# Journal of Consulting and Clinical Psychology

## Model Complexity Improves the Prediction of Nonsuicidal Self-Injury

Kathryn R. Fox, Xieyining Huang, Kathryn P. Linthicum, Shirley B. Wang, Joseph C. Franklin, and Jessica D. Ribeiro

Online First Publication, June 20, 2019. http://dx.doi.org/10.1037/ccp0000421

American Psychological Association

Journal of Consulting and Clinical Psychology


0022-006X/19/$12.00

http://dx.doi.org/10.1037/ccp0000421

# Model Complexity Improves the Prediction of Nonsuicidal Self-Injury

Kathryn R. Fox
Harvard University

Xieyining Huang and Kathryn P. Linthicum
Florida State University

Shirley B. Wang
Harvard University

Joseph C. Franklin and Jessica D. Ribeiro
Florida State University

***Objective:*** Efforts to predict nonsuicidal self-injury (NSSI; intentional self-injury enacted without suicidal intent) to date have resulted in near-chance accuracy. Incongruence between theoretical understanding of NSSI and the traditional statistical methods to predict these behaviors may explain this poor prediction. Whereas theoretical models of NSSI assume that the decision to engage in NSSI is relatively complex, statistical models used in NSSI prediction tend to involve simple models with only a few theoretically informed variables. The present study tested whether more complex statistical models would improve NSSI prediction. ***Method:*** Within a sample of 1,021 high-risk self-injurious and/or suicidal individuals, we examined the accuracy of three different model types, of increasing complexity, in predicting NSSI across 3, 14, and 28 days. Univariate logistic regressions of each predictor and multiple logistic regression with all predictors were conducted for each timepoint and compared with machine learning algorithms derived from all predictors. ***Results:*** Results demonstrated that model complexity was associated with predictive accuracy. Multiple logistic regression models (AUCs 0.70–0.72) outperformed univariate logistic models (average AUCs 0.56). Machine learning models that produced algorithms modeling complex associations across variables produced the strongest NSSI prediction across all time points (AUCs 0.87–0.90). These models outperformed all multiple logistic regression models, including those involving identical study variables. Machine learning algorithm performance remained strong even after the most important factor across algorithms was removed. ***Conclusions:*** Results parallel recent findings in suicide research and highlight the complexity that underlies NSSI.

***What is the public health significance of this article?***
When predicting individuals' short-term risk for engaging in nonsuicidal self-injury, this study demonstrated the importance of considering the complex relationships among a large number of predictors. The study strongly suggests that algorithms better equipped at modeling complexity predict nonsuicidal self-injury more accurately.

*Keywords:* nonsuicidal self-injury, prediction

*Supplemental materials:* http://dx.doi.org/10.1037/ccp0000421.supp



Nonsuicidal self-injury (NSSI) includes intentional self-harming behaviors enacted without suicidal intent (Nock, 2010). These behaviors are common, affecting upward of 5.5% of adults within their lifetimes (Swannell, Martin, Page, Hasking, & St John, 2014). The high prevalence of NSSI is concerning for several reasons, including the physical harm it causes (e.g., Cloutier, Martin, Kennedy, Nixon, & Muehlenkamp, 2010), its association with various forms of psychopathology (e.g., Gollust, Eisenberg, & Golberstein, 2008; Nock, Joiner, Gordon, Lloyd-Richardson, & Prinstein, 2006; Selby, Bender, Gordon, Nock, & Joiner, 2012), and its concurrent and prospective association with suicidal behaviors (for reviews, see Andover, Morris, Wren, & Bruzzese,

Kathryn R. Fox, Department of Psychology, Harvard University; Xieyining Huang and Kathryn P. Linthicum, Department of Psychology, Florida State University; Shirley B. Wang, Department of Psychology, Harvard University; Joseph C. Franklin and Jessica D. Ribeiro, Department of Psychology, Florida State University.

This work was supported in part by funding from the Military Suicide Research Consortium (MSRC), an effort supported by the Office of the Assistant Secretary of Defense for Health Affairs (Award W81XWH-10–2-0181). Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the MSRC or the Department of Defense. Kathryn R. Fox and Xieyining Huang contributed equally to this article.

Correspondence concerning this article should be addressed to Jessica D. Ribeiro, Department of Psychology, Florida State University, 1107 West Call Street, Tallahassee, FL 32306. E-mail: ribeiro@psy.fsu.edu

2012; Hamza, Stewart, & Willoughby, 2012; Ribeiro et al., 2016). Given its prevalence and dangerousness, researchers and clinicians are eager to identify those at risk for future NSSI to help direct treatment and prevention resources to those individuals most in need.

To facilitate NSSI risk detection, many studies have sought to longitudinally predict NSSI over the past two decades. These studies primarily used computational methods requiring researchers to identify factors they perceive as the most theoretically important, and then to use those small sets of factors to predict future NSSI (e.g., Fox, O'Sullivan, Wang, & Hooley, 2019; Franklin, Puzia, Lee, & Prinstein, 2014). Although this intuitive approach produced many statistically significant predictors of NSSI, these models generally failed to provide accurate or clinically significant NSSI prediction (Fox et al., 2015; see Yarkoni & Westfall, 2017 for a discussion of explanation vs. prediction models).

In contrast to these relatively simple statistical models of NSSI prediction, theoretical models of NSSI typically assume that the decision to engage in NSSI is complex (Muehlenkamp, Kerr, Bradley, & Adams Larsen, 2010; Nock, 2010; Selby, Franklin, Carson-Wong, & Rizvi, 2013; Taylor et al., 2018). This theoretical assumption is based on evidence that several states (e.g., suicidal, upset, self-critical) may lead to NSSI episodes when specific conditions are met. The mismatch between traditional statistical models and theoretical models may help explain observed inaccurate NSSI prediction. If NSSI is complex and multidetermined, statistical approaches including only a small number of variables are likely to result in weak and inaccurate NSSI prediction.

