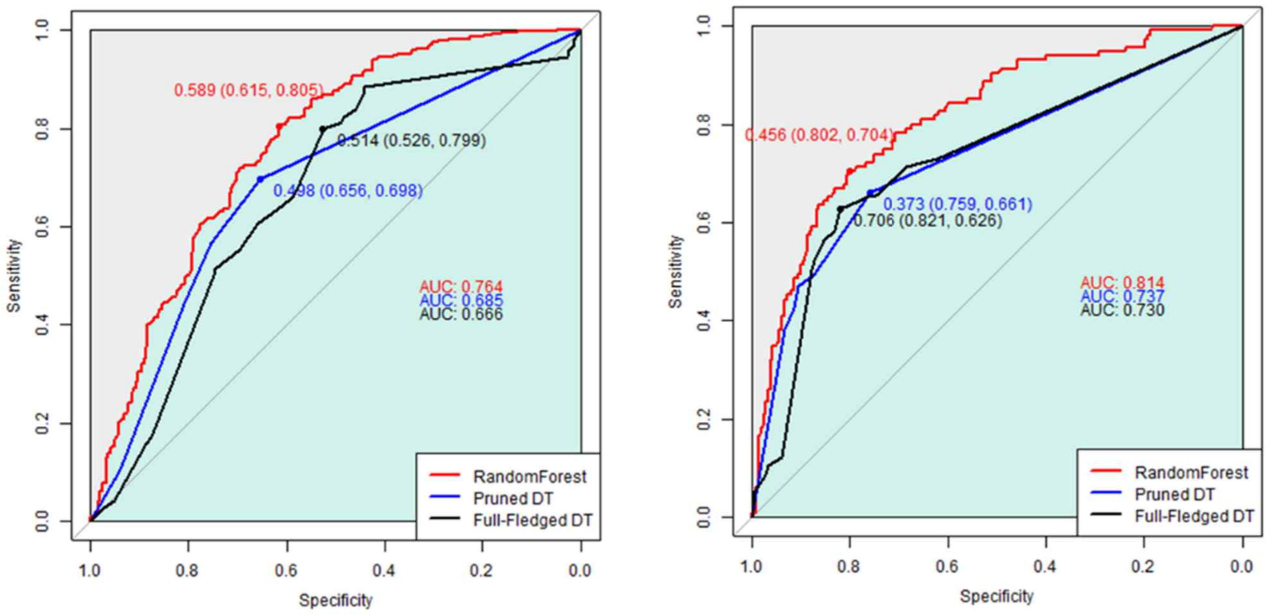


Figure 6a. Receiver operating characteristic curves and areas under the curve (AUCs) of classification and regression tree model (full-fledged and pruned tree) and random forest model for suicide death attempt in total sample.



a) Male

b) Female

Figure 6b. Receiver operating characteristic curves and areas under the curve (AUCs) of classification and regression tree model (full-fledged and pruned tree) and random forest model for suicide death in a) male and b) female sample

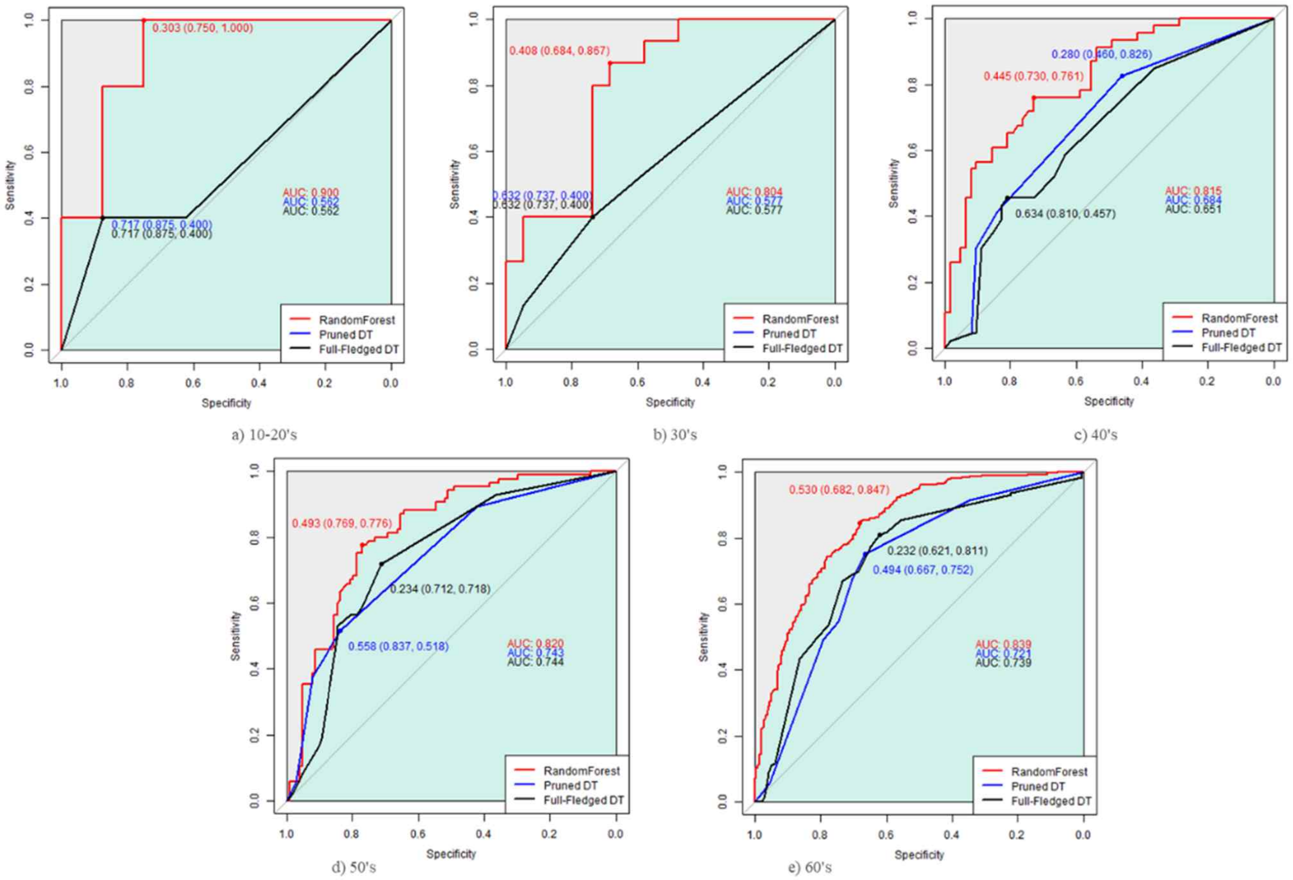
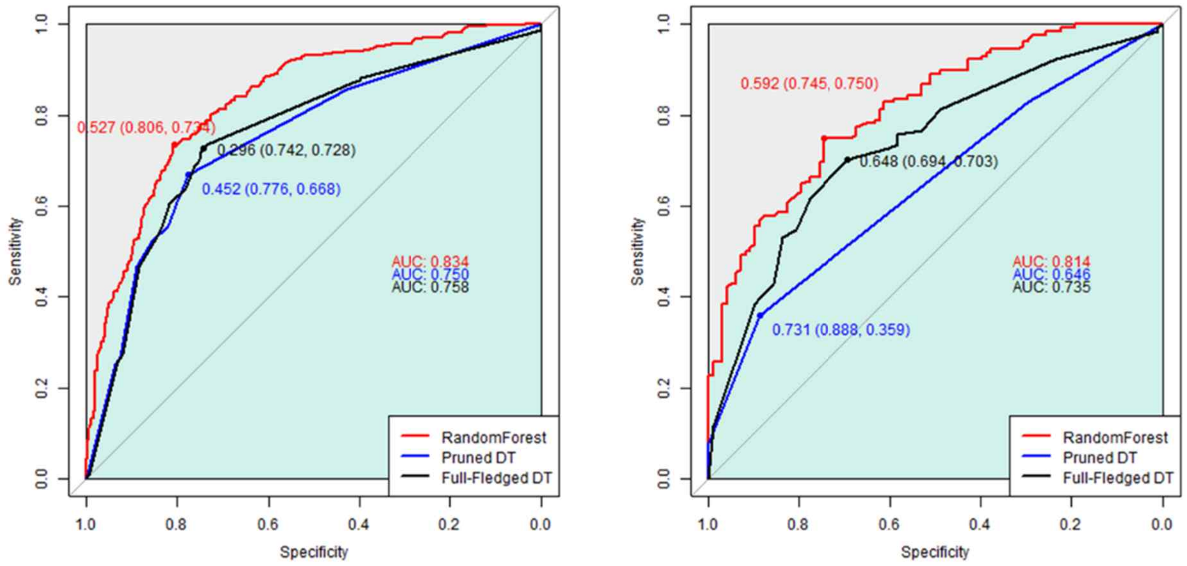


Figure 6c. Receiver operating characteristic curves and areas under the curve (AUCs) of classification and regression tree model (full-fledged and pruned tree) and random forest model for suicide death in a) 10-20's, b) 30's, c) 40's, d) 50's and e) 60's sample



a) C18-C19

b) C20

Figure 6d. Receiver operating characteristic curves and areas under the curve (AUCs) of classification and regression tree model (full-fledged and pruned tree) and random forest model for suicide death in a) C18-19 and b) C20

7. Nested case-control analysis

In the analysis using the above machine learning technique, major predictive factors for suicide could be identified in the total sample and stratified sub-samples of each age, gender and cancer type. The most prominent predictors included age and gender, as well as variables on whether to treat mental disorders such as psychotherapy or outpatient visits to psychiatric hospitals, and the use of psychiatric drugs such as sleeping pills, mood stabilizers, and antipsychotics prescription. Enteral nutrition, enema (e.g., rectal-tube insertion), urinary catheterization, and CRC-related surgical procedures were also among the discovered top-ranked predictors. Digestive and respiratory complications were also major predictors. Among these, a nested case-control study was performed to identify the magnitude of association among variables such as psychotherapy, psychiatric outpatient, hospitalization period, number of psychiatric emergency room visits and use of psychiatric drugs, surgical history, and urinary catheterization as proxy variables for inpatient treatment.

Table 5 summarizes the general characteristics (n=10,996) of the suicide group (n=1,839) and the control group (n=9,157) matched with age and gender one to five at each time point of the suicide death. There was no significant difference in age, gender, and insurance premium decile between the case and control groups. There was a significant difference in cancer type. In terms of cancer stage, the proportion of later stage was higher in the control group than in the suicide group. Sleep disorder

(F51), schizophrenia, schizotypal and delusional disorders (F20-F25), and major depressive disorder (F32) were significantly higher in the suicide group.

Table 5. Demographic characteristics of suicidal death in colorectal cancer (ICD-10 code: C18-20) patients from Korean National Health Insurance Service (2002-2018) in age- and sex-matched nested case-control study

		Suicide death, N(%)		
Variables	Total, N(%)	No suicide	Suicide	
	10,996 (100)	9,157 (83.28)	1,839 (16.72)	
	Mean (STD)	Mean (STD)	Mean (STD)	
Age (in years)	63.88 (13.00)	63.95 (13.08)	63.54 (12.77)	0.86
Insurance premium decile (0-10)	6.43 (2.80)	6.42 (2.79)	6.48 (2.86)	0.20
Variables	N (%)	N (%)	N (%)	p-value *
Colon cancer type				<.0001
Proximal colon	6491 (59.03)	5292 (57.79)	1199 (65.2)	
Distal colon	846 (7.69)	718 (7.84)	128 (6.96)	
Rectum	3659 (33.28)	3147 (34.37)	512 (27.84)	
Sex				0.99
Male	8232 (74.86)	6855 (74.86)	1377 (74.88)	
Female	2764 (25.14)	2302 (25.14)	462 (25.12)	
Colorectal cancer stage				<.0001
Stage unknown	4761 (43.3)	3528 (39.39)	1233 (67.05)	
Stage 1	4325 (39.33)	3712 (41.44)	413 (22.46)	
Stage 2	1127 (10.25)	1032 (11.52)	95 (5.17)	
Stage 3	40 (0.36)	33 (0.37)	7 (0.38)	
Stage 4	743 (6.76)	652 (7.28)	91 (4.95)	
Comorbidity				
Sleep disorder (F51)				<.0001
not diagnosed	10306 (93.72)	8465 (92.44)	1404 (76.35)	
diagnosed	690 (6.28)	692 (7.56)	435 (23.65)	
Schizophrenia, schizotypal and delusional disorders (F20-F25)				<.0001
not diagnosed	10891 (99.05)	8715 (95.17)	1591 (86.51)	
diagnosed	105 (0.95)	442 (4.83)	248 (13.49)	
Major depressive disorder (F32)				<.0001
not diagnosed	9869 (89.75)	9104 (99.42)	1787 (97.17)	
diagnosed	1127 (10.25)	53 (0.58)	52 (2.83)	

* Significant test is evaluated with t-test, ANOVA and chi-square test.

Table 6a and figure 7a summarize the results of conditional logistic regression analysis of nested case-control study design. Each drug use and psychotherapy prescription are arranged as an odds ratio per 10 prescriptions increment. The number of uses of psychiatric drugs such as sleeping pills, mood stabilizers, antipsychotics(atypical) and antidepressants showed a significant association with the suicide death. Psychotherapy also showed a significant association with suicide. The magnitude of the association tended to be larger as the prescription time was closer to the onset of suicide. For example, the number of prescriptions for sleeping pills 0-6 months ago (Odds ratio [95%CI]; 7.47 [6.04-9.23]) showed a greater association with suicide than the number of prescriptions for sleeping pills 48 months ago (1.11 [1.08-1.13]).

