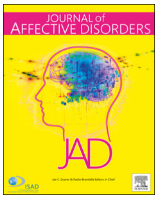
Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad



Research paper

Using machine learning to classify suicide attempt history among youth in medical care settings

Taylor A. Burke[a,*], Ross Jacobucci[b], Brooke A. Ammerman[b], Lauren B. Alloy[c], Guy Diamond[d]

[a] Alpert Medical School of Brown University, Department of Psychiatry and Human Behavior, Providence, RI, USA
[b] University of Notre Dame, Department of Psychology, Notre Dame, IN, USA
[c] Department of Psychology, Temple University, Philadelphia, PA, USA
[d] Center for Family Intervention Science, Drexel University, Philadelphia PA, USA

ARTICLE INFO

Keywords:
Suicide attempt
Medical setting
Pediatric
Emergency department
Primary care
Suicide risk screening
Machine learning
Exploratory data mining

ABSTRACT

Background: The current study aimed to classify recent and lifetime suicide attempt history among youth presenting to medical settings using machine learning (ML) as applied to a behavioral health screen self-report survey.

Methods: In the current study, 13,325 (mean age = 17.06, SD = 2.61) pediatric primary care patients from rural, semi-urban, and urban areas of Pennsylvania and 12,001 (mean age = 15.79, SD = 1.40) pediatric patients from an urban children's hospital emergency department were included in the analyses. We used two methods of ML (decision trees, random forests) to (a) generate algorithms to classify suicide attempt history, and (b) validate generated algorithms within and across samples to assess model performance. We also employed ridge regression to evaluate performance of the ML approaches.

Results: Our findings demonstrate that ML approaches did not enhance our ability to classify lifetime or recent suicide attempt history among youth across medical care settings, suggesting that relationships may be mainly linear and non-interactive. In line with prior research, a history of suicide planning, active suicidal ideation, passive suicidal ideation, and nonsuicidal self-injury emerged as relatively important correlates of suicide attempt.

Limitations: The cross-sectional nature of the current study prevents us from determining the extent to which the important variables identified confer risk for future suicidal behavior.

Conclusions: The present study underscores the importance of suicide risk screenings that focus on the assessment of active and passive suicidal ideation and suicide planning, in addition to nonsuicidal self-injury, across pediatric medical settings.

1. Introduction

Suicide is a significant public health concern, ranking as the second leading cause of death among children and adolescents aged 10–24 (Centers for Disease Control and Prevention, 2019). Nonfatal youth suicide attempts are also a significant public health problem, given that they are one of the most robust indicators of subsequent death by suicide (Franklin et al., 2017; Ribeiro et al., 2016). Indeed, rates of suicide and nonfatal suicide attempts have climbed drastically from 1999 to 2017 among youth (Hedegaard et al., 2017). Based on rising youth suicide rates, there has been a growing national interest in implementing screening for suicide in medical care settings (Horowitz et al., 2014; The Joint Commission on Accreditation of Healthcare Organizations; JCAHO, 2016; Williams et al., 2009). Medical settings offer a unique opportunity to broadly screen youth for suicide risk and refer those at heightened risk to prevention and intervention services. In fact, the Joint Commission on Accreditation of Healthcare Organizations recommends suicide surveillance and prevention with an emphasis on screening youth seen in all ambulatory and inpatient settings (JCAHO, 2016). However, screening efforts continue to be hampered by prediction limitations.

A recent meta-analysis suggested that despite mounting research on the predictors of nonfatal suicide attempts and suicide, our predictive accuracy of these behaviors has remained only slightly above chance levels (Franklin et al., 2017). Franklin et al. (2017) suggested that this stagnancy in our prediction of suicide may be due, in part, to our largely ineffective traditional statistical approach to prediction, which does not permit the testing of adequately complicated or complex

[*] Corresponding author.
E-mail address: taylor_burke@brown.edu (T.A. Burke).

https://doi.org/10.1016/j.jad.2020.02.048
Received 20 December 2019; Received in revised form 12 February 2020; Accepted 27 February 2020
Available online 28 February 2020
0165-0327/ Published by Elsevier B.V.

models. Such marginal levels of prediction have not informed clinical decision-making in a meaningful way.

Machine learning (ML) techniques may help researchers address the limitations of traditional statistical approaches (McArdle and Ritschard, 2014) and make progress in the face of rising suicide rates. Indeed, these methods permit the simultaneous testing of numerous factors and their complex interactions (McArdle and Ritschard, 2014). They allow for non-linearity in producing predictive algorithms, as opposed to imposing linearity on relationships. Additionally, ML techniques do not rely on researchers to specify hypothesized relationships (McArdle and Ritschard, 2014), but instead, in a blind and unbiased fashion, test every possible relationship to identify the superior performing algorithm. Such benefits particularly may be well suited for complex psychological phenomena, such as suicidality, as their occurrence are likely the result of numerous factors lacking linear relationships.

Investigators have begun to apply ML techniques to predict suicide, developing models that have augmented prediction well above chance levels (e.g., Kessler et al., 2017, 2015; Walsh et al., 2017). However, the application of ML to predict nonfatal suicide attempts remains limited. In a recent systematic review, it was found that limited studies have used ML techniques to predict suicide attempts and even fewer have focused on youth (Burke et al., 2019). For example, Bae et al. (2015) employed decision tree analysis to create an algorithm to classify past year history of nonfatal suicide attempts among a sample of 2754 Korean middle and high school students. Results suggested that the decision tree predictive model evidenced 90% classification accuracy. Findings demonstrated that depression severity was the primary predictive variable; however, unique predictors as well as interactions also were identified. Indeed, decision tree findings suggested that adolescents most likely to have attempted suicide within the past year were females with both high depression levels and frequent delinquent behavior (Bae et al., 2015). This study highlights the possible opportunity for ML to identify unique and complex relationships among predictor variables, and ultimately, enhance predictive accuracy.

1.1. The current study

The current study aimed to use ML techniques to develop algorithms that classify youth with recent and lifetime suicide attempt history based on their larger behavioral health symptom profile. This study addressed major research gaps in youth suicide prevention. Foremost, the current study is the first to apply ML to advance the detection of suicide risk in both pediatric primary care and emergency department settings, two entry points where providers have a unique opportunity to identify suicidal youth. The current study additionally aimed to assess the generalizability of the algorithms through cross-validation *within* both the emergency department and primary care samples and *across* these independent samples. This research design presents a unique opportunity to extend the science of ML research, as few to no ML studies have used a second, independent sample to validate models (for review, see Burke et al., 2019). It is hypothesized that our ML models will evidence superior performance in classifying suicide attempt history in both emergency and primary care pediatric health care settings compared to linear regression models. An exploratory aim was to identify the most important predictors, and interactions between predictors, of recent and lifetime suicide attempts among youth presenting to medical settings.

