Evaluating Machine Learning for Predicting Youth Suicidal Behavior Up to One Year after Contact with Mental Health Specialty Care

Lauren M. O'Reilly, BS[1]*
Seena Fazel, MBChB, MD, FRCPsych[2]
Martin E. Rickert, PhD[3]
Ralf Kuja-Halkola, PhD[4]
Martin Cederlof, PhD[4,5]
Clara Hellner, MD, PhD[6]
Henrik Larsson, PhD[4,5]
Paul Lichtenstein, PhD[4]
Brian M. D'Onofrio, PhD[3,4]

[1] Indiana University School of Medicine, Indianapolis, IN, USA
[2] Department of Psychiatry, University of Oxford, Oxford, UK
[3] Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA
[4] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden;
[5] School of Medical Sciences, Örebro University, Örebro, Sweden
[6] Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

*Corresponding Author
410 W. 10th St.
Indianapolis, IN 46202
loreilly@iu.edu

Abstract

This paper assessed the performance of several predictive modeling algorithms of suicide attempt resulting in inpatient hospitalization or suicide among youth aged 9-18 (n=34,528) after contact (6-12 months) with a mental health specialist in Stockholm, Sweden from 2006-2012. Using 209 predictors across domains (e.g., clinical, demographic, family, neighborhood, social) identified from national registers, we applied standard logistic regression, regularized logistic regression, and machine learning algorithms (i.e., random forests, gradient boosting, support vector machines). Standard logistic regression (0.77 [95% CI, 0.72-0.82]) and random forest models (0.80 [95% CI, 0.74-0.86]) demonstrated the highest AUCs. Sensitivities ranged from 0.33 (support vector machines)-0.91 (standard logistic regression). While the study was underpowered to detect a difference between logistic regression and machine learning algorithms (outcome prevalence=0.7%), performance metrics were similar across models. Logistic regression is not clearly worse than machine learning approaches. Ongoing research is needed to examine how prediction models can augment clinical decision making.

The rate of suicide have been trending upwards over the past 20 years in the United States (Center for Disease Control, 2018). In 2021, the US Surgeon General and National Action Alliance for Suicide Prevention (NAASP) jointly released a Call to Action on the 20th anniversary of the seminal report recognizing suicide as a major public health problem in the US. In the 2021 Call to Action, the Surgeon General and NAASP emphasized the importance of improving risk identification in health care settings (US Surgeon General & National Action Alliance for Suicide Prevention, 2021), which has been challenging for suicide research for two primary reasons: 1) the base rate of suicidal behavior is low in community samples (Nock et al., 2008), and 2) a large body of research informing our understanding of risk prediction includes long follow-up periods (i.e., 5-10 years). Research suggests that suicide risk can vary across short time periods (e.g., hours or days) (Witte et al., 2006). To aid suicidal behavior prediction, some researchers have shifted focus from predicting lifetime risk in the general population to acute periods of risk within higher-risk populations (Glenn & Nock, 2014).

Research has consistently demonstrated that adults (Stene-Larsen & Reneflot, 2019) and adolescents (Braciszewski et al., 2023) who die by suicide are often in contact with health care professionals within a proximal time frame prior to their death. It is important to aid suicidal prediction among those who contact the health care system to identify elevated risk, intervene, and triage care (Asarnow et al., 2016; Nock, 2012). Research commonly refers to the period of heighted risk after contact with healthcare system as short-term risk. While this definition is poorly defined (Simon, 2006), it represents an attempt to shorten follow-up windows relative to past epidemiological research. While “short-term” includes various definitions, we utilize “short-term” throughout the introduction to refer to no more than one-year post-contact.

Research examining short-term risk for suicidal behavior is growing, as the emphasis on a critical, targetable, and high-risk period may aid intervention efforts (Olfson et al., 2016). While the existing research has identified important settings for intervention (e.g., primary care, emergency departments, inpatient hospitalization) (Conwell et al., 2000; King et al., 2015; Olfson et al., 2016) and numerous short-term risk predictors (e.g., prior suicidality, depression, substance use) (Asarnow et al., 2016; Cassells et al., 2005; Favril et al., 2022), existing research must be interpreted in light of several limitations. First, most research has focused on adults. With the exception of examining short-term risk after emergency department contact (Asarnow et al., 2016; Greenfield et al., 2008; Horwitz et al., 2015; King et al., 2015; Spirito et al., 2003), children and adolescents have been relatively understudied compared to adult populations. Second, research has primarily examined short-term risk following psychiatric inpatient hospitalization (Cassells et al., 2005; Goldston et al., 1999; King et al., 2010; Large et al., 2011; Olfson et al., 2016; Prinstein et al., 2008), as compared to contact with other health care settings. Third, the majority of predictors in short-term risk studies have been derived from self-report measures or electronic health records (EHRs), such that the predictors included have been primarily demographic (Greenfield et al., 2008; Horwitz et al., 2015) and clinical (Olfson et al., 2016). Notably, prior research has identified predictors such as a history of suicidality, depression, borderline personality disorder, drug misuse, anxiety, schizophrenia, bipolar disorder, family history of psychopathology, medication non-adherence, and female sex at birth (Asarnow et al., 2016; Cassells et al., 2005; Goldston et al., 1999; Greenfield et al., 2008; Horwitz et al., 2015; King et al., 2015; King et al., 2010; Large et al., 2011; Olfson et al., 2016; Prinstein et al., 2008; Spirito et al., 2003). However, little is known about other non-psychiatric risk factors such as somatic medical problems, family-level factors, and neighborhood-level

factors, that may improve short-term prediction (Chekroud, 2018). Fourth, the utility of prediction continues to be limited by the examination of one or a handful of predictors even though prior research has identified numerous predictors for clinicians to consider when estimating short-term risk. Additionally, only a few studies explored potential interactions among risk factors (Fazel et al., 2023; Goldston et al., 1999; Horwitz et al., 2015; King et al., 2010). Researchers propose that suicidal behavior is a phenomena that is the culmination of an array of risk factors conferring risk over time (Nock et al., 2008). Ribeiro and colleagues (2016), for example, have suggested that computational models must reflect the complexity of suicidality, yet more research is needed on complex models and their clinical feasibility and acceptability remains uncertain (Glenn & Nock, 2014; Nock, 2012; Ribeiro et al., 2016).

The current paper contributes to the field in four ways. First, we focus on a large-scale, youth population (i.e., approximately 34,000 youth aged 9-18). Late childhood/adolescence is a crucial developmental period to study to deepen our understanding of suicidal behavior because the prevalence of suicidal behavior dramatically increases between the ages of 9 and 12 and continues to increase through late adolescence (Nock et al., 2013). Second, we examine short-term risk for suicidal behavior after contact with primarily outpatient mental health care. This is significant, as prior research examining short-term risk has concentrated on either the emergency department or inpatient psychiatric hospitalization. Different predictors may be associated with short-term risk depending on the setting, and, thus, prediction of short-term risk in outpatient settings requires further research. Third, the current paper links individuals to population-based registers, allowing the inclusion of predictors in a variety of domains including demographic (e.g., familial income), psychiatric (e.g., prior hospitalizations, medication), medical (e.g.,

chronic conditions), psychosocial (e.g., sexual abuse history), familial (e.g., psychopathology), and neighborhood (e.g., deprivation). The use of these various registers broadens the scope of predictors beyond those in medical claims/charts, thereby allowing for identification of potential risk factors not previously considered in prior research or routine clinical screening.

Finally, we utilize multiple machine learning algorithms to predict short-term risk among youth after contact with specialist mental health care. Machine learning aims to improve outcome prediction by balancing the trade-off between minimizing bias and maximizing generalizability through iterative algorithm training and testing (Kuhn & Johnson, 2013). Researchers have increasingly applied machine learning to clinical outcomes in attempts to improve prediction above clinician judgment (Deo, 2015). The power of machine learning comes from the ability to include multiple model predictors simultaneously (Claassen et al., 2014), consider potential interactions (Chekroud et al., 2016), and examine non-linear associations (Walsh et al., 2017). These are advantages compared to prior research focused on bivariable associations (Nock et al., 2008). While bivariable associations are valuable in interpretability, they lack clinical significance (Walsh et al., 2017), as evidenced by consistent chance-level prediction from existing suicidality models from using single predictors (Franklin et al., 2017).

Over the decade, there has been a proliferation of machine learning applications to predict suicidal behavior. In fact, in 2017, the Veterans Health Administration implemented a suicide prediction model strategy, the Recovery Engagement and Coordination for Health-Veterans Enhanced Treatment (REACH VET), after a pilot program developed and tested the model (Kessler, Hwang, et al., 2017). Despite the potential promise of machine learning and implementation in certain health care settings, some researchers remain skeptical about the integration and sustainability of machine learning techniques in behavioral health settings (Mohr

et al., 2018). Recent reviews have found that while the machine learning algorithms have demonstrated good accuracy, they have limitations in some performance measures in simulations (Belsher et al., 2019). Additionally, while some studies have demonstrated the superior performance of machine learning models compared to bivariable or multivariable logistic regression (Ribeiro et al., 2019; Walsh et al., 2017; Walsh et al., 2018), a systematic review of machine learning algorithms (i.e., classification trees, random forests, artificial neural networks, support vector machines) found no difference in the area under the receiver operating characteristic curve (AUC) compared to logistic regression (both maximum likelihood and regularized) for numerous clinical outcomes (Christodoulou et al., 2019). Research examining machine learning algorithms should, therefore, compare such models to evaluate the added benefit of models of increasing complexity.

