Empirical Article

Clinical Psychological Science
1–13

DOI: 10.1177/2167702617691560
www.psychologicalscience.org/CPS

Predicting Risk of Suicide Attempts Over Time Through Machine Learning

Colin G. Walsh[1,2,3], Jessica D. Ribeiro[4], and Joseph C. Franklin[4]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center; [2]Department of Medicine, Vanderbilt University Medical Center; [3]Department of Psychiatry, Vanderbilt University Medical Center; and [4]Department of Psychology, Florida State University

Abstract

Traditional approaches to the prediction of suicide attempts have limited the accuracy and scale of risk detection for these dangerous behaviors. We sought to overcome these limitations by applying machine learning to electronic health records within a large medical database. Participants were 5,167 adult patients with a claim code for self-injury (i.e., ICD-9, E95x); expert review of records determined that 3,250 patients made a suicide attempt (i.e., cases), and 1,917 patients engaged in self-injury that was nonsuicidal, accidental, or nonverifiable (i.e., controls). We developed machine learning algorithms that accurately predicted future suicide attempts (AUC = 0.84, precision = 0.79, recall = 0.95, Brier score = 0.14). Moreover, accuracy improved from 720 days to 7 days before the suicide attempt, and predictor importance shifted across time. These findings represent a step toward accurate and scalable risk detection and provide insight into how suicide attempt risk shifts over time.

Keywords

suicide prevention, prediction, prevention, classification

Received 8/9/16; Revision accepted 1/10/17

Suicide attempts are a major public health problem, with an estimated 25 million nonfatal suicide attempts occurring each year worldwide (Centers for Disease Control and Prevention [CDC], 2016; World Health Organization, 2016). Beyond considerable economic and societal burdens associated with nonfatal attempts (Shepard, Gurewich, Lwin, Reed, & Silverman, 2016), nonfatal suicide attempts are among the strongest predictors of suicide death—a leading cause of death worldwide (Ribeiro et al., 2016a). The scope and seriousness of the problem have prompted substantial research attention (Franklin et al., 2017). Yet our ability to predict nonfatal attempts remains marginally above chance levels (Bentley et al., 2016; Chang et al., 2016; Franklin et al., 2017; Ribeiro et al., 2016a). Rates of nonfatal suicide attempts remain intractable, with recent estimates suggesting that nonfatal attempts may be on the rise (CDC, 2016). The purpose of the present study was to evaluate the accuracy and temporal variation of a potentially scalable suicide attempt risk detection strategy: machine learning applied to electronic health records (EHRs).

Recent meta-analyses have shown that the ability to predict suicide attempts has been near chance for decades (Franklin et al., 2017). A major reason for this poor prediction is that the majority of studies tested predictors in isolation (e.g., a depression diagnosis), and even the best isolated predictors are inaccurate (Franklin et al., 2017; Ribeiro et al., 2016a). Accurate suicide attempt prediction may require complex combinations of hundreds of risk factors. Traditional statistical techniques are not ideal for such analyses; fortunately, machine learning (ML) techniques are well suited for such problems. These techniques can test a wide range of complex associations among large numbers of potential factors to produce algorithms that optimize prediction. Retrospective ML studies suggest that this approach may be promising

Corresponding Author:
Colin G. Walsh, Departments of Biomedical Informatics, Medicine, and Psychiatry, Vanderbilt University Medical Center, 2525 West End Ave., Ste. 1475, Nashville, TN 37203
E-mail: colin.walsh@vanderbilt.edu

(Delgado-Gomez, Blasco-Fontecilla, Sukno, Socorro Ramos-Plasencia, & Baca-Garcia, 2012; Mann, Ellis, Waternaux, & Liu, 2008), producing discriminative accuracies for suicide attempters (AUCs = 0.60 to 0.80) that exceed those of isolated risk factors (AUCs = 0.58; Franklin et al., 2017). Although the retrospective design of these studies precluded any conclusions about prediction, a recent prospective ML study produced similar findings (AUC = 0.76; Kessler, Stein, et al., 2016).

Brief Overview of ML

Although ML has been a part of the computer science field for many decades, it has only recently been applied to clinical psychology. Here, we provide a brief overview to orient readers to what ML is, its advantages over traditional statistical approaches in clinical psychology, and the metrics used to evaluate the performance of ML algorithms.

Many problems in clinical psychology can be framed as classification problems. That is, much of clinical psychological science is aimed at classifying what psychopathologies exist, who possess a particular form of psychopathology, when or if someone will develop a form of psychopathology, who will engage in a problematic behavior in the future, and who will respond to a particular form of treatment. Many of these classification problems are complex, requiring the simultaneous consideration of tens or hundreds of factors to produce accurate classification. Classification problems are solved with algorithms, which are a sets of steps for solving a problem. Algorithms can be simple (e.g., 2 + 2) or highly complex (e.g., the Google search algorithm, which considers more than 200 factors); simple algorithms are best for simple classification problems, but complex algorithms are necessary for complex classification problems. Due to convention and the limitations of most traditional statistical approaches, clinical psychological science has often attempted to use simple algorithms to solve complex classification problems. This approach can produce statistical significance but has a limited ability to produce clinical significance. For example, as noted earlier, recent meta-analyses on hundreds of studies from the past 50 years indicate that the ability to predict future suicide attempts has always been at near chance levels. The primary reason for this lack of progress is that researchers have almost always used a single factor (i.e., a simple algorithm) to predict future suicide attempts (i.e., a complex classification problem; see Ribeiro et al., 2016b). Fortunately, ML represents a potentially effective approach for the development of complex algorithms capable of solving (or making substantial progress toward solving) complex classification problems.

There is a wide range of ML methods, with each method possessing advantages and disadvantages (see Kotsiantis, 2007); however, all possess at least four notable advantages over traditional statistical approaches for the development of complex algorithms. First, ML methods determine the most effective and parsimonious algorithm on their own. This is where ML gets its name from: The *machine* itself automatically progresses through many iterations as it *learns* the ideal set of operations for classifying data into the desired groups. Traditional statistical approaches in clinical psychology generally require that the researcher determine the ideal algorithm a priori. This traditional approach typically produces a simple algorithm (i.e., fewer than 10 factors, primarily additive or multiplicative operations) that is unlikely to be the most effective or parsimonious classification algorithm. Second, ML algorithms are able to consider complex combinations of factors in terms of both number and type. Such combinations are unlikely to arise from traditional statistical approaches in clinical psychology.

