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

DOI:
10.1111/acps.13629

Publication date:
2023

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

Document license:
CC BY

Citation for published version (APA):
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 Nielsen1,2 | Rune H. B. Christensen2 | Trine Madsen3,4 | Karen-Ingé Karstoft5 | Line Clemmensen1 | Michael E. Benros2,6

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.
Funding information

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

Keywords

machine learning, prediction, psychiatry, suicide, suicide attempt

1 INTRODUCTION

Suicide is a major public health problem with 800,000 deaths annually in the world. 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, predictive models with sufficient accuracy and clinical applicability have not been successfully put forward. 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.

Prior studies predicting suicide and suicide attempt are limited in several ways: Some studies use a case-control sampling design (leading to overrepresentation of suicides where the predicted risk will not reflect the population risk), some apply retrospective designs looking backwards in time from the point of suicide (resulting in studies that cannot be used for prediction at time of discharge), some use small and biased populations, and some focus on suicides and suicide attempt combined to increase the number of cases, and will thus be skewed toward predicting non-fatal suicidal behavior. 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; 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 and suicide attempt, 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.

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. 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,\(^{31}\) and 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 odds42 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 scores43 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 (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
### 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

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>Threshold</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0.0789</td>
<td>0.9966</td>
<td>0.2625</td>
<td>0.9860</td>
<td>0.2473</td>
<td>216</td>
<td>177,449</td>
<td>607</td>
<td>2523</td>
</tr>
<tr>
<td>20%</td>
<td>0.1855</td>
<td>0.9901</td>
<td>0.2240</td>
<td>0.9875</td>
<td>0.1557</td>
<td>508</td>
<td>176,296</td>
<td>1760</td>
<td>2231</td>
</tr>
<tr>
<td>30%</td>
<td>0.2811</td>
<td>0.9803</td>
<td>0.1803</td>
<td>0.9888</td>
<td>0.1035</td>
<td>770</td>
<td>174,555</td>
<td>3501</td>
<td>1969</td>
</tr>
<tr>
<td>40%</td>
<td>0.3804</td>
<td>0.9658</td>
<td>0.1459</td>
<td>0.9902</td>
<td>0.0686</td>
<td>1042</td>
<td>171,958</td>
<td>6098</td>
<td>1697</td>
</tr>
<tr>
<td>50%</td>
<td>0.4681</td>
<td>0.9429</td>
<td>0.1121</td>
<td>0.9914</td>
<td>0.0451</td>
<td>1282</td>
<td>167,897</td>
<td>10,159</td>
<td>1457</td>
</tr>
<tr>
<td>60%</td>
<td>0.5688</td>
<td>0.9050</td>
<td>0.0843</td>
<td>0.9927</td>
<td>0.0270</td>
<td>1558</td>
<td>161,136</td>
<td>16,920</td>
<td>856</td>
</tr>
<tr>
<td>70%</td>
<td>0.6875</td>
<td>0.8417</td>
<td>0.0626</td>
<td>0.9943</td>
<td>0.0162</td>
<td>1883</td>
<td>149,869</td>
<td>28,187</td>
<td>551</td>
</tr>
<tr>
<td>80%</td>
<td>0.7988</td>
<td>0.7243</td>
<td>0.0427</td>
<td>0.9957</td>
<td>0.0096</td>
<td>2188</td>
<td>128,969</td>
<td>49,087</td>
<td>551</td>
</tr>
<tr>
<td>90%</td>
<td>0.8890</td>
<td>0.5299</td>
<td>0.0283</td>
<td>0.9968</td>
<td>0.0062</td>
<td>2435</td>
<td>94,357</td>
<td>83,699</td>
<td>304</td>
</tr>
</tbody>
</table>

