Prediction models of suicide and non-fatal suicide attempt after discharge from a psychiatric inpatient stay: A machine learning approach on nationwide Danish registers

Nielsen, Sara Dorthea; Christensen, Rune H B; Madsen, Trine; Karstoft, Karen-Inge; Clemmensen, Line; Benros, Michael E

Published in:
Acta Psychiatrica Scandinavica

Link to article, DOI:
10.1111/acps.13629

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Received: 22 July 2023 | Revised: 6 October 2023 | Accepted: 13 October 2023

DOI: 10.1111/acps.13629

Acta Psychiatrica Scandinavica WILEY

ORIGINAL ARTICLE

Prediction models of suicide and non-fatal suicide attempt after discharge from a psychiatric inpatient stay: A machine learning approach on nationwide Danish registers

Sara Dorthea Nielsen[1,2] | **Rune H. B. Christensen**[2] | **Trine Madsen**[3,4] | **Karen-Inge Karstoft**[5] | **Line Clemmensen**[1] | **Michael E. Benros**[2,6]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

[2]Copenhagen Research Center for Biological and Precision Psychiatry, Mental Health Centre Copenhagen, Copenhagen University Hospital, Copenhagen, Denmark

[3]Danish Research Institute of Suicide Prevention, Mental Health Center Copenhagen, Copenhagen, Denmark

[4]Section of Epidemiology, University of Copenhagen, Faculty of Health and Medical Sciences, Copenhagen, Denmark

[5]Department of Psychology, Faculty of Social Sciences, University of Copenhagen, Copenhagen, Denmark

[6]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Correspondence
Line Clemmensen, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark.
Email: lkhc@dtu.dk

Michael Eriksen Benros, Copenhagen Research Center for Biological and Precision Psychiatry, Mental Health Centre Copenhagen, Copenhagen University Hospital, Copenhagen, Denmarks.
Email: michael.eriksen.benros@regionh.dk

Funding information
Danish Foundation TrygFonden, Grant/Award Numbers: 153635, 145720; Lundbeckfonden, Grant/Award Number: R278-2018-1411

Abstract

Introduction: To develop machine learning models capable of predicting suicide and non-fatal suicide attempt as separate outcomes in the first 30 days after discharge from a psychiatric inpatient stay.

Methods: Prospective cohort study using nationwide Danish registry data. We included individuals who were 18 years or older, and all discharges from psychiatric hospitalizations in Denmark from 1995 to 2018. We trained predictive models using 10-fold cross validation on 80% of the data and did testing on the remaining 20%.

Results: The best model for predicting non-fatal suicide attempt was an ensemble of predictions from gradient boosting (XGBoost) and categorical boosting (catBoost). The ROC-AUC for predicting suicide attempt was 0.85 (95% CI: 0.84–0.85). At a risk threshold of 4.36%, positive predictive value (PPV) was 11.0% and sensitivity was 47.2%. The best model for predicting suicide was an ensemble of predictions from random forest, XGBoost and catBoost. For suicide, the ROC-AUC was 0.71 (95% CI: 0.70–0.73). At a risk threshold of 0.15%, PPV was 0.34% and sensitivity was 56.0%. The most contributing predictors differed when predicting suicide and suicide attempt, indicating that separate models are needed. The ensemble model was fair across sex and age, and more so than the penalized logistic regression model.

Conclusions: We achieved good performance for predicting suicide attempts and demonstrated a clinical application of ensemble models. Our results indicate a difference in predictive performance for models predicting suicide and suicide attempt, respectively. Thus, we recommend that suicide and suicide attempt are treated as two separate endpoints, in particular for clinical application. We demonstrated that the ensemble model is fairer across sex and age compared with a penalized logistic regression, and therefore we recommend the use of well-tested ensembles despite a more complex explainability.

KEYWORDS
machine learning, prediction, psychiatry, suicide, suicide attempt

Line Clemmensen and Michael E. Benros share the last authorship.

1 | INTRODUCTION

Suicide is a major public health problem with 800,000 deaths annually in the world.[1] Suicides are often preventable deaths, and identification of those at risk are of utmost importance. However, while several risk factors for suicidal behavior have been identified,[2] predictive models with sufficient accuracy and clinical applicability have not been successfully put forward.[3–17] Prediction of suicide is challenged by suicide being a rare outcome even in psychiatric populations. Suicide is difficult to predict for both clinicians and machine learning (ML) models, where the positive predictive value (PPV) is often low (<1%), even if a model has high classification accuracy due to the low prevalence (the population rate of suicides). Therefore, focusing on high suicide risk populations is important for accurate prediction of suicide. The highest suicide rates are immediately after discharge from a psychiatric admission. Thus, there is a large potential for interventions to prevent suicidal behavior if the ones at highest risk are identified at discharge.[18] Studies have indicated high usability of ML for rare events like suicidal behavior.[19,20]

Prior studies predicting suicide and suicide attempt are limited in several ways: Some studies use a case–control sampling design[5,9,15] (leading to overrepresentation of suicides where the predicted risk will not reflect the population risk),[21,22] some apply retrospective designs looking backwards in time from the point of suicide[9,15] (resulting in studies that cannot be used for prediction at time of discharge), some use small and biased populations,[5,10,14] and some focus on suicides and suicide attempt combined to increase the number of cases,[6,8] and will thus be skewed toward predicting non-fatal suicidal behavior.[23] Moreover, prediction of suicide following psychiatric hospitalization in the overall psychiatric inpatient population has been relatively under-studied compared with predictions in other settings. A prior study, on a sample of US Army members and Veterans, applied data available at the day of discharge to prospectively predict suicide risk in the period after a psychiatric admission[10,14]; however, due to this highly selective sample, these findings may not be generalizable to the broader population of people who experience a psychiatric hospitalization. Some prior studies have utilized the Danish register data to predict both suicide[9,15] and suicide attempt,[24,25] respectively, after psychiatric hospitalization discharge. They show important findings in the difference in predictors among men and women; however, due to their case-cohort study design they have an overrepresentation of suicides, and the predicted risk will thus not reflect the population risk, and moreover also include predictors from after the point of discharge and therefore cannot be used for prediction at time of discharge.