Third, most ML approaches are enacted with clinical significance in mind. Most traditional statistical approaches focus on explaining a statistically significant proportion of variance in a particular data set; most ML approaches are primarily concerned with raw classification performance (see the discussion later) and how accurately the ML algorithm will classify new data points. Fourth, ML approaches are well suited to classification problems involving "high-dimensional" data, or data in which there are a large number of potential predictors. Compared to traditional approaches, ML approaches are generally more resistant against "overfitting" high-dimensional data. Overfitting occurs when the model capitalizes on the idiosyncratic noise of a particular data set, producing a model that is highly accurate in one data set but performs poorly when applied to other data sets. This is particularly likely to occur when a model includes a large number of predictors because higher numbers of predictors mean more opportunity to account for unique variance. As described later (see the Method section), ML approaches typically integrate strategies to guard against overfitting. These four advantages are especially pronounced in larger data sets, where a greater number of factors and cases/controls increases the potential for accurate and robust classification.

Both traditional statistical and ML algorithms can be measured with respect to discrimination or classification performance—the ability of these algorithms to separate data into different classes, often denoted "cases" and "controls" in biomedical investigations. There are several ways to quantify classification performance, each with its own trade-offs and biases. As described in greater detail later, discriminative accuracy—in the literature, frequently

measured with an area under the receiver operating characteristic curve statistic (i.e., AUC or AUROC)—is often presented as a metric of how well an ML algorithm distinguishes cases from controls. AUC scores can range from 0.5 (accuracy no better than chance) to 1.0 (perfect accuracy). This metric can be misleading when true negatives (i.e., controls) far outnumber true positives (i.e., cases) because algorithms can achieve high AUCs with a good identification of controls but poor identification of cases. Due to low base rates, such case-control imbalances are common in suicide research. It is accordingly important for ML work in this area to also consider precision-recall statistics, which jointly consider positive predictive value (the ratio of true positives compared to the sum of true and false positives) and the recall or sensitivity of an algorithm (i.e., in essence, how accurately an algorithm identifies cases).

Although an algorithm may perform well in terms of AUC/precision recall, its predictions may not reflect the real-world probability of a particular phenomenon; that is, the algorithm may not be *calibrated* to the real world. A well-calibrated model that suggests a 10% risk of an outcome would find that 10 out of 100 similar data points would actually have that outcome in practice. Calibration can be particularly important for suicide-related outcomes because (a) there are dire consequences when a false negative occurs, (b) there can also be major consequences for false positives, including use of valuable and limited health care resources, and (c) low real-world base rates make it difficult to translate study information to the clinic. In such instances it is important to also evaluate calibration with metrics such as Brier scores. Considering all three of these metrics together provides a well-rounded assessment of classification performance.

The Present Study

The first aim of the present study was to build on preliminary studies to evaluate a novel suicide risk detection strategy. First, this study was longitudinal, examining risk at time points ranging from 1 week to 2 years. Second, it included a larger number of suicide attempters (N = 3,250) than the existing longitudinal ML suicide attempt study ($N \sim 40$), and each suicide attempt was validated by suicide researchers with years of experience in classifying suicidal behaviors. Third, analyses included a stringent comparison group: individuals whose records contained International Classification of Diseases (ICD-9) codes for self-injury but were judged not to have made a suicide attempt after comprehensive chart review. These individuals typically received these codes for unintentional drug overdoses, accidental injury, nonsuicidal self-injury, or injury with unclear intent. Given the substantial overlap between risk factors for these behaviors and suicidal behaviors (Fox et al., 2015; Franklin et al., 2017; Neeleman, 2001; Wenzel & Beck, 2008), we expected this control group to provide a rigorous test of ML algorithm performance. Fourth, to empirically examine the comparative rigor of this control group, we also conducted secondary analyses with a random sample of patients from the general hospital population as the control group. We expected these secondary analyses to produce better ML performance, but focus on analyses involving the more rigorous control group of nonsuicidal self-injurers. Fifth, to rule out the possibility that the present approach would primarily be effective for individuals with a prior history of self-injurious behavior, we conducted secondary analyses that separately utilized repeat and single attempters as cases. Sixth, ML analyses drew from a much broader set of potential predictors compared to previous studies. Seventh and finally, we employed several strategies to prevent overfitting and relied on well-accepted measures of performance including precision/recall and calibration in evaluation.

A second major aim was to investigate how risk estimates change as suicide attempts become more imminent. Although short-term risk is a major part of many suicide theories (e.g., O'Connor, 2011; Wenzel & Beck, 2008) and a major focus of clinical guidelines (e.g., Rudd et al., 2006), few studies have examined short-term risk for suicide attempts. The shortest term study to our knowledge included a follow-up interval of 1 month (Ribeiro et al., 2012); the second shortest included a follow-up interval of 6 months (Cha, Najmi, Park, Finn, & Nock, 2010). Moreover, few studies have examined how risk for suicidal behaviors shifts over time, and none have examined how suicide risk shifts from years out to days out. In the present study we developed separate ML algorithms with information from up to 7, 14, 30, 60, 90, 180, 365, or 720 days before the suicide attempt. We were primarily interested in whether prediction improved as the suicide attempt became more imminent (i.e., from 720 to 7 days before the attempt).

An exploratory aim of this study was to examine how specific factors contributed to changes in risk over time. We investigated whether the most important factors in ML algorithms changed over time, and we sought to identify which factors produced the largest spikes in ML-risk estimates during the 2 years before the suicide attempt. Given difficulties with interpreting specific factor contributions within ML algorithms (Hastie, Tibshirani, & Friedman, 2001; Van Calster et al., 2013) and that present analyses were specific to information found in EHRs, these analyses were exploratory. However, this exploration has the potential to provide a foundation for studies that more directly address this question.

We hypothesize that the present approach will produce accurate prediction of suicide attempts, improved

model performance as the attempt approaches, and shifts in risk factor importance as the attempt approaches. This investigation has the potential to provide an important step toward scalable and accurate suicide attempt risk detection and to provide new information about temporal variation in suicide attempt risk.

Method

Participants

Data were drawn from the BioVU Synthetic Derivative (SD), a deidentified data repository of clinical EHR data at Vanderbilt University Medical Center (Roden et al., 2008). The SD incorporates data across two decades and millions of patients. Claims data identifying self-injury codes (ICD9 E95x) produced a candidate data set of 5,543 records of adult patients. Their records were reviewed comprehensively by two suicide experts (J.D.R. and J.C.F.) and labeled when applicable as suicide attempts. Specifically, a suicide attempt was defined as direct nonfatal self-injury enacted with nonzero suicidal intent. The date of the most recent documented attempt was recorded for reference in temporal analyses. Experts independently reviewed records and agreed on 95% of cases; subsequent discussion produced agreement on the remaining 5%. Individuals deemed to have made a nonfatal suicide attempt were classified as cases; individuals for whom this determination could not be made were classified as controls. An exception to this rule was suicide decedents; individuals who died by suicide were not included in the present analyses.