We utilized prospectively collected data on approximately 43,000 visits to examine short-term risk for suicidal behavior among youth after contact with a mental health specialist within 6-12 months based on predictors across domains. Informed by the inconsistent research surrounding the improved performance of machine learning algorithms compared to logistic regression, the primary aim of this paper is to examine several algorithms and assess their performance (i.e., maximum likelihood "standard" logistic regression, regularized logistic regression, and machine learning methods). We hypothesized that the machine learning models (i.e., random forest, gradient boosting, and support vector machines) would capitalize on novel combinations of predictors and outperform the standard and regularized logistic regression models across numerous metrics (notably, AUCs, sensitivity, specificity).

Transparency and Openness

The study was not preregistered. Data are not made publicly available. Analytic code is posted to a github repository (github.com/LaurenMOreilly/STR). Data were deidentified and exempt from IRB approval.

Methods

Sample

We derived the sample from the Clinical Database for Child and Adolescent Psychiatry in Sweden (Pastill), which is a longitudinal, prospective dataset that records all mental health care specialist visits among children and adolescents in Stockholm County, Sweden. The majority (approximately 80%) of visits are outpatient visits; of the 118 specialty mental health clinics contributing to the Pastill dataset, 19 (16.10%) are inpatient clinics, 1 (0.85%) is an emergency ward, and 4 (3.39%) are designated as "other" (primarily psychiatry consultation clinics). The current dataset ranges from 2006 through 2012, forming a cohort of 34,528 patients (52.9% female) with 43,278 unique treatment periods (**Table 1**). Treatment periods were defined as non-overlapping periods comprised of at least one treatment (see below for more details). Given that suicidal behavior data ended in 2013, we restricted the dataset to end in 2012 allowing one year of follow-up. We restricted ages to 9-18 years (mean=13.90 at first treatment period), as suicidal behavior is rare among young children. Pastill includes demographic data (i.e., age at visit), number of treatment visits, length of treatment provided, mental disorders as per *International Classification of Disease*-10th Edition (*ICD*-10) codes, and standardized assessment of psychosocial problems (e.g., parental neglect, sexual abuse, peer problems). All specialists also completed a checklist at intake as part of clinical routine, which includes information regarding self-harm (unspecified intent to die).

In addition to Pastill, we merged information from seven longitudinal, population-based registers via individuals' unique identification codes. First, the Multi-Generation Register links individuals to family members since 1961 (Statistics Sweden, 2010), which allows for the inclusion of familial risk factors. Second, the Medical Birth Register captured nearly all (>99%) live births in Sweden since 1973 and collected information on youth sex assigned at birth. Third, the Education Register recorded information on parental educational attainment since 1970 (D'Onofrio et al., 2010). Fourth, the National Patient Register collected information on youth and parental inpatient visits since 1969 and outpatient visits since 2001 for psychiatric and medical diagnoses using *ICD* codes (Jacobsson, 2009). Fifth, the Prescribed Drug Register collected information on prescription use since 2005. Medications are classified per the Anatomical Therapeutic Chemical (ATC) classification system (Wallerstedt et al., 2016). Sixth, the National Crime Register contained criminal conviction data for all individuals over the age of 15 since 1973 (Fazel & Grann, 2006). Finally, the Integrated Database for Labor Market Research recorded information on familial income, familial receipt of income support or sickness/injury benefits, and neighborhood residence (i.e., Small Area Market Statistics areas) since 1990 (Statistics Sweden, 2011). We derived a neighborhood deprivation score and determined the presence of suicidal behavior in the same neighborhood (Sariaslan et al., 2013).

Candidate predictors

We chose the predictor domains given prior research demonstrating that demographic factors (Lodebo et al., 2017; Nock et al., 2012), educational factors (Björkenstam et al., 2011), psychopathology (Simon et al., 2007), familial psychopathology (Brent et al., 1996), medical conditions (Tang & Crane, 2006), criminal convictions (Sahlin et al., 2017), and neighborhood-

level factors (Cox et al., 2012) are associated with suicidal behavior among youth. Predictor definitions and coding were based on prior machine learning research using Swedish registers (Chen et al., 2020). For youth diagnostic and medication predictors from the National Patient Register and Prescribed Drug Register, respectively, we categorized variables into three windows varying in proximity to the start of the treatment period (≤ 1 month, 1-12 months, >1 year) in attempts to capture lifetime versus recent risk factors. There were 209 predictors (see **Supplemental Online Material [SOM] 1**). **SOM 2** and **3** provide information regarding the *ICD* and ATC codes used to derive diagnostic and medication candidate predictors. To characterize the clinical features of the cohort, **SOM 4** presents select diagnostic and psychosocial information on the sample.

Outcome

We defined short-term suicidal behavior as the presence or absence of either suicide attempt resulting in inpatient hospitalization or suicide which were derived from the National Patient Register and the National Death Register, respectively, in the follow-up window (i.e., 6-12 months post-contact) (Claassen et al., 2014; Simon, 2006). Previous research investigating short-term risk after an emergency department visit or hospitalization has defined post-contact as beginning either at intake (Czyz et al., 2016; Horwitz et al., 2015) or at discharge (Large et al., 2011; Olfson, 2017). Given that outpatient treatment can extend for a longer period compared to an acute hospital visit, the definition of post-contact can have considerable impact on prediction for short-term risk. The identification of suicidal behavior from ICD codes is limited; it is unclear whether the presence of suicidal behavior codes are distinct events, as codes may be repeated and/or individuals may seek care in various settings resulting after one event. In attempts to

increase the likelihood of capturing distinct suicidal events, we defined the follow-up period as beginning 6 months after the end of a treatment period. There were 284 events of suicidal behavior between 6-12 months (0.66%) after the last visit in a treatment period.

Additionally, patients could be enrolled in multiple clinics/treatments at once and therapy could persist over numerous sessions; the start of the short-term period was not as discrete as prior research examining inpatient psychiatric hospitalizations. To characterize the overlapping nature of treatment visits, we created periods of non-overlapping treatment periods, after which the short-term period started. Stated differently, if treatments overlapped based on their first and last visit dates, these periods were subsumed into one treatment period non-overlapping with other treatment periods. This was chosen to inform clinical decision-making regarding termination of treatment and risk for suicidal behavior without active clinician monitoring. Treatment periods were separated by at least one day. Most candidate predictors were referenced to the first date of the non-overlapping treatment period (i.e., occurred before treatment initiation), although we also included information captured over the course of the treatment period as predictors (e.g., inpatient hospitalization, total number of days in the treatment period, and total number of outpatient visits). The outcome had to occur after the last treatment date.

Analytic rationale

Given extant reviews suggesting that machine learning may not outperform logistic regression and the lack of research for short-term prediction within a youth outpatient sample, we devised three aims to examine model performance ranging on a continuum from human-driven to machine-driven (Beam & Kohane, 2018).

Aim 1 evaluates "standard" logistic regression models based on maximum likelihood estimation. Aim 1 is designed to mimic past research that has considered only a few risk factors, as well as serve as a baseline set of models against which to compare the machine learning models and make inferences about relative model performance and clinical utility. These models, however, were multivariable models, as compared to prior literature that has focused on bivariable associations. The primary advantage is interpretability.

Aim 2 is designed to use machine learning algorithms that select a subset of salient predictors. Also referred to as variable selection models, these models are data-driven approaches in which model complexity is penalized to reduce the number of predictors, potentially at the cost of predictive power. Aim 2 serves as the middle of the continuum in terms of model interpretability versus complexity. The added benefit of Aim 2 compared to Aim 1 is that Aim 2 may determine important predictors for short-term suicidal behavior, which could be further investigated in basic research or integrated into clinical care. Both Aims 1 and 2 simplify the computational complexity compared to the machine learning models in Aim 3, thereby improving the interpretability of results.

In contrast to Aims 1 and 2, random forest and gradient boosting combine "weaker" base estimators into one "stronger" prediction model that optimize fit through the reduction of bias and improvement of precision. Support vector machines identify linear and non-linear hyperplanes differentiating classes. Therefore, researchers regard these methods as valuable in prediction, although limited in the interpretability (Cortez & Embrechts, 2013). The comparison of models across aims will help to inform the extent to which Aim 2 and/or Aim 3 improve upon predictive power, as well as inferences about the clinical utility of different machine learning methods within an adolescent specialty mental health care sample. For example, if models from

Aim 1 perform equally well as those in Aims 2 and 3, the use of complex prediction models in an outpatient clinical setting may not provide incremental utility. However, if Aim 3 performs substantively better than Aims 1 and 2, the employment of risk algorithms in clinical settings and training clinicians on how to use them may be helpful. Of note, we use the term "substantively better" not in the statistical sense, but rather to capture examining performance across metrics.