Table 6a. The associations between suicide and the number of prescriptions for psychiatric medication and number of prescriptions for procedure related to colorectal cancer by 0-6, 6-12, 12-24, 24-48, 48+ month time intervals identified by conditional logistic regression analysis (n=10,996)

Variable	OR [95% CI] *	Variable	OR [95% CI] *
sleeping pills (0-6m)	7.47 [6.04-9.23]	Atypical antipsychotic (0-6m)	6.75 [5.15-8.85]
sleeping pills (6-12m)	4.05 [3.34-4.91]	Atypical antipsychotic (6-12m)	3.74 [2.92-4.78]
sleeping pills (12-24m)	2.03 [1.82-2.25]	Atypical antipsychotic (12-24m)	1.71 [1.49-1.96]
sleeping pills (24-48m)	1.46 [1.37-1.55]	Atypical antipsychotic (24-48m)	1.26 [1.17-1.36]
sleeping pills (48m+)	1.11 [1.08-1.13]	Atypical antipsychotic (48m+)	1.09 [1.06-1.12]
Mood stabilizer (0-6m)	4.06 [3.36-4.91]	Opioids (0-6m)	1.58 [1.34-1.85]
Mood stabilizer (6-12m)	2.71 [2.29-3.20]	Opioids (6-12m)	1.53 [1.31-1.80]
Mood stabilizer (12-24m)	1.72 [1.55-1.90]	Opioids (12-24m)	1.29 [1.18-1.42]
Mood stabilizer (24-48m)	1.31 [1.24-1.38]	Opioids (24-48m)	1.17 [1.11-1.23]
Mood stabilizer (48m+)	1.06 [1.04-1.08]	Opioids (48m+)	1.03 [1.01-1.06]
Antidepressant (0-6m)	8.38 [5.55-12.65]	Sedative (0-6m)	3.38 [1.98-5.78]
Antidepressant (6-12m)	3.63 [2.52-5.21]	Sedative (6-12m)	2.20 [1.30-3.74]
Antidepressant (12-24m)	2.35 [1.89-2.93]	Sedative (12-24m)	1.55 [1.18-2.05]
Antidepressant (24-48m)	1.52 [1.35-1.71]	Sedative (24-48m)	1.29 [1.11-1.50]
Antidepressant (48m+)	1.1 [1.05-1.15]	Sedative (48m+)	1.08 [1.02-1.16]
Typical antipsychotic (0-6m)	87.41 [9.96-767.42]	Psychotherapy (0-6m)	12.91 [9.58-17.41]
Typical antipsychotic (6-12m)	8.24 [0.95-71.41]	Psychotherapy (6-12m)	6.44 [4.91-8.44]
Typical antipsychotic (12-24m)	3.33 [0.81-13.78]	Psychotherapy (12-24m)	2.94 [2.52-3.43]
Typical antipsychotic (24-48m)	1.83 [0.96-3.50]	Psychotherapy (24-48m)	1.79 [1.63-1.96]
Typical antipsychotic (48m+)	1.17 [0.93-1.47]	Psychotherapy (48m+)	1.21 [1.15-1.27]
MDD (0-6m)	35.98 [20.35-63.62]	Foley Catheterization (0-6m)	844.3 [320.4-999.9]
MDD (6-12m)	16.73 [10.24-27.34]	Foley Catheterization (6-12m)	8.33 [3.05-22.69]
MDD (12-24m)	5.46 [4.00-7.44]	Foley Catheterization (12-24m)	3.37 [1.54-7.35]
MDD (24-48m)	2.79 [2.27-3.42]	Foley Catheterization (24-48m)	1.24 [0.66-2.32]
MDD (48m+)	1.20 [1.13-1.27]	Foley Catheterization (48m+)	0.05 [0.03-0.10]

* The model was adjusted with colorectal cancer type and stage, insurance premium percentile, and comorbidity such as sleep disorder (F51), schizophrenia, schizotypal and delusional disorder (F20-25), and major depressive disorder (F32).

Table 6b. The associations between suicide and the number of prescriptions for psychiatric medication for entire follow-up period identified by conditional logistic regression analysis (n=10,996)

Variable	OR [95% CI] *	Variable	OR [95% CI] *
Sleep pill	1.11 [1.09-1.12]	Opioid	1.04 [1.02-1.05]
Mood Stabilizer	1.07 [1.05-1.08]	Sedatives	1.09 [1.04-1.14]
Antidepressant	1.11 [1.08-1.15]	Psychotherapy	1.17 [1.14-1.21]
Typical antipsychotics	1.16 [0.97-1.38]	Colostomy	1.02 [0.87-1.20]
Atypical antipsychotics	1.09 [1.07-1.12]	Foley catheterization	1.01 [0.99-1.03]

* The model was adjusted with colorectal cancer type and stage, insurance premium percentile, and comorbidity such as sleep disorder (F51), schizophrenia, schizotypal and delusional disorder (F20-25), and major depressive disorder (F32).

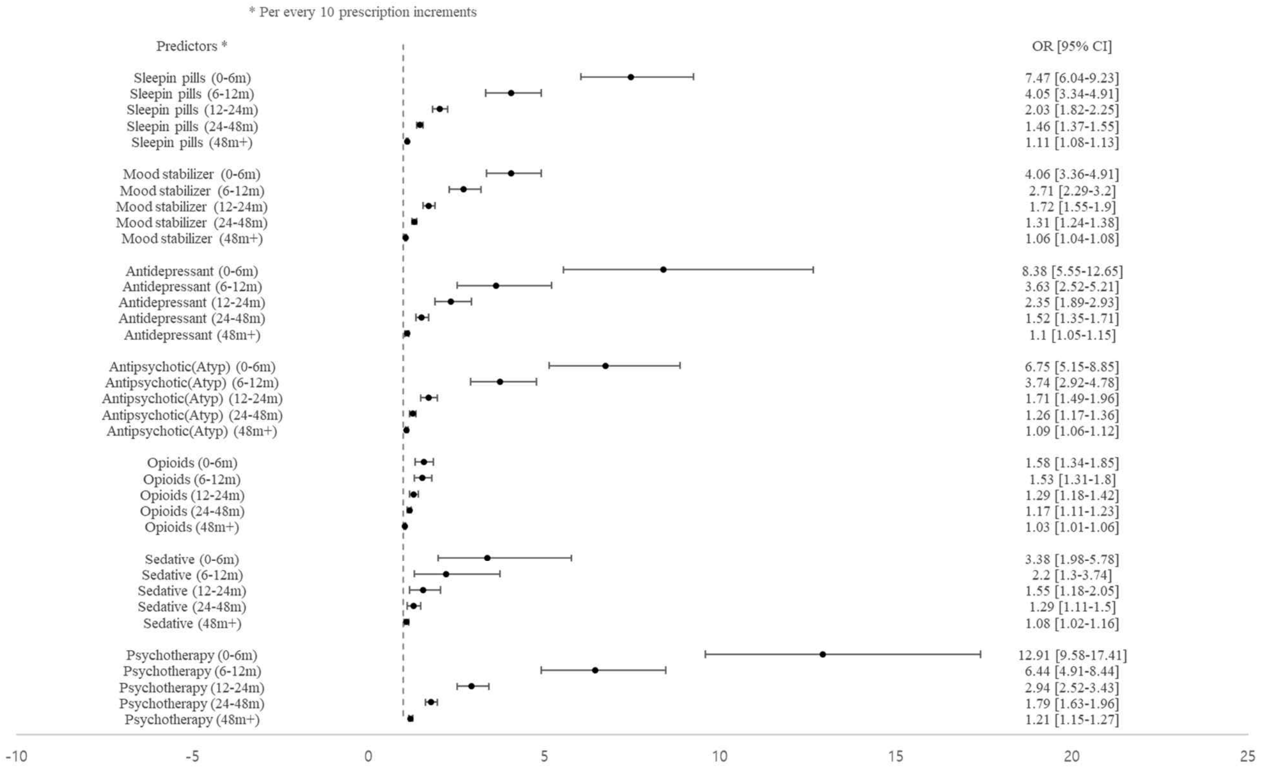


Figure 7a. The associations between suicide and the number of prescriptions for psychiatric medication and psychotherapy by 0-6, 6-12, 12-24, 24-48, 48+ month time intervals identified by conditional logistic regression analysis

Table 6a and figure 7b summarizes the results of measuring the overall effect size of association between variables such as psychiatric drugs, psychotherapy, colostomy, and Foley catheterization (as proxy for inpatient treatment). As a result of conditional logistic regression analysis by summing all prescriptions for each variable before the onset of suicide, all psychiatric drugs except typical antipsychotic had significant positive associations. colostomy surgery and Foley catheterization were not significantly associated with suicide when temporal distance of variables was not considered (Table 6b).

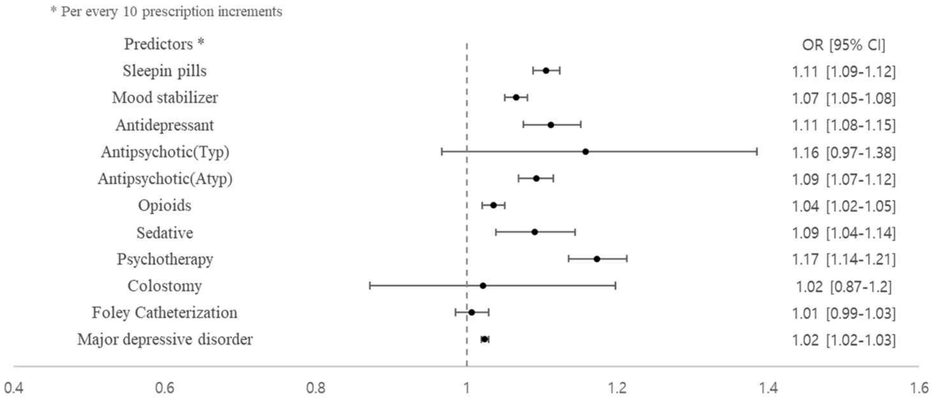


Figure 7b. The associations between suicide and procedures and psychiatric medications over the total follow-up period (n=10,996)

8. Operating characteristics of high-risk thresholds

Cross-validated RF predicted probabilities were rank ordered, and operating characteristics were calculated among individuals in the top 50% of the predicted risk distribution. CRC patients in the top 1%, 5%, 10%, and 20% of predicted risk accounted for 10.68%, 34.64%, 50.76% and 71.02% of all cases of suicide death, respectively (specificity = 98.55%, 94.62%, 89.71% and 79.73%, respectively). The sensitivity among individuals in the top 1% and 5% of predicted risk was 10.68 and 6.93 times, respectively, higher than the expected value among total CRC patient (10.68/1% and 34.64%/5% respectively) (Table 7). The PPV was 5.06% in the top 1%, 3.30% in the 5%, and 2.44% in the 10% (Figure 8).

Table 7. Operating characteristics of high-risk thresholds for total sample (n=95,625)

threshold	TP	FP	FN	TN	Positive	Negative	Suicide	No suicide	PPV	SN	SP
Top 1%	49	920	410	93,787	969	94,656	459	95,166	5.06	10.68	98.55
Top 5%	159	4,663	300	90,044	4,822	90,803	459	95,166	3.30	34.64	94.62
Top 10%	233	9,333	226	85,374	9,566	86,059	459	95,166	2.44	50.76	89.71
Top 20%	326	18,831	133	75,876	19,157	76,468	459	95,166	1.70	71.02	79.73
Top 25%	329	23,717	130	70,990	24,046	71,579	459	95,166	1.37	71.68	74.60
Top 30%	372	28,437	87	66,270	28,809	66,816	459	95,166	1.29	81.05	69.64
Top 50%	428	47,493	31	47,214	47,921	47,704	459	95,166	0.89	93.25	49.61

Abbreviation: TP, true positive; FP, false positive; TN, true negative; FN, false negative; PPV, positive predictive value; SN, sensitivity; SP, specificity

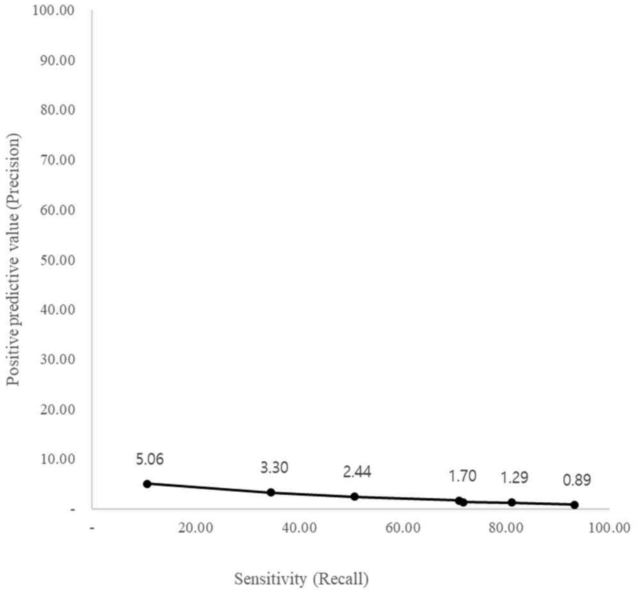


Figure 8. Precision-recall curve (n=95,625)

9. Number needed to screen

Due to unknown effectiveness of the intervention due to screening of the prediction model, the number needed to screen was resented for each variable rate of reduction in mortality (Table 8 and Figure 9). Unscreened mortality of suicide death (i.e., 0.48%) and the sensitivity of the prediction model are fixed values in all variable mortality situations. As the sensitivity of the tool increased, the size of the absolute risk reduction increased, and it was confirmed that the value of NNS, which is the reciprocal, decreased. In addition, the greater the intervention effect due to the variably assumed screening test (i.e., 25%, 50%, 75%, 100%), the greater the absolute risk reduction value, thereby decreasing the number needed to screen.