Significant outcomes

- We achieved good performance when predicting suicide attempt and acceptable performance when predicting suicide within the first 30 days after discharge from a psychiatric inpatient stay.
- Predictors for suicide and suicide attempt differed.
- We demonstrated clinical applicability due to sufficient accuracy, robustness in terms of validation on unseen data, representative data, and prospective approach.
- Sensitivity analysis on sex and age showed that the ensemble model had better equalized odds compared to a penalized logistic regression

Limitations

- The nationwide registers do not contain detailed information on specific symptoms and behavior, which could have further improved the model.
- The risk of recurrence was not considered in this study as it was restricted to incident episodes of suicide attempt.

In this study, we used nationwide Danish registry data to develop predictive models of all recorded suicides and non-fatal suicide attempts, respectively, following discharge from a psychiatric hospital. We investigate differences in predictors between suicide and non-fatal suicide attempts, and we report performance measures on an unseen test set. Afterwards, we conduct a sensitivity analysis on sex and age since previous studies have found a difference in predictors.[9,24] Moreover, we pave the way for a clinical decision support tool that can flag subjects at high risk of suicide or suicide attempts in the first 30 days after discharge from a psychiatric hospital.

2 | MATERIALS AND METHODS

2.1 | Data sources

All registered residents in Denmark are assigned a unique personal identity number, enabling accurate linkage between Danish national registers. All psychiatric hospital contacts and diagnoses are obtained in the

Psychiatric Central Research Register.[26] Diagnoses are defined according to the 10th version of the International Classification of Diseases (ICD-10) used in Denmark since January 1, 1994. Furthermore, we obtained information from The Medical Birth Register,[27] The Danish Civil Registration System Register,[28] The Danish National Patient Register,[29] The Danish Work Classification Register,[30] The Danish Income Register,[31] The Danish Highest Completed Education Register.[32]

2.2 | Study population

The study population consists of individuals with a psychiatric inpatient stay from January 1, 1995 to December 31, 2018, age 18 and above. Inpatient stays with less than 1 day between discharge from the first admission to the start date of the next were combined as it most likely was a transfer between inpatient wards. See Figure S1 for a flowchart.

2.3 | Outcome

1. *Suicides* were identified in the Cause of Death Register with an ICD-10 diagnosis code of intentional self-harm (X60-X80) or sequelae of intentional self-harm (Y87.0).
2. *Suicide attempts* were identified as a non-fatal hospital contact, both psychiatric and somatic, with an ICD-10 diagnosis code (X60-84) or where the "reason for contact" was "suicide attempt."

This study population was chosen to reflect the actual population to which a potential screening procedure based on the resulting algorithm would be applied. Following the same argument, we chose a prediction time window of 30 days as we intent to apply the algorithm to detect suicide attempts and suicides in the high-risk period immediately after discharge. While increasing the risk window may improve model performance slightly by increasing the number of cases, it would limit the clinical applicability of the model.

2.4 | Predictors

We selected clinical and sociodemographic predictors based on the findings in previous studies and the availability in the Danish registers. Predictors at time of admission included demographic characteristics (sex, age), socio-economic (income, highest completed education, marital status, occupation status), and medical records (acute/not acute admission). Predictors at discharge included month of discharge and specific psychiatric and somatic diagnosis given. Furthermore, psychiatric and somatic history of disease for both patient and parents were included (see ICD-10 codes in Table S1). Patient history of disease was represented by binary indicators for 7 time periods up to 5 years before the inpatient stay. Methods used in prior suicidal attempt were categorized using the first three digits in the ICD codes (ICD-10: X60-X84; see Table S2). A full list of the predictors included in the model derivation can be found in the Table S3. For missing data in categorial variables, a separate category "unknown" was created. We had 472 variables.

2.5 | Model development and model selection

Data were split in 80% training data consisting of 731,323 visits by 206,484 patients and 20% hold-out test data consisting of 180,795 visits by 51,622 patients. This split was stratified by the classes in the outcome variables, and individual patients only appear in either the training or the test data.

Previous studies predicting suicide and/or suicide attempt[5–17,33] have shown that elastic net penalized logistic regression,[34] random forest,[35] and gradient boosting (XGBoost)[36] have good performance. Boosting has shown good results for imbalanced and categorical data,[37] therefore, we also included a categorical boosting (catBoost) model.[38] To address the high imbalance in the data, a hyperparameter assigning different weights to the positive and negative class was tuned.[37] The optimal hyperparameters for each model were found using 10-fold cross validation (CV) and using area under the receiver operating characteristic curve (ROC-AUC) as evaluation metric. Using these hyperparameters, an ensemble of two or more models were chosen based on the argmax of the sums of the predicted probabilities, again using the same 10 CV folds and ROC-AUC as evaluation metric. The area under the precision recall curve (PR-AUC) was also reported and compared with the fraction of positives as a baseline. Significant differences between models were tested using McNemar's test at a significance level of 5%, including Bonferroni correction for multiple testing.[39]

After selecting the best model, sensitivity, specificity, PPV, and negative predictive value (NPV) were reported for a series of risk thresholds. To select an optimal threshold, the selected model was refitted to the full training

data and applied to the test data. The test scores for the selected threshold then provide an unbiased evaluation of the final model enabling unbiased selection of thresholds and expected performance measures. See Supporting Information on threshold setting for a more detailed explanation. Figure S2 shows a flow chart of the full modeling procedure.

To produce an illustrative risk chart displaying the differences for the most important variables for the models of suicide and suicide risk, respectively, a simple logistic regression with only the three most important variables for predicting suicide attempt and five variables for predicting suicide was modeled.[40,41] Three and five variables were selected, since they achieved acceptable prediction power. The models were trained and tested on the same 80%/20% split, mimicking the procedure for the ML models.

A sensitivity analysis on sex and age was performed on the test set both for suicide attempt and suicide for the final model. This was compared to a penalized logistic regression. Predictions before and after 2015 were also compared to assess whether the algorithm trained across 23 years of information perform as well in recent years as it does averagely across the whole interval. We use Equalized odds[42] as a definition of fairness, as there is a strong emphasis on predicting the positive outcome correctly and on minimizing false positives. To obtain Equal odds, we want to minimize both the difference in sensitivity and specificity between the classes.

