JAMA Psychiatry | Original Investigation

Clinician Suicide Risk Assessment for Prediction of Suicide Attempt in a Large Health Care System

Kate H. Bentley, PhD; Chris J. Kennedy, PhD; Pratik N. Khadse, MSBA; Jasmin R. Brooks Stephens, PhD; Emily M. Madsen, BSc; Matthew J. Flics, BA; Hyunjoon Lee, MSc; Jordan W. Smoller, MD, ScD; Taylor A. Burke, PhD

Supplemental content

IMPORTANCE Clinical practice guidelines recommend suicide risk screening and assessment across behavioral health settings. The predictive accuracy of real-world clinician assessments for stratifying patients by risk of future suicidal behavior, however, remains understudied.

OBJECTIVE To evaluate routine clinical suicide risk assessment for prospectively predicting suicide attempt.

DESIGN, SETTING, AND PARTICIPANTS This electronic health record–based, prognostic study included 89 957 patients (≥5 years of age) with a structured suicide risk assessment (based on the Suicide Assessment Five-step Evaluation and Triage framework) that was documented by 2577 clinicians during outpatient, inpatient, and emergency department encounters at 12 hospitals in the Mass General Brigham health system between July 2019 and February 2023.

MAIN OUTCOMES AND MEASURES The primary outcome was an emergency department visit with a suicide attempt code recorded in the electronic health record within 90 days or 180 days of the index suicide risk assessment. The predictive performance of suicide risk assessments was evaluated on a temporal test set first using stratified prevalence (clinicians' overall risk estimates from a single suicide risk assessment item indicating minimal, low, moderate, or high risk) and then using machine learning models (incorporating all suicide risk assessment items).

RESULTS Of the 812 114 analyzed suicide risk assessments from the electronic health record, 58.81% were with female patients and 3.27% were with patients who were Asian, 5.26% were Black, 3.02% were Hispanic, 77.44% were White, and 11.00% were of Other or Unknown race. After suicide risk assessments were conducted during outpatient encounters, the suicide attempt rate was 0.12% within 90 days and 0.22% within 180 days; for inpatient encounters, the rate was 0.79% within 90 days and 1.29% within 180 days; and for emergency department encounters, the rate was 2.40% within 90 days and 3.70% within 180 days. Among patients evaluated during outpatient encounters, clinicians' overall single-item risk estimates had an area under the curve (AUC) value of 0.77 (95% CI, 0.72-0.81) for 90-day suicide attempt prediction; among patients evaluated during inpatient encounters, the AUC was 0.64 (95% CI, 0.59-0.69); and among patients evaluated during emergency department encounters, the AUC was 0.60 (95% CI, 0.55-0.64). Incorporating all clinician-documented suicide risk assessment items (87 predictors) via machine learning significantly increased the AUC for 90-day risk prediction to 0.87 (95% CI, 0.83-0.90) among patients evaluated during outpatient encounters, 0.79 (95% CI, 0.74-0.84) among patients evaluated during inpatient encounters, and 0.76 (95% CI, 0.72-0.80) among patients evaluated during emergency department encounters. Performance was similar for 180-day suicide risk prediction. The positive predictive values for the best-performing machine learning models (with 95% specificity) ranged from 3.6 to 10.1 times the prevalence for suicide attempt.

CONCLUSIONS AND RELEVANCE Clinicians stratify patients for suicide risk at levels significantly above chance. However, the predictive accuracy improves significantly by statistically incorporating information about recent suicidal thoughts and behaviors and other factors routinely assessed during clinical suicide risk assessment.

JAMA Psychiatry. 2025;82(6):599-608. doi:10.1001/jamapsychiatry.2025.0325
Published online April 9, 2025.

Author Affiliations: Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital, Boston (Bentley, Kennedy, Khadse, Brooks Stephens, Madsen, Flics, Lee, Smoller, Burke); Department of Psychiatry, Harvard Medical School, Harvard University, Boston, Massachusetts (Bentley, Kennedy, Smoller, Burke); Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston (Madsen, Lee, Smoller); Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee (Lee).

Corresponding Author: Kate H. Bentley, PhD, Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital, 185 Cambridge St, Boston, MA 02114 (kbentley@mgh.harvard.edu).

Suicide is the second leading cause of death among people between the ages of 10 and 14 years and between the ages of 20 and 34 years; it is the fifth leading cause of death among people between the ages of 10 and 64 years.[1] More than 90% of people who die by suicide saw a health care professional in the year prior (>50% within the prior month).[2,3] Clinicians have a key role in identifying individuals at risk for suicide.[4,5]

Estimating suicide risk has been described as the "quintessential clinical judgment"[6] of mental health clinicians. However, there is criticism that the ability of clinicians to predict suicide risk may be too inaccurate for clinical utility.[7-9] Numerous studies have tested the predictive accuracy of well-validated suicide risk screening tools[10-14] that are used to identify patients requiring comprehensive suicide risk assessment (SRA) by clinicians.[15,16] However, little research has evaluated the accuracy of SRAs by clinicians for predicting suicidal behavior in real-world settings.[7,17,18]

A study assessing the predictive accuracy of clinicians' judgments for patient suicide risk alone reported[19] an area under the curve (AUC) of 0.76 for future suicide attempt within 1 month and another study reported[20] an AUC of 0.73 for future suicide attempt within 6 months. Other studies assessing the accuracy of information routinely collected by clinicians during SRA (eg, suicidal thoughts and behaviors, risk and protective factors[21,22]) with or without clinicians' overall judgments reported AUCs from 0.53[23] to 0.60[8] for 6-month suicide attempt prediction. The prior studies had limitations such as small sample sizes, restriction to a single clinical setting or population, and being conducted in research contexts and not in real-world clinical practice.[8,19,20,24] Robust evaluations of the predictive accuracy of routine clinician SRA across care settings are needed.[18,25]

This retrospective, electronic health record (EHR)-based, prognostic study assessed the predictive accuracy of SRAs by clinicians in real-world practice. Using data from more than 800 000 SRAs routinely conducted across 3 clinical settings (outpatient, inpatient, and emergency department [ED]) in a large health care system, we aimed to: (1) describe clinician risk stratification based on routine SRA, (2) determine the predictive accuracy of clinicians' overall risk estimates (including differences by patient sociodemographic groups and clinician credentials), and (3) evaluate whether statistical models incorporating other routinely collected information during SRAs by clinicians (including suicidal thoughts and behaviors and risk and protective factors) improve predictive accuracy.

Key Points

Question How accurate are clinician assessments for risk stratification of future suicide attempt?

Findings In this electronic health record–based, prognostic study, clinicians' overall single-item risk estimates predicted 90- and 180-day suicide attempt at significantly above chance levels. Incorporating all suicide risk assessment items via machine learning significantly increased predictive accuracy.