Analytic approach

We conducted all dataset preparation in SAS 9.4 (SAS Institute Inc., 2016). Prior to modeling, we applied several pre-processing techniques to the data; this vital step to optimize the performance of machine learning algorithms (Kuhn & Johnson, 2013) by aiding convergence and reducing potential bias introduced in machine learning models. We also: (a) dropped 14 predictors with a variance of zero (see **SOM 1**), (b) dropped 24 variables with high multicollinearity with other variables as determined by tetrachoric correlations > 0.95 (**SOM 5**), and (c) combined variables with low variance and of a similar construct (i.e., nine predictors combined into dichotomized "autoimmune disorder" variable; see **SOM 1**); this resulted in a total of 209 predictors. All subsequent analyses were derived from a complete case sample (2,317 treatment periods from a dataset of 45,595 treatment periods were dropped due to missing data on the predictor variables to a final dataset of 43,278 treatment periods). The following steps were completed in R 4.0.5 (R Core Team, 2022). We randomly split the dataset 70-30 into training and holdout datasets, respectively (Fazel et al., 2017). Data splitting was performed prior to variable transformation and oversampling to avoid any bias introduced in the holdout sample through these steps. Additionally, because individuals would be repeated in the dataset, data splitting was done respective to individual identification numbers. This ensured that individuals

were represented in either the training or holdout dataset, but not both. We also stratified the outcome between the training and holdout sets. For a summary of the number of observations and outcomes in the full, training, and holdout datasets, refer to **Table 1**.

We continued pre-processing to address the use of both binary and continuous variables in prediction. To reduce bias introduced by variables with differing ranges specifically for support vector machines (Edwards & Raskutti, 2004), we rescaled two continuous variables (age and neighborhood deprivation) to a range of 0-1, and quartiled and subsequently dummy coded two continuous variables (total number of outpatient visits over the treatment period and total number of days over the treatment period). Note all other variables were dichotomized as not present (0) or present (1). Additionally, due to the small proportion of suicidal behavior events within the follow-up window of 6-12 months after the last visit, class imbalance posed a potential issue. Class imbalance refers to differences in prior class probabilities and can lead to performance inflation by solely predicting the majority class (no suicidal behavior event) compared to minority class (suicidal behavior events) (Guo et al., 2008). There are various approaches to handle class imbalance, including (1) addressing the distribution of the majority and minority classes, (2) comparing a wide variety of algorithms, and (3) examining various performance metrics (Kessler, Stein, et al., 2017). We adopted all three approaches. In regards to distribution techniques, we oversampled suicidal behavior events using the ROSE (Random Over-Sampling Examples) R package (Lunardon et al., 2014), which randomly selects cases from the minority class with replacement. We chose a 40% oversampling distribution. We conducted all pre-processing steps equivalently on the holdout dataset except for oversampling. A TRIPOD checklist of model development and validation can be found in **SOM 6**.

Performance metrics

Performance was evaluated by applying the algorithm trained on the training dataset to the holdout dataset. To evaluate performance, we first used the total area under the receiver operating characteristic curve (AUC), which provides a robust and informative way to evaluate a binary classification model's ability to accurately distinguish between two classes. AUC ranges from 0 to 1 with higher AUC (closer to 1) indicating better performance; 0.5 corresponds to random guessing. The receiver operating characteristic curve traditionally plots sensitivity (y-axis) against 1-specificity (x-axis), thereby demonstrating the trade-off between these values along varying decision thresholds. Confidence intervals were generated from 10,000 bootstrapped samples from the training set and models were applied on the holdout set (Robin et al., 2011). Second, we reported sensitivity (the proportion of individuals with suicidal behavior who were correctly classified) and specificity (proportion of individuals without suicidal behavior who were correctly classified). Third, we reported the positive predictive value (PPV; proportion of individuals who were classified by the model to have suicidal behavior and who had suicidal behavior) and negative predictive value (NPV; proportion of individuals who were classified by the model to not have suicidal behavior and who did not have suicidal behavior). Fourth, to capture potential clinical utility, we included the net benefit, which can be interpreted as the proportion of those correctly classified (true positives) minus those incorrectly classified (false positives) weighted by that decision threshold (Vickers et al., 2016). Positive net benefit values suggest the benefits of classification outweigh harms; a net benefit of zero would be equivalent to no intervention/classification of individuals.

While AUC, sensitivity, specificity, PPV, and NPV are valuable classification metrics, we also present additional discrimination and calibration metrics to inform a more robust understanding of model performance. First, PRAUC (area under the precision-recall curve, where "precision" is equivalent to PPV and "recall" is equivalent to sensitivity) calculates the total area under the precision-recall curve. Second, partial AUC calculates the AUC for a specific range of a given metric; we calculated the AUC for the sensitivity over 0.80 given the importance of correctly classifying those who attempt or die by suicide. Compared to the AUC, the PRAUC and partial AUC may be more appropriate metrics when evaluating performance affected by class imbalance, as neither precision nor recall consider true negatives (Carrington et al., 2020). Third, accuracy calculates the proportion of cases that were correctly classified. Fourth, balanced accuracy is the mean of sensitivity and specificity and is often more appropriate than accuracy when class imbalance is present, as accuracy is often inflated when cases are predominately the majority class and, thereby, correctly classified. Fifth, we present additional classification metrics derived from the confusion matrix: kappa, detection rate, and detection prevalence. Sixth, we reported the F1 score, which calculates the harmonic mean of the precision and recall. Seventh and eighth, compared to discrimination metrics, which measure classification performance, we calculated the Brier score and integrated calibration index. The Brier score measures the accuracy of the predicted probabilities, and range from zero to one, with the scores closer to zero representing greater accuracies of predicted probabilities. The integrated calibration is the weighted average of the absolute difference between the observed and predicted probabilities, with scores closer to zero representing no difference between observed and predicted values (Austin & Steyerberg, 2019).

Note that the initial confusion matrix for all performance metrics was based on 10% threshold for predicted values, we present additional thresholds of 5% and 25% in the SOM.

Aim 1: Standard logistic regression

After addressing all relevant data preparation, we conducted two sets of analyses. First, we identified a set of nine *a priori* predictors based on prior literature that captured domains of depression, anxiety, substance use, prior self-harm or suicide attempt, and sex. The predictors were (1) antidepressant medication within one month prior to start of the treatment period, (2) major depression diagnosis at first visit of the treatment period, (3) self-reported chief complaint of depression at first visit, (4) anxiety diagnosis at first visit, (5) self-reported chief complaint of anxiety at first visit, (6) substance use diagnosis at first visit, (7) female sex assigned at birth, (8) suicide attempt diagnosis (intentional self-injury) at first visit, and (9) self-reported chief complaint of self-harm at first visit. Each outcome was predicted from the nine predictors in the training set, after which the model was applied to the holdout set. We aimed to choose seven predictors to prioritize interpretability and, relatedly, maintain a set of predictors that could more easily be remembered by clinicians due to working memory constraints (Miller, 1956). However, nine predictors were retained in the *a priori* model as we had multiple measures of depression, anxiety, and self-harm domains.

Second, we conducted bivariable logistic regression between each of the 209 candidate predictors and the outcome and cross-validated the results using 10-fold, internal cross validation (i.e., the training dataset is subsequently divided and validated on 10 equal components). We then ranked each bivariable association based on AUC. From there, we selected the top seven predictors and predicted the outcome in a multivariable logistic regression model in the training

set. The resulting model was then used to create predictions in the holdout set. Given that model interpretability is a primary motivation of Aim 1, we reported the odds ratios (ORs) and 95% confidence intervals (CIs) for each of the predictors from the corresponding model.

Aim 2: Regularized logistic regression

We tested the performance of three types of variable selection algorithms that determine and regularize a subset of variables: (1) lasso, (2) ridge, and (3) elastic net (Tay et al., 2023). The lasso, ridge, and elastic net algorithms all impose a penalty during the estimation process on the size of regression parameters while simultaneously minimizing error (i.e., residual sum of squares). Each differ in the type and size of the penalty for parameter values (Kotsiantis, 2007). Additionally, the above methods all involve tuning parameters, which aim to optimize performance and are determined via a 10-fold cross validation. We tuned the models via a grid search of hyperparameters chosen based on minimum mean squared error for ridge and lasso and accuracy for elastic net. The models were then applied to the holdout set. We used the glmnet (v4.1-8) R package (Friedman et al., 2021) for lasso and ridge, and caret (v6.0-94) package (Kuhn et al., 2020) for elastic net. For Aims 2 and 3, we included all 209 predictors.