2.6 | Variable importance

Before assessing variable importance, variables were grouped into 23 categories such as suicide history, hospitalization history, and diagnosis groups (F0, F1, ..., F9). We evaluated the importance by calculating the decrease in ROC-AUC when each of the 23 categories of variables (see Table S3) were removed from the candidate variable set. This was done both for the four considered models separately and for the best ensemble model. We also included the 30 most important single variables for the different models. For logistic regression, we reported the coefficients. For the random forest, XGBoost and catBoost, we reported the Gini importance, the average gain across all splits where the feature was used.

2.7 | Learning curve analysis and calibration

A learning curve analysis was performed by assessing a potential increase in ROC-AUC with increasing sample size to assess whether the study could benefit from a larger sample size. A nonparametric approach based on isotonic regression was used to calibrate the model. Brier scores[43] and calibration plots were used to assess the model calibration on the test set.[44]

R software 4.1.3 (https://www.r-project.org/) were used for constructing the datasets and descriptive analyses. Scikit-learn (https://scikit-learn.org) and XGBoost

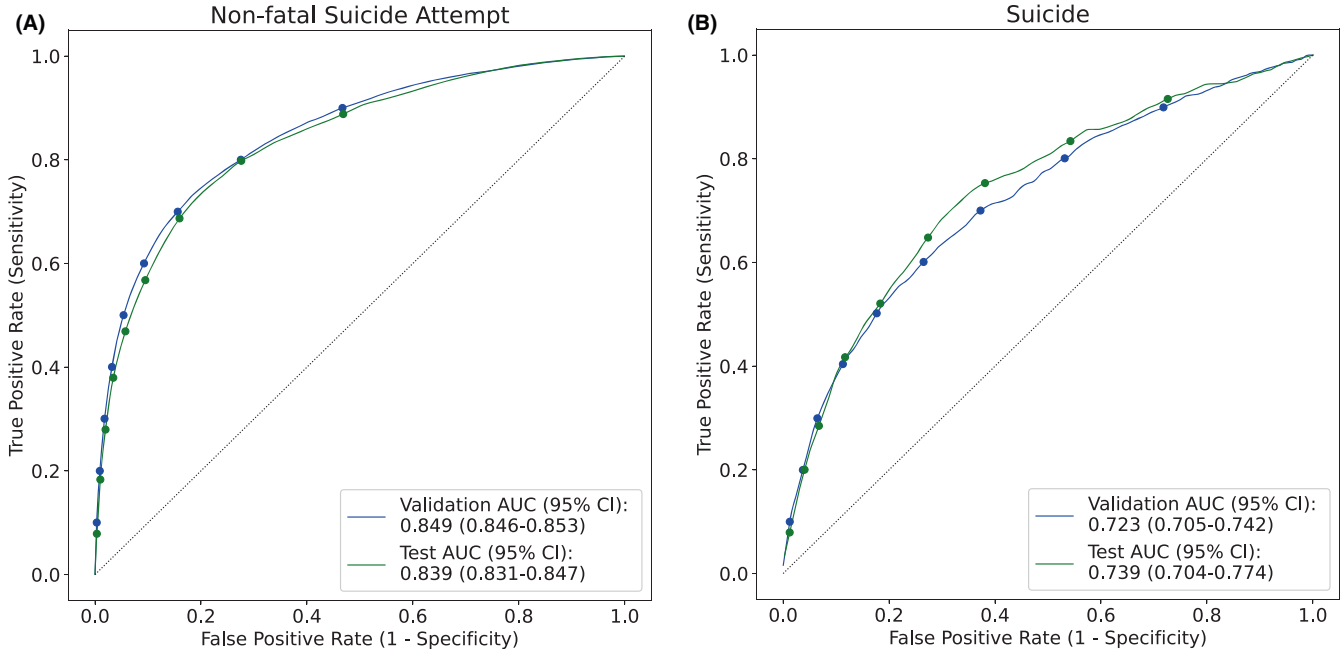

FIGURE 1 Validation (blue) and test (green) ROC-AUC curves for suicide attempt (A) and suicide (B).

(https://xgboost.readthedocs.io/en/latest/python/) packages for Python programming language 3.7.4 (https://www.python.org/) were used for the ML analyses during model derivation and evaluation. The default method, Gini, in scikit-learn is used for computing variable importance.

3 | RESULTS

During the study period from 1995 to 2018, a total of 258,106 individuals had admissions to psychiatric hospitals with a total of 912,118 admissions. Among these, 1053 (0.12%) died by suicide and 6371 (0.70%) had a suicide attempt within 30 days after discharge (see Table S4 and Figure S3 for characteristics).

3.1 | Model selection

The hyperparameters are listed in Table S5, and the mean training and validation ROC-AUCs for the best models and for the different combinations of the ensemble of models are listed in Table S6. The ensemble of the random forest, XGBoost, and catBoost yielded the highest validation ROC-AUC when predicting suicide and was thus selected. For suicide attempts, the ensemble consisting of XGBoost and catBoost yielded the highest validation ROC-AUC and was thus selected. The ROC-AUC of the best and the individual models are depicted in Figure S4. The ROC-AUC curves are depicted in Figure 1. The individual models in the ensemble have similar variable importance and can be seen in Figure S5. The top 30 individual variable importance for the individual models can be found in Figure S6.

3.2 | Model evaluation

3.2.1 | Prediction of non-fatal suicide attempts

The test ROC-AUC for suicide attempts was 0.85 (95% CI: 0.84–0.85). The corresponding training and test

TABLE 1 Performance measures for different thresholds for suicide and suicide attempt on the test set. The performance measures on the validation sets can be found in Table S9.