Meaning Clinicians stratify patients for suicide risk at significantly above chance levels; however, predictive accuracy is significantly enhanced by statistically incorporating information about recent suicidal thoughts and behaviors and other risk and protective factors routinely assessed during suicide risk assessment.

Methods

Study Population

Data from clinical SRAs conducted between July 1, 2019, and February 6, 2023, were extracted from the Mass General Brigham health system EHR. The requirements for conducting SRAs within this health system vary across hospitals, clinical settings, and departments. Additional information about the study methods appear in the eMethods in Supplement 1. A protocol was not prepared and this study was not preregistered. There was no patient or public involvement during the design, conduct, reporting, interpretation, or dissemination of the study.

Extracted SRAs were included if they were (1) documented and (2) collected during a clinical encounter with a patient aged 5 years or older in (3) an outpatient setting (general medical or psychiatric), an inpatient setting (general medical or psychiatric), or an ED. In this health system, SRAs are documented in a structured form (eTable 1 in Supplement 1). Suicide risk assessments were excluded if they were from unknown or uncategorizable clinical settings (n = 72 313; 8.17% of total).

The included SRAs were linked to the Mass General Brigham research patient data registry to extract patient and clinician characteristics and outcome variables.[26] The study procedures were approved by the Mass General Brigham institutional review board, which granted a waiver of informed consent. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) + AI reporting guideline[43] was used for this investigation.

Study Measures

Primary Outcome

The primary outcome was a subsequent ED visit with an *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10)* diagnostic code for suicide attempt assigned in the EHR within 90 or 180 days after the index SRA.[27,28] Any *ICD-10* codes not assigned in an ED or preceded by an earlier suicide attempt code within 5 days were excluded, reducing duplicate suicide attempt entries.[29]

Given that EHR data only captured ED visits for suicide attempt within the Mass General Brigham health system, we conducted a set of prediction model sensitivity analyses on a subset of the sample with insurance claim data available. For these sensitivity analyses, we defined the outcome as an *ICD-10* diagnostic code of suicide attempt assigned in an ED per either EHR data or insurance claims data.

Predictors of Suicide Risk

The Mass General Brigham SRA is a structured EHR form based on the validated Suicide Assessment Five-step Evaluation and Triage (SAFE-T) framework,[30] which was developed by the Substance Abuse and Mental Health Services Administration. The

SRA form is used to assess recent suicidal thoughts and behaviors (eg, wish to be dead without thoughts of suicide; suicidal ideation without a plan or intent; suicide plan, intent, or prior suicide attempt) and nonsuicidal self-injurious or violent or destructive thoughts or behaviors. Additional SRA items assess the presence of other suicide risk factors (eg, depressed mood, recent loss, firearm access) or protective factors (eg, social support). A final item asks clinicians to document their overall estimate of patients' risk ("What is the patient's current, overall, acute risk of harm to self and/or others?") that was limited to the response options of "minimal," "low," "moderate," or "high" (all SRA items appear in eTable 1 in Supplement 1).

We evaluated 3 increasingly comprehensive structured predictor sets from each SRA in the EHR. The first predictor set included the single-item, clinician overall risk estimate (item 10 in eTable 1 in Supplement 1) and the clinical setting (outpatient, inpatient, or ED) in which the SRA was completed. The second predictor set included clinicians' overall risk estimates plus a subset of predictors (structured SRA items 1 through 6; 38 predictors) capturing suicidal or nonsuicidal self-injurious thoughts and behaviors and violent or destructive thoughts and behaviors. The third predictor set included the predictors from the first predictor set (clinicians' overall risk estimates) and the second predictor set (clinicians' overall risk estimates and the 38 predictors) plus structured SRA items 7 through 8 (total of 87 predictors), capturing the presence of a wide range of other risk and protective factors.

The risk and protective factors were only included in the third predictor set because we were interested in determining the potential incremental benefit of incorporating a broader range of risk and protective factors beyond those exclusively referring to self-harm and other-directed harm. In addition, the items including checkboxes to indicate specific risk and protective factors were only added to the EHR form for the SRA in May 2021. For the analyses using the third predictor set, a dataset containing a subset of 414 150 SRAs was used (the included SRAs were obtained after April 2021). Any missing values were imputed via mode imputation and missingness indicators were added.

Statistical Analysis

For the analyses on the full dataset, a temporal partition[31] was created that separated the training set (containing SRAs conducted before May 6, 2022) and the test set (containing 170 356 [21% of the total sample] SRAs collected after May 6, 2022).

Stratified Prevalence Model Testing Clinicians' Overall Risk Estimates

For the stratified prevalence models using the clinical setting (outpatient, inpatient, and ED) and clinicians' overall risk estimates, the SRA predictors were converted to absolute risk estimates by calculating the stratified suicide attempt prevalence on the training set for each combination of values (3 settings × 4 potential risk levels = 12 prevalence estimates).

Machine Learning Models

The larger 2 predictor sets were modeled using random forests with 2000 trees.[32,33] We also evaluated the predictor sets using LASSO (least absolute shrinkage and selection operator) regression,[34] naive Bayes, and a stacked ensemble (additional information appears in the eMethods in Supplement 1).[35,36]

Performance Evaluation

The primary performance metric was AUC. The SEs were calculated analytically.[37] The statistical significance of the differences in the AUC values was calculated using the method of DeLong et al.[38] Predicted probabilities were converted to binary suicide attempt predictions by selecting the threshold closest to a target sensitivity of 70%, and then we assessed for the positive predictive value (PPV), the negative predictive value (NPV), and specificity. The best-performing machine learning models were also evaluated at a target specificity of 95% for comparison with other studies that used EHR-based machine learning models for suicide risk prediction.[27,28,39]

The differences in predictive performance across patient sociodemographic groups and clinician credentials were evaluated by examining the AUCs for stratified prevalence and using machine learning models for patient sex, patient age, patient race and ethnicity, insurance type (public vs private), and the credentials of the clinicians who conducted the SRAs. The Benjamini-Hochberg method was used to adjust for multiple comparisons of model performance between patient and clinician subgroups. Predictor importance for the best-performing machine learning models was estimated using mean absolute SHAP (Shapley Additive Explanation) values.[40,41]

The analyses were conducted with R version 4.2.2 (R Foundation for Statistical Computing).[42] $P < .05$ indicated statistical significance.