2. Method

2.1. Participants

Emergency department and primary care patients were eligible to complete the Behavioral Health Screen if they were physically able (not precluded based on illness acuity) and were between the ages of 14 and 24 years old. Approximately 13,325 youth (mean age = 17.06, SD = 2.61) were included in this sample from pediatric primary care settings from rural, semi-urban, and urban areas of Pennsylvania. Of the primary care sample, 55.3% were female, 14.9% identified as Hispanic, 71.1% as White, 9.2% as Black/African American, 0.6% as American Indian/Alaskan Native, 3.0% as Asian, 0.8% as Native Hawaiian/Other Pacific Islander, 8.4% as more than one race, and 6.4% as “not sure”. An additional 12,001 youth (mean age = 15.79, SD = 1.40) patients at the Children's Hospital of Philadelphia's emergency department were included in this analysis. Of the emergency department sample, 65.1% were female, 9.4% identified as Hispanic, 31.5% as White, 51.6% as Black/African American, 0.7% as American Indian/Alaskan Native, 2.5% as Asian, 0.5% as Native Hawaiian/Other Pacific Islander, and 10.4% as more than one race.

2.2. Procedures

Patients completed the Behavioral Health Screen (BHS; Diamond et al., 2010) electronically on laptops or tablets in the waiting or exam rooms prior to meeting with their medical provider. Patients were provided instructions stating that all information that they provided on the BHS would be confidential unless they reported harm to self or others. All patients give consent at the beginning of the screen to allow their clinic and partners (i.e., Drexel University) to use their de-identified data for research purposes. This method has been approved by the Drexel University IRB. The BHS software automatically scores the questionnaire and generates a clinical report that can download into the electronic medical record. With patient consent, referral sources can be given access to the report on an encrypted web site. The primary care version takes about 12 min to complete, whereas the emergency department version (fewer items; see Measures) takes about 7 min to complete. The medical staff uses the report to guide assessment and triage to mental health service, if indicated. None of the sites screened all patients. The regularity of BHS implementation evolved over time as providers valued the tool more. As a result, the current samples cannot be used as epidemiological measures of prevalence.

2.3. Measures

The Behavioral Health Screen (BHS; Diamond et al., 2010) is a well-validated electronic behavioral health risk screening tool. It assesses all the domains recommended as best practice by the American Academy of Pediatrics for a pediatric well visit (e.g., risk factors and psychiatric syndromes) (American Academy of Pediatrics Committee, 2014). Using 61 main items and 46 follow up items, the BHS covers 14 domains: demographics, medical, school, family, safety, substance use, sexual risk, nutrition and eating, anxiety, depression, suicide and self-harm, psychosis, trauma, bullying, and gun access. The tool has been modified and used in different service environments (e.g., schools, emergency department, crisis services, and residential treatment) and translated into several languages. The emergency department version only contains items that emergency room doctors would be compelled to address urgently, and not general psychosocial factors (e.g. school attendance) (see Fein et al., 2010). Implementation feasibility has been evaluated in primary care and emergency departments and is acceptable to medical providers, parents, and adolescents (Fein et al., 2010; Pailler et al., 2009). The BHS is psychometrically robust for adolescents (Bevans et al., 2012; Diamond et al., 2010; Fein et al., 2010; Pailler et al., 2009).

The BHS is deployed through a health science-based web technology platform developed by Medical Decision Logic, a health informatics software company in Baltimore (www.mdlogix.com). Therefore, the data can be aggregated for quality improvement and research reports. The tool has been operational for over 10 years in 14 states, with an aggregate database of over 100,000 patients. Data for this study were collected between 2008 and 2012. In order to compare the primary care

Table 1
Model indicators.

Model indicators
Age at screen
Gender
Hispanic
Race
Gun in home
Gun access
Lifetime tobacco use
Past 30 days tobacco frequency
Average cigarettes per day
Lifetime alcohol use
Past 30 days alcohol frequency
Lifetime marijuana use
Past 30 days marijuana frequency
Lifetime any other substance presence
Drugs/alcohol: interference in responsibilities
Drugs/alcohol: while driving car or bike
Drugs/alcohol: approached by police
Drugs/alcohol: interference in relationships
Past year depression
Past 2 weeks depression
Past year anhedonia
Past 2 weeks depression sx: anhedonia
Past 2 weeks depression sx: eating
Past 2 weeks depression sx: sleeping
Past 2 weeks depression sx: irritable
Past 2 weeks depression sx: fatigue
Past 2 weeks depression sx: difficulty making decisions
Past 2 weeks depression sx: lonely
Past 2 weeks depression sx: worthless
Past 2 weeks depression interference
Past year physical fight
Past year physical or sexual hurt by romantic partner
Lifetime sexual abuse
Past year sexual abuse
Lifetime physical or sexual abuse from someone in home
Past year physical or sexual abuse from someone in home
Past 2 weeks PTSD sx: nightmares or unwanted thoughts
Past 2 weeks PTSD sx: avoidance of thoughts
Past 2 weeks PTSD sx: on guard or easily startled
Past 2 weeks PTSD sx: numb or detached
During free time at school, how often do you spend time with friends or are you mostly alone?
How often do you feel kids tease you, make fun of you, or ignore you?
How often do kids physically hurt you or threaten to hurt you?
How often are you cyber bullied?
Current mental health treatment presence
Lifetime NSSI
Recent NSSI
Lifetime passive SI
Recent passive SI
Lifetime active SI
Recent active SI
Lifetime suicide plan
Recent suicide plan

and emergency department data sets, we only use overlapping items from both versions. Thus, a total of 53 BHS items covering the risk factor domains were used as indicators for the ML predictive models. For a full list of indicators, please see Table 1. The outcome variables were recent nonfatal suicide attempts (*"In the past week, including today, have you tried to kill yourself"*) and lifetime nonfatal suicide attempts (*"Have you ever tried to kill yourself"*) (Diamond et al., 2010).

2.4. Analytic plan

The current study employed two ML techniques (decision trees and random forests) in addition to an extension of linear regression, ridge regression. These three methods capture relationships that span from linear (ridge regression) to nonlinear with the inclusion of interactions (decision trees and random forests).

Ridge regression. *Ridge regression* is an extension of linear regression that includes a penalty for the size of coefficients (Hoerl and Kennard, 1970). In contrast to lasso regression (Tibshirani, 1996), which performs variable selection through penalization, ridge regression does not shrink the coefficients all the way to zero, with the aim of producing more generalizable estimates and reducing collinearity among the predictors. We did not include lasso regression as the large sample sizes make it unlikely to actually perform variable selection with this method. The ridge coefficients can be interpreted as standardized (albeit shrunken) coefficients. To run the models, we used the caret package (Kuhn, 2008) as a wrapper to use repeated cross-validation (10 repeats) and output the performance metrics, whereas the glmnet package (Friedman et al., 2010) ran the ridge regression models. We used 100 penalty values, ranging from 0.001 to 0.32, with checking the best fitting penalties to make sure they were not at the extremes of this range. To make the results more comparable to those from random forests (detailed below), we report a variable importance metric, calculated based on the magnitude of the absolute standardized coefficients[1] (according to the model with the best fitting ridge penalty). The large size of the datasets precluded the use of more informative variable importance metrics for ridge regression (Grömping, 2006).