#### Suicide

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>Threshold</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0.0741</td>
<td>0.9870</td>
<td>0.0068</td>
<td>0.9989</td>
<td>0.0048</td>
<td>16</td>
<td>178,228</td>
<td>2351</td>
<td>200</td>
</tr>
<tr>
<td>20%</td>
<td>0.1991</td>
<td>0.9604</td>
<td>0.0060</td>
<td>0.9990</td>
<td>0.0032</td>
<td>43</td>
<td>173,422</td>
<td>7157</td>
<td>173</td>
</tr>
<tr>
<td>30%</td>
<td>0.2917</td>
<td>0.9312</td>
<td>0.0050</td>
<td>0.9991</td>
<td>0.0024</td>
<td>63</td>
<td>168,163</td>
<td>12,416</td>
<td>153</td>
</tr>
<tr>
<td>40%</td>
<td>0.4167</td>
<td>0.8826</td>
<td>0.0042</td>
<td>0.9992</td>
<td>0.0019</td>
<td>90</td>
<td>159,371</td>
<td>21,208</td>
<td>126</td>
</tr>
<tr>
<td>50%</td>
<td>0.5231</td>
<td>0.8167</td>
<td>0.0034</td>
<td>0.9993</td>
<td>0.0015</td>
<td>113</td>
<td>147,474</td>
<td>33,105</td>
<td>103</td>
</tr>
<tr>
<td>60%</td>
<td>0.6528</td>
<td>0.7263</td>
<td>0.0028</td>
<td>0.9994</td>
<td>0.0012</td>
<td>141</td>
<td>131,147</td>
<td>49,432</td>
<td>75</td>
</tr>
<tr>
<td>70%</td>
<td>0.7500</td>
<td>0.6180</td>
<td>0.0023</td>
<td>0.9995</td>
<td>0.0010</td>
<td>162</td>
<td>111,594</td>
<td>68,985</td>
<td>54</td>
</tr>
<tr>
<td>80%</td>
<td>0.8333</td>
<td>0.4579</td>
<td>0.0018</td>
<td>0.9996</td>
<td>0.0008</td>
<td>180</td>
<td>82,682</td>
<td>97,897</td>
<td>36</td>
</tr>
<tr>
<td>90%</td>
<td>0.9167</td>
<td>0.2746</td>
<td>0.0015</td>
<td>0.9996</td>
<td>0.0007</td>
<td>198</td>
<td>49,583</td>
<td>130,996</td>
<td>18</td>
</tr>
</tbody>
</table>

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

![Variable importance for prediction of suicide attempts (A) and suicide (B).](image-url)

**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

<table>
<thead>
<tr>
<th>Model</th>
<th>Sex</th>
<th>Count</th>
<th>sa</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>All</td>
<td>180,795</td>
<td>2739</td>
<td>1286</td>
<td>167,680</td>
<td>10,376</td>
<td>1453</td>
<td>0.4695</td>
<td>0.9417</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>88,176</td>
<td>836</td>
<td>193</td>
<td>849,663</td>
<td>2377</td>
<td>643</td>
<td>0.2309</td>
<td>0.9972</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>92,619</td>
<td>1903</td>
<td>1093</td>
<td>82,717</td>
<td>7999</td>
<td>810</td>
<td>0.5744</td>
<td>0.9118</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>0.3435</td>
<td>0.0854</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>All</td>
<td>180,795</td>
<td>2739</td>
<td>1282</td>
<td>167,897</td>
<td>10,159</td>
<td>1457</td>
<td>0.4681</td>
<td>0.9429</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>88,176</td>
<td>836</td>
<td>401</td>
<td>83,757</td>
<td>3583</td>
<td>435</td>
<td>0.4800</td>
<td>0.9590</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>92,619</td>
<td>1903</td>
<td>880</td>
<td>84,940</td>
<td>5776</td>
<td>1023</td>
<td>0.4623</td>
<td>0.9363</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>0.0177</td>
<td>0.0226</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Fatal suicide

<table>
<thead>
<tr>
<th>Model</th>
<th>Sex</th>
<th>Count</th>
<th>sui</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>All</td>
<td>180,795</td>
<td>216</td>
<td>111</td>
<td>148,157</td>
<td>32,422</td>
<td>105</td>
<td>0.5139</td>
<td>0.8205</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>88,176</td>
<td>134</td>
<td>84</td>
<td>65,531</td>
<td>22,511</td>
<td>50</td>
<td>0.6269</td>
<td>0.7443</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>92,619</td>
<td>82</td>
<td>55</td>
<td>82,626</td>
<td>9911</td>
<td>27</td>
<td>0.6707</td>
<td>0.8929</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>0.0439</td>
<td>0.1486</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>All</td>
<td>180,795</td>
<td>216</td>
<td>113</td>
<td>147,474</td>
<td>33,105</td>
<td>103</td>
<td>0.5231</td>
<td>0.8167</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>88,176</td>
<td>134</td>
<td>68</td>
<td>69,902</td>
<td>18,140</td>
<td>66</td>
<td>0.5082</td>
<td>0.7940</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>92,619</td>
<td>82</td>
<td>45</td>
<td>76,596</td>
<td>15,941</td>
<td>37</td>
<td>0.5475</td>
<td>0.8277</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>0.0393</td>
<td>0.0338</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (B) Non-fatal suicide attempt