Table 8. Number need to screen calculation table for variable mortality reduction rate due to screening by machine learning prediction model

25% mortality decrease assumed due to screen

Threshold	SN	Relative mortality decrease due to screen	Unscreened mortality (%)	Mortality after screen (%)	Absolute risk reduction (%)	NNS
Top 1%	10.68	2.67	0.48	0.47	0.013	7,803
5%	34.64	8.66	0.48	0.44	0.042	2,406
10%	50.76	12.69	0.48	0.42	0.061	1,642
20%	71.02	17.76	0.48	0.39	0.085	1,174
25%	71.68	17.92	0.48	0.39	0.086	1,163
30%	81.05	20.26	0.48	0.38	0.097	1,029
50%	93.25	23.31	0.48	0.37	0.112	894

50% mortality decrease assumed due to screen

Threshold	SN	Relative mortality decrease due to screen	Unscreened mortality (%)	Mortality after screen (%)	Absolute risk reduction (%)	NNS
Top 1%	10.68	5.34	0.48	0.45	0.026	3,902
5%	34.64	17.32	0.48	0.40	0.083	1,203
10%	50.76	25.38	0.48	0.36	0.122	821
20%	71.02	35.51	0.48	0.31	0.170	587
25%	71.68	35.84	0.48	0.31	0.172	582
30%	81.05	40.53	0.48	0.29	0.195	515
50%	93.25	46.63	0.48	0.26	0.224	447

75% mortality decrease assumed due to screen

Threshold	SN	Relative mortality decrease due to screen	Unscreened mortality (%)	Mortality after screen (%)	Absolute risk reduction (%)	NNS
Top 1%	10.68	8.01	0.48	0.44	0.038	2,601
5%	34.64	25.98	0.48	0.36	0.125	802
10%	50.76	38.07	0.48	0.30	0.183	548
20%	71.02	53.27	0.48	0.22	0.256	392
25%	71.68	53.76	0.48	0.22	0.258	388
30%	81.05	60.79	0.48	0.19	0.292	343
50%	93.25	69.94	0.48	0.14	0.336	298

100% mortality decrease assumed due to screen

Threshold	SN	Relative mortality decrease due to screen	Unscreened mortality (%)	Mortality after screen (%)	Absolute risk reduction (%)	NNS
Top 1%	10.68	10.68	0.48	0.43	0.051	1,951
5%	34.64	34.64	0.48	0.31	0.166	602
10%	50.76	50.76	0.48	0.24	0.244	411
20%	71.02	71.02	0.48	0.14	0.341	294
25%	71.68	71.68	0.48	0.14	0.344	291
30%	81.05	81.05	0.48	0.09	0.389	258
50%	93.25	93.25	0.48	0.03	0.448	224

Abbreviation: SN, sensitivity; NNS, number needed to screen

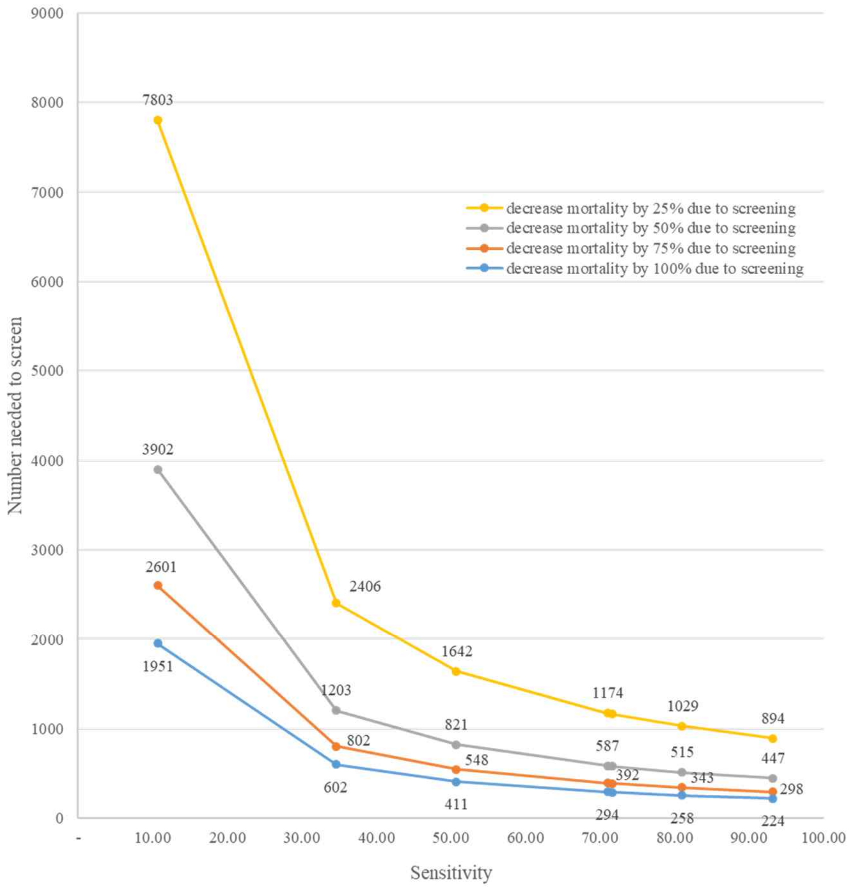


Figure 9. Number needed to screen curves (mortality reduction assumed by 25%, 50%, 75%, and 100%)

10. Suicide risk score-card

The predictors selected for the risk score-card were the top 10 prominent predictors found in the random forest model. Table 9b shows the regression coefficient for each factor identified through logistic regression analysis. The risk score-card obtained by standardizing the weight of each coefficient is shown in Table 9a. For example, a female patient is given -42 points while a male patient gets zero, and 14 points are added for each prescription history if they have received psychotherapy within the last 6 months. Use of sleeping pills in the last 6 months adds 17 points per count. Finally, each patient's risk score can be evaluated by multiplying the scores of all factors and the number of factors appearing in the medical history. Table 9c shows operating characteristics according to each cutoff of the evaluated risk score. For example, the cutoff for the top 50% of risk scores was -20 points, and the sensitivity was 70.04, the specificity was 69.98, and the PPV was 70.

Table 9a. Score Table of suicide death prediction scale for colorectal cancer patients

		a) score	b) times (years)	a x b
Is the patient's gender female?	No	□ 0	0	0
	Yes	□ -42	n/a	
How old is the patient?	No	□ 0	0	0
	Yes	□ 3		
Does the patient have a history of first diagnosis of colorectal cancer within the past 6 months?	No	□ 0	0	0
	Yes	□ 1		
Does the patient have a history of first diagnosis of colorectal cancer within the last 12 months?	No	□ 0	0	0
	Yes	□ 2		
Does the patient have a history of first diagnosis of colorectal cancer within the last 24 months?	No	□ 0	0	0
	Yes	□ 2		
Has the patient had a history of being prescribed sleeping pills within the past 6 months?	No	□ 0	0	0
	Yes	□ 17		
Does the patient have a history of prescribed sleeping pills in the past 24-48 months?	No	□ 0	0	0
	Yes	□ 2		
Has the patient had a history of prescribed psychotherapy within the past 6 months?	No	□ 0	0	0
	Yes	□ 14		
Does the patient have a history of prescribed psychotherapy in the past 12-24 months?	No	□ 0	0	0
	Yes	□ 4		
Does the patient have a history of prescribed urinary catheterization within the past 6 months?	No	□ 0	0	0
	Yes	□ 78		
Sum score	Add the item scores to obtain the sum score			

The above predictors were composed of the top 10 most prominent predictors (not in order of importance) from the results of prediction model. To construct an easily used clinical score table, the regression coefficients of the predictors from the final model were standardized, dividing all regression coefficients by the smallest coefficient and rounding off the results. The total scores were linked to the risk of suicide death.

Table 9b. Multivariable Logistic Regression Model for candidate predictors as results of RF model

Predictors	Odds Ratio (95% CI)	Coefficient	SE	*P* Value
female gender	0.36 (0.31-0.43)	-0.507	0.0404	<.0001
age (per year)	1.05 (0.99-1.11)	0.049	0.0282	0.0841
first diagnosis of colorectal cancer within the past 6 months	1.01 (0.99-1.04)	0.012	0.0114	0.2851
first diagnosis of colorectal cancer within the past 6-12 months	1.03 (1.00-1.06)	0.028	0.0141	0.0491
first diagnosis of colorectal cancer within the past 12-24 months	1.04 (1.02-1.05)	0.034	0.0085	<.0001
sleeping pill within the past 6 months *	1.23 (1.18-1.29)	0.207	0.0224	<.0001
sleeping pills in the past 24-48 months *	1.03 (1.01-1.04)	0.026	0.0075	0.0005
psychotherapy within the past 6 months *	1.20 (1.13-1.26)	0.178	0.0278	<.0001
psychotherapy within the past 6-12 months *	1.05 (1.02-1.08)	0.051	0.0150	0.0007
urinary catheterization within the past 6 months *	2.60 (2.14-3.16)	0.956	0.0990	<.0001

* per each appearance in prescribed history

Table 9c. Operating characteristics of high-risk thresholds (n=3,678)

threshold	Cutoff	TP	FP	FN	TN	Positive	Negative	suicide	no suicide	PPV	SN	SP
Top 1%	≥378	35	1	1,804	1,838	36	3,642	1,839	1,839	97.22	1.90	99.95
5%	≥177	167	16	1,672	1,823	183	3,495	1,839	1,839	91.26	9.08	99.13
10%	≥110	318	49	1,521	1,790	367	3,311	1,839	1,839	86.65	17.29	97.34
20%	≥56	610	126	1,229	1,713	736	2,942	1,839	1,839	82.88	33.17	93.15
25%	≥32	745	175	1,094	1,664	920	2,758	1,839	1,839	80.98	40.51	90.48
30%	≥13	870	234	969	1,605	1104	2,574	1,839	1,839	78.80	47.31	87.28
50%	≥-20	1,288	552	551	1,287	1840	1,838	1,839	1,839	70.00	70.04	69.98

Abbreviation: TP, true positive; FP, false positive; TN, true negative; FN, false negative; PPV, positive predictive value; SN, sensitivity; SP, specificity

IV. DISCUSSION

1. Summary of the main findings

This study examined age-, sex-, and type-specific models of suicide death using a machine-learning algorithm and the NHID. Variables included as potential predictors were demographics, psychiatric and physical diagnoses, and prescription of medication and procedures related to CRC. As a result of the analysis, the classification tree and the random forest model with potential predictors of death due to suicide prescribed and diagnosed at 0–6, 6–12, 12–24, 24–48, and 48+ months before the onset of suicide achieved good predictive accuracy (i.e., an AUC >0.80). In addition, to measure the effect size within the cohort sample for the discovered predictors, a nested case-control study was designed, and a conditional logistic regression analysis was performed. The analysis showed that the size of the effect of the predictors tended to be in line with the results of the variable importance obtained from the machine-learning analysis.

As a result, various variables predicting suicide were confirmed by the variable importance of the classification tree and random forest models. Among them, we discuss the important variables based on the variable importance of the random forest model, which, in general, is a model with higher predictive power. These variables can be divided into several categories, including demographic variables, diagnostic-related variables, inpatient treatment-related variables, psychiatric drugs, and psychotherapy. Demographic variables included age, sex, and health insurance premium deciles. As age and gender are variables in the top 10 in the model for the total sample, it has become a data-driven rationale for additional machine-learning analysis with a sub-sample divided by age and gender. The premium decile, which indirectly measures the level of income, was also found to be an important predictor. Existing literature also found that among socioeconomic variables, living alone and unemployment had a large effect on suicide.[46]

In diagnosis-related variables, injuries to the abdomen, back, and reproductive system were found to be major variables in the age group of 10–20 years. According to the existing literature, previous attempts at self-harm are one of the strongest predictors of suicidal death.[47] Given the fact that X60–84 is rarely claimed, it can be interpreted that damage to the abdomen, back, and reproductive system may be a diagnosis that includes self-injurious behavior. In addition, circulatory- and respiratory-related diseases were classified as important predictors. They include "symptoms and signs involving the circulatory and respiratory systems (R00–R09)," "other acute lower respiratory infections (J20–J22)," and "acute upper respiratory infections (J00–J06)". The frequency of acute upper and lower respiratory tract infections and the occurrence of severe circulatory and respiratory symptoms caused by the infection are important predictors of suicide. According to several existing studies, an association between infection-induced inflammatory responses and suicidal behavior has been found. High levels of inflammation may be a tool to predict suicidal thoughts and depression in adults, and several studies have shown that systemic inflammation associated with depression may be associated with a leaky gut (i.e., abnormal permeability of the intestinal wall).[48] There are many factors predisposing to infection in this patient population, including local factors due to the tumor, specific deficiencies in host defense mechanisms due to certain malignant processes, and deficiencies in host defense mechanisms secondary to cancer chemotherapy.[49] It would be reasonable to interpret the incidence of complications in CRC patients as an important predictor of suicide.