Decision trees. Decision trees were developed to explore the dynamics between predictors and an outcome variable in large datasets, without imposing an explanatory structure between predictors and outcomes. A generalization of earlier tree-based methods (Morgan and Sonquist, 1963), *classification and regression trees* (CART; Breiman et al., 1984) is an inductive, sequential partitioning model that allows researchers to identify non-linearities, non-additive relationships, and to automatically identify interactions of many factors in the prediction of an outcome. Each model included all 53 indicators as possible splits for selection in the analyses.

Random forests. *Random forests* (Breiman, 2001) were employed as a complementary ML approach utilizing the same 53 indicators. Random forests was developed to overcome the disadvantages of single decision tree algorithms, such as variable selection bias, parameter instability, and over-fitting. Random forests generates a large number (e.g., 100, 1000) of decision trees by using a bootstrap (or subset) of the sample and a subset of the predictors to create each individual tree. After trees are generated, predictions are aggregated (by majority, for categorical outcomes) across each created tree. Variable importance estimates are calculated based on the total decrease in node impurities (measured by the Gini index; e.g., how much each variable decreases misfit) from splitting on the variable, averaged across all trees. These quantify the marginal effect of each variable (Grömping, 2009), while other methods exist that include the conditional effects as well (Strobl et al., 2008). The caret package as well as the randomForest package (Wiener, 2003) were used to run the random forests models and to enable testing of model fit on 20 bootstrap samples. Although random forests is advantageous with respect to performance in comparison to decision trees, it is much less interpretable than a single decision tree generated from CART or than interpreting a single coefficient from ridge regression. Thus, including two ML methods and a linear regression method allowed us to benefit from the advantages of each of these approaches as well as to compare outcomes across approaches. In using this model comparison approach, decision trees is used mostly for inferential purposes, and we would expect random forests to outperform ridge regression if there are interactions or other nonlinear relationships in the data.

Data management. Given the missingness in the data, and the requirement of complete cases for some ML software packages, we used multiple imputation to create multiple complete datasets. We used the

[1] There are limitations to this approach (Grömping, 2006). However, other importance metrics were too computationally demanding to be run given the size of our data.

mice package (Buuren and Groothuis-Oudshoorn, 2011) to facilitate filling in the missingness using decision trees, as this method better preserves any nonlinearities in the data, if they exist (Carrig et al., 2015). Additionally, as there are no guidelines for how many imputations to create when using ML methods, we created 200 imputed datasets.

Before carrying out each of the statistical methods (and after imputing), the samples were split into training data (75%) and testing data (25%). The testing data were saved to determine model performance using predictions from the best fitting model derived from the training data. Testing was carried out in three steps: 1) Training was carried out on the emergency department (ED) training data and tested on the ED testing data, 2) Training was carried out on the primary care (PC) training data and tested on the PC testing data, 3) The model trained on the PC training data was tested on the ED testing data. This resulted in three sets of model performance metrics (i.e., ED–ED, PC–PC, and PC–ED). However, variable importance is reported only from the two training samples (i.e., ED–ED and PC–PC).

To select among the hyperparameters for each of the methods, we used the F1 score[2] (the harmonic mean of precision and recall) as evaluated with repeated cross-validation or bootstrap sampling. After selecting final models to be tested on the test sample, we used the area under the precision recall curve (AUPRC; Davis and Goadrich, 2006) for comparison across algorithms. In comparison to the area under the receiver operating characteristic curve (AUC), the AUPRC is a more sensitive and realistic assessment when classes are imbalanced (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015). AUPRC does not take into account True Negatives, instead focusing on the classification of Positives (those with lifetime or recent suicide attempts). With the AUC, a value of 0.50 denotes chance level prediction. However, with the AUPRC, chance depends on the distribution of the outcome, with minimum values ranging from 0 to 0.10 depending on how imbalanced the outcome is. We report the AUPRC as calculated on the test sample, as well as accuracy, sensitivity, specificity, F1 score, precision, recall, and AUC.[3] Note that accuracy and specificity will be high, purely as a result of the class imbalance. As one strategy to overcome the class imbalance, we tested using a synthetic minority over-sampling technique (SMOTE; Chawla et al., 2002), paired with random forests. However, this did not improve model performance; therefore, we do not report these results.

3. Results

Among the ED sample, 9.3% of youth ($n = 1113$) reported a lifetime suicide attempt and 1.7% ($n = 209$) reported a recent (past week) suicide attempt. Among the PC sample, 4.6% of youth ($n = 608$) reported a lifetime suicide attempt and 0.3% ($n = 39$) reported a recent (past week) suicide attempt.

The model performance metrics are displayed in Tables 2 and 3. Note that we do not display results from the decision trees (i.e., CART). The tree models produced overly simplistic models (only two splits in most of the analyses) and evidenced performance far below those of ridge regression or random forests. In comparing Tables 2 and 3, the performance across methods was very similar. This can be interpreted that the majority of effects were linear, as the existence of interaction and nonlinear effects would result in higher performance for random forests. As such, the evaluation of variable importance (below) mostly will focus on the main effects in the model.

Given the class imbalance, which skews the interpretation of both accuracy and AUC (towards thinking our models did better), we devote most of our performance interpretation to AUPRC. Although our AUPRC values were not close to one, denoting less than optimal performance, in the case of classifying lifetime suicide attempts, training a model on PC youth did not result in much performance loss when evaluated on ED youth. This was not true for the classification of recent suicide attempts, with AUPRC values close to 0.1, which is lower than as trained and evaluated within the same patient population. The lower AUPRC values speak more to the ability of the algorithms to classify those with a suicide attempt history as having high estimated probabilities. More simply, few respondents were classified as having a high probability of having a lifetime or recent suicide attempt (estimated probabilities > 0.8). For example, in using the ED data to classify lifetime suicide attempts, the 75th percentile for estimated probabilities, as assessed on the ED test set, was 0.03. Of those cases that did receive high estimated probabilities, most had a lifetime history of suicide attempt. Fig. 1 displays the relationship between binned estimated probabilities from ridge regression and the actual proportion of ED youth with a lifetime suicide attempt history. Additionally, as denoted by the AUPRC, those with a history of lifetime or recent attempts tended to be assigned higher probabilities of suicide attempt per the models, as opposed to those without a history. This can be seen in the density plots from the ED trained-ED test evaluation in Supplementary Figure 1.