As noted earlier, two other types of group analyses were performed. First, to empirically test the rigor of the nonsuicidal control group noted earlier, in secondary analyses we compared suicide attempt cases to a random sample of hospital cases. This group was derived from the same clinical data repository by randomly selecting 12,695 adults with no documented history of suicide attempts. Age- and gender-matching were not performed given the well-documented predictive importance of age and gender in predicting risk of suicidal behaviors; matching, if performed, generates a uniform and therefore noninformative distribution in the data. These controls were selected to represent a broader segment of the population, and represent an "extreme groups" design that we expected to produce better discriminative performance. However, a major trade-off of this performance is the lowered rigor of ML performance evaluation and the correspondingly diminished quality of the information obtained by the test. We accordingly focused on the more rigorous nonsuicidal self-injury control group analyses.

Second, given the importance of prior suicidal behaviors in future suicidal behavior (see Ribeiro et al., 2016a), one possibility is that the present approach would be more effective for individuals with a documented history of prior suicidal behaviors. To examine this possibility, we conducted secondary analyses that respectively included only repeat or single attempters as cases. Similar performance across these algorithms would indicate that, at least within the present ML approach, knowledge about prior suicidal behavior does not substantially improve predictive accuracy for future suicide attempts. All methods were approved by the Vanderbilt University Medical Center Institutional Review Board.

Statistical modeling

Modeling setup. Data were preprocessed using Python with statistical analyses performed in R (R Development Core Team, 2012). Relevant Python libraries included the SciPy ecosystem, NumPy, Pandas, iPython (Pedregosa et al., 2011; Perez & Granger, 2007; van der Walt, Colbert, & Varoquaux, 2011). In R, random forests were implemented via the ranger package (Wright & Ziegler, 2015). Logistic regression was implemented via glm in base R.

Modeling approach. Random forests have been broadly accepted in the ML community for performance—in both accuracy and ease of implementation—and robustness (Amalakuhan et al., 2012; Austin, Lee, Steyerberg, & Tu, 2012; Futoma, Morris, & Lucas, 2015; Harrell & Slaughter, 2008; Kessler, van Loo, et al., 2016). The random forest represents an ensemble learning method that comprises a set of decision trees that are generated via recursive sampling of bootstrapped samples of predictor data. Decision trees are constructed via recursive splitting of random subsets of predictors to form "parent" and "child" nodes (Malley, Kruppa, Dasgupta, Malley, & Ziegler, 2011; Wright & Ziegler, 2015). "Splits" in the decision trees reflect binary (i.e., yes/no) questions phrased with respect to predictors. Several parameters are set by the user, including the number of predictors at each node and how each predictor is selected. In the present study, the number of predictors selected at each node was set as the square root of the total number of predictors. Predictors were selected via an error minimization approach, which selects the predictor that results in minimum mean squared error across all other randomly selected predictors in a particular node. The process is iterated until a "terminal node" (i.e., a node that does not have a child node) is achieved, yielding a single decision tree. The overall process is then repeated a set number of times, in turn producing a multitude (i.e., "forest") of decision trees. For this study, the process was repeated 500 times. Risk estimates are determined based on the proportion of trees that predict an outcome will occur versus not.

Predictor importance was quantified by evaluating decrease in "node impurity" at each split across all decision trees in the forest (Wright & Ziegler, 2015). In its simplest case, node impurity can be considered the proportion of cases versus controls at a particular node. The random forest module used here measures estimated variance in variables across trees; the variables that maximize variance of responses between nodes are those that track more closely with case or control status. Thus, they are more "important" for model performance.

An advantage of random forests is an ability to tolerate categorical/nominal variables such as race or gender (M = male, F = female) in the modeling paradigm. Random forests handle nominal variables via "split points" as a general approach. Splitting categorical variables means that at various points in the decision trees, some categories will be placed on the left of a split and the remainder on the right. This process, iterated thousands of times over hundreds of trees, relies on the same optimization parameters used for continuous variables to determine the optimal splits to distinguish cases and controls and, in doing so, can assign weights to the categories within nominal variables as output for future predictions.

To compare the performance of the present ML approach to a traditional logistic regression approach, we conducted a secondary set of logistic regression analyses. Specifically, traditional, nonregularized logistic regression was performed on the same data and candidate predictor set to estimate usual performance. Predictors were modeled as first-order terms without interactions in multiple logistic regression models constructed at each study time point of interest.

Bootstrapping was used to assess, quantify, and adjust for model optimism (i.e., spurious inflation in performance; overfitting; Harrell, 2006; Miao, Francisco, Boscardin, Francisco, & Francisco, 2013). In this approach, we first train a predictive model using all the available data. We also create a set of bootstrap replicates based on the original data; in this set, we used 100 bootstrap replicates. Models are then generated on the bootstrap replicates. The models derived from the bootstrapped data are then applied to the original nonbootstrapped data. This provides an estimate of "out of bag" performance. Differences between "out of bag" performance and bootstrapped performance are then calculated and averaged. The resultant estimate reflects the degree of optimism of the original model. The original model is then corrected by subtracting the degree of optimism from the original model performance. Although cross-validation and holdout sets are also viable alternatives to guard against overfitting, bootstrap optimism has been shown to provide more conservative estimates of model performance and lower absolute and mean squared errors (Hastie et al., 2001; Smith et al., 2014). Of note, the bootstrap replicates generated at this step are unique from those discussed earlier that were used to generate decision trees within random forest.

Model performance was evaluated using AUC and precision and recall metrics. To evaluate calibration in this study, we visually inspected the calibration plots and also considered Brier scores, which reflect the accuracy of probabilistic predictions. Brier scores range from 0 to 1; a score of 0 represents perfect calibration and discrimination. Brier scores can be calculated with the following formula,

$$B = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - o_i\right)^2$$

where B is the Brier score, N is the sample size of predicted instances or individuals, p_i is the forecast for individual i, and o_i is the outcome status for this same individual. A lower Brier score accordingly indicates less discrepancy between the *predicted probability* of an outcome and *actual* outcome for each individual. Consider, as an example, weather forecasting: Assume an algorithm is used to predict the chance of rain, and it predicts a rain chance of 10% tomorrow. Further assume that it does rain that day. The resulting Brier score would be 0.90, indicating a poorly calibrated model. However, if it does not rain that day, the Brier score would be 0.10, indicating good calibration. Extending this analogy to multiple days (or data points), a model predicting a 10% chance of rain on Monday would be well calibrated if it rained on one out of ten days that are similar to the Monday in question. We note in this example that the Brier score acts also as a reflection of accuracy in addition to pure calibration. Similarly, if an ML algorithm predicts that a given individual has an 80% chance of making a suicide attempt, and the individual makes an attempt, the (individual) Brier score would be 0.20. However, if this individual did not make an attempt, the (individual) Brier score would be 0.80. We include the Brier scores for all primary analyses later.