Digestive system-related disease variables account for the largest portion of diagnostic variables with high predictive power. They were diseases of the esophagus, stomach, and duodenum (K20–K31); other diseases of the intestines (K55–K64); diseases of the liver (K70–K77); and malignant neoplasms of digestive organs (C15–C26). Given that the study sample consisted of CRC patients, this could be interpreted as a visit to a medical institution for symptomatic treatment due to cancer. The number of medical visits due to diseases related to the digestive system can be interpreted as pain or the severity of symptoms due to CRC (e.g., diarrhea, constipation, bloody stool). According to the existing literature, pain due to chronic disease is known to be a predictor of suicidal behavior, so this interpretation may have validity.[50]

Variables with particularly significant predictive power among diagnostic-related variables were malignant neoplasms of the colon, rectosigmoid junction, and rectum (C18–C20). This variable is related to the time of diagnosis of CRC, and the closer the time of diagnosis of CRC is to suicide, the higher the risk of suicide. To interpret this, the incidence of most suicide deaths was highest immediately after the diagnosis of CRC and decreased over time. According to several previous studies, it was concluded that the relative risk of suicide increased within the first month of diagnosis and significantly decreased over time in both men and women after diagnosis.[51]

As factors related to hospitalization, the frequency of procedures such as enema, urinary catheterization, and enteral nutrition were found to be major predictors of suicide. Patients with CRC are often hospitalized for long periods of time for postoperative care, chemotherapy, radiation therapy, and several symptomatic treatments. Procedures such as enema, urinary catheterization, and enteral nutrition are performed as needed, causing considerable discomfort to patients. According to the existing literature, the ability to tolerate discomfort is strongly associated with suicide.[52] Therefore, the above factors performed following hospitalization can be interpreted as reasonable predictors of suicide.

Psychiatric drug prescriptions were among the most important predictors. Sleeping pills, mood stabilizers, and antipsychotics (atypical) were the most important psychiatric drugs. Substance abuse, or substance use disorder, is known to be an important factor in suicide.[53,54] According to existing literature, decreased sleep time, insomnia, and nightmares are associated with the risk of suicidal behavior,[55,56] and it is interpreted that sleeping pills are prescribed for CRC patients with sleep disorders. Bipolar disorder and schizophrenia are among the most common psychiatric diagnoses among individuals who die by suicide.[53] It is interpreted that antipsychotics and mood stabilizers were prescribed for CRC patients at high risk of suicide with symptoms suggestive of bipolar disorder and schizophrenia. The number of psychiatric outpatient visits and the psychotherapy prescriptions were also found to be strongly predictive variables. The number of psychiatric visits was the sum of outpatient visits until suicide onset, and psychotherapy prescription was a dummy variable that considered the temporal distance. In this study, suicide was also associated with the total number of psychiatric outpatient visits, and the more recent the prescription for psychotherapy, the greater the association. In the existing literature, it has been reported that the number of visits to a psychiatrist has a very strong association with suicide.[30] However, it would be more reasonable to interpret that the use of psychiatric drugs and psychiatric visits found in the claim data were made through the prescription of primary care and psychiatric specialists. Therefore, it would be more justified to interpret the association between psychiatric treatments (e.g., drug use and psychiatric visits) and suicide in this study as a result of treatment performed for CRC patients with psychiatric problems rather than a causal inference that psychiatric treatment may increase suicide risk.

Risk factors related to suicide in the general population include chronic illness, pain, depression, being elderly and young, living alone, and unemployment. Gender also affects suicide rates, with men being four times more likely to commit suicide than women. Individuals over 65 years of age have a higher suicide rate.[58] More than 90% of individuals who commit suicide in the general population have depression, mental illness, or substance abuse problems. Many studies have reported high suicide rates across a wide range of cancer categories. A strong predictor of future suicide in the general population was "previous suicidal behavior."[47] Anxiety disorders, impulse control disorders, post-traumatic stress disorder, eating disorders, suicidal exposure by others, alcohol and substance abuse, physical abuse, or dependence on others were also strong predictors.[47] Chronic disease is an important risk factor for suicide in the general population. Recently, a cancer diagnosis has been reported as a very important predictor of suicide.[51] Major depressive disorder or bipolar disorder are the most common mental disorders in individuals who commit suicide.[53] Recent (within 6 months) use of sedatives was also associated with suicide,[54] and there was a study that inflammatory diseases and physical diseases such as disorders or pain in important organs significantly contributed to suicide risk.[46] In the younger generation, factors such as impulsive-aggressive personality traits contributed significantly to suicide.[59] Conversely, in the elderly group, depression, accompanying physical disease, sleep disturbance, and cognitive impairment were predictive factors contributing to suicide.[55]

2. Risk profile in colorectal cancer

In this study on CRC, it was confirmed that the above predictors still made a significant contribution to suicide prediction, except for some differences. As the most important factors found in the CRC population of the study, recent diagnosis of CRC, prescription of sleeping pills, psychotherapy, and urinary catheterization were the main risk predictors. Psychiatric disorders, such as major depressive disorder, were also major predictors, regardless of the underlying disease or complications. A specific suicide risk profile in the CRC population was that it included hospitalization-related treatments not commonly observed in the general population, such as enemas, urinary catheterization, and enteral nutrition. However, except for the above, sleeping pills, psychiatric visits, and recent cancer diagnosis were the major factors shared, and the composition of the major factors for suicide between the general population and CRC patients was similar.

3. Findings of machine learning-based suicide prediction research

The predictive performance of this study appeared to be higher than that of existing single-scale-based traditional statistical analysis studies. Several existing suicide prediction studies have modeled suicide prediction scales by defining high-risk groups (e.g., suicide-related emergency room admissions, psychiatric hospitalization, and psychiatric hospital discharge),[14-16] and the general population.[17,18,19] Most early suicide prediction studies reported the performance of the developed model using self-reported single scales such as hopelessness, depression, overall psychopathological severity, suicidal intention, and attitude toward suicide as predictors.[20,21] Recently, there have been studies using a set of predictors consisting of clinical and social demographic data extracted from registry data and electronic medical records.[60,61] Critics argued that the predictive performance of single-scale studies is generally evaluated as not being used for clinical decision-making.[54] In recent years, the trend of suicide prediction research has begun to proceed with machine-learning studies based on real-world data collected from daily administration. However, most of the existing studies over the past 50 years have performed suicide prediction studies using a single scale (Beck Hopeless Scale, Suicide Intent Scale, etc.) for patients defined as high-risk.[24] In a meta-analysis that confirmed the predictive performance of single-scale-based studies using this conventional statistical methodology, the pooled sensitivity was 0.77 and the pooled specificity was reported to be 0.41.[54] The sensitivity and specificity of this study were 0.84 and 0.69 at the optimal threshold, and the AUC was 0.84, which was higher than the existing single-scale-based traditional statistical analysis studies.

Compared to recent machine-learning suicide prediction studies targeting the general population, this study 1) found suicide risk predictors in CRC patients, including some factors shared with the general population, and 2) had similar predictive performance, despite its limitations in the variety of variable selection.

The major predictors specific to CRC found in this study were as follows: inpatient treatment-related factors that cause discomforts, such as enema, enteral nutrition, catheter insertion, and the time of CRC diagnosis. A study of the general population using data from a Danish registry[25] revealed that in addition to known risk factors such as schizophrenia and depression, stress disorders and medications such as antidepressants, antipsychotics, hypnotics, and sedatives were data-driven risk factors. It has also been reported that these risk factor profiles differ according to sex. In this study, a suicide prediction study was conducted with a large sample of patients diagnosed with CRC, and the profile of suicide risk factors was presented separately according to sex as well as to the different age groups and types of CRC.

Suicide risk factor profiles showed sex differences, as in previous studies. In men, prescription drugs such as sleeping pills and mood stabilizers and socioeconomic variables such as age and insurance premium were important factors, but in women, outpatient psychiatric visits were more important variables. There was a difference in that the factors causing discomforts, such as enema, enteral nutrition, and urinary catheterization, had greater importance in women. In addition, there were differences in the profiles according to age. In the younger age group, damage to the abdomen, back, groin, upper and lower respiratory tract infections, diseases, and digestive system complications were important factors. Conversely, the higher the age, the more significant the variables were: the diagnosis of CRC, the prescription of psychiatric drugs, and the number of visits to the psychiatrist. The risk factor profile also differs depending on the type of cancer. In the CRC group (i.e., C18–19), the prescription of sleeping pills was a more important factor, whereas in the CRC group (i.e., C20), a treatment that caused discomfort, such as an enema, was more important than sleeping pills. In addition, this study analyzed the CRC variable timing by adding it to the predictor set, distinguishing it from general diagnostic variables, and found that CRC diagnosis timing was a very important variable in predicting suicide in the sample.

The prediction performance of this study reached a level similar to that of previous suicide prediction machine-learning studies, despite its limitations in a variety of variable selections, especially sociodemographic features. A study of the general population using data from the Danish Registry[25] reported its significant predictive performance (AUC: 0.83–0.91). The prediction performance presented as the AUC of this study also showed a similar level (AUC: 0.76–0.90). This level of performance was achieved almost exclusively with information about claimed prescriptions and diagnoses. The Danish Registry consists of a much larger population sample (n = 5,303,674) and provides more sociodemographic data, not only on gender and age but also on immigration status, citizenship, family suicidal behavior (suicide or attempted suicide deaths), data on parents and spouses, marital status, income, and employment status (including recent job loss). Although the NHID data used in this study provided only limited information in this area, it was possible to derive a similar level of predictive performance using only the claimed prescription and diagnosis information.

Additionally, this study solved the overfitting problem more dynamically compared with previous studies. In previous machine-learning studies,[25] owing to the overfitting issue of the classification tree, the hyperparameters of the classification tree and random forest were set as fixed values. In this study, the results of a parsimonious model obtained by setting data-driven complexity parameters through a 10-fold cross-validation hyperparameter tuning process are presented. In conclusion, there were no significant differences in the important predictors suggested by the parsimonious model.

4. Applicability of the prediction model in actual clinical settings

According to a systematic literature review on suicide prediction research,[30] a low PPV was mentioned as the greatest impediment to the use of the suicide prediction model in actual clinical settings. Suicidal death is a very rare outcome, and a low PPV is due to case imbalance. However, the low PPV of suicide prediction algorithms is well established,[62] but this does not preclude their clinical utility.[63]

Some critics argue that these predictive studies need to present precision-recall curves to evaluate their clinical utility.[59] In this study, a model trained on 3,678 individuals (1,839 cases) was used to predict the suicide probability of 380,569 individuals in the original cohort. The changes in PPV and sensitivity according to threshold changes are presented. When the suicide risk probability was in the top 5%, the PPV was 3.30%; when the suicide risk was in the top 1%, it was 5.06%. This means that when defining the top 1% of risk probability as a cutoff, 36,367 individuals are theoretically classified as a high-risk group for suicide to predict 1,839 suicides, which corresponds to 9.56% of all CRC patients.

Owing to the low PPV caused by case imbalance, the size of the high-risk classification group that needs to be investigated to intervene in the actual suicide group is still considerable. However, this is considered to be much more cost-effective than intervention in all CRC patients as a high-risk group. It is also widely understood by epidemiologists that even high-risk behaviors do not predict rare outcomes well.[64] It is reported that about 15% of current cigarette smokers will develop lung cancer during their lifetime.[64] This probability can be considered as the same value as the PPV of developing lung cancer for high-risk classification based on smoking status. However, the public health community still upholds tobacco control, in part based on lung cancer risk, despite its low PPV.

In addition, in this study, PPV and NNS were presented as indicators to evaluate the applicability of the clinical environment. To evaluate the effectiveness of treatment or testing, an evaluation index on an absolute scale is required. NNS is defined as the number of individuals who need to be screened to prevent death. According to the operating characteristics in Table 7, when the threshold was 30%, it was confirmed that it was one of the most optimal points, with a sensitivity of 81.05 and a specificity of 69.64. If NNS is calculated based on this sensitivity, it would have a value between 1,029 and 258 depending on the effect of the future intervention (Table 8). When the effect of the intervention was arbitrarily set to 50%, the NNS was approximately 515. According to an existing study that developed a suicide risk prediction model using the electronic health records of patients visiting tertiary hospitals, the NNS had values ranging from 3,448 to 450.[65] This can be interpreted to mean that the effect of the predictive model is likely to be improved more than in previous studies, depending on the effectiveness of suicide prevention interventions.

According to a systematic literature review,[30] a one-shot method is insufficient to predict suicide in a clinical environment and a multi-stage approach should be adopted. First, by using passively collected data, such as administratively collected claim data and electronic medical records, high-risk groups for suicide are selected without contact between investigators and potential patients. In the next stage, a survey using the structured suicide risk scale should be conducted for the selected high-risk group with minimal contact with medical service providers. Finally, clinicians perform unstructured, in-depth clinical psychosocial risk assessments. This study constitutes the first stage of this multi-step process and can play a role in limiting the cost and effort required for an in-depth assessment of suicide risk.

In addition, in this study, the results of the manual method of performing the first step above were obtained (Table 9c) and showed a fair level of predictive power (sensitivity, 71.0; specificity, 70.42). This method of classifying a high-risk group by calculating a patient's suicide risk score based on medical records using a suicide risk score table can be manually implemented in a relatively small medical institution.