Aim 3: Machine learning algorithms

We conducted three types of machine learning algorithms: (1) random forests, (2) gradient boosting, and (3) linear support vector machines. Random forest and gradient boosting methods are based on decision tree algorithms, in which a collection of trees are built and aggregated via their predicted values (Kotsiantis, 2007). Each differ in how trees are grown and the size of the trees. Decision tree algorithms are common approaches with imbalanced data. Support vector machine models aim to find the optimal hyperplane (either linear or nonlinear)

separating data points in a multidimensional space. We used the caret (v6.0-94) package for random forest and gradient boosting (Kuhn et al., 2020), and the e1071 (v1.7-13) package for support vector machines (Dimitriadou et al., 2009). Similar to Aim 2, the above methods all involve tuning parameters determined via internal, 10-fold, cross validation. The tuning grid of parameters was iteratively tested and determined to balance both performance and computational intensity (i.e., computation time, hardware resources, mathematical complexity); evaluation was based on accuracy metrics. We chose the Aims 2 and 3 algorithms due to robustness in research (Kessler, Stein, et al., 2017; Kessler et al., 2015; Walsh et al., 2017). For the random forest and boosting models, we reported the top 20 variables based on variable importance, which is a measure of the relative contribution of each variable. Hyperparameters tuned for Aims 2 and 3 are provided in **SOM 7**.

Sensitivity Analyses

We conducted two sensitivity analyses. First, the main prediction window was within 6 to 12 months after the last visit of the treatment period, however predicting suicidal behavior within a closer period may be more clinically relevant. Therefore, we conducted a sensitivity analysis predicting suicidal behavior within 1 and 6 months after the last date of the visit period (0.71% of treatment periods). Second, we evaluated two support vector machine kernels (radial and polynomial) in separate models to examine nonlinear hyperplanes. Support vector machines operate through a kernel function, which is a data transformation technique that determines the similarities between data points. Kernel functions are computationally efficient methods in high-dimensional predictor spaces that map predictor coordinates. Note that these kernels were examined separately from the linear kernel in the main analysis.

Results

Aim 1

Logistic regression with *a priori* predictors (e.g., nine variables indexing depression, anxiety, substance use, prior suicide attempt, and sex) demonstrated acceptable performance when predicting suicidal behavior within 6-12 months with an AUC of 0.77 (95% CI, 0.72-0.82). Using a 10% decision threshold, sensitivity was high (0.91) and specificity was low (0.40). Performance was similar when modeling logistic regression with predictors derived from the seven highest AUCs; the AUC was acceptable (0.80 [95% CI, 0.75-0.85]), sensitivity was high (0.91), and specificity remained low (0.40) Across both models, PPV remained low (0.01 for both) and NPV remained high (estimates were rounded) for each model. The net benefit was negative for each model (-0.06). The classification metrics of each model can be found in **Table 2** (and additional metrics in **SOM 10)**.

The odds ratios (ORs) and confidence intervals for each predictor in the multivariable logistic regression models are reported in **SOM 8** and **9**. For *a priori*-chosen predictors, four predictors demonstrated large magnitudes when predicting suicidal behavior within 6-12 months: antidepressant prescription within one month prior to treatment initiation (OR, 9.60 [95% CI, 8.11-11.41]), suicide attempt diagnosis at intake (OR, 8.59 [95% CI, 7.90-9.35]), female sex (OR, 5.81 [95% CI, 5.48-6.16]), and substance use disorder diagnosis at intake (OR, 5.03 [95% CI, 4.00-6.35]). For predictors derived from the seven highest AUCs, age, female sex, inpatient admission within the treatment period, self-reported chief complaint of self-harm at the first visit date, anxiety diagnosis, suicide attempt diagnosis, and prescription for mood stabilizer between 1-12 months prior to the treatment period was identified. The largest predictor based on

magnitudes was prescription for mood stabilizing drug within 1-12 months prior to the first visit date (OR, 9.76 [95% CI, 8.85-10.77]).

Aim 2

Ridge, lasso, and elastic net regression demonstrated poor or acceptable model performance when predicting within 6-12 months. Ridge and lasso demonstrated poor AUCs (0.66 [95% CI, 0.58-0.74] and 0.70 [95% CI, 0.63-0.77], respectively), and elastic net demonstrated an acceptable AUC (0.77 [95% CI, 0.72-0.83]). Sensitivities were moderate (range from 0.65-0.86), specificities were low to moderate (range from 0.50-0.62), and PPVs were 0.01 based on 10% decision threshold. The net benefit was -0.04 for ridge and lasso, and -0.05 for elastic net. Metrics are found in **Table 2** and **SOM 10**.

Aim 3

For Aim 3 models, AUCs ranged from 0.64-0.80; random forest demonstrated the highest AUC (0.80 [95% CI, 0.74-0.86]). Using a 10% decision threshold, sensitivities from 0.33-0.84, and specificities from 0.48-0.89. Net benefit was slightly positive for random forest (0.05) and negative for gradient boosting and support vector machines (-0.04 and -0.03, respectively). Metrics are found in **Table 2** and **SOM 10**. The variable importance of the top 20 variables in the random forest and boosting models for each outcome are reported in **SOM 11**. The top variables were similar to those in Aim 1 (i.e., inpatient admission, female sex, prior suicide attempt or self-harm, psychiatric medication receipt), however both random forest and boosting utilized demographic variables (e.g., neighborhood deprivation, paternal educational attainment, length of treatment), parental information (e.g., maternal criminal convictions) and stressful life

event/trauma indicators (e.g., neglect, physical abuse, child separation). For random forest, the variable of highest importance was a mood stabilizing drug prescription between 1-12 months prior to the treatment period when predicting suicidal behavior within 6-12 months. For gradient boosting, it was receiving a suicide attempt diagnoses at time of intake. For all models across aims, the confusion matrix with the 10% threshold can be found in **SOM 12**. Thresholds were updated to include 5% and 25% (confusion matrices presented in **SOM 13** and **14;** model estimation results presented in **SOM 15** and **16**). Across models, lowering the threshold resulted in increased sensitivity and decreased specificity. **Figure 1** presents the ROC plot.

Sensitivity Analyses

Given clinical interest to predict suicidal behavior more closely to the end of treatment, we added an additional outcome of suicidal behavior within 1 and 6 months after the last visit in the non-overlapping treatment period. There was a total of 306 events of suicidal behaviors within 1 and 6 months (**SOM 17**). Classification metrics using 5, 10, and 25% probability thresholds are presented in **SOM 15, 18**, and **16**, respectively. Compared to suicidal behavior prediction within 6-12 months in the main analyses, AUCs increased (range 0.78-0.85), with the highest AUC observed for AUC-derived logistic regression model. Based on a 10% decision threshold, sensitivities ranged from 0.46-0.97 and specificities ranged from 0.40-0.92 (**SOM 18**).

To examine support vector machines that develop nonlinear hyperplanes, we conducted two sensitivity analyses of support vector machines with radial and polynomial kernels. Results from the support vector machine radial and polynomial kernel demonstrated poor overall performance for both the 1-6- and 6-12-month outcomes (AUCs around 0.50; **SOM 19**).

Discussion

The primary objective of the current paper was to examine performance across models in the classification of suicidal behavior within 6-12 months after contact with mental health specialty care among adolescents using predictors from multiple domains. Two primary findings can be concluded: 1) machine learning methods (random forest, gradient boosting, and support vector machines) did not substantially outperform standard logistic regression and elastic net regularized regression across the majority of metrics, and more broadly, performed poor in prediction of suicidal behavior within 6-12 months, and 2) based on sensitivity, the models that demonstrated the overall best performance were the standard logistic regression models, followed by elastic net regularized regression. We expand on each finding below.

First, random forest, boosting, and support vector machine demonstrated comparable performance as Aim 1 and 2 models when examining AUCs. Results are consistent with a prior study (van Mens et al., 2020), in which various models predicting suicidal ideation and suicide attempts within 12 months of baseline among young adults were compared. Model performance was similar across logistic regression, random forests, gradient boosting, and support vector machines, however random forest had superior PPV than logistic regression when predicting suicidal ideation and attempt. Interestingly, van Mens et al. (2020) utilized numerous psychological measures (e.g., entrapment, defeat) which differed than our study's use of demographic, diagnostic, and psychosocial information. Within our study, random forest and AUC-derived logistic regression models had the same AUC (0.80) and PPV, and the AUC-derived logistic regression model had superior sensitivity than random forest (0.91 vs. 0.84) at the 10% decision threshold. In addition to assessing model performance based on AUCs, sensitivities, and specificities, we included numerous additional evaluation metrics. Notably, we

included the net benefit in efforts to determine potential clinical impact. All net benefits were negative at the 10% threshold (indicating the benefits of classification do not outweigh potential harms) except one: random forests predicting suicidal behavior between 6-12 months post-treatment period (0.05). The positive net benefit compared to the other models may suggest that random forest offers *slightly* improved clinical utility. Net benefit is an increasingly reported metric that can aid in comparing the potential clinical benefit between machine learning algorithms to standard regression models (Ehrmann et al., 2023), as well as compare models to no intervention (i.e., not classifying individuals). While sensitivity was high for numerous models, sensitivity remained low to moderate, highlighting the potential cost to false positives.