3.1. Variable importance

The variable importance metrics are displayed in Tables 4 and 5 for random forests and ridge regression, respectively. Of the 53 predictors, these tables were set to include those with non-negligible effects, which was quantified differently across methods due to differences in calculating importance. Specifically, only variables that had importance scores above 10 were presented for random forests (Table 4) and only variables that had importance scores above 5 were presented for ridge regression (Table 5). Generally, across methods, a history of active and passive suicidal ideation, suicide planning, and nonsuicidal self-injury emerged as important in classifying suicide attempt history. Given that the performance was similar across ridge and random forests, it is worth devoting more time to interpreting the results from ridge, as this was the least complex model. For the ridge results, we observed a large degree of variability in some importance values across imputations (SD), and across outcomes.

4. Discussion

The current study utilized ML techniques to classify recent (past week) and lifetime suicide attempt history among youth presenting to the emergency department and to primary care clinics. In considering results across methods, our hypothesis that ML models would augment classification beyond traditional linear regression methods was not supported. Models fit similarly, regardless of utilizing an extension of linear regression (i.e., ridge regression) or ML (i.e., random forests) methods; further, CART produced less accurate models than both ridge regression and random forests. The resultant CART trees only contained a limited number of splits (1–4), which resulted in only small increases in performance above and beyond chance. It is likely that ML methods did not increase overall predictive validity above chance and traditional linear regression methods because of the predominance of linear relationships between our predictors and suicide attempt outcomes, in

[2] We had originally planned on using the AUPRC to select among hyperparameters. However, this produced weird performance and tended to select suboptimal models with respect to precision and recall.

[3] For each of these metrics, estimated probabilities of class membership must be converted into class labels, based on a cutoff value. Although 0.5 is traditionally used, this would be inappropriate given the low endorsement of lifetime and recent suicidal ideation. We calculated the optimal cutoff on the test sample using the ROCR package (Sing, Sander, Beerenwinkel, & Lengauer, 2005) and the cutoff that corresponded to the highest F1 score. If our goal was the creation of a model to be used in an external setting for the purposes of screening, it would be more appropriate to calculate the optimal cutoff on the training sample, and then use this cutoff to calculate model performance on a test sample.

Table 2
Random forests model performance across the three sets of evaluations.

	Lifetime							
	AUPRC	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	AUC
ED	0.602	0.921	0.699	0.944	0.569	0.699	0.624	0.944
PC	0.602	0.962	0.700	0.974	0.553	0.700	0.618	0.973
PC -> ED	0.604	0.963	0.697	0.975	0.557	0.697	0.618	0.937
	Recent							
	AUPRC	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	AUC
ED	0.501	0.980	0.642	0.987	0.506	0.642	0.565	0.975
PC	0.563	0.996	0.520	0.999	0.693	0.520	0.582	0.969
PC -> ED	0.403	0.994	0.746	0.995	0.433	0.746	0.543	0.955

Table 3
Ridge regression performance across the three sets of evaluations.

	Lifetime							
	AUPRC	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	AUC
ED	0.626	0.914	0.716	0.934	0.533	0.716	0.610	0.944
PC	0.585	0.963	0.764	0.972	0.559	0.764	0.645	0.968
PC -> ED	0.602	0.954	0.822	0.960	0.489	0.822	0.610	0.946
	Recent							
	AUPRC	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	AUC
ED	0.507	0.980	0.665	0.987	0.511	0.665	0.575	0.984
PC	0.595	0.997	0.530	0.999	0.720	0.530	0.606	0.991
PC -> ED	0.494	0.994	0.706	0.996	0.460	0.706	0.555	0.975

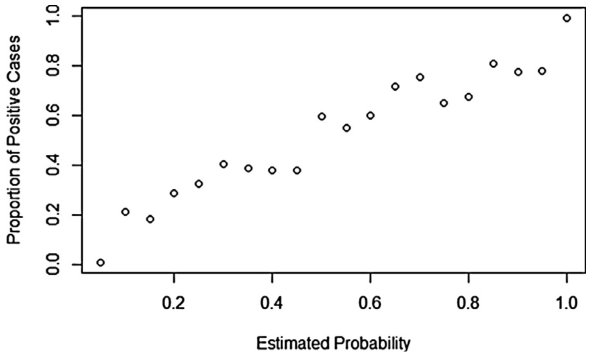


Fig. 1. Averaged proportions (across imputations) of positive cases (lifetime) were binned (20 bins) in the ED-ED test set evaluation with ridge regression Note. In accordance with performance metrics that denote that performance was better than chance, as the model's estimated probabilities increase, so do the proportions of positive cases.

Table 4
Random forests variable importance.

Variable	ED				PC			
	Recent		Lifetime		Recent		Lifetime	
	M	SD	M	SD	M	SD	M	SD
Age	15.6	1.7	14.8	0.3	54.3	3.9	15.8	0.3
Race	12.9	2.0	11.3	0.3	22.1	2.9	12.4	0.3
Tobacco frequency – Past 30 days					12.4	2.3	5.2	0.2
Average cigarettes per day					19.5	4.3	6.4	0.6
Alcohol frequency – Past 30 days					47.8	6.6	6.7	0.2
Marijuana frequency – Past 30 days					11.0	3.5	3.5	0.2
Drugs/alcohol interference in responsibilities					13.2	2.4	0.5	0.1
Anhedonia – Past 2 wks					11.1	1.4	5.4	0.2
Irritability – Past 2 wks					11.4	2.7	5.1	0.2
Sexual abuse – Past Yr					15.9	3.8	0.2	0.1
Physical/sexual abuse someone in home – Life					14.9	2.3	1.1	0.1
Physical/sexual abuse someone in home – Past Yr					11.5	2.6	0	0.1
Nightmares or unwanted thoughts – Past 2 wks					10.3	1.4	1.5	0.1
Eating – Past 2 wks	10.3	1.1	8.3	0.2				
Friend connectedness at school	10.9	1.6	8.6	0.2				
Bullied (teased/made fun of/ ignored)					18.9	6.1	5.8	0.3
Bullied (physically hurt or threatened)					11.9	4.7	1.8	0.2
Cyberbullied					23.6	6.9	3.1	0.3
NSSI – Life	2.9	0.3	17	0.8	4.0	1.0	21.5	0.9
NSSI – Rec	10.5	2.5	2.0	0.1	13.1	4.0	0.7	0.1
Passive SI – Life	1.1	0.1	15.8	1.1	0.3	0.3	28.0	1.3
Passive SI – Rec	0.1	0.1	51.6	2.3	14.4	7.1	2.0	0.2
Active SI – Life	21.1	12.5	3.8	0.2	0.6	0.6	100	0
Active SI – Rec	3.7	2.1	100	0	100	0	0.3	0.1

Note: Subset to just those variables with at least one importance value greater than 10 in any mean column. SD refers to the standard deviation of the importance values across imputed datasets.

addition to the lack of interactions between predictors. Generally, we found better prediction accuracy in classifying lifetime suicide attempt history as opposed to recent suicide attempt history. Across models, and across both emergency department and primary care settings, several important variables in differentiating those with and without suicide attempt history were identified. We found that a history of active and passive suicidal ideation, suicide planning, and nonsuicidal self-injury emerged as important indicators associated with suicide attempt history among youth.