According to a meta-analysis of existing studies, it has been argued that combined interventions from various providers in multiple areas are required for effective intervention in high-risk groups selected as predictive models.[66] This means integrating at the community and primary care levels. At the community level, examples include campaigns and media guidelines that use public relations campaigns to present information directly helpful to suicide and provide personnel, such as teachers and religious individuals, to help raise awareness of the potential risk of suicide. At the primary care level, the general practitioner may develop a pharmacological and non-pharmacological treatment plan for suicidal thoughts and behaviors and may request a referral to a higher-level care provider.

However, CRC patients, the subjects of this study, generally receive treatment at a tertiary hospital level; therefore, more direct intervention can be attempted. Inpatient treatment for patients at high risk of suicide attempts is known to be highly effective.[65]

The record of defined psychiatric treatment variables was an important predictor of the high-risk group for suicide, and this was interpreted by practitioners as the result of treatment for psychiatric symptoms. In other words, oncologists need to recognize the need for comprehensive intervention to prevent the progression of CRC in patients with psychiatric symptoms to psychiatric problems, including suicide, rather than just symptomatic treatment. This should be accompanied by a community-level approach that can improve the quality of life of symptomatic patients.

5. Limitation

This study has several limitations. First, the study relied on only two machine-learning classifiers: CART and RF. Other classifiers or meta-classifiers (such as super-learners) have the potential to improve prediction performance. Extending the classifiers used to investigate suicide prediction is an important area for future research. However, this study attempted to use a predictive model that generated a nonlinear model using a set of thousands of predictors. Tree-based algorithms are frequently used in classification problems because of their high accuracy in mapping nonlinear relationships, stability of implementation, and ease of interpretation.[67] In addition, the random forest model was used to solve the disadvantage of overfitting. Random forest is an algorithm that effectively offsets the bias-variance trade-off using a technique called bagging, and its predictive power is superior to that of existing tree-based models.[68]

Second, insurance claim data may misclassify suicide attempts. No patients were found in the data claimed with ICD-10 code X60–84. Claim data are not prepared according to the guidelines for collecting disease statistics but are being recorded as an auxiliary tool to justify the reimbursement of services provided to patients. Therefore, "External causes of morbidity and mortality (V01–Y98)" rarely appear in the claim data, and it was almost impossible to confirm suicidality, such as suicidal thoughts and attempts, which is a series of processes leading to death due to suicide.[69] According to the suicide prevention white paper published in 2021 by the Ministry of Health and Welfare and the Korea Foundation for Suicide Prevention,[70] suicide thoughts among adults were 4.6% in 2019, suicide plans were 1.3%, suicide attempts were 0.4%, and deaths due to suicide were reported at 0.027%. It can be said that the probability of having an outcome related to suicidality in CRC patients who did not die by suicide could be quite high. Suicidal thoughts or attempts cannot be measured in patients with CRC who do not die by suicide; therefore, the results of this study may have been somewhat diluted considering the above facts.

Third, the scope of these findings' international application may be unclear. However, many results are consistent with previous studies conducted externally.[25-26,71] Fourth, the psychiatric medications used in our study were classified as a class of medications. Drugs in the same class can be associated with different symptoms or diseases. For example, quetiapine and clozapine belong to the same class of atypical antipsychotics, but while clozapine is mainly used for treatment-resistant schizophrenia patients, quetiapine can also be used for patients with sleep disorders or bipolar disorder. Since this study did not distinguish between medications as predictor variables, the predictive power of each drug on suicide was unknown. This could be an area for further analysis in future studies. Finally, the data used in this study used claims information paid by the insurer, and non-insured services were not included in the study's data. In actual medical practice, since patients with CRC often use uninsured medical services, it should be considered that the predicted results reflecting this may differ from the results of this study.

V. CONCLUSION

The ability to clinically predict suicide remains poor despite abundant research in this area. This study developed a predictive model for suicidal death using machine-learning techniques based on data from the NHID tailored to CRC patients, a high-suicidal-risk population that can be used as a basis for further research and interventions. The top predictors and predictive performance of the model were confirmed by stratifying age, sex, and type of cancer using supervised machine-learning techniques (CART and RF) using more than 1,000 predictors such as demographic, diagnostic, medication, and prescription data. The proportion of patients with CRC was determined, and a nested case-control study design was used to determine the magnitude of the association of the predictors. Prescribed procedures and medications used to treat the quality of life, complications, and psychiatric disorders in patients with CRC have been identified as key predictors. In addition, the size of the association increased as such prescriptions occurred recently. The results of this study confirmed that interventions for suicide are needed not only in the field of psychiatry but also in fields related to physical diseases, and close monitoring of suicide is necessary after identifying important predictive factors. In addition, the results of this study are necessary as the first step in the multi-staged approach for suicide intervention described previously and have implications as a preliminary process necessary for cost-effective intervention before in-depth clinical psychosocial risk assessment.

REFERENCES

1. Misono S, Weiss NS, Fann JR, Redman M, Yueh B. Incidence of suicide in persons with cancer. Journal of Clinical Oncology. 2008;26(29):4731.

2. Bray F, Laversanne M, Weiderpass E, Soerjomataram I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer. 2021;127(16):3029-30.

3. Cengiz B, Bahar Z. Perceived barriers and home care needs when adapting to a fecal ostomy. Journal of Wound, Ostomy and Continence Nursing. 2017;44(1):63-8.

4. Krouse RS, Herrinton LJ, Grant M, Wendel CS, Green SB, Mohler MJ, et al. Health-related quality of life among long-term rectal cancer survivors with an ostomy: manifestations by sex. J Clin Oncol. 2009;27(28):4664-70.

5. Lorenz KA, Lynn J, Dy SM, Shugarman LR, Wilkinson A, Mularski RA, et al. Evidence for improving palliative care at the end of life: a systematic review. Annals of internal medicine. 2008;148(2):147-59.

6. Massie MJ, Gagnon P, Holland JC. Depression and suicide in patients with cancer. Journal of pain and symptom management. 1994;9(5):325-40.

7. Davidson JR, MacLean AW, Brundage MD, Schulze K. Sleep disturbance in cancer patients. Social science & medicine. 2002;54(9):1309-21.

8. Breitbart W, Alici Y. Evidence-based treatment of delirium in patients with cancer. Journal of Clinical Oncology. 2012;30(11):1206.

9. Dodds TJ. Prescribed benzodiazepines and suicide risk: a review of the literature. The primary care companion for CNS disorders. 2017;19(2):22746.

10. Hengartner MP, Amendola S, Kaminski JA, Kindler S, Bschor T, Plöderl M. Suicide risk with selective serotonin reuptake inhibitors and other new-generation antidepressants in adults: a systematic review and meta-analysis of observational studies. J Epidemiol Community Health. 2021;75(6):523-30.

11. Sharma V. Atypical antipsychotics and suicide in mood and anxiety disorders. Bipolar disorders. 2003;5:48-52.

12. Samawi H, Shaheen A, Tang P, Heng D, Cheung W, Vickers M. Risk and predictors of suicide in colorectal cancer patients: a Surveillance, Epidemiology, and End Results analysis. Current Oncology. 2017;24(6):513-7.

13. Park H-C, Shin A, Kim B-W, Jung K-W, Won Y-J, Oh JH, et al. Data on the characteristics and the survival of korean patients with colorectal cancer from the Korea central cancer registry. Annals of Coloproctology. 2013;29(4):144-9.

14. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry. 2017;210:387–95.

15. Chan KCG, Yam SCP, Zhang Z. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. J R Stat Soc Ser B Stat Methodol. 2016;78:673–700.

16. Katz C, Randall JR, Sareen J, Chateau D, Walld R, Leslie WD, et al. Predicting suicide with the SAD PERSONS scale. Depress Anxiety. 2017;34:809–16.

17. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, et al. Predicting suicidal behavior from longitudinal electronic health records. Am J Psychiatry. 2017;174:154–62.

18. Ben-Ari A, Hammond K. Text mining the EMR for modeling and predicting suicidal behavior among US veterans of the 1991 Persian gulf war. 2015 48th Hawaii International Conference on System Sciences. Kauai, HI; 2015;3168–75. https://doi.org/10.1109/HICSS.2015.

19. Choi SB, Lee W, Yoon J-H, Won J-U, Kim DW. Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. J Affect Disord. 2018;231:8–14.

20. Beck A, Steer R. BHS, Beck Hopelessness Scale: manual. San Antonio TX: Psychological Corporation; 1988.

21. Dozois DJ, Dobson KS, Ahnberg JL. A psychometric evaluation of the Beck Depression Inventory–II. Psychological assessment. 1998;10(2):83.

22. Large M, Kaneson M, Myles N, Myles H, Gunaratne P, Ryan C. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS ONE. 2016;11:e0156322.

23. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychological bulletin. 2017;143(2):187.

24. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychological bulletin. 2017;143(2):187.

25. Gradus JL, Rosellini AJ, Horváth-Puhó E, Street AE, Galatzer-Levy I, Jiang T, et al. Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. JAMA psychiatry. 2020;77(1):25-34.

26. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science. 2017;5(3):457-69.

27. Kessler RC, Warner CH, Ivany C, Petukhova MV, Rose S, Bromet EJ, et al. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). JAMA psychiatry. 2015;72(1):49-57.

28. Kessler RC, Stein MB, Petukhova MV, Bliese P, Bossarte RM, Bromet EJ, et al. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Molecular psychiatry. 2017;22(4):544-51.

29. Bossarte RM, Kennedy CJ, Luedtke A, et al. Invited commentary: new directions in machine learning analyses of administrative data to prevent suicide-related behaviors. Am J Epidemiol. 2021;190(12):2528–2533.

30. Kessler RC, Bossarte RM, Luedtke A, et al. Suicide prediction models: a critical review of recent research with recommendations for the way forward. Mol Psychiatry. 2020;25(1):168–179.

31. Rembold CM. Number needed to screen: development of a statistic for disease screening. Bmj. 1998;317(7154):307-12.

32. Ribeiro J, Franklin J, Fox K, Bentley K, Kleiman E, Chang B, et al. Suicide as a complex classification problem: machine learning and related techniques can advance suicide prediction-a reply to Roaldset (2016). Psychological medicine. 2016;46(9):2009-10.

33. Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Cai T, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Molecular psychiatry. 2016;21(10):1366-71.

34. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: Applications and ethics. Behavioral sciences & the law. 2019;37(3):214-22.

35. Gradus JL, Rosellini AJ, Horváth-Puhó E, Jiang T, Street AE, Galatzer-Levy I, et al. Predicting sex-specific nonfatal suicide attempt risk using machine learning and data from Danish national registries. American journal of epidemiology. 2021;190(12):2517-27.

36. Anguiano L, Mayer DK, Piven ML, Rosenstein D. A literature review of suicide in cancer patients. Cancer nursing. 2012;35(4):E14-E26.

37. Yoon-jung C, Kang J-g, Sang-hyun L, Yong-seok C, Youk TM. Colorectal cancer early screening compliance analysis using national health checkup data. Report. Ilsan Hospital Research Institute, Institute IHR; 2015 2015-12-01. Report No.: 2015-20-017.

38. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. University of California, Berkeley. 2004;110(1-12):24.

39. Huang BF, Boutros PC. The parameter sensitivity of random forests. BMC bioinformatics. 2016;17(1):1-13.

40. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological methods. 2009;14(4):323.

41. Platt R, Hutcheon J, Suissa S. Immortal time bias in epidemiology. Current Epidemiology Reports. 2019;6(1):23-7.

42. Gleiss A, Oberbauer R, Heinze G. An unjustified benefit: immortal time bias in the analysis of time-dependent events. Transplant international. 2018;31(2):125-30.

43. Wang M-H, Shugart YY, Cole SR, Platz EA. A simulation study of control sampling methods for nested case-control studies of genetic and molecular biomarkers and prostate cancer progression. Cancer Epidemiology Biomarkers & Prevention. 2009;18(3):706-11.

44. Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Stampfer MJ, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology (Cambridge, Mass). 2008;19(6):766.

45. de Man-van Ginkel JM, Hafsteinsdóttir TB, Lindeman E, Ettema RG, Grobbee DE, Schuurmans MJ. In-hospital risk prediction for post-stroke depression: development and validation of the Post-stroke Depression Prediction Scale. Stroke. 2013;44(9):2441-5.

46. Qin P, Agerbo E, Mortensen PB. Suicide risk in relation to socioeconomic, demographic, psychiatric, and familial factors: a national register–based study of all suicides in Denmark, 1981–1997. American journal of psychiatry. 2003;160(4):765-72.

47. Nock MK, Green JG, Hwang I, McLaughlin KA, Sampson NA, Zaslavsky AM, et al. Prevalence, correlates, and treatment of lifetime suicidal behavior among adolescents: results from the National Comorbidity Survey Replication Adolescent Supplement. JAMA psychiatry. 2013;70(3):300-10.

48. Brundin L, Bryleva EY, Thirtamara Rajamani K. Role of inflammation in suicide: from mechanisms to treatment. Neuropsychopharmacology. 2017;42(1):271-83.

49. Bodey GP. Infection in cancer patients: a continuing association. The American journal of medicine. 1986;81(1):11-26.

50. Racine M. Chronic pain and suicide risk: A comprehensive review. Progress in Neuro-Psychopharmacology and Biological Psychiatry. 2018;87:269-80.

51. Hem E, Loge JH, Haldorsen T, Ekeberg O. Suicide risk in cancer patients from 1960 to 1999. J Clin Oncol. 2004; 22 (20): 4209–4216.

52. Pennings SM, Anestis MD. Discomfort intolerance and the acquired capability for suicide. Cognitive therapy and research. 2013;37(6):1269-75.