Results

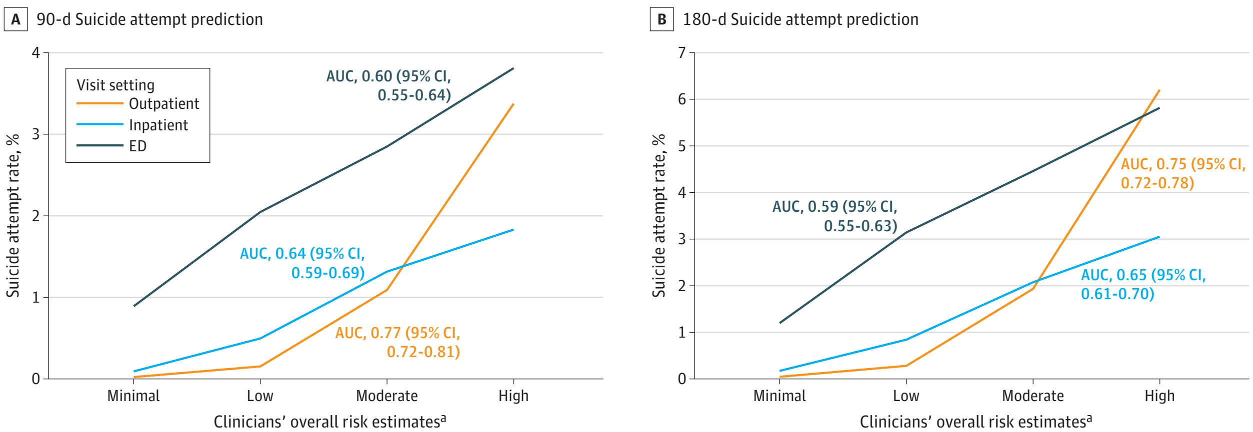
There were 812 114 SRAs conducted by 2577 clinicians at 12 hospitals among 89 957 patients. Of the 812 114 SRAs, 699 483 (86.13%) were from outpatient encounters, 76 723 (9.45%) were from inpatient encounters, and 35 908 (4.42%) were from ED encounters. Of the 812 114 SRAs, 477 628 (58.81%) were with female patients, 26 588 (3.27%) were Asian patients, 42 739 (5.26%) were Black patients, 24 564 (3.02%) were Hispanic patients, 628 897 (77.44%) were White patients, and 89 326 (11.00%) were patients of Other or Unknown race.

Clinicians' Overall Risk Estimates Alone

The distribution of clinicians' overall risk estimates for SRAs in each setting appears in eTable 2 in Supplement 1. Following SRAs conducted during outpatient encounters, the suicide attempt rate was 0.12% within 90 days and 0.22% within 180 days; for inpatient encounters, the rate was 0.79% within 90 days and 1.29% within 180 days; and for ED encounters, the rate was 2.40% within 90 days and 3.70% within 180 days (eTable 3 in Supplement 1). Outcome prevalence by patient sociodemographic characteristics across settings is detailed in eTables 4A-4C in Supplement 1.

For the final item of the SRA, clinicians were asked to document their overall estimate of the patient's suicide risk using the response options of "minimal," "low," "moderate," or

Figure 1. Suicide Attempt Prevalence Stratified by Clinicians' Overall Risk Estimates



[a]For the final item of the suicide risk assessment, clinicians were asked to document their overall estimate of the patient's suicide risk using the response options of "minimal," "low," "moderate," or "high."

ED indicates emergency department. The area under the curve (AUC) values were added for reference and were not derived from the data in these graphs.

"high." In outpatient settings, 3.38% of the SRAs with a high-risk estimate by clinicians were followed by a suicide attempt within 90 days compared with 1.09% of SRAs with a moderate-risk estimate, 0.15% with a low-risk estimate, and 0.02% with a minimal-risk estimate (**Figure 1** and eTable 3 in Supplement 1). In inpatient settings, 1.83% of SRAs with a high-risk estimate by clinicians were followed by a suicide attempt within 90 days compared with 1.32% of SRAs with a moderate-risk estimate, 0.50% with a low-risk estimate, and 0.09% with a minimal-risk estimate. In ED settings, 3.81% of SRAs with a high-risk estimate by clinicians were followed by a suicide attempt within 90 days compared with 2.85% of SRAs with a moderate-risk estimate, 2.05% with a low-risk estimate, and 0.89% with a minimal-risk estimate. The same pattern was present for suicide attempt within 180 days after the clinician risk estimate. The distribution of SRAs by the professional credentials of clinicians across settings (outpatient, inpatient, and ED) appears in eTable 5 in Supplement 1.

The stratified prevalence models using only clinicians' overall single-item risk estimates across all SRAs conducted in each clinical setting significantly outperformed chance-level prediction (**Table 1** and Figure 1). Among patients evaluated during outpatient encounters, clinicians' overall risk estimates for 90-day suicide attempt prediction had an AUC of 0.77 (95% CI, 0.72-0.81); among patients evaluated during inpatient encounters, the AUC was 0.64 (95% CI, 0.59-0.69); and among patients evaluated during ED encounters, the AUC was 0.60 (95% CI, 0.55-0.64). Suicide attempt prediction accuracy was similar for 180 days after an SRA. Based on a target sensitivity of 70%, the PPVs ranged from 0.22% (for 90-day suicide attempt prediction based on SRAs conducted during outpatient encounters) to 3.92% (for 180-day suicide attempt prediction based on SRAs conducted during ED encounters).

The performance statistics for the stratified prevalence models by patient sociodemographic groups (age, sex, race and ethnicity, insurance type) appear in eTables 6-9 in Supplement 1. The AUC point estimates were not statistically significantly better than chance for children (across all clinical settings) or for adolescents, older adults, and Asian patients, Black patients, and patients of Other or Unknown race in certain clinical settings. The performance statistics for the stratified prevalence models by clinician credentials appear in eTable 10 in Supplement 1. Suicide attempt predictive accuracy was variable by clinician credentials with AUC point estimates that were not statistically significantly better than chance for certain subgroups of clinicians in certain clinical settings.

Clinicians' Overall Risk Estimates Combined With Other SRA Items Using Machine Learning

Suicide attempt predictive accuracy was significantly higher for both machine learning models incorporating other SRA items than for the stratified prevalence models using only clinician risk estimates (Table 1). For the first set of machine learning models incorporating only the self-injurious and violent or destructive SRA items, the AUC improved by 0.11 (95% CI, 0.08-0.14) for SRAs conducted during outpatient encounters; by 0.12 (95% CI, 0.08-0.15) during inpatient encounters, and by 0.14 (95% CI, 0.10-0.19) during ED encounters.