Non-fatal suicide attempt									
Sensitivity	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Threshold**	**TP**	**TN**	**FP**	**FN**
10%	0.0789	0.9966	0.2625	0.9860	0.2473	216	177,449	607	2523
20%	0.1855	0.9901	0.2240	0.9875	0.1557	508	176,296	1760	2231
30%	0.2811	0.9803	0.1803	0.9888	0.1035	770	174,555	3501	1969
40%	0.3804	0.9658	0.1459	0.9902	0.0686	1042	171,958	6098	1697
50%	0.4681	0.9429	0.1121	0.9914	0.0451	1282	167,897	10,159	1457
60%	0.5688	0.9050	0.0843	0.9927	0.0270	1558	161,136	16,920	1181
70%	0.6875	0.8417	0.0626	0.9943	0.0162	1883	149,869	28,187	856
80%	0.7988	0.7243	0.0427	0.9957	0.0096	2188	128,969	49,087	551
90%	0.8890	0.5299	0.0283	0.9968	0.0062	2435	94,357	83,699	304
Suicide									
Sensitivity	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Threshold**	**TP**	**TN**	**FP**	**FN**
10%	0.0741	0.9870	0.0068	0.9989	0.0048	16	178,228	2351	200
20%	0.1991	0.9604	0.0060	0.9990	0.0032	43	173,422	7157	173
30%	0.2917	0.9312	0.0050	0.9991	0.0024	63	168,163	12,416	153
40%	0.4167	0.8826	0.0042	0.9992	0.0019	90	159,371	21,208	126
50%	0.5231	0.8167	0.0034	0.9993	0.0015	113	147,474	33,105	103
60%	0.6528	0.7263	0.0028	0.9994	0.0012	141	131,147	49,432	75
70%	0.7500	0.6180	0.0023	0.9995	0.0010	162	111,594	68,985	54
80%	0.8333	0.4579	0.0018	0.9996	0.0008	180	82,682	97,897	36
90%	0.9167	0.2746	0.0015	0.9996	0.0007	198	49,583	130,996	18

curves are depicted in Figure 1A. PPV was 11.0%, NPV was 99.1%, sensitivity was 47.2%, and specificity was 94.1% at a risk threshold at 4.51%. This means that 47.2% of the subsequent suicide attempts 30 days after discharge will successfully be predicted by the model and that 11.0% of the ones identified to be screened by the model would otherwise have conducted a suicide attempt. Additionally, sensitivity, specificity, PPV, and NPV are reported for sensitivity levels at 10%, 20%, ..., 90% (see Table 1). The PR-AUC was 0.128 (95% CI: 0.127–0.130), which is more than 18 times higher than the baseline (fraction of positives) at 0.0070.

Regarding variable importance, a history of previous suicide attempt shows the highest importance for suicide attempt, which is 10 times more important than the second most important variable (Figure 2A). Age, history of previous hospitalizations, stress-related disorders (ICD-10: F4), and mood disorders (ICD-10: F3) are also important variables.

3.2.2 | Prediction of suicide

The test ROC-AUC for suicide was 0.71 (95% CI: 0.70–0.73). The corresponding training and test curves

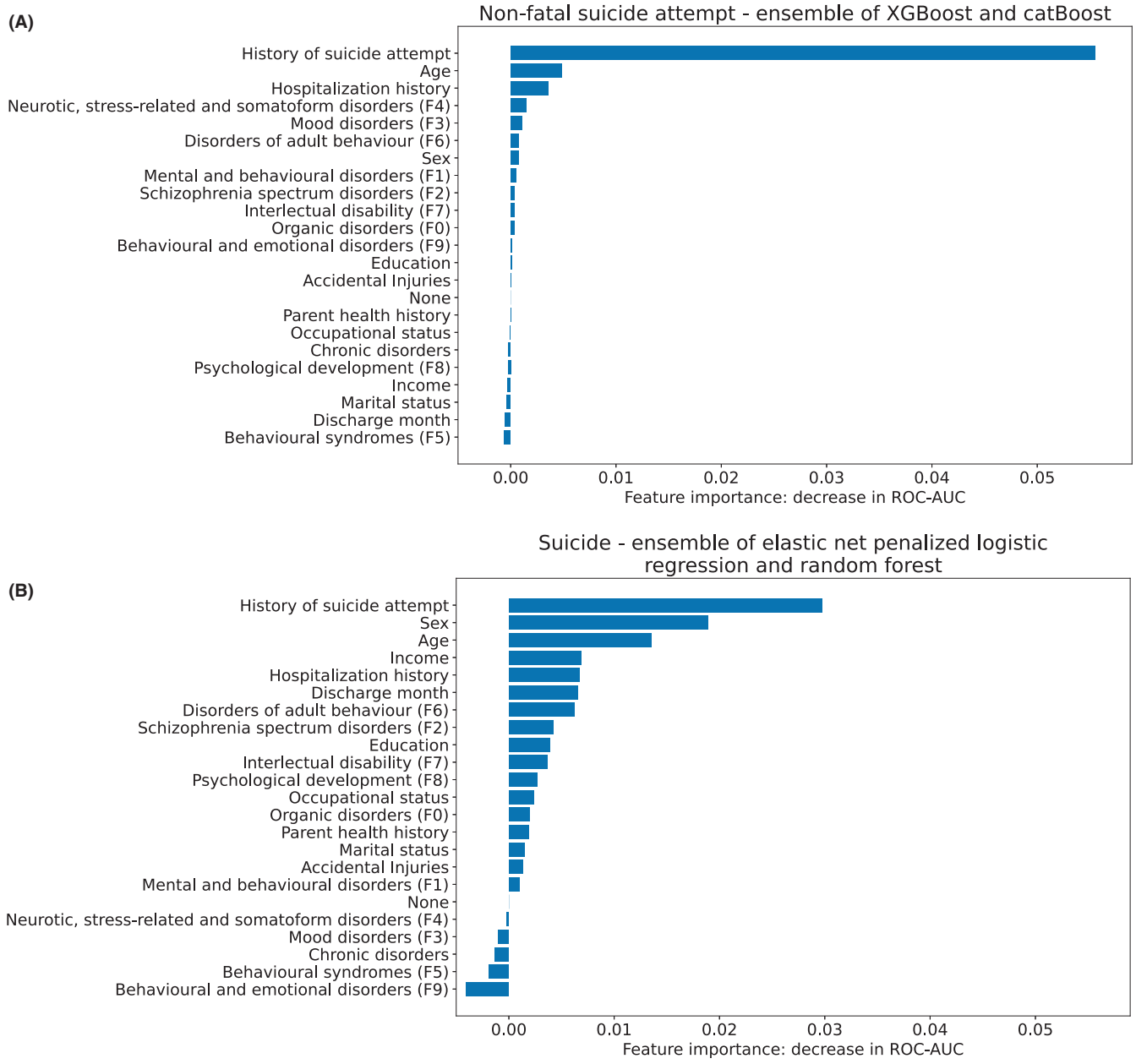

FIGURE 2 Variable importance for prediction of suicide attempts (A) and suicide (B).