Across both ridge regression and random forests, model performance was similar. Similar performance across these methods suggests that the relationships between the included predictors and recent/ lifetime suicide attempt history consist of primarily linear effects (Hong et al., 2020, in prep). That is, it appears that the variables' main effects (i.e., recent, active suicidal ideation; past, active suicidal ideation) may be prominent drivers in the predictive ability, as opposed to their interaction (i.e., younger age and recent, active suicidal ideation). Although there are limitations to this conclusion, we base this on the inability of ridge regression to model interactions or nonlinear effects (if not manually entered). The presence of largely linear relationships is surprising given the inherently complex nature of suicidal behavior. Indeed, previous research has demonstrated the importance of

Table 5
Ridge regression variable importance.

Variable	ED				PC			
	Recent		Lifetime		Recent		Lifetime	
	M	SD	M	SD	M	SD	M	SD
Gender (Transgender-Female to Male)	8.1	3.5	22.5	4.1				
Gender (Transgender-Male to Female)	56.7	6.6	30.8	7.1				
Race (American Indian/ Alaskan Native)	8.8	3.1	40.0	4.6	7.0	1.6	48.8	1.5
Race (Native Hawaiian/Other Pacific Islander)	8.4	10.6	35.4	8.0	2.3	3.8	38.6	2.7
Other substance use – Life	13.5	1.4	26.4	0.5				
Drugs/alcohol while driving car or bike	30.2	1.3	12.0	1.0				
Sexual abuse – Life	4.1	1.7	28.8	1.0				
Sexual abuse – Past Yr	24.9	2.3	12.8	1.9				
Physical/aexual abuse someone in home – Life	22.3	0.7	26.6	0.8				
Physical/aexual abuse someone in home – Past Yr					39.5	2.2	37.5	5.8
Bullied (physically hurt or threatened) – Often	19.5	1.2	34.0	0.9				
Cyberbullied – Often	22.1	2.2	17.7	1.2				
Psych tx – Current					2.5	1.0	22.6	2.1
NSSI – Life	6.9	0.7	48.5	0.5	5.6	0.5	58.5	0.8
NSSI – Rec	34.9	1.3	8.3	0.8	35.8	3.0	7.1	2.2
Passive SI – Life	7.5	0.4	45.6	0.4	3.8	0.2	34.3	0.5
Passive SI – Rec	26.5	0.8	5.6	0.6				
Active SI – Life	30.8	1.1	100	0	6.3	1.8	72.3	0.8
Active SI – Rec	87.6	2.1	3.9	0.7	35.1	3.6	15.1	1.8
Suicide Plan – Life	1.2	0.8	93.1	0.6	7.9	1.7	100	0
Suicide Plan – Rec	100	0.2	8.9	1.1	100	0	14.6	2.4

Note: Subset to just those variables with at least one importance value greater than 20 in any mean column. Polytomous variables were dummy coded prior to analysis. These variables are denoted with the category coded as one, with all other categories serving as a comparison. Dichotomous variables are coded 1 for Yes. SD refers to the standard deviation of the importance values across imputed datasets.

interactive effects in predicting suicide (Ilgen et al., 2009) and suicide attempt (Bae et al., 2015) history. However, it is possible this discrepancy may be related to the nature of included predictors. For example, Ilgen et al. (2009) found an interaction between race, substance use disorder diagnosis, and past year psychiatric hospitalization. These predictors differ from predictors in the present study, which focus more predominately on current distress. As such, it may be possible that ML techniques are most advantageous when the aim is to examine the interaction of distal and proximal risk factors. The present study also differs from previous research, particularly the study conducted by Bae and colleagues utilizing ML among a youth sample (2015), in that the current algorithms were trained using individual items from the Behavioral Health Screen, in contrast to scale summed scores. Further still, it is possible that nonlinear relationships might be more likely to be found when including psychological constructs previously found to be important in classifying suicide risk (e.g., hopelessness, distress tolerance, emotion regulation, perceived burdensomeness) and supported by numerous theories of suicide (e.g., Joiner, 2005). Additionally, the present sample was selected from medical care settings, which may be intrinsically different than a nationally representative sample of youth. Study-specific conjectures aside, it also may be possible that machine learning simply does not offer advantages consistently, even when predicting complex conditions. Indeed, the current findings are in line with those from a recent systematic review that concluded that the use of machine learning did not demonstrate superior prediction accuracy as compared to logistic regression when considering the prediction of a range of comparably complex clinical conditions (e.g., diabetes) (Christodoulou et al., 2019).

As an exploratory aim, the current study also considered the most important predictors in classifying recent versus lifetime suicide attempt history. We largely highlight predictors that demonstrated high importance values across both ridge regression and random forest results to ensure increased confidence in, and generalizability of, our findings. Consistent across models, and largely across samples (emergency department and primary care), was the identification of suicidal ideation and suicidal planning as important predictors. These findings are highly consistent with research demonstrating prior suicidal thoughts and planning as moderate to strong predictors of future suicide behavior (Ribeiro et al., 2016). The current study also highlights a differing importance in the presence of active (e.g., a desire to kill oneself) versus passive suicidal ideation (e.g., feeling like life is not worth living, a desire to no longer be alive). Across models, both recent and lifetime active suicidal ideation demonstrated greater importance in classifying suicide attempt history as compared to passive suicidal ideation. Although some previous research has supported the role of passive ideation, defined as a general desire for death as opposed to actively thinking about taking one's own life, in the prediction of suicide attempts (e.g., Baca-Garcia et al., 2011), current findings underscore the potential greater importance of assessing for active ideation specifically. As such, the present study emphasizes the importance of recent calls for increased screening of suicidal ideation (JCAHO, 2016), which may be of most benefit when including both active *and* passive ideation, among adolescents across healthcare settings.

Beyond the presence of suicidal ideation and suicidal planning, nonsuicidal self-injury also was an important predictor in classifying suicide attempt history across models. Consistent with prior theoretical (Hamza et al., 2012) and empirical work (Franklin et al., 2017; Ribeiro et al., 2016), the current study supports the link between nonsuicidal self-injury engagement and suicidal behavior, even as it occurs across the lifetime. Indeed, there is an increased need for clinicians and medical providers to screen for the presence of nonsuicidal self-injury particularly among youth, as the peak age of onset for the behavior is 13–14 years old (Ammerman et al., 2018).