The lack of clear inferior performance of Aim 1 models compared to Aim 3 suggests the machine learning algorithms may not be necessary to model the data utilized in this study. The benefit of the machine learning algorithms in highly complex prediction problems may, indeed, be *too* complex for the current classification. These findings align with prior reviews comparing logistic regression and machine learning methods, which fail to support the improvement in performance of machine learning methods above logistic regression (Christodoulou et al., 2019). Franklin et al.'s (2017) meta-analytical finding that there has been little improvement in suicide prediction over the past 50 years prompted their statement that "the additive or interactive effects of a small number of risk factors would also produce inaccurate prediction." Findings such as ours call into question this statement. We note that detecting significant improvements of machine learning algorithms over standard regression may be hindered by the low statistical power. Prior research has outlined approaches to determine sample size requirements to improve precision (Riley et al., 2020), which are higher for machine learning than for standard statistical models.

While there is not a direct comparison between predictors utilized in Aims 1 and 3, random forest and boosting utilized variables outside the domains included in the standard logistic regression models, such as neighborhood, parental health diagnoses, and stressful life events/trauma. The comparable performance of Aim 3 models, however, may reduce the impact of these variables to inform future modeling. Future research that includes predictors outside of demographic and psychiatric factors traditionally derived from EHRs can continue to examine the added benefit of these variables.

In the current study, prior suicidality, psychiatric disorders/medications, and demographics (age, female sex) may be sufficient to develop well-performing models. While we included clinician-reported questionnaires related to the reason an individual was seeking care, which could capture time-varying risk factors, the majority of our candidate predictors likely captures more stable factors, such as psychiatric diagnoses, medical diagnoses, and family history. Therefore, the current study speaks to the ability of predictors derived from EHRs and registers to predict suicidal behavior within a short time period after a given visit/treatment period. While a prior study concluded that the inclusion of imminent risk factors (e.g., hopelessness, loneliness, agitation, insomnia, life dissatisfaction, acquired capability, self-injurious behavior) did not improve machine learning model performance (Ribeiro et al., 2019), future research may benefit from more thoroughly examining the combination of self-reported measures of imminent factors, EHR-derived data (e.g., diagnostic codes), family-level factors, and neighborhood-level factors in machine learning models.

Second, based on sensitivity, standard logistic regression models performed the best. At the 10% threshold, sensitivity for 6-12 month prediction was 0.91 for both logistic regression with *a priori* predictors and AUC-derived predictors. This was closely followed by elastic net and

random forest at sensitivities of 0.86 and 0.84, respectively. A prior study using similar Swedish registry data predicting suicidal behavior within 30 and 90 days following psychiatric specialty care found similar performance between elastic net, random forest, gradient boosting, and neural network algorithms (Chen et al., 2020). Another study predicting suicide after outpatient visits using Veterans Administration data also found an elastic net model with 10-14 predictors optimized performance (compared to naïve Bayes, random forest, support vector machines) (Kessler, Stein, et al., 2017). Among high-risk adolescents, one study found that logistic regression performed similarly as machine learning techniques (Jung et al., 2019), which is mirrored in the current study.

Clinical implications

The findings suggest that a few key variables may be particularly powerful in predicting suicidal behavior 6-12 months after specialty mental health care, including depression and antidepressant medication use, anxiety, substance use disorder, sex, inpatient hospitalization history, and self-harm/suicide attempt history. These are not novel predictors and are often included in routine clinical assessment. However, if clinical assessment is time-limited and focuses primarily on suicidality (e.g., ideation, plan, and intent), clinicians may consider prioritizing additional assessment of depression, anxiety, and substance use if time allots.

Partially catalyzed by the rise of machine learning models and subsequently doubting their utility, researchers have stressed the need to re-focus attention on suicide *prevention*, not prediction, through increasing the implementation of evidence-based suicide intervention. The sentiment behind this argument is motivated by prediction models: (a) often including predictive risk factors rather than causal risk factors, (b) demonstrating limited clinical utility due to poor

sensitivity and specificity metrics, and (c) over-emphasizing risk categorization (e.g., low, moderate, high risk), thereby potentially causing harm by forgoing thorough suicide assessment and intervention among individuals categorized as low or moderate risk (Pisani et al., 2016).

While the primary goal of prediction is not to determine causality, some researchers have questioned how machine learning models using non-causally implicated risk factors (i.e., those traditionally included in epidemiological studies often derived from EHRs, such as psychiatric disorders) perform compared to those using factors conceptualized in causal theories (e.g., perceived burdensomeness, pain, hopelessness, acquire capability to die). Theoretically, it would be expected that models including causally implicated factors would outperform those without such factors. However, a meta-analysis compared machine learning models to standard regression models with theoretically derived factors (e.g., Ideation-to-Action, biosocial, hopelessness models) and found that machine learning models outperformed theoretical models as measured by weighted odds ratios (Schafer et al., 2021). In terms of clinical utility, Simon et al. (2019) identified that risk identification models for suicide attempt are comparable to risk models for breast cancer and cardiovascular events that have informed their respective fields' prevention guidelines. To state that machine learning models are unacceptable due to performance metrics, such as positive predictive values or sensitivities that are too high or too low depending on the setting needs, is to misunderstand to what machine learning models are compared. Clinical judgment is often insufficient to guide decision making (Tucker et al., 2018); however, clinical judgment is the primary source of guidance when making risk decisions about future suicide risk within the minutes, hours, and days after contact. Knowing the limitations of clinical judgment, predictive models (broadly defined to include regression models) should be evaluated to consider how they can augment clinical decision making rather than replace it.

Additionally, risk categorization is not a defining feature of actuarial methods, which can provide probability scores such as OxMIS (Fazel et al., 2019) and OxSATS, (Fazel et al., 2023) rather than low/high risk categories based on specified thresholds. Arguably, clinicians are making risk judgments without being prompted by prediction models. The reality of false negatives speaks to the importance of comprehensive assessment, informed by clinical judgment and actuarial models. Patient-centered research, both quantitative and qualitative, is critical in exploring the real-world implications of how and when to use predictive modeling. For example, a qualitative study of patient feedback around the REACH VET program suggested that 78% of individuals did not feel hopeless after high-risk identification (Reger et al., 2021).

The apprehension around prediction, both its clinical utility and how to do so ethically (Tucker et al., 2018), has implications for the current findings. Given that lack of inferior logistic regression performance, one may conclude that machine learning models are insufficient to predict suicidality in the short-term. While perhaps grounded in necessary skepticism, we caution against this conclusion for two reasons. First, machine learning is often assumed to be a data-driven approach without input from expert-opinion/theory-driven approaches. This dichotomization is overly simplistic. Arguably, the logistic regression model with *a priori* predictors could be considered a data-driven approach (albeit with more theory input), as a prediction model is developed and applied to a holdout dataset. On the other hand, theory informed the choice of predictors, which were included in machine learning algorithms. Second, machine learning used to predict suicidality, relative to prior research examining a limited selection of predictors, is a new field. Ongoing comparison against standard logistic regression, and even clinical judgment, is necessary to contextualize machine learning results. This is particularly relevant when considering metrics such as net benefit to assess clinical utility.

While the current study questions the added benefit of machine learning models to predict adolescent suicidal behavior in specialty mental health settings, it does not necessarily negate their use entirely. There are a variety of subpopulations, settings, and outcomes that have yet to be explored. While we compared numerous models, it may be beneficial to compare logistic regression models to clinician prediction based on a given set of predictors, which may more accurately reflect a "baseline" model.

Additionally, we highlight that the aim of the current paper was not to develop a deliverable clinical product, despite the goal of many machine learning models to improve precision medicine. Clinical deployment requires iterative testing within the clinical sample and quality improvement over time to examine potential biases. This is particularly relevant as we used a 10% threshold; clinical thresholds should be determined by the unique constraints of the clinical setting. For example, in one study, model thresholds were based on obtained 90% or 95% specificity (Barak-Corren et al., 2016). Importantly, even if prediction modeling achieves acceptable evaluation metrics, clinician uptake may be poor (Bentley et al., 2022). Ethical considerations will guide implementation and will need to be iteratively applied based on results of prediction modeling, the changing healthcare landscape, and patient and physician feedback (Tucker et al., 2018).

Strengths and limitations

Key strengths of this study include (1) a large, representative sample of youth seeking mental health care (primarily outpatient), and (2) the use of multiple variable selection and predictive modeling approaches. The majority of prior suicide machine learning studies have been conducted on adults presenting in emergency departments or inpatient hospitals (Bernert et

al., 2020). The use of a youth, primarily outpatient sample provides a more robust understanding of how these models perform across groups and settings. We leveraged national registers to include variables across numerous domains, including neighborhood and parental variables, which have not been commonly included in machine learning suicidality studies. In addition to the advantages provided by the expansive data linkages, we structured the aims of the paper to examine model performance from logistic regression to machine learning models. Finally, we presented both discrimination and calibration statistics to address a gap in prior machine learning studies. Model comparisons are often based solely on AUCs, and a meta-analysis found that less than a third of studies reported accuracy, sensitivity, and specificity statistics (Bernert et al., 2020). Model performance interpretation is improved when numerous metrics are considered given potential limitations of a single metric, especially given clinical implications.