The random forest models had the best performance for both 90-day and 180-day suicide attempt prediction (more than LASSO regression, naive Bayes, and a stacked ensemble; additional details appear in the eMethods in Supplement 1), reaching statistical significance in most model comparisons (Table 1 and eTables 11A-11B in Supplement 1). For 90-day suicide attempt prediction, the random forest models including 38 predictors yielded an AUC of 0.88 (95% CI, 0.84-0.91) based on SRAs conducted during outpatient encounters, 0.76 (95% CI, 0.71-0.81) during inpatient encounters, and 0.74 (95% CI, 0.70-0.79) during ED encounters. Suicide attempt prediction accuracy was similar for 180 days after an SRA. Based on a target sensitivity of 70%, the PPVs ranged from 0.56% (for 90-day suicide attempt prediction based on SRAs conducted during outpatient encounters) to 5.78% (for 180-day suicide attempt prediction based on SRAs conducted during ED encounters).

Table 1. Performance Statistics for Stratified Prevalence and Best-Performing Machine Learning Models

	Model description[a]	Risk stratification of future suicide attempt: AUC (95% CI)	Outcome prevalence in test set, %	Sensitivity, %[b]	Specificity, %	PPV, %	NPV, %	*P* value[c]	Model comparison
90-d Suicide attempt prediction									
Clinicians' overall risk estimates									
Outpatient	Stratified prevalence	0.77 (0.72-0.81)	0.12	90.06	49.90	0.22	99.98	<.001	Compared with null model[d]
Inpatient		0.64 (0.59-0.69)	0.75	60.83	62.84	1.23	99.53	<.001	
Emergency department		0.60 (0.55-0.64)	1.80	63.52	53.16	2.43	98.76	<.001	
Clinicians' overall risk estimates plus a subset of SRA items (38 predictors)									
Outpatient	Best-performing machine learning[e]	0.88 (0.84-0.91)	0.12	70.17	84.37	0.56	99.96	<.001	Compared with clinicians' overall risk estimates model
Inpatient		0.76 (0.71-0.81)	0.75	67.50	68.61	1.61	99.64	<.001	
Emergency department		0.74 (0.70-0.79)	1.80	71.07	65.94	3.69	99.20	<.001	
Clinicians' overall risk estimates plus all SRA items (87 predictors)									
Outpatient	Best-performing machine learning[f]	0.87 (0.83-0.90)	0.12	69.61	88.13	0.72	99.96	.06	Compared with clinicians' overall risk estimates plus subset of SRA items model
Inpatient		0.79 (0.74-0.84)	0.75	69.17	72.44	1.87	99.68	.01	
Emergency department		0.76 (0.72-0.80)	1.80	69.81	68.64	3.92	99.20	.36	
180-d Suicide attempt prediction									
Clinicians' overall risk estimates									
Outpatient	Stratified prevalence	0.75 (0.72-0.78)	0.21	91.39	49.93	0.38	99.96	<.001	Compared with null model[d]
Inpatient		0.65 (0.61-0.70)	1.15	62.30	62.96	1.92	99.31	<.001	
Emergency department		0.59 (0.55-0.63)	2.98	61.98	53.32	3.92	97.86	<.001	
Clinicians' overall risk estimates plus a subset of SRA items (38 predictors)									
Outpatient	Best-performing machine learning[e]	0.85 (0.82-0.87)	0.21	71.19	80.82	0.77	99.93	<.001	Compared with clinicians' overall risk estimates model
Inpatient		0.76 (0.70-0.80)	1.15	69.95	68.70	2.53	99.49	<.001	
Emergency department		0.74 (0.71-0.78)	2.98	72.62	63.66	5.78	98.70	<.001	
Clinicians' overall risk estimates plus all SRA items (87 predictors)									
Outpatient	Best-performing machine learning[f]	0.88 (0.86-0.91)	0.21	69.21	89.45	1.34	99.93	<.001	Compared with clinicians' overall risk estimates plus subset of SRA items model
Inpatient		0.79 (0.75-0.83)	1.15	69.95	70.87	2.72	99.51	.004	
Emergency department		0.77 (0.74-0.80)	2.98	69.96	69.80	6.64	98.70	.008	

Abbreviations: AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

[a] Clinicians' overall risk estimates from the suicide risk assessment (SRA) and clinical setting of the index SRA were included as predictors in all models.

[b] Thresholds were selected closest to a target of 70% sensitivity.

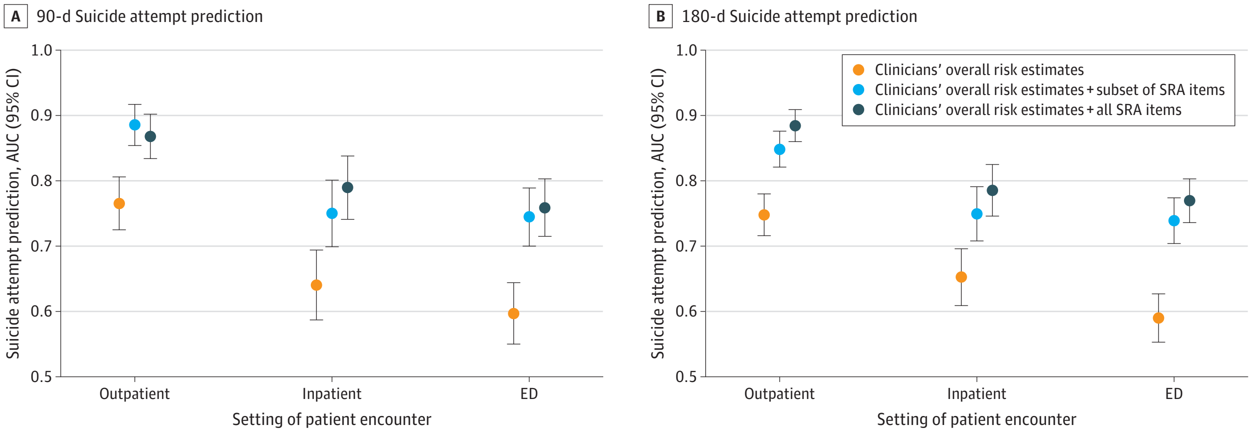
[c] The statistical significance for the model comparisons was calculated using the method of DeLong et al.[38]

[d] The null model refers to chance-level prediction (an AUC of 0.50).

[e] Built on a subset of the SRA dataset, given that the risk and protective factor SRA items were added in May 2021.

[f] The performance statistics from this set of models were built on the entire dataset.

Figure 2. Area Under the Curve (AUC) for Suicide Risk Assessments Predicting Suicide Attempt Within 90 and 180 Days by Setting of Index Encounter



The AUC values are based on models assessing (1) clinicians' overall risk estimates, (2) clinicians' overall risk estimates plus 38 suicide risk assessment (SRA) items, and (3) clinicians' overall risk estimates plus all 87 SRA items. ED indicates emergency department.