TABLE 2 Sensitivity analysis on sex and age.

(A)									
Non-fatal suicide attempt									
Model	**Sex**	**Count**	**sa**	**TP**	**TN**	**FP**	**FN**	**Sensitivity**	**Specificity**
EN	All	180,795	2739	1286	167,680	10,376	1453	0.4695	0.9417
	Male	88,176	836	193	849,663	2377	643	0.2309	0.9972
	Female	92,619	1903	1093	82,717	7999	810	0.5744	0.9118
	Difference							0.3435	0.0854
Ensemble	All	180,795	2739	1282	167,897	10,159	1457	0.4681	0.9429
	Male	88,176	836	401	83,757	3583	435	0.4800	0.9590
	Female	92,619	1903	880	84,940	5776	1023	0.4623	0.9363
	Difference							0.0177	0.0226
Fatal suicide									
Model	**Sex**	**Count**	**sui**	**TP**	**TN**	**FP**	**FN**	**Sensitivity**	**Specificity**
EN	All	180,795	216	111	148,157	32,422	105	0.5139	0.8205
	Male	88,176	134	84	65,531	22,511	50	0.6269	0.7443
	Female	92,619	82	55	82,626	9911	27	0.6707	0.8929
	Difference							0.0439	0.1486
Ensemble	All	180,795	216	113	147,474	33,105	103	0.5231	0.8167
	Male	88,176	134	68	69,902	18,140	66	0.5082	0.7940
	Female	92,619	82	45	76,596	15,941	37	0.5475	0.8277
	Difference							0.0393	0.0338
(B)									
Non-fatal suicide attempt									
Model	**Age**	**Count**	**sa**	**TP**	**TN**	**FP**	**FN**	**Sensitivity**	**Specificity**
EN	All	180,795	2739	1282	167,897	10,159	1457	0.4681	0.9429
	18–30	38,266	1116	714	32,632	4518	402	0.6398	0.8784
	30–50	77,141	1165	504	70,972	5004	661	0.4326	0.9341
	>50	65,388	458	68	64,076	854	390	0.1485	0.9868
	Difference							0.4913	0.1085
Ensemble	All	180,795	2739	1282	167,897	10,159	1457	0.4681	0.9429
	18–30	38,266	1116	500	34,136	3014	616	0.4483	0.9189
	30–50	77,141	1165	545	71,641	4335	620	0.4681	0.9429
	>50	65,388	458	219	62,922	2008	239	0.4790	0.9691
	Difference							0.0306	0.0502
Suicide									
Model	**Age**	**Count**	**sui**	**TP**	**TN**	**FP**	**FN**	**Sensitivity**	**Specificity**
EN	All	180,795	216	113	147,474	33,105	103	0.5231	0.8167
	18–30	38,266	18	3	35,640	2608	15	0.1667	0.9318
	30–50	77,141	93	50	64,753	12,295	43	0.5376	0.8404
	>50	65,388	105	58	47,764	17,519	47	0.5524	0.7316
	Difference							0.3857	0.2002

(Continues)

TABLE 2 (Continued)

Suicide									
Model	**Age**	**Count**	**sui**	**TP**	**TN**	**FP**	**FN**	**Sensitivity**	**Specificity**
Ensemble	All	180,795	216	113	147,474	33,105	103	0.5231	0.8167
	18–30	38,266	18	10	32,489	5759	8	0.5787	0.8494
	30–50	77,141	93	49	62,923	14,125	44	0.5231	0.8167
	>50	65,388	105	55	50,504	14,779	50	0.5231	0.7736
	Difference							0.0556	0.0758

Note: Performance measures for the final ensemble model and a penalized logistic regression (EN), both for suicide attempt (sa) and suicide (sui), on different subgroups: (A) Sex, divided into male and female. (B) Age, divided into 18–30, 30–50, and >50. The measure for equalized odds is highlighted in gray in each category. We see that the differences are smaller for the ensemble model compared with the penalized logistic regression.

are depicted in Figure 1B. With a sensitivity threshold around 50%, the model resulted in a PPV of 0.34%, NPV of 99.9%, sensitivity at 52.3%, and specificity at 81.7% for a risk threshold of 0.15%. The PR-AUC was 0.0047 and thereby more than four times that of a baseline classifier (0.0012).

Regarding variable importance for suicide prediction, history of suicide attempt, sex, age, and hospitalization history are important variables (Figure 2B). Other important variables include income, discharge month, level of education, disorders of adult personality and behavior (ICD-10: F6), schizophrenia spectrum disorders (ICD-10: F2), and intellectual disabilities (ICD-10: F7). Sex is more important when predicting suicide compared to predicting suicide attempt. Opposite suicide attempt, mood disorders (ICD-10: F3), and neurotic, stress-related, and somatoform disorders (ICD-10: F4) are less important for suicide.

3.2.3 | Sensitivity analyses

Sensitivity analyses were conducted on sex, age, and time. We saw close to no difference between male and female (0.02 and 0.04), between different age groups (0.05 and 0.08), and between prior to and after 2015 (0.02 and 0.02) in equalized odds for the final ensemble models (see Table 2). The equalized odds for the elastic net penalized logistic regression was more than three times higher for both sex and age. Performance measures for the different subgroups are found in Table 2 and Table S7.

3.2.4 | Risk charts

To illustrate the difference in predictor contributions, a simple 3- or 5-level risk chart based on a logistic regression was performed. A variable overview can be found in Table S8. The 3-level risk chart for suicide attempts achieved a ROC-AUC of 0.83 (95% CI: 0.82–0.84) and the 5-level risk chart for suicides achieved a ROC-AUC 0.70 (95% CI: 0.69–0.71). There was a significant difference between these and the best ML models (p-value 0.006), and there are considerable differences between the risk chart for suicide compared to the risk chart for suicide attempt. The risk charts are depicted in Figure 3.