Predictor types. Several data classes were modeled in this study: (a) Demographic data including age, gender, and race/ethnicity were included in candidate predictors; (b) diagnoses based in claims data were mapped through the Center for Medicare and Medicaid Services Hierarchical Condition Categories (CMS-HCC), which aggregate more than 14,000 ICD9 and 65,000 ICD10 claims codes into 189 clinically meaningful categories; (c) past health care utilization including numbers of outpatient clinic, inpatient admission, and emergency department visits were summed for each single year over the preceding 5-year period; (d) evidence of prior suicide attempts was

captured solely through diagnostic codes (often, E95X ICD9 codes) as described earlier (note: determination of single and repeat attempters was made by expert evaluations of EHRs, but these evaluations were not included as predictors because they would not typically be available and accordingly would not be a scalable predictor); (e) vital signs were not included given the temporally variant nature of this work, with the exception of body mass index, which was available and extractable at relevant study time points; (f) socioeconomic status was partially captured via the Area Deprivation Index, a census-based index incorporating education, property value, employment data, and others to assign a numeric deprivation index to zip codes (Singh, 2003);[1] and (g) medication data were extracted from clinical notes including problem lists, prescription events, and clinical documentation using natural language processing (Xu et al., 2010). Given the tremendous feature dimensionality of medication data including names, brand names, and dosages, dimensionality reduction was accomplished via mapping medication generic and brand names to the Anatomic Therapeutic Classification, Level V (ATC, Level 5)—clinically closest to the medication class level. The numeric values for each medication class were similar conceptually to Term Frequency–Inverse Document Frequency, a well-known metric in the natural language processing community intended to capture term importance compared to all words (in this case, medication names) in a corpus of text documents (Robertson, 2004).

Missing data. Because of the data preprocessing strategy to aggregate visits, medications, and diagnoses into counts of classes, missing data in these categories were minimized. Demographics were recorded in the study data source at the individual patient level and therefore were not missing. Exceptions were date of birth, which was missing in 13.7% of training data for general controls and 0% in the suicidal control cohort, and body mass index, which was missing in 53.9% of cases in suicidal controls and in 59% of data in the general control cohort. In general, we adhered to multiple imputation to address missingness in each iteration of the algorithm (Deeks, 2011; Harrell, 2006; Rubin, 2004). Specifically, the *aregImpute* algorithm in R was implemented to impute missing data via additive regression, bootstrapping, and predictive mean matching. Five imputations were created per bootstrap of the 100 bootstraps implemented in this study. The bootstraps created by *aregImpute* to impute missing data were in addition to those described earlier. Thus, bootstraps were created from original study data sets and the entire predictive modeling pipeline was iterated including multiple imputation for each bootstrap sample. The performance and predictor importance results were subsequently pooled across imputations for each study bootstrap and then pooled again across the 100 study bootstraps. This imputation package also imputes nominal variables without complication as it relies on regression and predictive mean matching to predict missing values. The package converts categorical variables into integers as a transformation to accomplish the regression itself, which is a common step in the calculation of regression models incorporating nominal or categorical variables. The imputed results, however, reflect the original categories in the final data set.

Temporal variance of prediction windows. To better understand temporal patterns of suicide attempt risk over time, an identical modeling setup as described earlier was applied to data censored at multiple time points of interest from either the most recent suicide attempt (for cases) or the last recorded clinical encounter (for controls). Data were censored at the following: 7 days, 14 days, 30 days, 60 days, 90 days, 180 days, 365 days, and 720 days from the most recent suicide attempt or last recorded encounter (for nonattempters). Each of these prediction windows led to a unique set of algorithms, which were assessed for optimism-corrected discrimination and calibration. All modeling efforts at each prediction window were processed using all available data after time censoring and through 100 bootstrap samples.

Results

A total of 5,543 patients with ICD codes for suicide and self-inflicted injury (i.e., E950–E959) formed the training data for this work. After expert chart review, only 3,250 (58.63%) had expert-confirmed histories of nonfatal suicide attempts. A total of 1,917 patients were judged not to have any definitive evidence of a suicide attempt in their EHRs; these patients formed a stringent control group (35.58% of the E95x candidate set). This finding suggests that data analytic strategies assuming all E95x codes are indicative of true suicide attempts may contain a large proportion of nonattempts. An additional 376 patients were confirmed by expert review to have died by suicide; these were excluded from analyses because, based on several conceptual (see Franklin et al., 2017) and empirical differences (e.g., methods used, peak age, gender differences, frequency; see CDC, 2016), nonfatal and fatal attempts are considered to be qualitatively distinct phenomena. Of note, 1,367 of the 3,250 patients who engaged in a nonfatal suicide attempt (42.06%) had records indicating prior attempts, as determined by expert review. These latter cases composed the repeat attempter subgroup (see the discussion later); cases with no evidence of a prior attempt composed the single/first attempter subgroup.

Table 1. Baseline Patient Characteristics

Characteristic	Cases (proportion), n = 3,250	Controls (proportion), n = 1,917	p value
Gender			
Male	1,696 (0.52)	1,010 (0.53)	.36
Female	1,534 (0.47)	898 (0.47)	.59
Unknown	20 (0.006)	9 (0.005)	.75
Race			
White	2,706 (0.832)	1,499 (0.78)	.051
Black	375 (0.115)	340 (0.18)	<.001
Asian	18 (0.006)	15 (0.008)	.16
Alaskan/Native American	6 (0.002)	18 (0.009)	<.001
Pacific Islander	2 (0.0006)	0 (0)	.86
Declined to respond	5 (0.002)	1 (0.005)	.85
Unknown/not recorded	38 (0.01)	19 (0.01)	.72
Age			
Mdn (*SD*)	37.1 (13.0)	39.1 (14.5)	

Note: p values were generated via differences in pooled sample proportions.

Baseline patient characteristics in cases and controls are shown in Table 1. As this is an observational cohort study and demographics such as age and gender are known risk factors for suicidal behaviors, statistically significant differences between these groups were expected.

Model performance

General model performance. Discrimination performance was good across all models and improved as suicide attempts became more imminent (Table 2). Recall (i.e., sensitivity) was high at the outcome prevalence threshold throughout; this result indicates true positives are likely to be captured with this approach. Optimism-corrected AUC values with confidence intervals, precision, and recall are shown in Table 2 and Figure 2, respectively. A contingency table at each time point is shown (Table 3) to provide raw information about classification accuracy. Assessed graphically and with the Brier score, calibration was consistent across time periods with scores in the range of 0.14 to 0.16. There is no strict cutoff for acceptable Brier scores, and 0 is the score for a "perfect" model.