53. Arsenault-Lapierre G, Kim C, Turecki G. Psychiatric diagnoses in 3275 suicides: a meta-analysis. BMC psychiatry. 2004;4(1):1-11.

54. Artenie AA, Bruneau J, Roy É, Zang G, Lespérance F, Renaud J, et al. Licit and illicit substance use among people who inject drugs and the association with subsequent suicidal attempt. Addiction. 2015;110(10):1636-43.

55. Bernert RA, Turvey CL, Conwell Y, Joiner TE. Association of poor subjective sleep quality with risk for death by suicide during a 10-year period: a longitudinal, population-based study of late life. JAMA psychiatry. 2014;71(10):1129-37.

56. Bernert RA, Kim JS, Iwata NG, Perlis ML. Sleep disturbances as an evidence-based suicide risk factor. Current psychiatry reports. 2015;17(3):1-9.

57. Control CfD, Prevention. Web-based injury statistics query and reporting system (WISQARS). www cdc gov/ncipc/wisqars. 2002.

58. Anguiano L, Mayer DK, Piven ML, Rosenstein D. A literature review of suicide in cancer patients. Cancer nursing. 2012;35(4):E14-E26.

59. Séguin M, Beauchamp G, Robert M, DiMambro M, Turecki G. Developmental model of suicide trajectories. The British Journal of Psychiatry. 2014;205(2):120-6.

60. Randall JR, Rowe BH, Dong KA, Nock MK, Colman I. Assessment of self-harm risk using implicit thoughts. Psychol Assess. 2013;25:714–21.

61. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. J Am Stat Assoc. 2015;90:122–9.

62. Belsher BE, Smolenski DJ, Pruitt LD, et al. Prediction models for suicide attempts and deaths: a systematic review and simulation. JAMA Psychiat. 2019;76(6):642–651.

63. Simon GE, Shortreed SM, Coley RY. Positive predictive values and potential success of suicide prediction models. JAMA Psychiat. 2019;76(8):868–869.

64. Bruder C, Bulliard J-L, Germann S, Konzelmann I, Bochud M, Leyvraz M, et al. Estimating lifetime and 10-year risk of lung cancer. Preventive medicine reports. 2018;11:125-30.

65. Walsh CG, Johnson KB, Ripperger M, Sperry S, Harris J, Clark N, et al. Prospective

validation of an electronic health record–based, real-time suicide risk model. JAMA network open. 2021;4(3):e211428-e.

66. Hofstra E, Van Nieuwenhuizen C, Bakker M, Özgül D, Elfeddali I, de Jong SJ, et al. Effectiveness of suicide prevention interventions: a systematic review and meta-analysis. General hospital psychiatry. 2020;63:127-40.

67. Breiman, Leo, et al. Classification and regression trees. Routledge, 2017.

68. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

69. Bae S-O, Kang G-W. A comparative study of the disease codes between Korean national health insurance claims and Korean national hospital discharge in-depth injury survey. Health Policy and Management. 2014;24(4):322-9.

70. Deok-cheol K. 2021 White paper on suicide prevention. report. Ministry of health and welfare, Welfare MOH; 2021 2021-07. Report No.: ISSN 2508-2485.

71. Breslau N, Davis GC, Andreski P. Migraine, psychiatric disorders, and suicide attempts: an epidemiologic study of young adults. Psychiatry research. 1991;37(1):11-23.

Appendix 1. Composition of dummy variables for machine learning model (n=1,608)

Code	Variable Label (each code with (*) asterisk has 5 temporal dummy variables)
Demographic Variables (n = 5)	
CC	Colon cancer type (i.e., c18, 19, 20)
Sex	gender
Age	age (in year)
Premium	insurance premium
Stage	colorectal cancer stage
Diagnostic variables (A00-T88) (n = 1,120) (each code with (*) asterisk has 5 temporal dummy variables)	
A00_A09 *	Intestinal infectious diseases
A15_A19 *	Tuberculosis
A20_A28 *	Certain zoonotic bacterial diseases
A30_A49 *	Other bacterial diseases
A50_A64 *	Infections with a predominantly sexual mode of transmission
A65_A69 *	Other spirochetal diseases
A70_A74 *	Other diseases caused by chlamydiae
A75_A79 *	Rickettsioses
A80_A89 *	Viral and prion infections of the central nervous system
A90_A99 *	Arthropod-borne viral fevers and viral hemorrhagic fevers
B00_B09 *	Viral infections characterized by skin and mucous membrane lesions
B10_B10 *	Other human herpesviruses
B15_B19 *	Viral hepatitis
B20_B20 *	Human immunodeficiency virus [HIV] disease
B25_B34 *	Other viral diseases
B35_B49 *	Mycoses
B50_B64 *	Protozoal diseases
B65_B83 *	Helminthiases
B85_B89 *	Pediculosis, acariasis and other infestations
B90_B94 *	Sequelae of infectious and parasitic diseases
B95_B97 *	Bacterial and viral infectious agents
B99_B99 *	Other infectious diseases
C18_C20 *	**Malignant neoplasms of colon, recto-signmoid junction, and rectum**
C00_C14 *	Malignant neoplasms of lip, oral cavity and pharynx
C15_C26 *	Malignant neoplasms of digestive organs
C30_C39 *	Malignant neoplasms of respiratory and intrathoracic organs
C40_C41 *	Malignant neoplasms of bone and articular cartilage
C43_C44 *	Melanoma and other malignant neoplasms of skin
C45_C49 *	Malignant neoplasms of mesothelial and soft tissue
C50_C50 *	Malignant neoplasms of breast
C51_C58 *	Malignant neoplasms of female genital organs
C60_C63 *	Malignant neoplasms of male genital organs
C64_C68 *	Malignant neoplasms of urinary tract
C69_C72 *	Malignant neoplasms of eye, brain and other parts of central nervous system
C73_C75 *	Malignant neoplasms of thyroid and other endocrine glands
C76_C80 *	Malignant neoplasms of ill-defined, other secondary and unspecified sites
C81_C96 *	Malignant neoplasms of lymphoid, hematopoietic and related tissue
D00_D09 *	In situ neoplasms
D10_D36 *	Benign neoplasms, except benign neuroendocrine tumors
D37_D48 *	Neoplasms of uncertain behavior, polycythemia vera and myelodysplastic syndromes
D49_D49 *	Neoplasms of unspecified behavior
D50_D53 *	Nutritional anemias
D55_D59 *	Hemolytic anemias
D60_D64 *	Aplastic and other anemias and other bone marrow failure syndromes
D65_D69 *	Coagulation defects, purpura and other hemorrhagic conditions
D70_D77 *	Other disorders of blood and blood-forming organs
D78_D78 *	Intraoperative and postprocedural complications of the spleen
D80_D89 *	Certain disorders involving the immune mechanism
E00_E07 *	Disorders of thyroid gland
E08_E13 *	Diabetes mellitus

E15_E16 *	Other disorders of glucose regulation and pancreatic internal secretion
E20_E35 *	Disorders of other endocrine glands
E36_E36 *	Intraoperative complications of endocrine system
E40_E46 *	Malnutrition
E50_E64 *	Other nutritional deficiencies
E65_E68 *	Overweight, obesity and other hyperalimentation
E70_E88 *	Metabolic disorders
E89_E89 *	Postprocedural endocrine and metabolic complications and disorders, not elsewhere classified
G00_G09 *	Inflammatory diseases of the central nervous system
G10_G14 *	Systemic atrophies primarily affecting the central nervous system
G20_G26 *	Extrapyramidal and movement disorders
G30_G32 *	Other degenerative diseases of the nervous system
G35_G37 *	Demyelinating diseases of the central nervous system
G40_G47 *	Episodic and paroxysmal disorders
G50_G59 *	Nerve, nerve root and plexus disorders
G60_G65 *	Polyneuropathies and other disorders of the peripheral nervous system
G70_G73 *	Diseases of myoneural junction and muscle
G80_G83 *	Cerebral palsy and other paralytic syndromes
G89_G99 *	Other disorders of the nervous system
H00_H05 *	Disorders of eyelid, lacrimal system and orbit
H10_H11 *	Disorders of conjunctiva
H15_H22 *	Disorders of sclera, cornea, iris and ciliary body
H25_H28 *	Disorders of lens
H30_H36 *	Disorders of choroid and retina
H40_H42 *	Glaucoma
H43_H44 *	Disorders of vitreous body and globe
H46_H47 *	Disorders of optic nerve and visual pathways
H49_H52 *	Disorders of ocular muscles, binocular movement, accommodation and refraction
H53_H54 *	Visual disturbances and blindness
H55_H57 *	Other disorders of eye and adnexa
H59_H59 *	Intraoperative and postprocedural complications and disorders of eye and adnexa
H60_H62 *	Diseases of external ear
H65_H75 *	Diseases of middle ear and mastoid
H80_H83 *	Diseases of inner ear
H90_H94 *	Other disorders of ear
H95_H95 *	Intraoperative and postprocedural complications and disorders of ear and mastoid process,
I00_I02 *	Acute rheumatic fever
I05_I09 *	Chronic rheumatic heart diseases
I10_I16 *	Hypertensive diseases
I20_I25 *	Ischemic heart diseases
I26_I28 *	Pulmonary heart disease and diseases of pulmonary circulation
I30_I52 *	Other forms of heart disease
I60_I69 *	Cerebrovascular diseases
I70_I79 *	Diseases of arteries, arterioles and capillaries
I80_I89 *	Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
I95_I99 *	Other and unspecified disorders of the circulatory system
J00_J06 *	Acute upper respiratory infections
J09_J18 *	Influenza and pneumonia
J20_J22 *	Other acute lower respiratory infections
J30_J39 *	Other diseases of upper respiratory tract
J40_J47 *	Chronic lower respiratory diseases
J60_J70 *	Lung diseases due to external agents
J80_J84 *	Other respiratory diseases principally affecting the interstitium
J85_J86 *	Suppurative and necrotic conditions of the lower respiratory tract
J90_J94 *	Other diseases of the pleura
J95_J95 *	Intraoperative and postprocedural complications and disorders of respiratory system
J96_J99 *	Other diseases of the respiratory system
K00_K14 *	Diseases of oral cavity and salivary glands
K20_K31 *	Diseases of esophagus, stomach and duodenum
K35_K38 *	Diseases of appendix

K40_K46 *	Hernia
K50_K52 *	Noninfective enteritis and colitis
K55_K64 *	Other diseases of intestines
K65_K68 *	Diseases of peritoneum and retroperitoneum
K70_K77 *	Diseases of liver
K80_K87 *	Disorders of gallbladder, biliary tract and pancreas
K90_K95 *	Other diseases of the digestive system
L00_L08 *	Infections of the skin and subcutaneous tissue
L10_L14 *	Bullous disorders
L20_L30 *	Dermatitis and eczema
L40_L45 *	Papulosquamous disorders
L49_L54 *	Urticaria and erythema
L55_L59 *	Radiation-related disorders of the skin and subcutaneous tissue
L60_L75 *	Disorders of skin appendages
L76_L76 *	Intraoperative and postprocedural complications of skin and subcutaneous tissue
L80_L99 *	Other disorders of the skin and subcutaneous tissue
M00_M02 *	Infectious arthropathies
M04_M04 *	Autoinflammatory syndromes
M05_M14 *	Inflammatory polyarthropathies
M15_M19 *	Osteoarthritis
M20_M25 *	Other joint disorders
M26_M27 *	Dentofacial anomalies [including malocclusion] and other disorders of jaw
M30_M36 *	Systemic connective tissue disorders
M40_M43 *	Deforming dorsopathies
M45_M49 *	Spondylopathies
M50_M54 *	Other dorsopathies
M60_M63 *	Disorders of muscles
M65_M67 *	Disorders of synovium and tendon
M70_M79 *	Other soft tissue disorders
M80_M85 *	Disorders of bone density and structure
M86_M90 *	Other osteopathies
M91_M94 *	Chondropathies
M95_M95 *	Other disorders of the musculoskeletal system and connective tissue
M96_M96 *	Intraoperative and postprocedural complications and disorders of musculoskeletal system, not elsewhere classified
M97_M97 *	Periprosthetic fracture around internal prosthetic joint
M99_M99 *	Biomechanical lesions, not elsewhere classified
N00_N08 *	Glomerular diseases
N10_N16 *	Renal tubulo-interstitial diseases
N17_N19 *	Acute kidney failure and chronic kidney disease
N20_N23 *	Urolithiasis
N25_N29 *	Other disorders of kidney and ureter
N30_N39 *	Other diseases of the urinary system
N40_N53 *	Diseases of male genital organs
N60_N65 *	Disorders of breast
N70_N77 *	Inflammatory diseases of female pelvic organs
N80_N98 *	Noninflammatory disorders of female genital tract
N99_N99 *	Intraoperative and postprocedural complications and disorders of genitourinary system, not elsewhere classified
O00_O08 *	Pregnancy with abortive outcome
O09_O09 *	Supervision of high risk pregnancy
O10_O16 *	Edema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium
O20_O29 *	Other maternal disorders predominantly related to pregnancy
O30_O48 *	Maternal care related to the fetus and amniotic cavity and possible delivery problems
O60_O77 *	Complications of labor and delivery
O80_O82 *	Encounter for delivery
O85_O92 *	Complications predominantly related to the puerperium
P00_P04 *	Newborn affected by maternal factors and by complications of pregnancy, labor, and delivery
P05_P08 *	Disorders of newborn related to length of gestation and fetal growth
P09_P09 *	Abnormal findings on neonatal screening
P10_P15 *	Birth trauma
P19_P29 *	Respiratory and cardiovascular disorders specific to the perinatal period