Several limitations should be considered. First and foremost, our original datasets were severely imbalanced (0.65% of treatment periods were followed by a suicidal behavior event). We addressed this imbalance through oversampling, including various models, and evaluating along different metrics. These approaches may be insufficient to handle such imbalanced data, as echoed in a recent review (Nordin et al., 2022). Research suggests that oversampling approaches show improvement in prediction (Malhotra & Kamal, 2019), which will likely artificially inflate the AUC estimates. Thus, our AUC estimates are better viewed in comparison to each other, rather than with non-oversampled datasets. Future studies should retain imbalance in the dataset (and subsequently conduct internal cross-validation to avoid data splitting). Relatedly, given the low base rate of suicidal events, we may lack statistical power affecting model fit. Modeling choices (e.g., rescaling continuous variable) may also exacerbate power concerns. Second, we used random splitting, which is not the ideal splitting technique as it may demonstrate

performance that is too optimistic (Altman et al., 2009). Temporal splitting may be more appropriate to help prevent inflated model performance and to capture prospective prediction akin to real-world settings; however, research suggests the impact of splitting technique lessens when the dataset is larger (Birba, 2020). Third, as noted above, the models performed better for prediction within 1-6 compared to 6-12 months. This may be reflective of how suicidal behavior is captured in this study. Apart from self-reported reason for treatment initiation, suicidal behaviors were gathered from ICD codes in medical records. It is unclear whether codes are carried over to future encounters and, therefore, one event is inaccurately being represented as multiple. We attempted to resolve this issue by adding a month gap between last treatment visit and prediction in the sensitivity analysis and being conservative in our prediction by focusing on 6-12 month prediction in the main analysis. Fourth, inpatient and outpatient clinics were included in the study; approximately 80% of clinics were outpatient, however prediction from inpatient may be different than outpatient. We included a predictor of whether an individual visited an inpatient clinic during the treatment period in attempts to capture differences in severity. Fifth, many predictors had low variance, which may prove problematic in machine learning models (Guo et al., 2008). In our data pre-processing steps, we eliminated variables with zero variance or combined related disorders. We did not further eliminate variables based on low variance as this decision would have been arbitrary and we wanted to examine how these variables would perform within a multivariable space. Sixth, we conducted a complete case analysis. While this resulted in excluding only 5.1% of treatment periods, multiple imputation approaches may be more appropriate. Seventh, one of the logistic models was generated by conducting bivariable analyses among each candidate predictor and the outcome and then choosing the top seven. Future research should compare bivariable to multivariable models to

choose the top performing predictors. Eighth, we focused on the prediction window of 6-12 months after the treatment period (with a sensitivity analysis of 1-6 months). However, clinical decision making can operate under shorter windows (e.g., hours, days). Prior studies have included windows as short as 0-3 days (Walsh et al., 2017; Walsh et al., 2018). However, with shortened prediction windows, PPVs often decrease (Simon et al., 2019). Some researchers have therefore called for longer follow-up windows (Zhong et al., 2019; Zhu & Zheng, 2018), which contrasts the push for short-term windows. This discrepancy among researchers reflects the trade-off between clinical utility and statistical constraints affecting performance metrics. Nineth, we did not have access to candidate predictors in the following areas: hopelessness, connectedness, or perceived burdensomeness that are causally implicated; biological assays; neuroimaging; data from wearable devices; and natural language processing (Bernert et al., 2020; Schafer et al., 2021). Tenth, we did not have access to racial/ethnic or gender identity information, which are critical variables to inspect when determining potential biases of machine learning algorithms. Finally, we were not aware of the reason for treatment termination (e.g., symptom remission, referral to a different provider, or frequent cancellations/no-shows), which could be a predictor for suicidal behavior.

Conclusion

The aim of the current paper was to assess the performance of several prediction modeling algorithms of suicidal behavior (defined as 6-12 months) among youth after contact with a mental health specialist in Stockholm, Sweden. We did not find support for the outperformance of random forest, gradient boosting, or support vector machine algorithms over

standard logistic regression or variable selection models). Future research is needed to examine how prediction models can augment clinical decision making in real-world settings.

Author Contributions: L.O. conceptualized, ran analyses, and wrote the manuscript. S.F., M.R., and R.K-H. provided significant methodological supervision and result interpretation. BD provided overall supervision. H.L., P.L., and B.D. facilitated data access. M.C. and C.H. provided codebook and variable interpretation support. S.F., M.R., R.K., M.C., C.H., H.L., P.L., and B.D. provided critical manuscript revisions. All authors approved of the final version of the paper for submission.

Conflicts of Interest: The authors have no conflicts of interest to declare.

Funding: This project was supported by grant F31MH121039 to Lauren O'Reilly from the National Institute on Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funders.

References

Altman, D. G., Vergouwe, Y., Royston, P., & Moons, K. G. (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ*, *338*.

Asarnow, J. R., Berk, M., Zhang, L., Wang, P., & Tang, L. (2016). Emergency Department Youth Patients With Suicidal Ideation or Attempts: Predicting Suicide Attempts Through 18 Months of Follow-Up. *Suicide and Life-Threatening Behavior*, n/a-n/a. https://doi.org/10.1111/sltb.12309

Austin, P. C., & Steyerberg, E. W. (2019). The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, *38*(21), 4051-4065. https://doi.org/https://doi.org/10.1002/sim.8281

Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., Nock, M. K., Smoller, J. W., & Reis, B. Y. (2016). Predicting Suicidal Behavior From Longitudinal Electronic Health Records. *American Journal of Psychiatry*, *174*(2), 154-162. https://doi.org/10.1176/appi.ajp.2016.16010077

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, *319*(13), 1317-1318. https://doi.org/10.1001/jama.2017.18391

Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., Morgan, R. L., Evatt, D. P., Tucker, J., & Skopp, N. A. (2019). Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation Prediction Models for Suicide Attempts and Death. https://doi.org/10.1001/jamapsychiatry.2019.0174

Bentley, K. H., Zuromski, K. L., Fortgang, R. G., Madsen, E. M., Kessler, D., Lee, H., Nock, M. K., Reis, B. Y., Castro, V. M., & Smoller, J. W. (2022). Implementing Machine Learning Models for Suicide Risk Prediction in Clinical Practice: Focus Group Study With Hospital Providers. *JMIR Form Res*, *6*(3), e30946. https://doi.org/10.2196/30946

Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnousi, F. (2020). Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *International Journal of Environmental Research and Public Health*, *17*(16), 5929.

Birba, D. E. (2020). *A comparative study of data splitting algorithms for machine learning model selection* KTH Royal Institute of Technology]. Stockholm, Sweden.

Björkenstam, C., Weitoft, G. R., Hjern, A., Nordström, P., Hallqvist, J., & Ljung, R. (2011). School Grades, Parental Education and Suicide—A National Register-Based Cohort Study. *Journal of Epidemiology & Community Health*, *65*(11), 993-998.

Braciszewski, J. M., Lanier, A., Yeh, H.-H., Sala-Hamrick, K., Simon, G. E., Rossom, R. C., Lynch, F. L., Waring, S. C., Lu, C. Y., & Owen-Smith, A. A. (2023). Health diagnoses and service utilization in the year before youth and young adult suicide. *Psychiatric Services*, *74*(6), 566-573.

Brent, D., Bridge, J., Johnson, B., & Connolly, J. (1996). Suicidal Behavior Runs in Families: A Controlled Family Study of Adolescent Suicide Victims. *Archives of General Psychiatry*, *53*(12), 1145-1152.

Carrington, A. M., Fieguth, P. W., Qazi, H., Holzinger, A., Chen, H. H., Mayr, F., & Manuel, D. G. (2020). A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making*, *20*(1), 4. https://doi.org/10.1186/s12911-019-1014-6

Cassells, C., Paterson, B., Dowding, D., & Morrison, R. (2005). Long- and Short-Term Risk Factors in the Prediction of Inpatient Suicide: A Review of the Literature. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *26*(2), 53-63. https://doi.org/10.1027/0227-5910.26.2.53

Center for Disease Control. (2018). *Suicide Rising Across the US* (CDC Vital Signs, Issue.

Chekroud, A. M. (2018). Anticipating Suicide Will Be Hard, But This Is Progress. *American Journal of Psychiatry*, *175*(10), 921-922. https://doi.org/10.1176/appi.ajp.2018.18060714

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *Lancet Psychiatry*, *3*(3), 243-250.

Chen, Q., Zhang-James, Y., Barnett, E. J., Lichtenstein, P., Jokinen, J., D'Onofrio, B. M., Faraone, S. V., Larsson, H., & Fazel, S. (2020). Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: A machine learning study using Swedish national registry data. *PLOS Medicine*, *17*(11), e1003416.