Repeat and single/first attempter subanalyses. Restricting the data to those without prior suicide attempts, that is, those with only a single suicide attempt, resulted in similar performance to that in the group of repeat and first attempters combined. For first attempters, optimism-adjusted AUC values ranged from 0.82 (95% CI [0.81, 0.83]) at 7 days prior to suicide attempts to 0.75 (95% CI [0.74, 0.76]) at 720 days prior to suicide attempts. For those with prior attempts or repeat attempters, AUC values ranged from 0.85 (95% CI [0.84, 0.86]) at 7 days prior to suicide attempts to 0.76 (95% CI [0.76, 0.78]) at 720 days prior to suicide attempts.

Subanalyses with a random sample of hospital patients as controls. As expected, model performance with this less stringent control group was improved. Optimism-adjusted AUC values ranged from 0.92 (95% CI

Table 2. Discriminative and Calibration Performance of Models by Time Period Before Suicide Attempts

Prediction window	AUC [95% CI]	Precision[a]	Recall[b]	Brier score[c]
7 days	0.84 [0.83, 0.85]	0.79	0.95	0.14
14 days	0.83 [0.82, 0.84]	0.79	0.95	0.15
30 days	0.82 [0.82, 0.83]	0.78	0.95	0.15
60 days	0.82 [0.81, 0.82]	0.77	0.95	0.15
90 days	0.81 [0.81, 0.82]	0.77	0.95	0.15
180 days	0.81 [0.80, 0.82]	0.76	0.94	0.16
365 days	0.83 [0.82, 0.84]	0.75	0.96	0.15
720 days	0.80 [0.80, 0.81]	0.74	0.95	0.16

Note: AUC = area under the receiver operating curve; CI = confidence interval. Cases were 3,250 patients with an expert-determined nonfatal suicide attempt; controls were 1,917 patients with a self-injury ICD code who could not be confirmed as having made a nonfatal suicide attempt.

[a]Precision ~ positive predictive value = the ratio of true positives divided by the sum of true positives and false positives. [b]Recall ~ sensitivity = the number of true positives divided by the sum of true positives and false negatives. [c]Brier score indexes the discrepancy between the predicted probability of a nonfatal suicide attempt and the actual outcome of a nonfatal suicide attempt for each individual. The metric ranges between 0 and 1, with scores closer to 0 indicating less discrepancy between predicted probability and actual outcome.

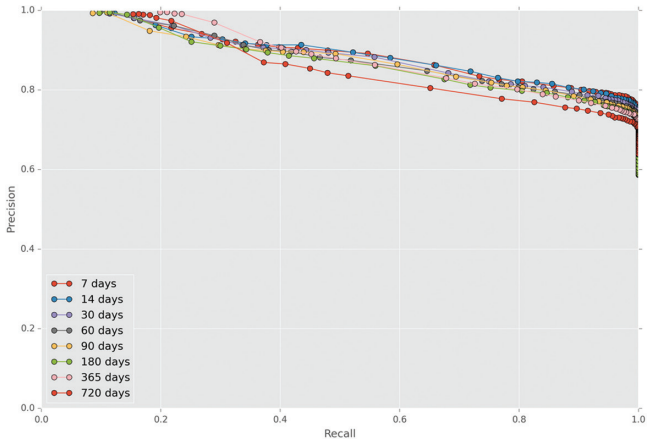

Fig. 1. Precision recall curves of predictive models of risk of suicide attempts.

[0.91, 0.92]) at 7 days prior to suicide attempts to 0.86 (95% CI [0.85, 0.86]) at 720 days prior to suicide attempts.

Comparison to analysis with a traditional method. Multiple logistic regression performed worse than random forests, with AUC values ranging from 0.66 (95% CI [0.58, 0.75]) at 7 days prior to suicide attempts to 0.68 (95% CI [0.66, 0.71]) at 720 days prior to suicide attempts.

Predictor importance

Predictor importance also varied across time points. Figure 1 summarizes importance of the top 50 predictors for each model. The size of the points in the plot indicates relative weight—larger points indicate greater importance to the relevant random forest. The full set of importance values over time is included in Table S1 (in the Supplemental Material available online). Further out from the nonfatal suicide attempt, hospital utilization history and visit tallies are the most important predictors. Demographics such as age, gender, and race are consistently informative.

Multiple CMS-HCC diagnoses (with names matching the 2014 release in publicly available data sets; Pope et al., 2006) were important in predicting nonfatal suicide attempts. Recurrent depression with psychosis, schizophrenia, and schizoaffective disorder were consistently ranked highly in importance. Age and diagnoses of dependence on opioids, sedative-hypnotics, and cannabis increased in relative importance as prediction windows shortened. Some codes that likely indicate prior suicide attempts were also consistently predictive: poisoning, the most common mechanism of prior nonfatal suicide attempts in these data; injuries by firearms; and injuries "NEC" or not elsewhere classifiable. A preponderance of NOS (not otherwise specified) and NEC codes are noted throughout; clinical claims are well known for provider reliance on these types of codes. The increase in granularity of ICD-10 is in part an attempt to improve diagnostic specificity in claims coding.

Medication classes such as selective serotonin reuptake inhibitors (SSRIs), benzodiazepines, anilides (such as acetaminophen), and propionic acid derivatives (such as ibuprofen) appear stronger within longer prediction windows. Melatonin receptor agonists such as melatonin supplements gain relative importance closer to the suicide attempt (i.e., shorter prediction windows).

Discussion

Accurate and scalable methods of suicide attempt risk detection are an important part of efforts to reduce these behaviors on a large scale. In an effort to contribute to the development of one such method, we applied ML to EHR data. Our major findings included the following: (a) This method produced more accurate prediction of suicide attempts than traditional methods (e.g., ML produced AUCs in the 0.80s, traditional regression in the 0.50s and 0.60s, which also demonstrated wider confidence intervals/greater variance than the ML approach), with notable lead time (up to 2 years) prior to attempts; (b) model performance steadily improved as the suicide attempt become more imminent; (c) model performance was similar for single and repeat attempters; and (d) predictor importance within algorithms shifted over time. Here, we discuss each of these findings in more detail.