Ridge regression and random forest models also indicated the relevance of other factors in the classification of suicide attempt history. Although these findings should be interpreted with caution until further replicated, a variety of adverse interpersonal experiences, including abuse at home (physical, sexual) and bullying/cyber bullying from peers, were demonstrated as important in some of the current models, associations that have been supported in prior research (Dube et al., 2001; Klomek et al., 2010). Further, substance use (alcohol, drug use) and associated risky behavior (i.e., behaviors engaged in under the influence) were highlighted as important across several of the models in identifying those with a history of suicide attempt. Substance use previously has been found to be a strong correlate and predictor of suicidal behavior (Darvishi et al., 2015; Mars et al., 2019). Findings also indicated that age and belonging to an underrepresented group may be important in classifying recent and lifetime suicide attempt history. This is consistent with literature demonstrating elevated suicide risk among transgendered individuals (e.g., Kuper et al., 2018) and individuals identifying with minority racial groups, American Indian, Alaskan Natives, and Pacific Islanders (e.g., Strayer et al., 2014; Wong et al., 2012). As such, these findings provide support to consider screening for and incorporating specific demographic and interpersonal factors in the assessment and quantification of suicide risk. Broad band screening tools such as the one used in this study (Diamond et al., 2010), query about these psychosocial risk factors to help providers assess level of risk.

Surprisingly, the developed algorithms evidenced better classification accuracy for the classification of lifetime suicide attempts as opposed to recent suicide attempts. Although meta-analyses of longitudinal studies predicting suicide attempts have demonstrated inconsistent to negligible effects of timeframe in prediction (Franklin et al., 2017; Ribeiro et al., 2016), a recent study in a medical setting highlighted increased predictive accuracy for more recent suicidal behaviors (Walsh et al., 2017). It is possible our finding is due to a decreased target sample size of individuals with a recent suicide attempt as opposed to a lifetime history of attempt. However, we focused on model performance metrics designed to help correct biases due to class imbalance. Nevertheless, future research conducted in a sample with a lower class imbalance is needed to ascertain whether the relatively low proportion of patients endorsing a recent suicide attempt in our ED and PC samples could be driving results.

We examined the performance of our models across an independent dataset, which is an important, although often overlooked, aspect of evaluating ML techniques (Efron, 2014). Indeed, one of the main concerns when generating algorithms is that overfitting will occur that will threaten generalizability, even if good performance is found on a holdout testing sample. Thus, a stringent test of overfitting is conducted by testing model fit on an (entirely) independent sample. We found that the emergency department model performed just as well on the primary care test sample as it did on the emergency department test sample when classifying lifetime suicide attempt history, suggesting that our model did not overfit the data. However, we found significantly worse performance of the emergency department model on the primary care test sample when classifying recent suicide attempts, suggesting model overfitting or that different factors are related to the outcome across both samples. It is possible that the setting itself may have contributed to these findings, where the emergency department may “pull for” positive answers on the screener (i.e., admissions of higher distress) than the primary care setting simply due to the more acute nature of emergency departments. On the other hand, this difference may have occurred due to the comparatively (and absolutely) low number of positive cases of recent suicide attempts in primary care as compared to the emergency department numbers. The number of positive cases can be depicted as the effective sample size (Vergouwe et al., 2005) and is related to the power of an algorithm to differentiate between classes. Simply put, the emergency department setting had higher power to differentiate between attempters and non-attempters. Given that recent suicide attempt history was measured over the previous one week, this outcome was rare in both the emergency department and primary care datasets. Thus, it is plausible that this same reasoning may explain the consistent finding that lifetime suicide attempt models were superior to recent suicide attempt models.

4.1. Limitations

Limitations of this study must be considered. First, the cross-sectional nature of the current study prevents us from determining the extent to which the important predictors identified confer risk for future suicidal behavior; thus, the findings must be replicated in a longitudinal design prior to informing screening. Utilizing the BHS was valuable in allowing us to capture a wide range of behavioral health indicators. It is surprising however that the rich clinical data from the BHS did not permit greater performance in classifying suicide attempt history. The BHS assesses many of the variables that have historically been found to be associated with high risk for suicidal behavior (e.g., substance use, depression, abuse), and previous studies employing the BHS have found that the assessed symptoms do differentiate high and low suicide risk profiles (Diamond et al., 2017; Herres et al., 2018). Nonetheless, given that the performance metrics of the current models were not optimal, there is much left to be explained in our outcomes of lifetime and recent suicide attempts. Future research may find that the BHS is most useful for informing suicide risk classification if used in conjunction with electronic medical record (EMR) data, which has been shown to be promising in the prediction of suicidal behavior when used with ML approaches (e.g., Walsh et al., 2017). To date, over 100,000 patients in 14 United States have been screened with the BHS, and thus, if paired with EMR data, may allow for a rich, diverse dataset for applying and benefiting from the advantages of ML to augment the prospective prediction of youth suicide risk. An additional limitation is that the current data were collected via self-report only. As with all methods of self-report sensitive data collection, it is possible that youth hesitation to disclose personal information, particularly related to abuse and drug use (e.g., Delaney-Black et al., 2010; Paine and Hansen, 2002), may have impacted our findings. Despite this potential limitation, it is important to acknowledge that prior research has indicated that patients may prefer providing sensitive information via self-report as compared to face-to-face interviews with clinicians (Kurth et al., 2004). Nevertheless, future research should consider whether attaining collateral report (e.g., caregiver-report) might enhance risk detection. A final limitation worth noting is that the importance of any given variable in the classification of suicide attempts must be interpreted within the context of the model. That is, the model results are dependent on all variables in the model; consequently, removing a variable of high importance may impact the importance of all subsequent variables. For example, it is possible that inclusion of strong correlates of suicide attempts, such as suicidal ideation, caused other BHS psychosocial risk variables to not emerge as important. Future research employing this dataset might consider examining suicide attempt classification without including suicide-related variables (e.g., suicidal ideation, suicide planning) to explore this possibility.

Despite these limitations, the current study had a number of significant strengths. It benefitted from the inclusion of two distinct medical setting samples of diverse community youth, which allowed us to examine the within and cross-sample validity of our models. The sample sizes were large, a significant advantage when studying relatively rare outcomes. Further, we employed ML methods to determine whether these approaches would enhance classification of history of nonfatal suicide attempts; these methods rarely have been employed in youth samples (Burke et al., 2019). Also, we increased the power of our study by imputing missing data as opposed to performing listwise deletion. As most ML methods cannot incorporate missing data, we imputed 200 datasets, using two different imputation algorithms, and then averaged the results. The large number of imputations should eliminate any possibility of systematic bias in our results.

The present study underscores the importance of suicide risk screenings that focus on the assessment of active and passive suicidal ideation and suicide planning, in addition to nonsuicidal self-injury, across pediatric medical settings. Given the importance of these variables in classifying suicide attempt history, assessing their lifetime and recent history may help medical staff in early detection of patients at risk for suicide in these clinical settings. Medical staff may then consider utilizing the additional BHS psychosocial risk variables in order to make appropriate disposition decisions.