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. *Journal of Clinical Epidemiology*, *110*, 12-22. https://doi.org/https://doi.org/10.1016/j.jclinepi.2019.02.004

Claassen, C. A., Harvilchuck-Laurenson, J. D., & Fawcett, J. (2014). Prognostic Models to Detect and Monitor the Near-Term Risk of Suicide: State of the Science. *American Journal of Preventive Medicine*, *47*(3, Supplement 2), S181-S185. https://doi.org/http://dx.doi.org/10.1016/j.amepre.2014.06.003

Conwell, Y., Lyness, J. M., Duberstein, P., Cox, C., Seidlitz, L., DiGiorgio, A., & Caine, E. D. (2000). Completed Suicide Among Older Patients in Primary Care Practices: A Controlled Study. *Journal of the American Geriatrics Society*, *48*(1), 23-29. https://doi.org/10.1111/j.1532-5415.2000.tb03024.x

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1-17. https://doi.org/https://doi.org/10.1016/j.ins.2012.10.039

Cox, G., Robinson, J., Williamson, M., Lockley, A., Cheung, Y. T., & Pirkis, J. (2012). Suicide Clusters in Young People: Evidence for the Effectiveness of Postvention Strategies. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *33*(4), 208-214.

Czyz, E. K., Horwitz, A. G., & King, C. A. (2016). Self-Rated expectations of suicidal behavior predict future suicide attempts among adolescent and young adult psychiatric emergency patients. *Depression and Anxiety*, n/a-n/a. https://doi.org/10.1002/da.22514

D'Onofrio, B. M., Singh, A., Neiderhiser, J., Iliadou, A., Lambe, M., Hultman, C., Långström, N., & Lichtenstein, P. (2010). A Quasi-Experimental Study of Maternal Smoking During Pregnancy and Offspring Academic Achievement. *Child Development*, *81*(1), 80-100.

Deo, R. C. (2015). Machine Learning in Medicine [10.1161/CIRCULATIONAHA.115.001593]. *Circulation*, *132*(20), 1920. http://circ.ahajournals.org/content/132/20/1920.abstract

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., & Leisch, M. F. (2009). Package 'e1071'. *R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html*.

Edwards, C., & Raskutti, B. (2004). The effect of attribute scaling on the performance of support vector machines. Australasian Joint Conference on Artificial Intelligence,

Ehrmann, D. E., Joshi, S., Goodfellow, S. D., Mazwi, M. L., & Eytan, D. (2023). Making machine learning matter to clinicians: model actionability in medical decision-making. *npj Digital Medicine*, *6*(1), 7. https://doi.org/10.1038/s41746-023-00753-7

Favril, L., Yu, R., Uyar, A., Sharpe, M., & Fazel, S. (2022). Risk factors for suicide in adults: systematic review and meta-analysis of psychological autopsy studies. *BMJ Ment Health*, *25*(4), 148-155.

Fazel, S., & Grann, M. (2006). The Population Impact of Severe Mental Illness on Violent Crime. *American Journal of Psychiatry*, *163*(8), 1397-1403.

Fazel, S., Vazquez-Montes, M. D., Molero, Y., Runeson, B., D'Onofrio, B. M., Larsson, H., Lichtenstein, P., Walker, J., Sharpe, M., & Fanshawe, T. R. (2023). Risk of death by suicide following self-harm presentations to healthcare: development and validation of a multivariable clinical prediction rule (OxSATS). *BMJ Ment Health*, *26*(1).

Fazel, S., Wolf, A., Larsson, H., Lichtenstein, P., Mallett, S., & Fanshawe, T. R. (2017). Identification of low risk of violent crime in severe mental illness with a clinical prediction tool (Oxford Mental

Illness and Violence tool [OxMIV]): a derivation and validation study. *The Lancet. Psychiatry*, *4*(6), 461-468. https://doi.org/10.1016/S2215-0366(17)30109-8

Fazel, S., Wolf, A., Larsson, H., Mallett, S., & Fanshawe, T. R. (2019). The prediction of suicide in severe mental illness: development and validation of a clinical prediction rule (OxMIS). *Translational Psychiatry*, *9*(1), 1-10.

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Jaroszewski, A. C., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, *143*(2), 187-232.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Qian, J. (2021). Package 'glmnet'. *CRAN R Repository*, *595*.

Glenn, C. R., & Nock, M. K. (2014). Improving the Short-Term Prediction of Suicidal Behavior. *American Journal of Preventive Medicine*, *47*(3, Supplement 2), S176-S180. https://doi.org/http://dx.doi.org/10.1016/j.amepre.2014.06.004

Goldston, D. B., Daniel, S. S., Reboussin, D. M., Reboussin, B. A., Frazier, P. H., & Kelley, A. E. (1999). Suicide Attempts Among Formerly Hospitalized Adolescents: A Prospective Naturalistic Study of Risk During the First 5 Years After Discharge. *Journal of the American Academy of Child & Adolescent Psychiatry*, *38*(6), 660-671.

Greenfield, B., Henry, M., Weiss, M., Tse, S. M., Guile, J., Dougherty, G., Zhang, X., Fombonne, E., Lapalme-Remis, S., & Harnden, B. (2008). Previously suicidal adolescents: Predictors of six-month outcome. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *17*(4), 197-201.

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008, 18-20 Oct. 2008). On the Class Imbalance Problem. 2008 Fourth International Conference on Natural Computation, Jinan, China.

Horwitz, A. G., Czyz, E. K., & King, C. A. (2015). Predicting Future Suicide Attempts Among Adolescent and Emerging Adult Psychiatric Emergency Patients. *Journal of Clinical Child & Adolescent Psychology*, *44*(5), 751-761. https://doi.org/10.1080/15374416.2014.910789

Jacobsson, A. (2009). In English - the National Patient Register *Socialstyrelsen*. https://www.socialstyrelsen.se/en/statistics-and-data/registers/national-patient-register/

Jung, J. S., Park, S. J., Kim, E. Y., Na, K.-S., Kim, Y. J., & Kim, K. G. (2019). Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. *PLoS ONE*, *14*(6), e0217639. https://doi.org/10.1371/journal.pone.0217639

Kessler, R., Hwang, I., Hoffmire, C. A., McCarthy, J. F., Petukhova, M. V., Rosellini, A. J., Sampson, N. A., Schneider, A. L., Bradley, P. A., & Katz, I. R. (2017). Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. *International Journal of Methods in Psychiatric Research*, *26*(3), e1575.

Kessler, R., 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., Bell, A. M., 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). *Molecular Psychiatry*, *22*(4), 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., III., Cai, T., Colpe, L. J., Cox, K. L., C.S., F., Gilman, S. E., Gruber, M. J., Heeringa, S. G., Lewandowski-Romps, L., Li, J., Millikan-Bell, A. M., Naifeh, J. A., Nock, M. K., . . . Ursano, R. J. (2015). Predicting Suicides After Psychiatric Hospitilization in US Army Soldiers: The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA Psychiatry*, *72*(1), 49-57.

King, C. A., Berona, J., Czyz, E., Horwitz, A. G., & Gipson, P. Y. (2015). Identifying Adolescents at Highly Elevated Risk for Suicidal Behavior in the Emergency Department. *Journal of Child and Adolescent Psychopharmacology*, *25*(2), 100-108. https://doi.org/10.1089/cap.2014.0049

King, C. A., Kerr, D. C. R., Passarelli, M. N., Foster, C. E., & Merchant, C. R. (2010). One-Year Follow-Up of Suicidal Adolescents: Parental History of Mental Health Problems and Time to Post-Hospitalization Attempt. *Journal of Youth and Adolescence*, *39*(3), 219-232. https://doi.org/10.1007/s10964-009-9480-2

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. In J. Breuker, R. Dieng-Kuntz, N. Guarino, J. N. Kok, J. Liu, R. López de Mántaras, R. Mizoguchi, M. Musen, & N. Zhong (Eds.), *Frontiers in Artificial Intelligence and Applications* (Vol. 160, pp. 3-24). IOS Press.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (5th ed.). Springer Science+Busines Media.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., & Team, R. C. (2020). Package 'caret'. *The R Journal*, *223*, 7. https://cran.radicaldevelop.com/web/packages/caret/caret.pdf

Large, M., Sharma, S., Cannon, E., Ryan, C., & Nielssen, O. (2011). Risk factors for suicide within a year of discharge from psychiatric hospital: a systematic meta-analysis. *Australian and New Zealand Journal of Psychiatry*, *45*, 619-628.

Lodebo, B. T., Möller, J., Larsson, J.-O., & Engström, K. (2017). Socioeconomic position and self-harm among adolescents: a population-based cohort study in Stockholm, Sweden. *Child and Adolescent Psychiatry and Mental Health*, *11*, 46. https://doi.org/10.1186/s13034-017-0184-1

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a package for binary imbalanced learning. *R journal*, *6*(1), 79-89.