ML models performed with acceptable accuracy using structured EHR data mapped to known clinical terminologies like CMS-HCC and ATC, Level 5. Recent meta-analyses indicate that traditional suicide risk detection approaches produce near-chance accuracy (Franklin et al., 2017), and a traditional method—multiple logistic regression—produced similarly poor accuracy in the present study. The lone longitudinal study that applied

Table 3. Classification Table Over Time in Days Before Suicide Attempts

Days	7	14	30	60	90	180	365	720
True positives	3,066	3,094	3,094	3,115	3,116	3,093	3,194	3,188
False positives	550	585	639	671	699	725	802	891
False negatives	184	156	156	135	134	157	56	62
True negatives	1,367	1,332	1,278	1,246	1,218	1,192	1,115	1,026

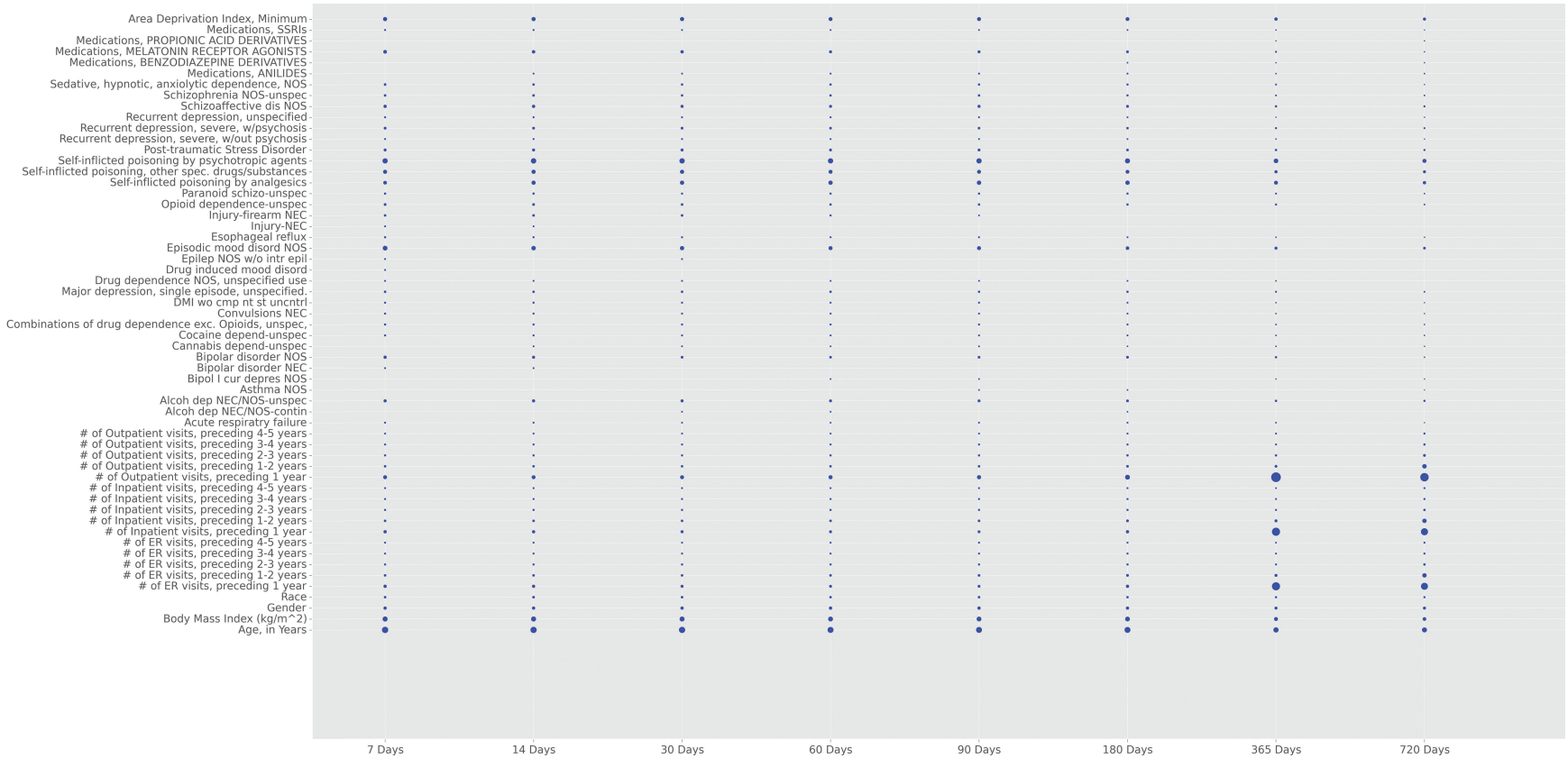

Fig. 2. Relative importance of predictors over time. Size of filled circle corresponds to relative predictor importance, with larger circles representing predictors with greater importance. NEC = not elsewhere classified; NOS = not otherwise specified; SSRIs = selective serotonin reuptake inhibitors. Number of days on the *x*-axis refers to the number of days preceding a suicide attempt.

ML to predict suicide attempts obtained greater discriminative accuracy than typically obtained with traditional approaches like logistic regression (i.e., AUC = 0.76; Kessler, Stein, et al., 2016). The present study extends this pioneering work with its use of a larger comparison group of self-injurers without suicidal intent, ability to display a temporally variant risk profile over time, scalability of this approach to any EHR data adhering to accepted clinical data standards, and performance in terms of discriminative accuracy (AUC = 0.84, 95% CI [0.83, 0.85]), precision recall, and calibration (see Table 1). This approach can be readily applied within large medical databases to provide constantly updating risk assessments for millions of patients based on an outcome derived from expert review.

Although short-term risk and shifts in risk over time are often noted in clinical lore, risk guidelines, and suicide theories (e.g., O'Connor, 2011; Rudd et al., 2006; Wenzel & Beck, 2008), few studies have directly investigated these issues. The present study examined risk at several intervals from 720 to 7 days and found that model performance improved as suicide attempts became more imminent. This finding was consistent with hypotheses; however, two aspects of the present study should be considered when interpreting this finding. First, this pattern was confounded by the fact that more data were available naturally over time; predictive modeling efforts at point of care should take advantage of this fact to improve model performance as additional data are collected. Second, due to the limitations of EHR data, we were unable to directly integrate information about potential precipitating events (e.g., job loss) or data not recorded in routine clinical care into the present models. Such information may have further improved short-term prediction of suicide attempts. Future studies should build on the present findings to further elucidate how risk changes as suicide attempts become more imminent.

Results were similar for single and repeat attempters. A prior suicide attempt is one of the strongest predictors of a future suicide attempt (Franklin et al., 2017; Ribeiro et al., 2016a), raising the possibility that algorithms for nonrepeat attempts would be far less accurate. Our findings did not support this possibility, suggesting that the present approach is similarly useful for both types of attempts. This finding is consistent with evidence that although prior attempt is a strong predictor *relative to* other predictors, it is a weak predictor in an absolute sense (Ribeiro et al., 2016a). Our results indicate that, for repeat attempters in particular, short-term predictions were comparable to the combined data set of repeat and first attempters, but distal prediction 2 years prior to suicide attempts may be more difficult to predict in repeat attempters specifically. Accordingly, knowledge about a prior suicide attempt may provide only a very small improvement in accuracy within the context of a large ML algorithm. We note, however, that this does not obviate the value of understanding more about the nature of initial versus repeat attempts (or single vs. multiple attempters). Future ML studies aimed at addressing the many questions surrounding potential differences between these phenomena are an important future direction for this work.