CRediT authorship contribution statement

Taylor A. Burke: Conceptualization, Formal analysis, Writing - original draft. **Ross Jacobucci:** Conceptualization, Formal analysis, Writing - original draft. **Brooke A. Ammerman:** Conceptualization, Formal analysis, Writing - original draft. **Lauren B. Alloy:** Conceptualization, Writing - original draft. **Guy Diamond:** Conceptualization, Data curation, Writing - original draft.

Declaration of competing interest

Dr. Taylor A. Burke, Dr. Ross Jacobucci, Dr. Brooke A. Ammerman, and Dr. Lauren Alloy report no conflicts of interest. Dr. Guy Diamond is the lead developer of the Behavioral Health Screen. The Children's Hospital of Philadelphia owns the tool and licenses it to Medical Decision Logic for distribution. Dr. Diamond will get a small royalty if and when the company makes a profit from the tool.

Acknowledgements

This research was supported by a University of Michigan James N.

Morgan Fund grant to Taylor A. Burke. The funding source had no role in study design, in the collection, analysis and interpretation of data, in the writing of the report, or in the decision to submit the article for publication. Taylor A. Burke was supported by NIMH T32 grant MH019927. Lauren B. Alloy was supported by MH079369 and MH101168.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2020.02.048.

References

American Academy of Pediatrics Committee on Practice and Ambulatory Medicine - Bright Futures Periodicity Schedule, 2014. Recommendations for pediatric preventive health care. Pediatrics 133, 568–570.

Ammerman, B.A., Jacobucci, R., Kleiman, E.M., Uyeji, L.L., McCloskey, M.S., 2018. The relationship between nonsuicidal self-injury age of onset and severity of self-harm. Suicide Life-Threat. Behav. 48, 31–37. https://doi.org/10.1111/sltb.12330.

Baca-Garcia, E., Perez-Rodriguez, M.M., Oquendo, M.A., Keyes, K.M., Hasin, D.S., Grant, B.F., Blanco, C., 2011. Estimating risk for suicide attempt: are we asking the right questions?: Passive suicidal ideation as a marker for suicidal behavior. J. Affect. Disord. 134, 327–332. https://doi.org/10.1016/j.jad.2011.06.026.

Bae, S.M., Lee, S.A., Lee, S.H., 2015. Prediction by data mining, of suicide attempts in Korean adolescents: a national study. Neuropsychiatr. Dis. Treat. 11, 2367–2375. https://doi.org/10.2147/NDT.S91111.

Bevans, K.B., Diamond, G., Levy, S., 2012. Screening for adolescents' internalizing symptoms in primary care: item response theory analysis of the behavior health screen depression, anxiety, and suicidal risk scales. J. Dev. Behav. Pediatr. 33, 283–290. https://doi.org/10.1097/DBP.0b013e31824eaa9a.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Burke, T.A., Ammerman, B.A., Jacobucci, R., 2019. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: a systematic review. J. Affect. Disord. https://doi.org/10.1016/j.jad.2018.11.073.

Buuren, S.V., Groothuis-Oudshoorn, K., 2011. Mice:multivariate imputation by chained equations in R. J. Stat. Softw. 45. https://doi.org/10.18637/jss.v045.i03.

Carrig, M.M., Manrique-Vallier, D., Ranby, K.W., Reiter, J.P., Hoyle, R.H., 2015. A nonparametric, multiple imputation-based method for the retrospective integration of data sets. Multivar. Behav. Res. 50, 383–397. https://doi.org/10.1080/00273171.2015.1022641.

Centers for Disease Control and Prevention, 2019. Web-based injury statistics query and reporting system (WISQARS) [WWW Document]. URL https://www.cdc.gov/injury/wisqars/LeadingCauses.html.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., Van Calster, B., 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J. Clin. Epidemiol. https://doi.org/10.1016/j.jclinepi.2019.02.004.

Darvishi, N., Farhadi, M., Haghtalab, T., Poorolajal, J., 2015. Alcohol-related risk of suicidal ideation, suicide attempt, and completed suicide: a meta-analysis. PLoS One. https://doi.org/10.1371/journal.pone.0126870.

Davis, J., Goadrich, M., 2006. The relationship between precision-recall and ROC curves. In: ACM International Conference Proceeding Series, . https://doi.org/10.1145/1143844.1143874.

Delaney-Black, V., Chiodo, L.M., Hannigan, J.H., Greenwald, M.K., Janisse, J., Patterson, G., Huestis, M.A., Ager, J., Sokol, R.J., 2010. Just say "I don't": lack of concordance between teen report and biological measures of drug use. Pediatrics 126, 887–893. https://doi.org/10.1542/peds.2009-3059.

Diamond, G., Levy, S., Bevans, K.B., Fein, J.A., Wintersteen, M.B., Tien, A., Creed, T., 2010. Development, validation, and utility of internet-based, behavioral health screen for adolescents. Pediatrics 126, 163–170. https://doi.org/10.1542/peds.2009-3272.

Diamond, G.S., Herres, J.L., Ewing, Krauthamer, E.S., Atte, T.O., Scott, S.W., Wintersteen, M.B., Gallop, 2017. Comprehensive screening for suicide risk in primary care. Am. J. Prev. Med. 53, 48–54. https://doi.org/10.1016/j.amepre.2017.02.020.

Dube, S.R., Anda, R.F., Felitti, V.J., Chapman, D.P., Williamson, D.F., Giles, W.H., 2001. Childhood abuse, household dysfunction, and the risk of attempted suicide throughout the life span: findings from the adverse childhood experiences study. J. Am. Med. Assoc. 286, 3089–3096. https://doi.org/10.1001/jama.286.24.3089.

Efron, B., 2014. Estimation and accuracy after model selection. J. Am. Stat. Assoc. https://doi.org/10.1080/01621459.2013.823775.

Fein, J.A., Pailler, M.E., Barg, F.K., Wintersteen, M.B., Hayes, K., Tien, A.Y., Diamond, G.S., 2010. Feasibility and effects of a web-based adolescent psychiatric assessment administered by clinical staff in the Pediatric Emergency Department. Arch. Pediatr. Adolesc. Med. 164, 1112–1117. https://doi.org/10.1001/archpediatrics.2010.213.

Franklin, J.C., Ribeiro, J.D., Fox, K.R., Bentley, K.H., Kleiman, E.M., Huang, X., Musacchio, K.M., Jaroszewski, A.C., Chang, B.P., Nock, M.K., 2017. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol. Bull. https://doi.org/10.1037/bul0000084.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33. https://doi.org/10.18637/jss.v033.i01.

Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. Am. Stat. 63, 308–319. https://doi.org/10.1198/tast.2009.08199.

Grömping, U., 2006. Relative importance for linear regression in *r*:the package relaimpo. J. Stat. Softw. https://doi.org/10.18637/jss.v017.i01.

Hamza, C.A., Stewart, S.L., Willoughby, T., 2012. Examining the link between nonsuicidal self-injury and suicidal behavior: a review of the literature and an integrated model. Clin. Psychol. Rev. https://doi.org/10.1016/j.cpr.2012.05.003.