Malhotra, R., & Kamal, S. (2019). An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data. *Neurocomputing*, *343*, 120-140. https://doi.org/ https://doi.org/10.1016/j.neucom.2018.04.090

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Mohr, D. C., Riper, H., & Schueller, S. M. (2018). A solution-focused research approach to achieve an implementable revolution in digital mental health. *JAMA Psychiatry*, *75*(2), 113-114. https://doi.org/10.1001/jamapsychiatry.2017.3838

Nock, M. K. (2012). Future Directions for the Study of Suicide and Self-Injury. *Child and Adolescent Psychology*, *41*(2), 255-259.

Nock, M. K., Borges, G., Bromet, E., Cha, C., Kessler, R., & Lee, S. (2008). Suicide and Suicidal Behavior. *Epidemiologic Reviews*, *30*(1), 133-154.

Nock, M. K., Borges, G., Bromet, E., Cha, C., Kessler, R., & Lee, S. (2012). The Epidemiology of Suicide and Suicidal Behavior. In M. Nock, G. Borges, & Y. Ono (Eds.), *Suicide: Global Perspectives from the WHO World Mental Health Surveys* (pp. 5-32). Cambridge University Press.

Nock, M. K., Green, J., Hwang, I., McLaughlin, K., Sampson, N., Zaslavsky, A., & Kessler, R. (2013). Prevalence, Correlates, and Treatment of Lifetime Suicidal Behavior Among Adolescents. *JAMA Psychiatry*, *70*(3), 300-310.

Nordin, N., Zainol, Z., Mohd Noor, M. H., & Chan, L. F. (2022). Suicidal behaviour prediction models using machine learning techniques: A systematic review. *Artificial Intelligence in Medicine*, *132*, 102395. https://doi.org/https://doi.org/10.1016/j.artmed.2022.102395

Olfson, M. (2017). Suicide risk after psychiatric hospital discharge. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2017.1043

Olfson, M., Wall, M., Wang, S., & et al. (2016). Short-term suicide risk after psychiatric hospital discharge. *JAMA Psychiatry*, *73*(11), 1119-1126. https://doi.org/10.1001/jamapsychiatry.2016.2035

Pisani, A. R., Murrie, D. C., & Silverman, M. M. (2016). Reformulating Suicide Risk Formulation: From Prediction to Prevention. *Academic psychiatry : the journal of the American Association of*

Directors of Psychiatric Residency Training and the Association for Academic Psychiatry, *40*(4), 623-629. https://doi.org/10.1007/s40596-015-0434-6

Prinstein, M. J., Nock, M. K., Simon, V., Aikins, J. W., Cheah, C. S. L., & Spirito, A. (2008). Longitudinal Trajectories and Predictors of Adolescent Suicidal Ideation and Attempts Following Inpatient Hospitalization. *Journal of Consulting and Clinical Psychology*, *76*(1), 92-103. https://doi.org/10.1037/0022-006X.76.1.92

R Core Team. (2022). *R: A language and environment for statistical computing.* . In http://www.R-project.org/

Reger, M. A., Ammerman, B. A., Carter, S. P., Gebhardt, H. M., Rojas, S. M., Lee, J. M., & Buchholz, J. (2021). Patient feedback on the use of predictive analytics for suicide prevention. *Psychiatric Services*, *72*(2), 129-135.

Ribeiro, J. D., Franklin, J. C., Fox, K. R., Bentley, K. H., Kleiman, E. M., Chang, B. P., & Nock, M. K. (2016). Letter to the editor: Suicide as a complex classification problem: Machine learning and related techniques can advance suicide prediction - A reply to Roaldset. *Psychological Medicine*, *46*(9), 2009-2010.

Ribeiro, J. D., Huang, X., Fox, K. R., Walsh, C. G., & Linthicum, K. P. (2019). Predicting Imminent Suicidal Thoughts and Nonfatal Attempts: The Role of Complexity. *Clinical Psychological Science*, 2167702619838464. https://doi.org/10.1177/2167702619838464

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Sahlin, H., Kuja-Halkola, R., Bjureberg, J., & et al. (2017). Association between deliberate self-harm and violent criminality. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2017.0338

Sariaslan, A., Långström, N., D'Onofrio, B. M., Hallqvist, J., Franck, J., & Lichtenstein, P. (2013). The Impact of Neighborhood Deprivation on Adolescent Violent Criminality and Substance Misuse: A Longitudinal, Quasi-Experimental Study of the Total Swedish Population. *International Journal of Epidemiology*, *42*(4), 1057-1066.

SAS Institute Inc. (2016).

Schafer, K. M., Kennedy, G., Gallyer, A., & Resnik, P. (2021). A direct comparison of theory-driven and machine learning prediction of suicide: A meta-analysis. *PLoS ONE*, *16*(4), e0249833.

Simon, G., Hunkeler, E., Fireman, B., Lee, J., & Savarino, J. (2007). Risk of suicide attempt and suicide death in patients treated for bipolar disorder. *Bipolar Disorders*, *9*(5), 526-530. https://doi.org/10.1111/j.1399-5618.2007.00408.x

Simon, G. E., Shortreed, S. M., & Coley, R. Y. (2019). Positive Predictive Values and Potential Success of Suicide Prediction Models. *JAMA Psychiatry*, *76*(8), 868-869. https://doi.org/10.1001/jamapsychiatry.2019.1516

Simon, R. I. (2006). Imminent Suicide: The Illusion of Short-Term Prediction. *Suicide and Life-Threatening Behavior*, *36*(3), 296-301.

Spirito, A., Valeri, S., Boergers, J., & Donaldson, D. (2003). Predictors of Continued Suicidal Behavior in Adolescents Following a Suicide Attempt [Article]. *Journal of Clinical Child & Adolescent Psychology*, *32*(2), 284. http://proxyiub.uits.iu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=9873272&site=ehost-live&scope=site

Statistics Sweden. (2010). *Multi-Generation Register 2009: A Description of Contents and Quality* Orebro, Sweden Retrieved from https://www.scb.se/en/finding-statistics/statistics-by-subject-area/other/other/other-publications-non-statistical/

Statistics Sweden. (2011). *Integrated Database for Labour Market Research 1990-2009*. Retrieved from https://www.scb.se/en/finding-statistics/statistics-by-subject-area/labour-market/

Stene-Larsen, K., & Reneflot, A. (2019). Contact with primary and mental health care prior to suicide: a systematic review of the literature from 2000 to 2017. *Scandinavian Journal of Public Health*, *47*(1), 9-17.
Tang, N., & Crane, C. (2006). Suicidality in chronic pain: a review of the prevalence, risk factors and psychological links. *Psychological Medicine*, *36*(5), 575-586.
Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic Net Regularization Paths for All Generalized Linear Models. *J Stat Softw*, *106*. https://doi.org/10.18637/jss.v106.i01
Tucker, R. P., Tackett, M. J., Glickman, D., & Reger, M. A. (2018). Ethical and Practical Considerations in the Use of a Predictive Model to Trigger Suicide Prevention Interventions in Healthcare Settings. *Suicide and Life-Threatening Behavior*, n/a-n/a. https://doi.org/10.1111/sltb.12431
US Surgeon General, & National Action Alliance for Suicide Prevention. (2021). *The Surgeon General's Call to Action: To Implement the National Strategy for Suicide Prevention*.
van Mens, K., de Schepper, C. W. M., Wijnen, B., Koldijk, S. J., Schnack, H., de Looff, P., Lokkerbol, J., Wetherall, K., Cleare, S., C O'Connor, R., & de Beurs, D. (2020). Predicting future suicidal behaviour in young adults, with different machine learning techniques: A population-based longitudinal study. *Journal of Affective Disorders*, *271*, 169-177. https://doi.org/https://doi.org/10.1016/j.jad.2020.03.081
Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, *352*, i6. https://doi.org/10.1136/bmj.i6
Wallerstedt, S. M., Wettermark, B., & Hoffmann, M. (2016). The First Decade with the Swedish Prescribed Drug Register – A Systematic Review of the Output in the Scientific Literature. *Basic & Clinical Pharmacology & Toxicology*, *119*(5), 464-469. https://doi.org/10.1111/bcpt.12613
Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science*, 2167702617691560. https://doi.org/10.1177/2167702617691560
Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *Journal of Child Psychology and Psychiatry*, *59*(12), 1261-1270. https://doi.org/10.1111/jcpp.12916
Witte, T. K., Fitzpatrick, K. K., Warren, K. L., Schatschneider, C., & Schmidt, N. B. (2006). Naturalistic evaluation of suicidal ideation: Variability and relation to attempt status. *Behaviour Research and Therapy*, *44*(7), 1029-1040. https://doi.org/https://doi.org/10.1016/j.brat.2005.08.004
Zhong, Q.-Y., Mittal, L. P., Nathan, M. D., Brown, K. M., Knudson González, D., Cai, T., Finan, S., Gelaye, B., Avillach, P., Smoller, J. W., Karlson, E. W., Cai, T., & Williams, M. A. (2019). Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. *European Journal of Epidemiology*, *34*(2), 153-162. https://doi.org/10.1007/s10654-018-0470-0
Zhu, L., & Zheng, W. (2018). Informatics, data science, and artificial intelligence. *JAMA*, *320*(11), 1103-1104. https://doi.org/10.1001/jama.2018.8211