P35_P39 *	Infections specific to the perinatal period
P50_P61 *	Hemorrhagic and hematological disorders of newborn
P70_P74 *	Transitory endocrine and metabolic disorders specific to newborn
P76_P78 *	Digestive system disorders of newborn
P80_P83 *	Conditions involving the integument and temperature regulation of newborn
P84_P84 *	Other problems with newborn
P90_P96 *	Other disorders originating in the perinatal period
Q00_Q07 *	Congenital malformations of the nervous system
Q10_Q18 *	Congenital malformations of eye, ear, face and neck
Q20_Q28 *	Congenital malformations of the circulatory system
Q30_Q34 *	Congenital malformations of the respiratory system
Q35_Q37 *	Cleft lip and cleft palate
Q38_Q45 *	Other congenital malformations of the digestive system
Q50_Q56 *	Congenital malformations of genital organs
Q60_Q64 *	Congenital malformations of the urinary system
Q65_Q79 *	Congenital malformations and deformations of the musculoskeletal system
Q80_Q89 *	Other congenital malformations
Q90_Q99 *	Chromosomal abnormalities, not elsewhere classified
R00_R09 *	Symptoms and signs involving the circulatory and respiratory systems
R10_R19 *	Symptoms and signs involving the digestive system and abdomen
R20_R23 *	Symptoms and signs involving the skin and subcutaneous tissue
R25_R29 *	Symptoms and signs involving the nervous and musculoskeletal systems
R30_R39 *	Symptoms and signs involving the genitourinary system
R40_R46 *	Symptoms and signs involving cognition, perception, emotional state and behavior
R47_R49 *	Symptoms and signs involving speech and voice
R50_R69 *	General symptoms and signs
R70_R79 *	Abnormal findings on examination of blood, without diagnosis
R80_R82 *	Abnormal findings on examination of urine, without diagnosis
R83_R89 *	Abnormal findings on examination of other body fluids, substances and tissues, without diagnosis
R90_R94 *	Abnormal findings on diagnostic imaging and in function studies, without diagnosis
R97_R97 *	Abnormal tumor markers
R99_R99 *	Ill-defined and unknown cause of mortality
S00_S09 *	Injuries to the head
S10_S19 *	Injuries to the neck
S20_S29 *	Injuries to the thorax
S30_S39 *	Injuries to the abdomen, lower back, lumbar spine, pelvis and external genitals
S40_S49 *	Injuries to the shoulder and upper arm
S50_S59 *	Injuries to the elbow and forearm
S60_S69 *	Injuries to the wrist, hand and fingers
S70_S79 *	Injuries to the hip and thigh
S80_S89 *	Injuries to the knee and lower leg
S90_S99 *	Injuries to the ankle and foot
T07_T07 *	Injuries involving multiple body regions
T14_T14 *	Injury of unspecified body region
T15_T19 *	Effects of foreign body entering through natural orifice
T20_T25 *	Burns and corrosions of external body surface, specified by site
T26_T28 *	Burns and corrosions confined to eye and internal organs
T30_T32 *	Burns and corrosions of multiple and unspecified body regions
T33_T34 *	Frostbite
T36_T50 *	Poisoning by, adverse effect of and underdosing of drugs, medicaments and biological substances
T51_T65 *	Toxic effects of substances chiefly nonmedicinal as to source
T66_T78 *	Other and unspecified effects of external causes
T79_T79 *	Certain early complications of trauma
T80_T88 *	Complications of surgical and medical care, not elsewhere classified
Psychiatric diagnostic variables (F01-F99) n = 360 (each code with (*) asterisk has 5 temporal dummy variables)	
F01_CD *	Vascular dementia
F02_CD *	Dementia in other diseases classified elsewhere
F03_CD *	Unspecified dementia
F04_CD *	Amnestic disorder due to known physiological condition
F05_CD *	Delirium due to known physiological condition

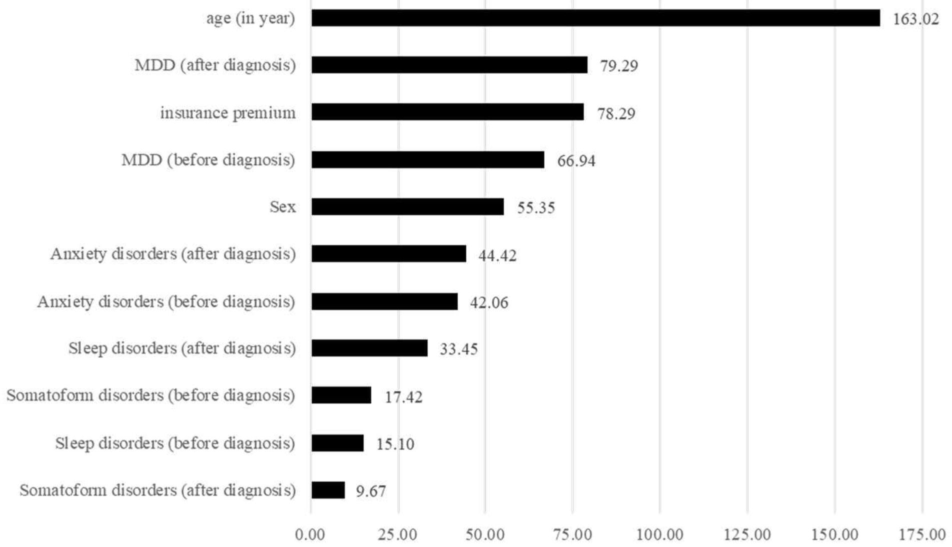
F06_CD *	Other mental disorders due to known physiological condition
F07_CD *	Personality and behavioral disorders due to known physiological condition
F09_CD *	Unspecified mental disorder due to known physiological condition
F10_CD *	Alcohol related disorders
F11_CD *	Opioid related disorders
F12_CD *	Cannabis related disorders
F13_CD *	Sedative, hypnotic, or anxiolytic related disorders
F14_CD *	Cocaine related disorders
F15_CD *	Other stimulant related disorders
F16_CD *	Hallucinogen related disorders
F17_CD *	Nicotine dependence
F18_CD *	Inhalant related disorders
F19_CD *	Other psychoactive substance related disorders
F20_CD *	Schizophrenia
F21_CD *	Schizotypal disorder
F22_CD *	Delusional disorders
F23_CD *	Brief psychotic disorder
F24_CD *	Shared psychotic disorder
F25_CD *	Schizoaffective disorders
F28_CD *	Other psychotic disorder not due to a substance or known physiological condition
F29_CD *	Unspecified psychosis not due to a substance or known physiological condition
F30_CD *	Manic episode
F31_CD *	Bipolar disorder
F32_CD *	Major depressive disorder, single episode
F33_CD *	Major depressive disorder, recurrent
F34_CD *	Persistent mood [affective] disorders
F39_CD *	Unspecified mood [affective] disorder
F40_CD *	Phobic anxiety disorders
F41_CD *	Other anxiety disorders
F42_CD *	Obsessive-compulsive disorder
F43_CD *	Reaction to severe stress, and adjustment disorders
F44_CD *	Dissociative and conversion disorders
F45_CD *	Somatoform disorders
F48_CD *	Other nonpsychotic mental disorders
F50_CD *	Eating disorders
F51_CD *	Sleep disorders not due to a substance or known physiological condition
F52_CD *	Sexual dysfunction not due to a substance or known physiological condition
F53_CD *	Puerperal psychosis
F54_CD *	Psychological and behavioral factors associated with disorders or diseases classified elsewhere
F55_CD *	Abuse of non-psychoactive substances
F59_CD *	Unspecified behavioral syndromes associated with physiological disturbances and physical factors
F60_CD *	Specific personality disorders
F63_CD *	Impulse disorders
F64_CD *	Gender identity disorders
F65_CD *	Paraphilias
F66_CD *	Other sexual disorders
F68_CD *	Other disorders of adult personality and behavior
F69_CD *	Unspecified disorder of adult personality and behavior
F70_CD *	Mild intellectual disabilities
F71_CD *	Moderate intellectual disabilities
F72_CD *	Severe intellectual disabilities
F73_CD *	Profound intellectual disabilities
F78_CD *	Other intellectual disabilities
F79_CD *	Unspecified intellectual disabilities
F80_CD *	Specific developmental disorders of speech and language
F81_CD *	Specific developmental disorders of scholastic skills
F82_CD *	Specific developmental disorder of motor function
F84_CD *	Pervasive developmental disorders
F88_CD *	Other disorders of psychological development
F89_CD *	Unspecified disorder of psychological development

F90_CD *	Attention-deficit hyperactivity disorders
F91_CD *	Conduct disorders
F93_CD *	Emotional disorders with onset specific to childhood
F94_CD *	Disorders of social functioning with onset specific to childhood and adolescence
F95_CD *	Tic disorder
F98_CD *	Other behavioral and emotional disorders with onset usually occurring in childhood and adolescence
F99_CD *	Mental disorder, not otherwise specified
Psychiatric drug variables (n = 35) (each code with (*) asterisk has 5 temporal dummy variables)	
YAK_AD *	**Antidepressant** ('107501ATB', '107502ATB', '161501ACH', '161502ACH', '162501ATB', '196201ATB', '209301ATB', '242901ATB', '242902ATB', '247502ACR', '247504ACR','428102ATR', '428301ATB', '474802ATB')
YAK_TYP_PSY *	**Typical antipsychotic** ('131901ATB', '131905ATB', '131908ATB', '132101ATB', '167903ATB', '167904ATB', '167905ATB', '167906ATB', '167908ATB', '183301ATB', '183302ATB', '183303ATB', '196901ATB', '196902ATB', '211401ATB', '212401ATB', '212402ATB', '167930BIJ', '168030BIJ')
YAK_ATYP_PSY *	**Atypical antipsychotic** ('183501ATB', '204001ATB', '204002ATB', '224201ATB', '224202ATB', '378601ATB', '378602ATB')
YAK_OPIOID *	**Opioids** ('120205CPC', '137703ATB', '185102ACH', '242301ATR', '242302ATR', '242305ACH', '242305ATB', '267400ATB', '268000ATB', '313400ACH', '480600ATB', '513000ATB', '513000ATR', '514100ATR')
YAK_AC *	**Anticonvulsant (Mood stabilizer)** ('101501ATB', '123102ATB', '123102ATR', '123104ATR', '135702ATB', '136401ATB', '137102ACH', '142902ATB', '142903ATB', '147702ATR', '160601ATB', '164201ACH', '164202ACH', '164203ACH', '164204ATB','181001ATB', '181002ATB', '181003ATB', '185501ATB', '185504ATB', '191701ATB', '206301ATB', '206302ATB', '206303ATB', '211701ATB', '221603ATB', '229705ACR', '229705ATB', '229705ATR', '229707ATR', '241801ATB', '241803ATB', '250601ATB', '301600ATB', '427800ACH', '480401ACH', '480402ACH', '488501ATB')
YAK_HYPN *	**Sleeping pills (Hypnotics)** ('105502ATB', '105504ATB', '105505ATB', '118501ATB', '131201ATB', '131202ATB', '137302ATB', '156201ATB', '156202ATB', '156501ATB', '156502ATB', '156503ATB', '161801ATB', '194201ATB', '243501ATB', '243502ATB', '250501ATB', '255800ATB')
YAK_SED *	**Sedative** ('113501ATB', '113504ATB', '138701ATB', '138702ACR', '192001ATB', '192003ATB', '192004ATB', '205203ATR', '205303ATB', '240701ATB')
Mental illness screening and treatment (n = 20) (each code with (*) asterisk has 5 temporal dummy variables)	
MH_DZ_SCREEN *	증상 및 행동 평가 척도 Symptomatic and Behavioral Evaluation Scale
PSY_TRM *	개인정신치료 (지지요법, 심층분석요법, 가족치료, 약물이용면담 등)
DMT_EXAM *	치매 척도 검사 (GDS, CDR)
DMT_SCREEN *	치매관련 척도 및 선별검사 (7-minute Screen(7-MS), Dementia Activity of Daily Living)
Cancer-related procedures (n = 25) (each code with (*) asterisk has 5 temporal dummy variables)	
CTX *	대장암 항암 치료 여부 (oxaliplatin, levoleucovorin, leucovorin,5-fluorouracil, irinotecan, bevacizumab, aflibercept, cetuximab)
RADIO *	체외조사 기본방사선치료
SURGERY *	수술 (결장경하 종양 수술, 결장및직장전절제술, 결장절제술, 직장및에스장절제술, 직장종양절제술)
Stomy *	수술 (장루조성술)
HEPA_META *	간전이 처치 (간절제, 고주파열치료)
Inpatient treatment-related variables (n = 40) (each code with (*) asterisk has 5 temporal dummy variables)	
EMG *	신경전도검사(H-Reflex, Bulbocavernous Reflex Test)
ENEMA *	Enema (Finger Enema, Cauterization of Umbilical Granuloma, Rectal Tube Insertion)
FOLEY *	Foley Catheterization
NELATON *	Nelaton Catheterization
REC_PRC *	Rectal Massage
STM_PRC *	**Post-colostomy care** ('M0131', 'B07030')
T_FEED *	Tubal nutrition
TPN *	Total Parenteral Nutrition (중심정맥영양법)
Psychiatric hospitalization records (n = 3)	
INPAT_CNT	Number of psychiatric inpatient visit
OUTPAT_CNT	Number of psychiatric outpatient visit
EMT_CNT	Number of psychiatric emergency visit