The performance statistics for the second set of machine learning models that incorporated all SRA items (87 predictors), including clinician-documented risk and protective factors, appear in Table 1. The best-performing machine learning models (random forests) for 90-day suicide attempt prediction had an AUC of 0.87 (95% CI, 0.83-0.90) for SRAs conducted during outpatient encounters, an AUC of 0.79 (95% CI, 0.74-0.84) during inpatient encounters, and an AUC of 0.76 (95% CI, 0.72-0.80) during ED encounters. Suicide attempt prediction accuracy was similar for 180 days after an SRA. Based on a target sensitivity of 70%, the PPVs ranged from 0.72% (for 90-day suicide attempt prediction based on SRAs conducted during outpatient encounters) to 6.64% (for 180-day suicide attempt prediction based on SRAs conducted during ED encounters).

The performance statistics for the best-performing machine learning models by patient sociodemographic groups and clinician credentials appear in eTables 12-16 in Supplement 1. Similar to the stratified prevalence models, predictive accuracy varied across subgroups. Due to risk of the temporal test set performance being inflated when the same patients and clinicians were also observed during the training set, the performance statistics for the best-performing machine learning models were stratified by whether individual patients and clinicians appeared in both the training and test sets or the test set only appear in eTables 17-18 in Supplement 1. The performance statistics were similar for the models in which individual patients and clinicians appeared in both the training and test sets vs only in the test set.

The machine learning models that included risk or protective factor items were built on the subset of SRAs conducted during or after May 2021. To compare performance of the models that included vs did not include risk and protective factors, we also built the first set of machine learning models (using the subset of SRA items, 38 predictors) in the subset of SRAs conducted during or after May 2021. Compared with the full dataset, model performance in the subset of SRAs was similar. The random forest models for 90-day suicide attempt prediction yielded an AUC of 0.89 (95% CI, 0.85-0.92) based on SRAs conducted during outpatient encounters, 0.75 (95% CI, 0.70-0.80) during inpatient encounters, and 0.74 (95% CI, 0.70-0.79) during ED encounters. The random forest models for 180-day suicide attempt prediction yielded an AUC of 0.85 (95% CI, 0.82-0.88) based on SRAs conducted during outpatient encounters, 0.75 (95% CI, 0.71-0.79) during inpatient encounters, and 0.74 (95% CI, 0.70-0.77) during ED encounters.

The stratified prevalence model (clinicians' overall risk estimates) and both machine learning models (the first model using clinicians' overall risk estimates plus the subset of SRA items [38 predictors] and the second model using clinicians' overall risk estimates plus all SRA items [87 predictors]) appear in **Figure 2**. The predictive accuracy of both machine learning models using random forests was significantly higher than the stratified prevalence model (clinicians' overall risk estimates alone). The second machine learning model (clinicians' overall risk estimates plus all SRA items [87 predictors]) outperformed the stratified prevalence model and the first machine learning model (clinicians' overall risk estimates plus a subset of SRA items [38 predictors]) for 90-day suicide attempt prediction based on SRAs conducted during inpatient encounters, but not based on SRAs conducted during ED or outpatient encounters (Table 1). The random forest models for 180-day suicide attempt prediction including risk and protective factors outperformed those without risk and protective factors across all 3 settings.

When using a target specificity threshold of 95%, the PPVs for the best-performing machine learning models ranged from 1.13% for 90-day suicide attempt prediction (based on SRAs conducted during outpatient encounters) to 6.89% (based on SRAs conducted during ED encounters) (**Table 2**). When using a target specificity threshold of 95%, the PPVs for the best-performing machine learning models ranged from 2.12% for 180-day suicide attempt prediction (based on SRAs conducted during outpatient encounters) to 11.57% (based on SRAs conducted during ED encounters). Selected performance statistics for the best-performing machine learning models across

Table 2. Performance Metrics for Best-Performing Machine Learning Models

	Best-performing machine learning models: risk stratification by overall clinician assessment plus all SRA items (87 predictors)						
	AUC (95% CI)	Outcome prevalence in test set, %	Sensitivity, %	Specificity, %[a]	PPV, %	NPV, %	Lift (PPV/prevalence)
90-d Suicide attempt prediction							
Outpatient	0.87 (0.83-0.90)	0.12	45.86	95.00	1.13	99.93	9.42
Inpatient	0.79 (0.74-0.84)	0.75	18.33	95.00	2.71	99.35	3.61
Emergency department	0.76 (0.72-0.80)	1.80	20.13	95.00	6.89	98.48	3.83
180-d Suicide attempt prediction							
Outpatient	0.88 (0.86-0.91)	0.21	51.99	95.00	2.12	99.90	10.10
Inpatient	0.79 (0.75-0.83)	1.15	19.13	95.00	4.26	99.02	3.70
Emergency department	0.77 (0.74-0.80)	2.98	21.29	95.00	11.57	97.52	3.88

Abbreviations: AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value; SRA, suicide risk assessment.

[a] A target specificity threshold of 95% was used.

4 different target sensitivity thresholds (10%, 30%, 50%, 70%) appear in eTable 19 in Supplement 1.

Predictor Importance

The best (top 25) 90-day suicide attempt predictors based on the best-performing machine learning models (stratified by outpatient, inpatient, and ED setting) appear in **Figure 3** (the best 180-day predictors appear in eFigure 1 in Supplement 1). In the outpatient setting, help-seeking behavior, clinicians' overall risk estimates, and prior suicidal behaviors or attempts (presence or absence) were the most important predictors. For inpatient and ED settings, the most important predictors were clinical setting of the index SRA, clinicians' overall risk estimates, recent discharge (within past 3 months) from a psychiatric facility, a history of prior suicide attempts, and presence or absence of prior suicidal behaviors or attempts.

Sensitivity Analyses

To assess the generalizability of these findings, we conducted sensitivity analyses in a subset of patients with available insurance claims data (8707 patients and 48 959 SRAs; 6.03% of the full SRA sample), which were combined with the data from the Mass General Brigham health system EHR, to assess suicide attempt outcomes. We observed a similar pattern of overall results (for the stratified prevalence and best-performing machine learning [random forest] models) when predicting suicide attempt based on diagnostic codes obtained from either insurance claims data or the EHR (eTable 20 and eFigure 2 in Supplement 1).