We also explored predictor importance within ML algorithms. Some predictors were consistently important (e.g., psychotic disorders, recurrent depression, poisoning), others were important only several months or years before the suicide attempt (e.g., prescriptions for SSRIs, benzodiazepines, acetaminophen; recent inpatient, outpatient, and emergency department visits), and others still were important only in the days or months directly preceding the suicide attempt (e.g., age, certain substance use diagnoses, prescription for melatonin receptor agonists). These findings are generally consistent with the widely held belief that risk factor importance shifts over time, but we caution that these findings should be regarded as tentative, associative, and exploratory. For example, medications can be indicative of diagnoses or latent variables of care (patients may accumulate lists of "as needed" medications that they may not take).

In addition to general difficulties with interpreting predictor importance within ML algorithms, the present findings are subject to the biases inherent in clinical EHR data. For example, the relative importance of health care utilization in longer prediction windows may reflect a relative paucity of other types of clinical data (e.g., psychiatric diagnoses) in those time periods. These limitations notwithstanding, the present findings suggest that predictor importance may change over time and lay a foundation for future studies aimed at more directly investigating such patterns.

The present findings should be considered in light of several general limitations. First, the present sample originated from a single medical center and both left- and right-censoring occur naturally as patients enter and leave health systems. Data mining and predictive modeling efforts based on EHR data are inherently limited by these potential biases. Although these limitations may hamper modeling efforts, they also confer ecological validity as real-world applications of the present approach would likely encounter similar biases. Second, one algorithm was implemented in this study, and although the random forest was chosen for its appealing properties for this use case, there may be other algorithms that perform as well or better, or perhaps ensemble methods that combine multiple algorithms that would perform better still. Future work would benefit from investigating this possibility. Third, we used claims information (i.e., ICD codes) to identify the initial cohort of patients with claims of

suicidal ideation and self-injury. Claims codes are inherently biased but remain a mainstay of predictive modeling work given their ubiquity, their collection as structured data, and the wealth of prior work in clinical practice and in the literature to make them more informative, such as CMS-HCC as was used here. Future studies may benefit from applying ML methods in conjunction with different recruitment and diagnostic strategies. Fourth, at first blush, given the strong performance metrics, it may be compelling to advocate that these algorithms be used as standalone determinants of imminent risk. The algorithms developed in this study can fairly accurately address the question of *who* will attempt by suicide, but not *when* someone will die. Although accurate knowledge of who is at risk of *eventual* suicide attempt is still critically important to inform clinical decisions about risk, it is not sufficient to determine *imminent risk*. The inability to speak to imminent risk prediction is in part a reflection of the limitations of the types of data available in EHRs. Applying ML methods to data derived from studies designed to predict short-term or imminent suicidal behavior would be valuable. Similarly, it is critical to consider the clinical usefulness of applying a model in a given use case; this requires an understanding of underlying clinical utilities and a decision analytic approach to support implementing predictive models into clinical practice where there can be both positive and negative consequences. The clinical informatics methodology required to determine the appropriate point at which to insert new predictions into the clinical workflow is another critical aspect of this type of work that must be considered before models are implemented in practice. Future work is planned to address these issues.

Fifth and finally, it is tempting to interpret individual predictor importance in any ML study; the risk of conflating correlation and causation is extremely high, however. The importance metric chosen here—mean decrease in impurity—can discard potentially informative predictors that may provide similar information but were not selected algorithmically (Schwarz, König, & Ziegler, 2010). For example, the presence of melatonin receptor agonists may be replaced by a well-modeled predictor of "sleep disturbance" in another study, or it may capture a unique aspect of clinical workflow whose latent variables are not well modeled in extant data. We also emphasize that each "important predictor" identified in the present study (see Fig. 1) should be considered in the context of the algorithm as a whole because this was the context in which importance was determined. The present findings may permit conclusions about broad patterns of importance (e.g., general hospitalization and prescription variables are primarily important several months or years before a suicide attempt), but we would caution against more specific conclusions (e.g., acetaminophen prescription status is a particularly important predictor several months or years before an attempt).

The present study represents an important step toward the development of an accurate and scalable suicide attempt risk detection. The present findings are promising, but further studies would be helpful for investigating the contribution of other predictors (e.g., life events), validating these algorithms on external data, and testing how this type of risk identification approach affects intervention usage and efficacy. For example, a clinical decision support trial implementing a temporal risk profile of suicidal behaviors at clinical encounters might enable providers to target interventions not just to the *right* individuals but also at the *right time*.

Author Contributions

All authors developed the study concept and contributed to the study design. Testing and data collection were performed by C. G. Walsh; expert validation was performed by J. D. Ribeiro and J. C. Franklin. C. G. Walsh performed the data analysis and machine learning. All authors contributed to interpretation of data analyses. C. G. Walsh drafted the manuscript, and J. D. Ribeiro and J. C. Franklin contributed to manuscript prose and to critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

The authors wish to thank Xieyining Huang and Katherine M. Musacchio-Schafer for their assistance on ensuring coding quality.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Note

1. The study data source, the BioVU Synthetic Derivative, does not categorically include health plan or insurance payer data as it is deidentified; thus, these data were not included.

Supplemental Material

Additional supporting information may be found at online.

References

Amalakuhan, B., Kiljanek, L., Parvathaneni, A., Hester, M., Cheriyath, P., & Fischman, D. (2012). A prediction model for COPD readmissions: Catching up, catching our breath, and improving a national problem. *Journal of Community Hospital Internal Medicine Perspectives*, *2*, 1–7.

Austin, P. C., Lee, D. S., Steyerberg, E. W., & Tu, J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? *Biometrical Journal*, *54*, 657–673. http://doi.org/10.1002/bimj.201100251

Bentley, K. H., Franklin, J. C., Ribeiro, J. D., Kleiman, E. M., Fox, K. R., & Nock, M. K. (2016). Anxiety and its disorders as risk factors for suicidal thoughts and behaviors: A meta-analytic review. *Clinical Psychology Review*, *43*, 30–46. http://doi.org/10.1016/j.cpr.2015.11.008

Centers for Disease Control and Prevention. (2016). *Injury prevention & control: Data & statistics (WISQARS)*. Retrieved from https://www.cdc.gov/injury/wisqars/

Cha, C. B., Najmi, S., Park, J. M., Finn, C. T., & Nock, M. K. (2010). Attentional bias toward suicide-related stimuli predicts suicidal behavior. *Journal of Abnormal Psychology*, *119*, 616–622. http://doi.org/10.1037/a0019710

Chang, B. P., Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., & Nock, M. K. (2016). Biological risk factors for suicidal behaviors: A meta-analysis. *Translational Psychiatry*, *6*, 887.