Hedegaard, H., Curtin, S.C., & Warner, M., 2017. Suicide mortality in the United States, 1999–2017 [WWW Document]. URL https://www.cdc.gov/nchs/products/databriefs/db330.htm.

Herres, J., Kodish, T., Fein, J., Diamond, G., 2018. Screening to identify groups of pediatric emergency department patients using latent class analysis of reported suicidal ideation and behavior and non-suicidal self-Injury. Arch. Suicide Res. 22, 20–31. https://doi.org/10.1080/13811118.2017.1283264.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Hong, M., Jacobucci, R., & Lubke, G., 2020 Deductive data mining.

Horowitz, L.M., Bridge, J.A., Pao, M., Boudreaux, E.D., 2014. Screening youth for suicide risk in medical settings: time to ask questions. Am. J. Prev. Med. https://doi.org/10.1016/j.amepre.2014.06.002.

Ilgen, M.A., Downing, K., Zivin, K., Hoggatt, K.J., Kim, H.M., Ganoczy, D., Austin, K.L., McCarthy, J.F., Patel, J.M., Valenstein, M., 2009. Exploratory data mining analysis identifying subgroups of patients with depression who are at high risk for suicide. J. Clin. Psychiatry 70, 1495–1500. https://doi.org/10.4088/JCP.08m04795.

Kessler, R.C., Stein, M.B., Petukhova, M.V., Bliese, P., Bossarte, R.M., Bromet, E.J., Fullerton, C.S., Gilman, S.E., Ivany, C., Lewandowski-Romps, L., Millikan Bell, A., Naifeh, J.A., Nock, M.K., Reis, B.Y., Rosellini, A.J., Sampson, N.A., Zaslavsky, A.M., Ursano, R.J., 2017. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Mol. Psychiatry 22, 544–551. https://doi.org/10.1038/mp.2016.110.

Kessler, R.C., Warner, C.H., Ivany, C., Petukhova, M.V., Rose, S., Bromet, E.J., Brown, M., Cai, T., Colpe, L.J., Cox, K.L., Fullerton, C.S., Gilman, S.E., Gruber, M.J., Heeringa, S.G., Lewandowski-Romps, L., Li, J., Millikan-Bell, A.M., Naifeh, J.A., Nock, M.K., Rosellini, A.J., Sampson, N.A., Schoenbaum, M., Stein, M.B., Wessely, S., Zaslavsky, A.M., Ursano, R.J., 2015. Predicting suicides after psychiatric hospitalization in US army soldiers: the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). JAMA Psychiatry 72, 49–57. https://doi.org/10.1001/jamapsychiatry.2014.1754.

Klomek, A.B., Sourander, A., Gould, M., 2010. The association of suicide and bullying in childhood to young adulthood: a review of cross-sectional and longitudinal research findings. Can. J. Psychiatry. https://doi.org/10.1177/070674371005500503.

Kuhn, M., 2008. caret package. J. Stat. Softw.

Kuper, L.E., Adams, N., Mustanski, B.S., 2018. Exploring cross-sectional predictors of suicide ideation, attempt, and risk in a large online sample of transgender and gender nonconforming youth and young adults. LGBT Heal. 5, 391–400. https://doi.org/10.1089/lgbt.2017.0259.

Kurth, A.E., Martin, D.P., Golden, M.R., Weiss, N.S., Heagerty, P.J., Spielberg, F., Handsfield, H.H., Holmes, K.K., 2004. A comparison between audio computer-assisted self-interviews and clinician interviews for obtaining the sexual history. Sex. Transm. Dis. 31, 719–726. https://doi.org/10.1097/01.olq.0000145855.36181.13.

Mars, B., Heron, J., Klonsky, E.D., Moran, P., O'Connor, R.C., Tilling, K., Wilkinson, P., Gunnell, D., 2019. Predictors of future suicide attempt among adolescents with suicidal thoughts or non-suicidal self-harm: a population-based birth cohort study. Lancet Psychiatry. https://doi.org/10.1016/S2215-0366(19)30030-6.

McArdle, J.J., Ritschard, G., 2014. Contemporary issues in exploratory data mining in the behavioral sciences. Contemp. Issues Explor. Data Min. Behav. Sci.

Morgan, J.N., Sonquist, J.A., 1963. Problems in the analysis of survey data, and a proposal. J. Am. Stat. Assoc. 58, 415–434. https://doi.org/10.1080/01621459.1963.10500855.

Pailler, M.E., Cronholm, P.F., Barg, F.K., Wintersteen, M.B., Diamond, G.S., Fein, J.A., 2009. Patients' and caregivers' beliefs about depression screening and referral in the emergency department. Pediatr. Emerg. Care 25, 721–727. https://doi.org/10.1097/PEC.0b013e3181bec8f2.

Paine, M.L., Hansen, D.J., 2002. Factors influencing children to self-disclose sexual abuse. Clin. Psychol. Rev. https://doi.org/10.1016/S0272-7358(01)00091-5.

Ribeiro, J.D., Franklin, J.C., Fox, K.R., Bentley, K.H., Kleiman, E.M., Chang, B.P., Nock, M.K., 2016. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. Psychol. Med. 46, 225–236. https://doi.org/10.1017/S0033291715001804.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10. https://doi.org/10.1371/journal.pone.0118432.

Strayer, H., Craig, J., Asay, E., Haakenson, A., Provost, E., 2014. Alaska native injury atlas: an update 150.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinform. 9. https://doi.org/10.1186/1471-2105-9-307.

The Joint Commission on Accreditation of Healthcare Organizations (JCAHO), 2016. . Jt. Comm. Sentin. Event Alert.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Vergouwe, Y., Steyerberg, E.W., Eijkemans, M.J.C., Habbema, J.D.F., 2005. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J. Clin. Epidemiol. https://doi.org/10.1016/j.jclinepi.2004.06.017.

Walsh, C.G., Ribeiro, J.D., Franklin, J.C., 2017. Predicting risk of suicide attempts over time through machine learning. Clin. Psychol. Sci. 5, 457–469. https://doi.org/10.1177/2167702617691560.

Wiener, A.L., M., 2003. Classification and regression by randomForest. R News 2. R News 3, 18–22.

Williams, S.B., O'Connor, E.A., Eder, M., Whitlock, E.P., 2009. Screening for child and adolescent depression in primary care settings: a systematic evidence review for the us preventive services task force. Pediatrics 123, e716–e735. https://doi.org/10.1542/peds.2008-2415.

Wong, S.S., Sugimoto-Matsuda, J.J., Chang, J.Y., Hishinuma, E.S., 2012. Ethnic differences in risk factors for suicide among American high school students, 2009: the vulnerability of multiracial and Pacific Islander adolescents. Arch. Suicide Res. 16, 159–173. https://doi.org/10.1080/13811118.2012.667334.