Discussion

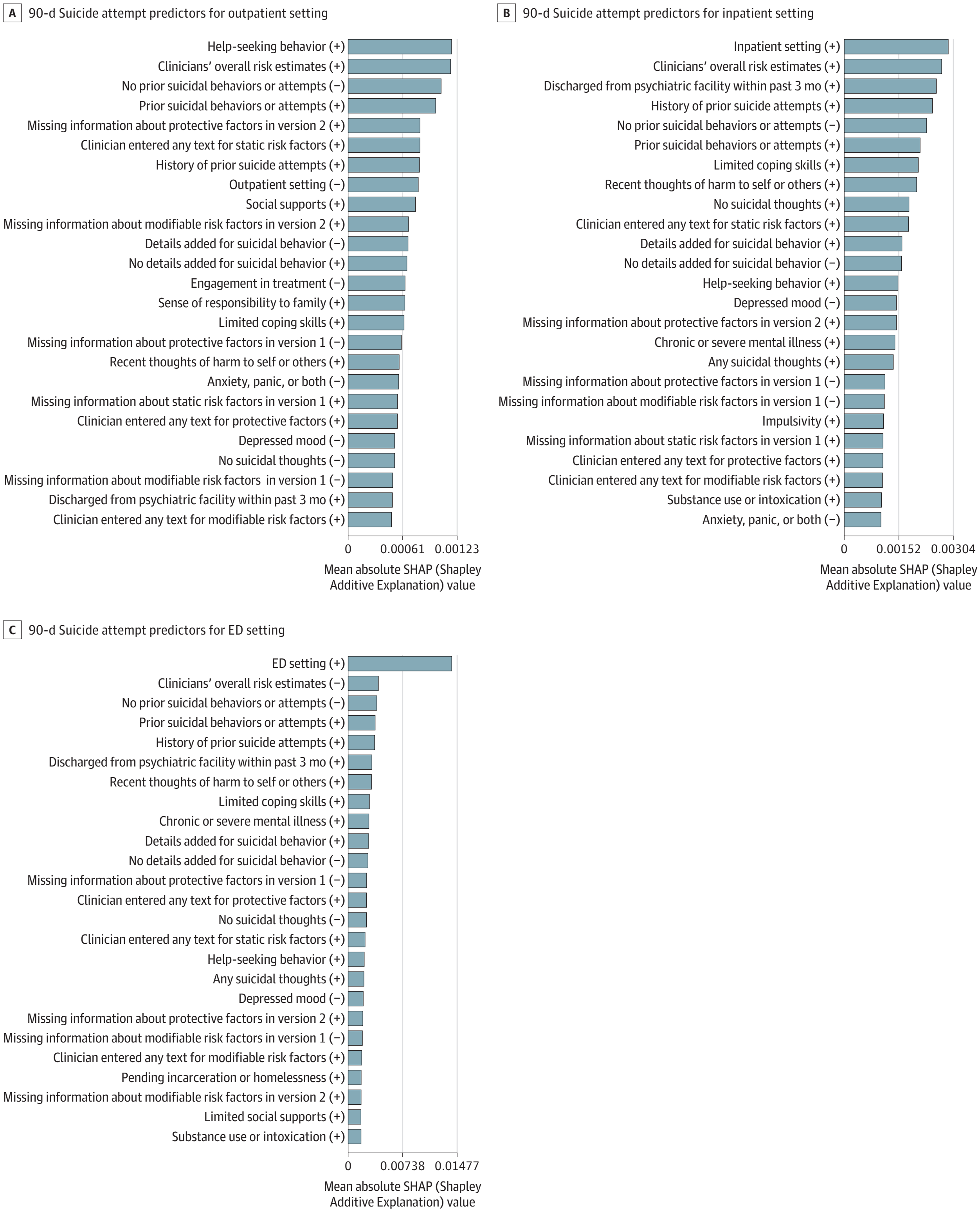
Clinicians estimated patients' suicide risk at levels significantly better than chance during routine SRA. This determination was made with high precision by extracting all SRAs from the EHR for a health system, resulting in a larger, more diverse clinical sample than prior work, and with a temporal validation strategy that parallels prospective clinical use.[31] Compared with using only clinicians' overall single-item risk estimates; however, we achieved significantly enhanced statistical predictions using machine learning and incorporating additional routinely documented information during SRA. The single-item, clinician overall risk estimate was among the top 2 most important predictors in all best-performing machine learning models. Although our findings suggest that clinician judgment alone should not be dismissed as having no predictive value, statistical models that use all clinician-documented information during SRAs are consistently superior to clinical judgment alone.

For certain patient and clinician subgroups in certain clinical settings, clinicians' risk estimates alone (and in some cases, the machine learning models incorporating other SRA items) did not exceed chance-level prediction. This may be due in part to much smaller sample sizes in these sets of stratified models, implicit biases, and variability in clinician training and experience.

The best-performing outpatient machine learning models incorporating all SRA data (88% specificity at 70% sensitivity) satisfy accuracy thresholds for determining whether a suicide risk prediction method is accurate enough to be cost-effective if used to target evidence-based interventions for suicide risk reduction to high-risk patients.[44] Our models, which relied solely on EHR-sourced clinician SRA data, leveraged a novel set of predictors which, to our knowledge, have not yet been incorporated in other EHR-based suicide risk prediction work. They demonstrate strong promise compared with other models using vast amounts of standard structured EHR data (eg, diagnostic or procedure codes, medications). For example, in the same health care system, Sheu et al[39] reported an AUC of 0.74 for psychiatric ED prediction of suicide attempt at 6 months and Nock et al[8] reported an AUC of 0.78 compared with an AUC of 0.77 for the best-performing model in the current study. Sheu et al[39] reported an AUC of 0.79 for psychiatric inpatient prediction of suicide attempt at 6 months compared with an AUC of 0.79 for the best-performing model in the current study for SRAs conducted in the inpatient setting.

Despite the overall low PPVs typical for low-prevalence events,[27,28,39,45,46] the PPVs for our best-performing models (when using a 95% specificity threshold) were up to 10.1 times the baseline suicide attempt prevalence for outpatient settings (up to 3.9-fold for SRAs conducted during ED encounters and up to 3.7-fold for SRAs conducted during inpatient encounters), and comparable with other EHR-based machine learning models in

Figure 3. Top 25 Predictors in the Best-Performing Machine Learning Models for Suicide Attempt Within 90 Days



A positive sign indicates the variable is positively associated with suicide risk and a negative sign indicates the variable is negatively associated with suicide risk (ie, is protective). SRA indicates suicide risk assessment.

the same health system.[27,39] However, caution is warranted when comparing studies due to variation in outcome definitions, thresholds, patient populations and cohorts, validation methods, and intended clinical use (eg, screening vs assessment[15]).

Our findings may hold substantial translational promise. Models using SRA data could be integrated into EHRs by health system analytics teams, enabling real-time risk estimation immediately after SRAs. Clinicians could receive alerts embedded within EHRs that contain risk thresholds tailored to health system priorities and resources, alongside suggested interventions.[47-49] Future research should compare machine learning models using clinician SRA alone vs combined with structured EHR data to rigorously evaluate incremental utility of standard EHR data,[50,51] and improve the understanding of how to optimally combine (and potentially sequence) clinical judgment or risk assessment with newer statistical modeling approaches.[18,50-53] Combining clinician SRA with powerful emerging EHR suicide risk algorithms has potential to augment suicide risk prediction, paving the way for more targeted and timely intervention.