3.2.5 | Learning curve analysis

Figure S7 shows the validation and test ROC-AUCs for increasing training sample size. For suicide attempt, the curves clearly converge and end up close to each other (0.85–0.87), indicating low variance and no overfitting. The test curve for suicide converges while the training curve does not.

3.2.6 | Calibration

Brier scores[51] were estimated to be 0.0139 and 0.0012 (both close to 0) for the models predicting suicide attempt and suicide, respectively, indicating good model calibration. The calibration plots (Figure S8) further illustrate high agreement between the observed proportion of positives and mean predicted risk of the outcomes. More details can be found in the Supporting Information.

4 | DISCUSSION

To the best of our knowledge, this is the first study which uses ML to predict suicide and suicide attempt as two separate outcomes in the high-risk period prospectively

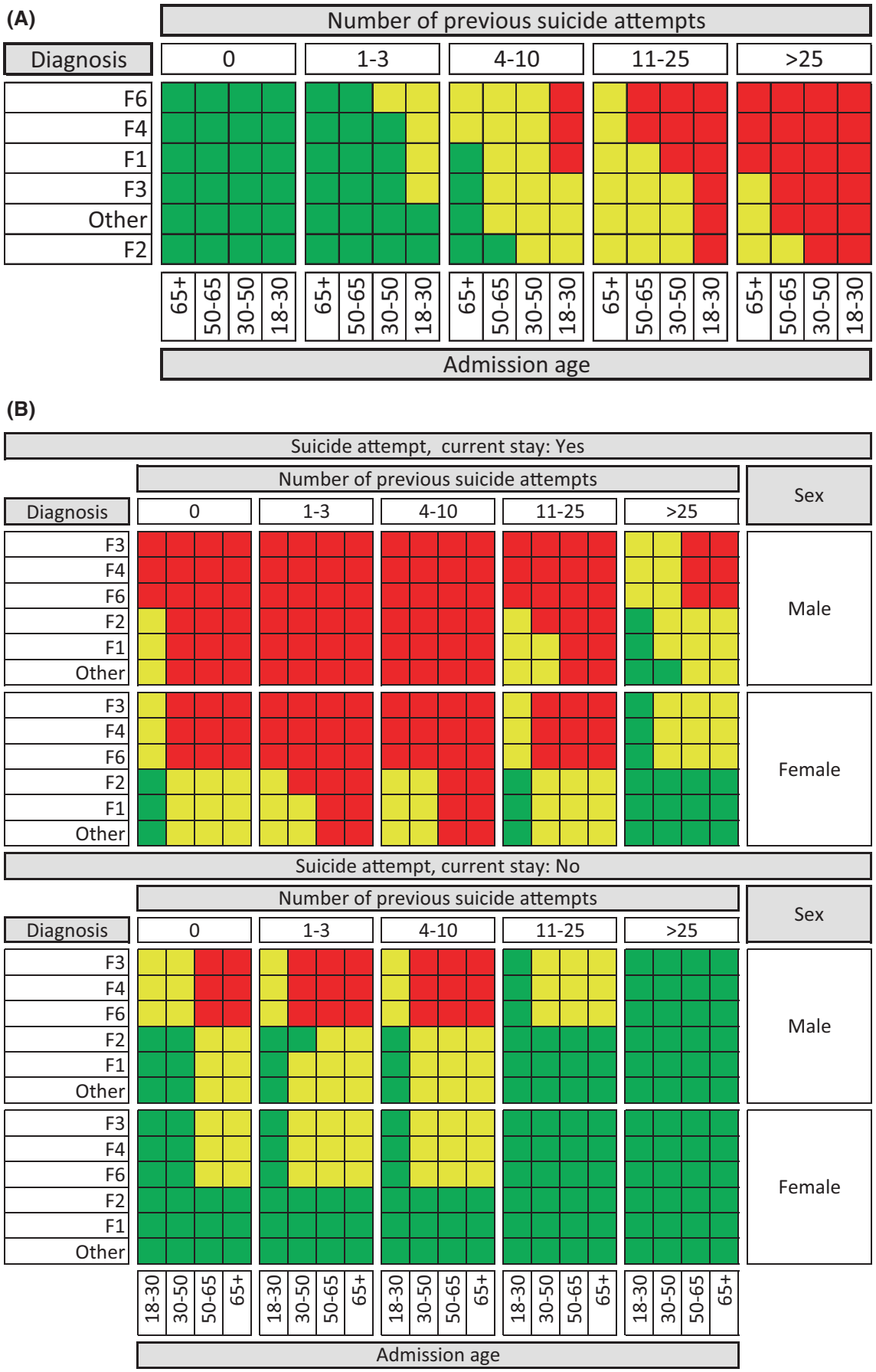

FIGURE 3 Legend on next page.

at discharge from a psychiatric inpatient stay. The test ROC-AUC for suicide attempt was 0.85 and 0.71 for suicide. At a sensitivity threshold around 50%, the model predicting suicide attempt had a test PPV of 11.0% and the model predicting suicide had a test PPV of 0.34%, NPV of 99.9%, sensitivity at 52.3%, and specificity at 81.7%. Choosing a higher threshold would yield a lower PPV, and choosing a lower threshold would yield a higher PPV. We, thus, have a tradeoff between sensitivity and PPV. The PR-AUC was more than 18 times the baseline (fraction of positives) for suicide attempt and more than four times the baseline for suicide. Thus, the model is better than a stratified random classifier.

Our study is the first to provide flexibility in providing risk threshold on a validation set and report scores on a separate test set. This is done by choosing one or more risk thresholds for the predicted probabilities based on the validation scores (sensitivity and PPV) and then reporting the test scores (sensitivity and PPV) for the given threshold(s). This increases the flexibility and practical relevance of the model and enables professionals to choose different models for implementation depending on a required cost/benefit evaluation.

We additionally constructed simpler risk charts in line with the Framingham risk chart for cardiovascular events.[45,46] These substantiate the different associations between predictors with suicide attempt and suicide. The risk of suicide attempt increases as the number of previous suicide attempts increases, whereas the risk of suicide first increases as the number of suicide attempts increases but decreases after 4–10 attempts. Likewise, we see that diagnosis categories influence differently for suicide attempt and suicide.