Appendix 2. Composition of dummy variables for sensitivity analysis (n=149)

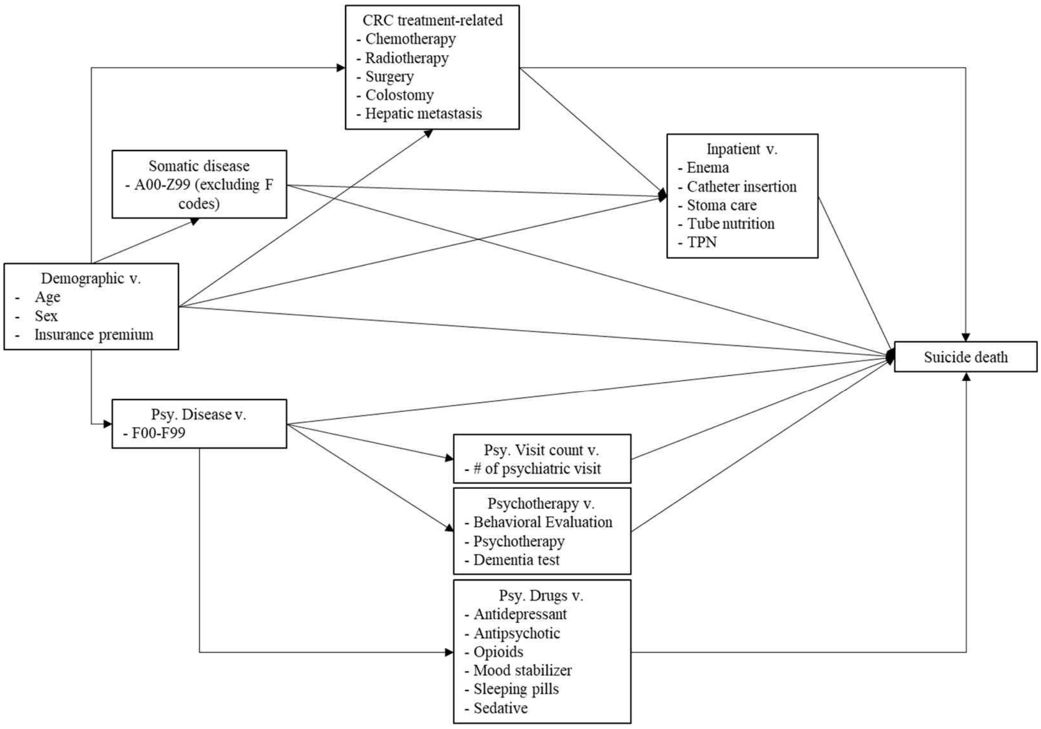
Code	Variable Label	Before diagnosis of colorectal cancer (underlying disease)	After colon cancer diagnosis (complications)
Psychiatric disorders (F01-F99) n = 144			
F01_CD	Vascular dementia	F01_CD_PRE	F01_CD_POST
F02_CD	Dementia in other diseases classified elsewhere	F02_CD_PRE	F02_CD_POST
F03_CD	Unspecified dementia	F03_CD_PRE	F03_CD_POST
F04_CD	Amnestic disorder due to known physiological condition	F04_CD_PRE	F04_CD_POST
F05_CD	Delirium due to known physiological condition	F05_CD_PRE	F05_CD_POST
F06_CD	Other mental disorders due to known physiological condition	F06_CD_PRE	F06_CD_POST
F07_CD	Personality and behavioral disorders due to known physiological condition	F07_CD_PRE	F07_CD_POST
F09_CD	Unspecified mental disorder due to known physiological condition	F09_CD_PRE	F09_CD_POST
F10_CD	Alcohol related disorders	F10_CD_PRE	F10_CD_POST
F11_CD	Opioid related disorders	F11_CD_PRE	F11_CD_POST
F12_CD	Cannabis related disorders	F12_CD_PRE	F12_CD_POST
F13_CD	Sedative, hypnotic, or anxiolytic related disorders	F13_CD_PRE	F13_CD_POST
F14_CD	Cocaine related disorders	F14_CD_PRE	F14_CD_POST
F15_CD	Other stimulant related disorders	F15_CD_PRE	F15_CD_POST
F16_CD	Hallucinogen related disorders	F16_CD_PRE	F16_CD_POST
F17_CD	Nicotine dependence	F17_CD_PRE	F17_CD_POST
F18_CD	Inhalant related disorders	F18_CD_PRE	F18_CD_POST
F19_CD	Other psychoactive substance related disorders	F19_CD_PRE	F19_CD_POST
F20_CD	Schizophrenia	F20_CD_PRE	F20_CD_POST
F21_CD	Schizotypal disorder	F21_CD_PRE	F21_CD_POST
F22_CD	Delusional disorders	F22_CD_PRE	F22_CD_POST
F23_CD	Brief psychotic disorder	F23_CD_PRE	F23_CD_POST
F24_CD	Shared psychotic disorder	F24_CD_PRE	F24_CD_POST
F25_CD	Schizoaffective disorders	F25_CD_PRE	F25_CD_POST
F28_CD	Other psychotic disorder not due to a substance or known physiological	F28_CD_PRE	F28_CD_POST
F29_CD	Unspecified psychosis not due to a substance or known physiological	F29_CD_PRE	F29_CD_POST
F30_CD	Manic episode	F30_CD_PRE	F30_CD_POST
F31_CD	Bipolar disorder	F31_CD_PRE	F31_CD_POST
F32_CD	Major depressive disorder, single episode	F32_CD_PRE	F32_CD_POST
F33_CD	Major depressive disorder, recurrent	F33_CD_PRE	F33_CD_POST
F34_CD	Persistent mood [affective] disorders	F34_CD_PRE	F34_CD_POST
F39_CD	Unspecified mood [affective] disorder	F39_CD_PRE	F39_CD_POST
F40_CD	Phobic anxiety disorders	F40_CD_PRE	F40_CD_POST
F41_CD	Other anxiety disorders	F41_CD_PRE	F41_CD_POST
F42_CD	Obsessive-compulsive disorder	F42_CD_PRE	F42_CD_POST
F43_CD	Reaction to severe stress, and adjustment disorders	F43_CD_PRE	F43_CD_POST
F44_CD	Dissociative and conversion disorders	F44_CD_PRE	F44_CD_POST
F45_CD	Somatoform disorders	F45_CD_PRE	F45_CD_POST
F48_CD	Other nonpsychotic mental disorders	F48_CD_PRE	F48_CD_POST
F50_CD	Eating disorders	F50_CD_PRE	F50_CD_POST
F51_CD	Sleep disorders not due to a substance or known physiological condition	F51_CD_PRE	F51_CD_POST
F52_CD	Sexual dysfunction not due to a substance or known physiological condition	F52_CD_PRE	F52_CD_POST
F53_CD	Puerperal psychosis	F53_CD_PRE	F53_CD_POST
F54_CD	Psychological and behavioral factors associated with disorders or diseases	F54_CD_PRE	F54_CD_POST
F55_CD	Abuse of non-psychoactive substances	F55_CD_PRE	F55_CD_POST
F59_CD	Unspecified behavioral syndromes associated with physiological	F59_CD_PRE	F59_CD_POST
F60_CD	Specific personality disorders	F60_CD_PRE	F60_CD_POST
F63_CD	Impulse disorders	F63_CD_PRE	F63_CD_POST
F64_CD	Gender identity disorders	F64_CD_PRE	F64_CD_POST
F65_CD	Paraphilias	F65_CD_PRE	F65_CD_POST
F66_CD	Other sexual disorders	F66_CD_PRE	F66_CD_POST
F68_CD	Other disorders of adult personality and behavior	F68_CD_PRE	F68_CD_POST
F69_CD	Unspecified disorder of adult personality and behavior	F69_CD_PRE	F69_CD_POST
F70_CD	Mild intellectual disabilities	F70_CD_PRE	F70_CD_POST
F71_CD	Moderate intellectual disabilities	F71_CD_PRE	F71_CD_POST

F72_CD	Severe intellectual disabilities	F72_CD_PRE	F72_CD_POST
F73_CD	Profound intellectual disabilities	F73_CD_PRE	F73_CD_POST
F78_CD	Other intellectual disabilities	F78_CD_PRE	F78_CD_POST
F79_CD	Unspecified intellectual disabilities	F79_CD_PRE	F79_CD_POST
F80_CD	Specific developmental disorders of speech and language	F80_CD_PRE	F80_CD_POST
F81_CD	Specific developmental disorders of scholastic skills	F81_CD_PRE	F81_CD_POST
F82_CD	Specific developmental disorder of motor function	F82_CD_PRE	F82_CD_POST
F84_CD	Pervasive developmental disorders	F84_CD_PRE	F84_CD_POST
F88_CD	Other disorders of psychological development	F88_CD_PRE	F88_CD_POST
F89_CD	Unspecified disorder of psychological development	F89_CD_PRE	F89_CD_POST
F90_CD	Attention-deficit hyperactivity disorders	F90_CD_PRE	F90_CD_POST
F91_CD	Conduct disorders	F91_CD_PRE	F91_CD_POST
F93_CD	Emotional disorders with onset specific to childhood	F93_CD_PRE	F93_CD_POST
F94_CD	Disorders of social functioning with onset specific to childhood	F94_CD_PRE	F94_CD_POST
F95_CD	Tic disorder	F95_CD_PRE	F95_CD_POST
F98_CD	behavioral and emotional disorders with onset usually occurring in childhood	F98_CD_PRE	F98_CD_POST
F99_CD	Mental disorder, not otherwise specified	F99_CD_PRE	F99_CD_POST



Abbreviation: MDD, Major depressive disorder

Appendix 3. Variable importance for random forest in colorectal cancer patients with variable set of psychiatric diagnosis (n=3,678)



Appendix 4. Directed acyclic graph for the predictive model of the study

국문요약

기계 학습 알고리즘을 이용한 대장직장암 환자의 자살 위험 예측 모델: 국민건강보험 2002-2018년 대장암 맞춤형 자료 분석

<지도교수 정선재>

연세대학교 대학원 의학과

이 영 롱

서론: 기계 학습을 이용한 자살 예측 모델에 대한 선행 연구들은 일관되게 일반 인구에서 높은 예측 성능을 보여주고 있으며, 기계 학습 자살 예측 연구를 대장암과 같은 고위험 인구 집단에 적용할 필요성에 대해서 제안하고 있다. 이 연구는 기계 학습을 사용하여 2002년부터 2018년까지 대장직장암 진단을 받은 환자의 맞춤형 청구 자료를 이용하여 자살에 대한 연령, 성별 및 암 유형별 위험요인 프로파일과 학습된 모델의 예측 성능을 확인하였다.

연구방법: 2002년부터 2018년 사이에 대장직장암을(C18-20) 진단받은 환자(n=380,569) 중, 자살로 사망한 환자를 사례군에 포함하였다. 자살 사망자 수는 1,839명(0.48%)이었으며, 사례 불균형 문제를 해결하기 위해 대조군을 사례군(총 n=3,678명)과 같은 수의 표본으로

과소추출(undersampling)하였다. 연령, 성별, 암 유형별로 계층화된 각 모델의 성능 및 위험 프로파일을 확인하였다. 각 모델은 인구통계학적 요인, 신체 및 정신질환의 검사 및 치료 관련 청구 요인, 암 병기, 대장암 관련 수술, 처방약, 외래, 응급실, 입원 횟수 등의 1,600개 이상의 예측 변수를 사용하여 훈련되었다. 기계 학습 모델 개발은 분류 트리와 랜덤 포레스트로 수행하였다. 모델에서 발견된 중요예측요인은 nested case control 연구 설계에서 조건부 로지스틱 회귀를 통해 평가되었다.

연구결과: 모든 연령과 성별, 암 종류로 나눈 집단 모두에서 정신치료 처방, 수면제 및 기분 안정제를 포함한 정신과 약물, 정신과 외래 방문 횟수가 자살 시도의 중요한 예측 인자였다. 대장암 특이적인 자살 위험 요인으로는, 최근 대장암 진단 시점과 관장, 도뇨관삽관, 장관 영양등의 입원 관련 처방 변수들이 있었다. 자살위험요인 프로파일은 연령, 성별, 암 유형에 따라 차이를 보였다. 대장직장암 환자에 대한 랜덤 포레스트 모델의 민감도는 0.84(84%), 특이도는 0.68(68%), 수용체 작동 곡선 아래 면적(AUC)은 0.84였습니다. 연령, 성별, 대장암 유형으로 나눈 그룹에 대한 모델의 AUC는 대부분 0.8에 근접한 값으로 산출되었다. 예측 위험도의 상위 1%, 5%, 10% 및 20%에 속하는 대장직장암 환자는 모든 자살 사망 사례의 각각 9.37%, 36.6%, 53.38% 및 70.81%를

차지했다. Nested case control 연구의 결과, 발견된 예측 변수와 자살 간의 연관성은 기계 학습 모델에서 식별된 변수 중요도 결과와 일치했다.

결론: 본 연구는 기계학습 기법을 통해 대장암 환자의 자살 사망을 예측할 수 있는 위험인자를 조명하고, 비용 효과적인 자살예방 중재를 위한 단계별 과정에서 본 자살 예측 모델의 임상적 활용 가능성을 제시하였다.

핵심되는 말: 기계학습 알고리즘, 자살, 대장암, 정신종양학