Deeks, J. J. (2011). Analysing data and undertaking meta-analyses. In *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0, updated March 2011). Retrieved from http://handbook.cochrane.org/chapter_9/9_analysing_data_and_undertaking_meta_analyses.htm

Delgado-Gomez, D., Blasco-Fontecilla, H., Sukno, F., Socorro Ramos-Plasencia, M., & Baca-Garcia, E. (2012). Suicide attempters classification: Toward predictive models of suicidal behavior. *Neurocomputing*, *92*, 3–8. http://doi.org/10.1016/j.neucom.2011.08.033

Fox, K. R., Franklin, J. C., Ribeiro, J. D., Kleiman, E. M., Bentley, K. H., & Nock, M. K. (2015). Meta-analysis of risk factors for nonsuicidal self-injury. *Clinical Psychology Review*, *42*, 156–167. http://doi.org/10.1016/j.cpr.2015.09.002

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., . . . Nock, M. K. (2017). *Psychological Bulletin, 143*, 187–232.

Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, *56*, 229–238. http://doi.org/10.1016/j.jbi.2015.05.016

Harrell, F. E., Jr. (2006). *Regression modeling strategies*. New York, NY: Springer.

Harrell, F. E., Jr., & Slaughter, J. C. (2001). *Introduction to biostatistics for biomedical research*. Retrieved from data.vanderbilt.edu/biosproj/CI2/handouts.pdf

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Kessler, R. C., Stein, M. B., Petukhova, M. V., Bliese, P., Bossarte, R. M., Bromet, E. J., . . . Keilp, J. (2016). Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Molecular Psychiatry*. Advance online publication. http://doi.org/10.1038/mp.2016.110

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., . . . Zaslavsky, A. M. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry*. Advance online publication. http://doi.org/10.1038/mp.2015.198

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249–268.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., & Ziegler, A. (2011). Probability machines. *Methods of Information in Medicine*, *51*, 74–81. http://doi.org/10.3414/ME00-01-0052

Mann, J., Ellis, S., Waternaux, C., & Liu, X. (2008). Classification trees distinguish suicide attempters in major psychiatric disorders: A model of clinical decision making. *Journal of Clinical Psychiatry*, *69*, 23–31.

Miao, Y., Francisco, S., Boscardin, W. J., Francisco, S., & Francisco, S. (2013). *SAS Global Forum 2013: Statistics and data analysis. Estimating Harrell's optimism on predictive indices using bootstrap samples*. Retrieved from http://support.sas.com/resources/papers/proceedings13/504-2013.pdf.

Neeleman, J. (2001). A continuum of premature death. Meta-analysis of competing mortality in the psychosocially vulnerable. *International Journal of Epidemiology*, *30*, 154–162. http://doi.org/10.1093/ije/30.1.154

O'Connor, R. C. (2011). Towards an integrated motivational–volitional model of suicidal behaviour. In R. C. O'Connor, S. Platt, & J. Gordon (Eds.), *International handbook of suicide prevention: Research, policy and practice* (pp. 181–198). New York, NY: Wiley.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning and Research*, *12*, 2825–2830.

Perez, F., & Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science and Engineering*, *9*(3), 21–29. http://doi.org/10.1109/MCSE.2007.53

Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Lezzoni, L. I., . . . Robst, L. (2006). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Finance Review*, *25*, 119–141.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

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

Ribeiro, J. D., Pease, J. L., Gutierrez, P. M., Silva, C., Bernert, R. A., Rudd, M. D., & Joiner, T. E. (2012). Sleep problems outperform depression and hopelessness as cross-sectional and longitudinal predictors of suicidal ideation and behavior in young adults in the military. *Journal of Affective Disorders*, *136*, 743–750. http://doi.org/10.1016/j.jad.2011.09.049

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*, 503–520. http://doi.org/10.1108/00220410410560582

Roden, D., Pulley, J., Basford, M., Bernard, G., Clayton, E., Balser, J., & Masys, D. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics*, *84*, 362–369. http://doi.org/10.1038/clpt.2008.89

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. *81*). New York, NY: John Wiley.

Rudd, M. D., Berman, A. L., Joiner, T. E., Nock, M. K., Silverman, M. M., Mandrusiak, M., . . . Witte, T. (2006). Warning signs for suicide: Theory, research, and clinical applications. *Suicide and Life-Threatening Behavior*, *36*, 255–262. http://doi.org/10.1521/suli.2006.36.3.255

Schwarz, D. F., König, I. R., & Ziegler, A. (2010). On safari to random Jungle: A fast implementation of random forests for high-dimensional data. *Bioinformatics*, *26*, 1752–1758. http://doi.org/10.1093/bioinformatics/btq257

Shepard, D. S., Gurewich, D., Lwin, A. K., Reed, G. A., & Silverman, M. M. (2016). Suicide and suicidal attempts in the United States: Costs and policy implications. *Suicide and Life-Threatening Behavior*, *46*, 352–362. http://doi.org/10.1111/sltb.12225

Singh, G. K. (2003). Area deprivation and widening inequalities in US mortality, 1969–1998. *American Journal of Public Health*, *93*, 1137–1143.

Smith, G. C. S., Seaman, S. R., Wood, A. M., Royston, P. and White, I. R. (2014). Correcting for optimistic prediction in small data sets. *American Journal of Epidemiology, 180,* 318–324. doi:10.1093/aje/kwu140.

Van Calster, B., Vickers, A., Pencina, M., Baker, S. G., Timmerman, D., & Steyerberg, E. (2013). Evaluation of markers and risk prediction models: Overview of relationships between NRI and decision-analytic measures. *Medical Decision Making*, *33*, 490–501. http://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*(2), 22–30. http://doi.org/10.1109/MCSE.2011.37

Wenzel, A., & Beck, A. T. (2008). A cognitive model of suicidal behavior: Theory and treatment. *Applied and Preventive Psychology*, *12*, 189–201. http://doi.org/10.1016/j.appsy.2008.05.001

World Health Organization. (2016). *Suicide data*. Geneva, Switzerland: Author.

Wright, M. N., & Ziegler, A. (2015). *Ranger: A fast implementation of random forests for high dimensional data in C++ and R*. Retrieved from https://arxiv.org/pdf/1508.04409.pdf

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, *17*, 19–24. http://doi.org/10.1197/jamia.M3378