Previous studies have shown a difference in predictors and prediction performance between male and female patients when predicting both suicide and suicide attempt.[9,25] Our ensemble models perform similarly for both sexes, and for different age groups. Moreover, we also performed a sensitivity analysis on different time periods. Again, we saw no difference in predictive performance between inpatient stays before and after 2015, indicating that the ensample prediction models are robust over time in the prediction of suicide and suicide attempts. This was however not the case for the penalized logistic regression where we saw a difference in prediction performance both in sex and age. This illustrates that the ensemble is not just superior in predictive performance but also regarding fairness between sex and age (here measured by equalized odds). This demonstrates the practical usefulness of ML for rare events like suicidal behavior.[19,20]

A history of suicide attempt was the top predictor both when predicting suicide and when predicting suicide attempt. However, in general, suicide and suicide attempt only share five out of the top 10 predictors. Mood disorders (ICD-10: F3) and neurotic, stress-related, and somatoform disorders (ICD-10: F4) have high importance when predicting suicide attempt, but on the contrary, they have negative importance when predicting suicide. We also see this difference in the risk charts. For suicide attempt, the risk increases with an increasing number of previous suicide attempts, but for suicide, the risk peaks at 4–10 previous suicide attempts and hereafter decreases.

Sociodemographic factors appear less important in our study compared to some previous studies on US Army member and veterans.[10,47] However, studies indicate that sociodemographic factors on suicide tend to be overestimated when clinical variables are not taken into account.[48] Thus, a possible explanation for the difference is that the effects of the sociodemographic variables are mediated by the clinical variables.[5] Future studies of this possible interaction effect could provide valuable insights for preventive intervention design.

4.1 | Strengths and limitations

A major strength of the study is that the analysis is prospective only using data available at the time the model is intended to be used. Further, the model is trained and tested on a representative sample of the population for which it is intended. This enables early identification of those at risk of suicide or suicide attempt and can potentially be used at the time of discharge to initiate different proactive strategies, and thus hopefully prevent several suicides and suicide attempts. Moreover, this is the first nationwide register-based study that separates suicide and suicide attempt. It also offers a way to pick different

FIGURE 3 Risk charts made based on a simple logistic regression. Green-yellow-red indicates low-medium-high risk. (A) Variables for non-fatal suicide attempt include number of previous suicide attempts, main diagnosis given at current stay, and admission age. The thresholds for non-fatal suicide attempt are 0.033 and 0.1. (B) Variables for suicide attempt include number of previous suicide attempts, main diagnosis given at current stay, admission age, sex and suicide attempt connected to current stay. The thresholds for suicide are 0.003 and 0.001. F1: Mental and behavioral disorders, F2: Schizophrenia spectrum disorders, F3: Mood disorders, F4: Neurotic, stress-related and somatoform disorders, F6: Disorders of adult behavior, Other: Other psychiatric disorders.

sensitivity levels for different intervention purposes through a table of performance metrics for varying sensitivity thresholds selected on the training data and re-evaluated on the test data. The selection of different sensitivity thresholds is only possible since the thresholds were evaluated on the training data and not on the test data. Finally, we also show that the ensemble model is fairer compared to a penalized logistic regression.

Limitations include uncertainty regarding the validity of the outcome of interest. The analysis only includes suicide attempts registered in a hospital contact, thus omitting suicide attempts that did not result in a hospital contact. This means that the actual number of reported suicide attempts in the 30 days post-discharge period is probably higher than the number of suicide attempts recorded in the outcome measure. Furthermore, some events of self-injury or self-harm that did not involve suicidal intent might be included in the outcome. However, a validation study indicated that 73% of ICD-10 codes X60-X84 did indeed represent acts with suicidal intents.[49] Therefore, we do not believe that this impacts the findings greatly. Another limitation is that the test set is not independent in time. Moreover, the model for suicide could benefit from an even larger sample size, which is indicated by the training ROC-AUC that is still decreasing with larger sample size. On the contrary, the training ROC-AUC for suicide attempt does converge, indicating no improvements with a bigger sample size. As the validation ROC-AUC no longer increases with increased training sample size, future improvements to the model may require more informative predictors and higher model capacity rather than a larger sample size. Kessler et al[16] have shown that clinical notes improve the prediction of suicide after discharge for their US Army members and Veteran, which should be further explored by future research efforts testing this on the general psychiatric population to see if natural language processing on medical notes could also further improve prediction here.

To Conclude, we achieved good performance for predicting suicide attempts and demonstrated a clinical application of ensemble models. Our results indicate a difference in predictive performance and most contributing predictors for models predicting suicide and suicide attempt, respectively. Thus, we recommend that suicide and suicide attempt are treated as two separate endpoints, in particular for clinical application. We produced simple risk-charts but also demonstrated that the ensemble model is fairer across sex and age compared to a penalized logistic regression, and therefore we recommend the use of well-tested ensembles despite a more complex explainability.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/acps.13629.

DATA AVAILABILITY STATEMENT

The utilized individual level data from the Nationwide Danish registers are available to all researchers in Denmark after appropriate approvals from the Danish Data Protection Agency, Denmark's Statistics and The Danish Health Data Authorities. The code is available upon request.

ORCID

Sara Dorthea Nielsen https://orcid.org/0000-0001-9272-7466
Trine Madsen https://orcid.org/0000-0003-4918-5862
Line Clemmensen https://orcid.org/0000-0001-5527-5798
Michael E. Benros https://orcid.org/0000-0003-4939-9465

REFERENCES

1. World Health Organization. *Preventing Suicide: A Global Imperative*. World Health Organization; 2014.
2. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull*. 2017;143(2):187-232.
3. Belsher BE, Smolenski DJ, Pruitt LD, et al. Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA Psychiatry*. 2019;76(6):642-651.
4. Adamou M, Antoniou G, Greasidou E, et al. Toward automatic risk assessment to support suicide prevention. *Crisis*. 2018;40:249-256.
5. Kessler RC, Stein MB, Petukhova MV, et al. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Mol Psychiatry*. 2017;22(4):544-551.
6. Barak-Corren Y, Castro VM, Javitt S, et al. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry*. 2017;174(2):154-162.
7. Carson NJ, Mullin B, Sanchez MJ, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. *PloS One*. 2019;14(2):e0211116.
8. Chen Q, Zhang-James Y, Barnett EJ, et al. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: a machine learning study using Swedish national registry data. *PLoS Med*. 2020;17(11):e1003416.
9. Gradus JL, Rosellini AJ, Horváth-Puhó E, 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.