Limitations

This study has limitations. First, the SRA item used for clinicians' overall risk estimates referred to patient "risk of harm to self and/or others." Although self- and other-directed harm can co-occur,[54,55] these are distinct behaviors with distinct risk profiles.

Second, this study examined SRAs conducted in routine care in a large health system spanning multiple hospitals in which clinician SRA workflows may vary across settings and patient encounters. The results apply only to patients who received SRAs administered by clinicians, and may reflect setting-specific subsets of patients rather than broader patient populations.

Third, for the vast majority of the study sample, we were only able to ascertain suicide attempt outcomes via the EHR from a single health system. Comprehensive access to insurance claims data capturing diagnoses occurring outside this health system, as well as adding patient-reported outcomes assessment,[8] would permit more robust and precise outcome measurement, potentially increasing the robustness and generalizability of the models.

Conclusions

Clinicians stratify patients for suicide risk at levels significantly above chance. However, the predictive accuracy improves significantly by statistically incorporating information about recent suicidal thoughts and behaviors and other factors routinely assessed during clinical SRA.

ARTICLE INFORMATION

Accepted for Publication: January 24, 2025.

Published Online: April 9, 2025. doi:10.1001/jamapsychiatry.2025.0325

Author Contributions: Dr Bentley had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Drs Bentley and Kennedy served as co–first authors.
Concept and design: Bentley, Kennedy, Smoller, Burke.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Bentley, Kennedy, Khadse, Burke.
Critical review of the manuscript for important intellectual content: All authors.
Statistical analysis: Kennedy, Khadse, Lee.
Obtained funding: Bentley, Kennedy.
Administrative, technical, or material support: Khadse, Brooks Stephens, Madsen, Flics.
Supervision: Bentley, Kennedy, Burke.

Conflict of Interest Disclosures: Dr Smoller reported serving on the scientific advisory board for Sensorium Therapeutics Inc and having stock options in and receiving grants from Biogen Inc. No other disclosures were reported.

Funding/Support: This research was supported by the Claflin Distinguished Scholar Award (awarded to Dr Bentley) from the Massachusetts General Hospital Executive Committee on Research; grants from the National Institute of Mental Health (K23MH120436 [awarded to Dr Bentley]; K01MH135131 [awarded to Dr Kennedy]; P50MH129699 Scholar Award [awarded to Dr Brooks Stephens]; P50MH129699 [awarded to Dr Smoller]; and K23MH126168 [awarded to Dr Burke]); and a gift from the Tommy Fuss Fund (awarded to Dr Smoller).

Role of the Funder/Sponsor: The funders/sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the sponsors/funders.

Data Sharing Statement: See Supplement 2.

Additional Contributions: We thank Sara B. Golas, MA (Mass General Brigham), for her assistance with extracting certain data from the electronic health records. Ms Golas was not compensated for her work on this project.

REFERENCES

1. US Centers for Disease Control and Prevention. WISQARS Leading Causes of Death Visualization Tool: 10 leading causes of death, United States, 2022. Accessed March 3, 2025. https://wisqars.cdc.gov/lcd/?o=LCD&y1=2022&y2=2022&ct=10&cc=ALL&g=00&s=0&r=0&ry=2&e=0&ar=lcd1age&at=groups&ag=lcd1age&a1=0&a2=199

2. Ahmedani BK, Simon GE, Stewart C, et al. Health care contacts in the year before suicide death. *J Gen Intern Med*. 2014;29(6):870-877. doi:10.1007/s11606-014-2767-3

3. Ahmedani BK, Westphal J, Autio K, et al. Variation in patterns of health care before suicide: a population case-control study. *Prev Med*. 2019;127:105796. doi:10.1016/j.ypmed.2019.105796

4. Joint Commission. Suicide prevention R3 report: national patient safety goal for suicide prevention. Published November 20, 2019. Accessed March 3, 2025. https://www.jointcommission.org/resources/patient-safety-topics/suicide-prevention/

5. Moran M. Joint commission issues update for suicide prevention. *Psychiatr News*. 2019;54(1). doi:10.1176/appi.pn.2019.1a19

6. Jacobs DG, Baldessarini RJ, Conwell Y, et al; Work Group on Suicidal Behaviors. Practice guideline for the assessment and treatment of patients with suicidal behaviors. Published online November 2003. Accessed March 7, 2025. https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/suicide.pdf

7. Woodford R, Spittal MJ, Milner A, et al. Accuracy of clinician predictions of future self-harm: a systematic review and meta-analysis of predictive studies. *Suicide Life Threat Behav*. 2019;49(1):23-40. doi:10.1111/sltb.12395

8. Nock MK, Millner AJ, Ross EL, et al. Prediction of suicide attempts using clinician assessment, patient self-report, and electronic health records. *JAMA Netw Open*. 2022;5(1):e2144373. doi:10.1001/jamanetworkopen.2021.44373

9. Carter G, Spittal MJ. Suicide risk assessment. *Crisis*. 2018;39(4):229-234. doi:10.1027/0227-5910/a000558

10. Riblet NB, Matsunaga S, Lee Y, et al. Tools to detect risk of death by suicide: a systematic review and meta-analysis. *J Clin Psychiatry*. 2022;84(1):21r14385. doi:10.4088/JCP.21r14385

11. Harris IM, Beese S, Moore D. Predicting future self-harm or suicide in adolescents: a systematic review of risk assessment scales/tools. *BMJ Open*. 2019;9(9):e029311. doi:10.1136/bmjopen-2019-029311

12. Thom R, Hogan C, Hazen E. Suicide risk screening in the hospital setting: a review of brief validated tools. *Psychosomatics*. 2020;61(1):1-7. doi:10.1016/j.psym.2019.08.009

13. O'Connor EA, Perdue LA, Coppola EL, Henninger ML, Thomas RG, Gaynes BN. Depression and suicide risk screening: updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA*. 2023;329(23):2068-2085. doi:10.1001/jama.2023.7787

14. Posner K, Brown GK, Stanley B, et al. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am J Psychiatry*. 2011;168(12):1266-1277. doi:10.1176/appi.ajp.2011.10111704

15. Horowitz LM, Ryan PC, Wei AX, Boudreaux ED, Ackerman JP, Bridge JA. Screening and assessing suicide risk in medical settings: feasible strategies for early detection. *Focus (Am Psychiatr Publ)*. 2023;21(2):145-151. doi:10.1176/appi.focus.20220086

16. Horowitz LM, Bridge JA, Teach SJ, et al. Ask Suicide-Screening Questions (ASQ): a brief instrument for the pediatric emergency department. *Arch Pediatr Adolesc Med*. 2012;166(12):1170-1176. doi:10.1001/archpediatrics.2012.1276