<table>
<thead>
<tr>
<th>Model</th>
<th>Age</th>
<th>Count</th>
<th>sa</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>All</td>
<td>180,795</td>
<td>2739</td>
<td>1282</td>
<td>167,897</td>
<td>10,159</td>
<td>1457</td>
<td>0.4681</td>
<td>0.9429</td>
</tr>
<tr>
<td></td>
<td>18–30</td>
<td>38,266</td>
<td>1116</td>
<td>714</td>
<td>32,632</td>
<td>4518</td>
<td>402</td>
<td>0.6398</td>
<td>0.8784</td>
</tr>
<tr>
<td></td>
<td>30–50</td>
<td>77,141</td>
<td>1165</td>
<td>504</td>
<td>70,972</td>
<td>5004</td>
<td>661</td>
<td>0.4326</td>
<td>0.9341</td>
</tr>
<tr>
<td></td>
<td>&gt;50</td>
<td>65,388</td>
<td>458</td>
<td>68</td>
<td>64,076</td>
<td>854</td>
<td>390</td>
<td>0.1485</td>
<td>0.9868</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>0.4913</td>
<td>0.1085</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>All</td>
<td>180,795</td>
<td>2739</td>
<td>1282</td>
<td>167,897</td>
<td>10,159</td>
<td>1457</td>
<td>0.4681</td>
<td>0.9429</td>
</tr>
<tr>
<td></td>
<td>18–30</td>
<td>38,266</td>
<td>1116</td>
<td>714</td>
<td>32,632</td>
<td>4518</td>
<td>402</td>
<td>0.6398</td>
<td>0.8784</td>
</tr>
<tr>
<td></td>
<td>30–50</td>
<td>77,141</td>
<td>1165</td>
<td>504</td>
<td>70,972</td>
<td>5004</td>
<td>661</td>
<td>0.4326</td>
<td>0.9341</td>
</tr>
<tr>
<td></td>
<td>&gt;50</td>
<td>65,388</td>
<td>458</td>
<td>68</td>
<td>64,076</td>
<td>854</td>
<td>390</td>
<td>0.1485</td>
<td>0.9868</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>0.0306</td>
<td>0.0502</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Suicide

<table>
<thead>
<tr>
<th>Model</th>
<th>Age</th>
<th>Count</th>
<th>sui</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>All</td>
<td>180,795</td>
<td>216</td>
<td>113</td>
<td>147,474</td>
<td>33,105</td>
<td>103</td>
<td>0.5231</td>
<td>0.8167</td>
</tr>
<tr>
<td></td>
<td>18–30</td>
<td>38,266</td>
<td>18</td>
<td>3</td>
<td>35,640</td>
<td>2608</td>
<td>15</td>
<td>0.1667</td>
<td>0.9318</td>
</tr>
<tr>
<td></td>
<td>30–50</td>
<td>77,141</td>
<td>93</td>
<td>50</td>
<td>64,753</td>
<td>12,295</td>
<td>43</td>
<td>0.5376</td>
<td>0.8404</td>
</tr>
<tr>
<td></td>
<td>&gt;50</td>
<td>65,388</td>
<td>105</td>
<td>58</td>
<td>47,764</td>
<td>17,519</td>
<td>47</td>
<td>0.5524</td>
<td>0.7316</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>0.3857</td>
<td>0.2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 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\textsuperscript{51} were estimated to be 0.0139 and 0.0012 (both close to 0) for the models predicting suicide attempt.

---

**Table 2 (Continued)**

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

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.
### FIGURE 3
Legend on next page.
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 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. 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. 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 ensemble 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.

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 members and veterans. However, studies indicate that sociodemographic factors on suicide tend to be overestimated when clinical variables are not taken into account. Thus, a possible explanation for the difference is that the effects of the sociodemographic variables are mediated by the clinical variables. 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 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.\textsuperscript{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\textsuperscript{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 \(\text{https://orcid.org/0000-0001-9272-7466}\)
Trine Madsen \(\text{https://orcid.org/0000-0003-4918-5862}\)
Line Clemmensen \(\text{https://orcid.org/0000-0001-5527-5798}\)
Michael E. Benros \(\text{https://orcid.org/0000-0003-4939-9465}\)

**REFERENCES**


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;148(6):525-537. doi:10.1111/acps.13629