10. Kessler RC, Warner CH, Ivany C, 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.
11. Simon GE, Shortreed SM, Johnson E, et al. What health records data are required for accurate prediction of suicidal behavior? *J Am Med Inform Assoc*. 2019;26(12):1458-1465.
12. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *J Child Psychol Psychiatry*. 2018;59(12):1261-1270.
13. Ryu S, Lee H, Lee DK, Park K. Use of a machine learning algorithm to predict individuals with suicide ideation in the general population. *Psychiatry Investig*. 2018;15(11):1030-1036.
14. Kessler RC, Bauer MS, Bishop TM, et al. Using administrative data to predict suicide after psychiatric hospitalization in the veterans health administration system. *Front Psych*. 2020;6(11):390.
15. Jiang T, Rosellini AJ, Horváth-Puhó E, et al. Using machine learning to predict suicide in the 30 days after discharge from psychiatric hospital in Denmark. *Br J Psychiatry*. 2021;219(2):440-447.
16. Kessler RC, Bauer MS, Bishop TM, et al. Evaluation of a model to target high-risk psychiatric inpatients for an intensive post-discharge suicide prevention intervention. *JAMA Psychiatry*. 2023;80(3):230-240.
17. Sanderson M, Bulloch AG, Wang J, Williams KG, Williamson T, Patten SB. Predicting death by suicide following an emergency department visit for parasuicide with administrative health care system data and machine learning. *EClinicalMedicine*. 2020;20:100281.
18. Madsen T, Erlangsen A, Hjorthøj C, Nordentoft M. High suicide rates during psychiatric inpatient stay and shortly after discharge. *Acta Psychiatr Scand*. 2020;142(5):355-365.
19. McHugh CM, Large MM. Can machine-learning methods really help predict suicide? *Curr Opin Psychiatry*. 2020;33(4):369-374.
20. Corke M, Mullin K, Angel-Scott H, Xia S, Large M. Meta-analysis of the strength of exploratory suicide prediction models; from clinicians to computers. *BJPsych Open*. 2021;7(1):e26.
21. Krautenbacher N, Theis FJ, Fuchs C. Correcting classifiers for sample selection bias in two-phase case-control studies. *Comput Math Methods Med*. 2017;24:2017-2018.
22. Rose S, van der Laan MJ. A note on risk prediction for case-control studies. 2008.
23. Carroll R, Metcalfe C, Gunnell D. Hospital presenting self-harm and risk of fatal and non-fatal repetition: systematic review and meta-analysis. *PloS One*. 2014;9(2):e89944.
24. Zelkowitz RL, Jiang T, Horváth-Puhó E, et al. Predictors of nonfatal suicide attempts within 30 days of discharge from psychiatric hospitalization: sex-specific models developed using population-based registries. *J Affect Disord*. 2022;306:260-268.
25. Gradus JL, Rosellini AJ, Horváth-Puhó E, et al. Predicting sex-specific nonfatal suicide attempt risk using machine learning and data from Danish national registries. *Am J Epidemiol*. 2021;190(12):2517-2527.
26. Mors O, Perto GP, Mortensen PB. The Danish psychiatric central research register. *Scand J Public Health*. 2011;39(7_suppl):54-57.
27. Bliddal M, Broe A, Pottegård A, Olsen J, Langhoff-Roos J. The Danish medical birth register. *Eur J Epidemiol*. 2018;33:27-36.
28. Pedersen CB. The Danish civil registration system. *Scand J Public Health*. 2011;39(7_suppl):22-25.
29. Lynge E, Sandegaard JL, Rebolj M. The Danish national patient register. *Scand J Public Health*. 2011;39(7_suppl):30-33.
30. Petersson F, Baadsgaard M, Thygesen LC. Danish registers on personal labour market affiliation. *Scand J Public Health*. 2011;39(7_suppl):95-98.
31. Baadsgaard M, Quitzau J. Danish registers on personal income and transfer payments. *Scand J Public Health*. 2011;39:103-105.
32. Jensen VM, Rasmussen AW. Danish education registers. *Scand J Public Health*. 2011;39(7_suppl):91-94.
33. De La Garza ÁG, Blanco C, Olfson M, Wall MM. Identification of suicide attempt risk factors in a national US survey using machine learning. *JAMA Psychiatry*. 2021;78(4):398-406.
34. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301-320.
35. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
36. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. Association for Computer Machinery; 2016:785-794.
37. Brownlee J. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery; 2020.
38. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Adv Neural Inf Process Syst*. 2018:6638-6648.
39. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947;12(2):153-157.
40. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22.
41. Jacobucci R, Littlefield AK, Millner AJ, Kleiman EM, Steinley D. Evidence of inflated prediction performance: a commentary on machine learning and suicide research. *Clin Psychol Sci*. 2021;9(1):129-134.
42. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. *Adv Neural Inf Process Syst*. 2016:3315-3323.
43. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78(1):1-3.
44. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning. Association for Computer Machinery*; 2005:625-632.
45. Conroy RM, Pyörälä K, Fitzgerald AE, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J*. 2003;24(11):987-1003.
46. D'Agostino RB Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117(6):743-753.
47. Gilman SE, Bromet EJ, Cox KL, et al. Sociodemographic and career history predictors of suicide mortality in the United States Army 2004–2009. *Psychol Med*. 2014;44(12):2579-2592.
48. 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. *Am J Psychiatry*. 2003;160(4):765-772.

49. Gasse C, Danielsen AA, Pedersen MG, Pedersen CB, Mors O, Christensen J. Positive predictive value of a register-based algorithm using the Danish National Registries to identify suicidal events. *Pharmacoepidemiol Drug Saf*. 2018;27(10): 1131-1138.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Nielsen SD, Christensen RHB, Madsen T, Karstoft K-I, Clemmensen L, Benros ME. Prediction models of suicide and non-fatal suicide attempt after discharge from a psychiatric inpatient stay: A machine learning approach on nationwide Danish registers. *Acta Psychiatr Scand*. 2023;1-13. doi:10.1111/acps.13629