17. Nock MK, Park JM, Finn CT, Deliberto TL, Dour HJ, Banaji MR. Measuring the suicidal mind: implicit cognition predicts suicidal behavior. *Psychol Sci*. 2010;21(4):511-517. doi:10.1177/0956797610364762

18. Simon GE, Matarazzo BB, Walsh CG, et al. Reconciling statistical and clinicians' predictions of suicide risk. *Psychiatr Serv*. 2021;72(5):555-562. doi:10.1176/appi.ps.202000214

19. Barzilay S, Yaseen ZS, Hawes M, et al. Determinants and predictive value of clinician assessment of short-term suicide risk. *Suicide Life Threat Behav*. 2019;49(2):614-626. doi:10.1111/sltb.12462

20. Randall JR, Sareen J, Chateau D, Bolton JM. Predicting future suicide: clinician opinion versus a standardized assessment tool. *Suicide Life Threat Behav*. 2019;49(4):941-951. doi:10.1111/sltb.12481

21. 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. doi:10.1037/bul0000084

22. Ribeiro JD, Franklin JC, Fox KR, et al. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychol Med*. 2016;46(2):225-236. doi:10.1017/S0033291715001804

23. Tran T, Luo W, Phung D, et al. Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. *BMC Psychiatry*. 2014;14(1):76. doi:10.1186/1471-244X-14-76

24. Barnes SM, Bahraini NH, Forster JE, et al. Moving beyond self-report: implicit associations about death/life prospectively predict suicidal behavior among veterans. *Suicide Life Threat Behav*. 2017;47(1):67-77. doi:10.1111/sltb.12265

25. Runeson B, Odeberg J, Pettersson A, Edbom T, Jildevik Adamsson I, Waern M. Instruments for the assessment of suicide risk: a systematic review evaluating the certainty of the evidence. *PLoS One*. 2017;12(7):e0180292. doi:10.1371/journal.pone.0180292

26. Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a research patient data repository. *AMIA Annu Symp Proc*. 2006;2006:1044.

27. 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. doi:10.1176/appi.ajp.2016.16010077

28. Barak-Corren Y, Castro VM, Nock MK, et al. Validation of an electronic health record–based suicide risk prediction modeling approach across multiple health care systems. *JAMA Netw Open*. 2020;3(3):e201262. doi:10.1001/jamanetworkopen.2020.1262

29. Bentley KH, Madsen EM, Song E, et al. Determining distinct suicide attempts from recurrent electronic health record codes: classification study. *JMIR Form Res*. 2024;8:e46364. doi:10.2196/46364

30. Fowler JC. Suicide risk assessment in clinical practice: pragmatic guidelines for imperfect assessments. *Psychotherapy (Chic)*. 2012;49(1):81-90. doi:10.1037/a0026148

31. de Hond AAH, Shah VB, Kant IMJ, Van Calster B, Steyerberg EW, Hernandez-Boussard T. Perspectives on validation of clinical predictive algorithms. *NPJ Digit Med*. 2023;6(1):86. doi:10.1038/s41746-023-00832-9

32. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

33. Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw*. 2017;77:1-17. doi:10.18637/jss.v077.i01

34. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc*. 1996;58(1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

35. Polley E, LeDell E, Kennedy C, Lendle S, van der Laan M. Package "SuperLearner." Accessed March 3, 2025. https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf

36. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1):e25. doi:10.2202/1544-6115.1309

37. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36. doi:10.1148/radiology.143.1.7063747

38. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845. doi:10.2307/2531595

39. Sheu YH, Sun J, Lee H, et al. An efficient landmark model for prediction of suicide attempts in multiple clinical settings. *Psychiatry Res*. 2023;323:115175. doi:10.1016/j.psychres.2023.115175

40. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA.

41. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst*. 2014;41(3):647-665. doi:10.1007/s10115-013-0679-x

42. R Project for Statistical Computing. R: A Language and Environment for Statistical Computing. Accessed March 3, 2025. https://www.R-project.org/

43. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378

44. Ross EL, Zuromski KL, Reis BY, Nock MK, Kessler RC, Smoller JW. Accuracy requirements for cost-effective suicide risk prediction among primary care patients in the US. *JAMA Psychiatry*. 2021;78(6):642-650. doi:10.1001/jamapsychiatry.2021.0089

45. 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. doi:10.1001/jamapsychiatry.2019.0174

46. Simon GE, Johnson E, Lawrence JM, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *Am J Psychiatry*. 2018;175(10):951-960. doi:10.1176/appi.ajp.2018.17101167

47. Bolton JM, Gunnell D, Turecki G. Suicide risk assessment and intervention in people with mental illness. *BMJ*. 2015;351:h4978. doi:10.1136/bmj.h4978

48. Simpson SA, Loh RM, Goans CRR. New data on suicide risk assessment in the emergency department reveal the need for new approaches in research and clinical practice. *Psychol Med*. 2023;53(3):1122-1123. doi:10.1017/S0033291721001653

49. Kline-Simon AH, Sterling S, Young-Wolff K, et al. Estimates of workload associated with suicide risk alerts after implementation of risk-prediction model. *JAMA Netw Open*. 2020;3(10):e2021189. doi:10.1001/jamanetworkopen.2020.21189

50. Wilimitis D, Turer RW, Ripperger M, et al. Integration of face-to-face screening with real-time machine learning to predict risk of suicide among adults. *JAMA Netw Open*. 2022;5(5):e2212095. doi:10.1001/jamanetworkopen.2022.12095

51. Bentley KH, Zuromski KL, Fortgang RG, et al. Implementing machine learning models for suicide risk prediction in clinical practice: focus group study with hospital providers. *JMIR Form Res*. 2022;6(3):e30946. doi:10.2196/30946

52. Brown LA, Benhamou K, May AM, Mu W, Berk R. Machine learning algorithms in suicide prevention: clinician interpretations as barriers to implementation. *J Clin Psychiatry*. 2020;81(3):10951. doi:10.4088/JCP.19m12970

53. Burke TA, Jacobucci R, Ammerman BA, Alloy LB, Diamond G. Using machine learning to classify suicide attempt history among youth in medical care settings. *J Affect Disord*. 2020;268:206-214. doi:10.1016/j.jad.2020.02.048

54. Sahlin H, Kuja-Halkola R, Bjureberg J, et al. Association between deliberate self-harm and violent criminality. *JAMA Psychiatry*. 2017;74(6):615-621. doi:10.1001/jamapsychiatry.2017.0338

55. Shafti M, Taylor PJ, Forrester A, Pratt D. The co-occurrence of self-harm and aggression: a cognitive-emotional model of dual-harm. *Front Psychol*. 2021;12:586135. doi:10.3389/fpsyg.2021.586135