Building on theoretical assumptions, NSSI prediction may benefit from machine learning methods that can better account for the complexity of these behaviors. Unlike traditional statistical approaches, machine learning methods minimize a priori assumptions regarding relationships among variables (e.g., Yarkoni & Westfall, 2017). Instead of relying on researchers to select variables of theoretical interest and determine their relationships (e.g., additive, polynomial), machine learning algorithms optimally identify variables and the relationships among those variables to maximize prediction accuracy. Emphasis on statistical accuracy, alongside evidence that these approaches accurately predict related outcomes including suicide death (Kessler et al., 2016) and suicide attempts (Walsh, Ribeiro, & Franklin, 2017), suggest that machine learning approaches may improve NSSI prediction.

The present study tested whether model complexity was associated with improved NSSI prediction. Within a sample of 1,021 high-risk self-injurious and/or suicidal individuals who responded to questions assessing a wide-range of variables related to NSSI and self-injurious thoughts and behaviors (SITBs) more broadly, we examined three different model types across three short-term prediction windows (3, 14, and 28 days). Each successive model type was intended to better approximate the complexity that may underlie NSSI prediction. First, to test simple models, we calculated the univariate predictive accuracy of a wide range of traditional risk factors for NSSI (e.g., prior NSSI, psychopathology). Based on recent meta-analytic evidence (Fox et al., 2015), we hypothesized that this approach would produce predictive accuracy only marginally better than random guessing. Second, to test whether additive combination among these univariate factors would improve prediction, we tested a multiple logistic regression analysis including all possible study variables. We hypothesized that this slightly more complex approach would improve prediction but would still produce poor-to-fair prediction.

Third, we applied machine learning models to examine whether greater model complexity would outperform the aforementioned traditional statistical approaches. Specifically, we employed random forests that considered all study variables (described in more detail below). Based on prior research and theory (see Ribeiro et al., 2016b), we hypothesized that random forest models, which can identify complex risk algorithms involving numerous variables when necessary, would produce greatly improved prediction. Finally, to further examine the potential complexity of NSSI prediction, we tested the effect of removing the most important predictor on machine learning performance. Consistent with prior work in suicide research (e.g., Walsh et al., 2017), we hypothesized that this removal would not affect predictive accuracy.

## Method and Procedures

All study procedures were approved by the Institutional Review Boards at Vanderbilt University and Florida State University. Participants were recruited from online forums related to suicide and psychopathology. Interested individuals were invited to complete a brief screening survey to determine eligibility. To increase participant anonymity, participants were contacted and paid via e-mail addresses that did not include identifiable information (e.g., surname, date of birth). Participant indicated consent by selecting a box affirming informed consent and providing e-mail aliases. Those who qualified and indicated consent were contacted via e-mail with a link to the baseline assessment. The baseline survey included approximately 50-min of computer-based tasks and questionnaires, described in more detail below. These questionnaires included a wide range of factors previously shown to be related, albeit weakly, to NSSI and SITBs more generally (Franklin et al., 2017; Fox, Millner, & Franklin, 2016). Participants were then contacted 3 days (T2), 14 days (T3), and 28 days (T4) later to complete the first, second, and third follow-up surveys. Participants were allowed 48-hr to complete each assessment. Within 24 hr of each survey completion, participants were provided compensation via electronic Amazon.com gift cards. The compensation schedule was specifically designed to increase retention rates (i.e., \$10 for each of the first three surveys, \$20 for the last assessment, and a \$20 bonus for completing all assessments).

Inclusion criteria included: (a) English fluency; (b) daily Internet access; (c) 18+ years of age; and (d) past year suicide attempt, at least two self-cutting episodes without wanting to die in the past 2 weeks, or frequent active suicide ideation within the last 2 weeks. We adopted multiple procedures to ensure higher data quality. First, to decrease the possibility of ineligible individuals participating in the study covertly, inclusion criteria were not posted in any study advertisements or provided to potential participants during the screening process. Second, filler questions irrelevant to the inclusion criteria and pairs of duplicate items, to ensure consistent responses across items, were included in the screening questionnaire. Third, the screening survey included brief, open-ended items (i.e., details of most severe suicide attempt). Before study entry, screening responses were read to ensure English fluency and guard against potential malingering. In cases where open-ended responses were unintelligible, extremely unlikely, or identical across screening surveys, individuals were excluded. Lastly, each IP address was only allowed to participate once; this step was employed to prevent the same individuals from completing the surveys more than once (see Appendix).

## Participants

A total of 1,021 participants completed the baseline assessment. Retention rates were high, with 974 (95.26%) participants completing T2, 950 (92.70%) completing T3, and 926 (90.19%) participants completing T4. These retention rates are between 5% and 35% higher than those observed in other longitudinal studies of NSSI engagement (e.g., Fox et al., 2015; Franklin et al., 2014; Garisch & Wilson, 2015; Glenn & Klonsky, 2011; Guan, Fox, & Prinstein, 2012; Marshall, Tilton-Weaver, & Stattin, 2013). The immediate compensation (i.e., within 24 hr) and the compensation schedule adopted by our study might have fostered the high retention rates.

Participants were primarily young adults (*M* age = 26.53, *SD* = 7.39) reporting female gender (66.19%). Participants were recruited internationally. The majority reported they were from the U.S. (65.89%), Canada (3.80%), Great Britain (3.70%), or some other country. Regarding race, most participants reported being White (79.09%), with remaining participants reporting Black or African American (3.86%), Asian (6.19%), or Other (10.86%) races. At baseline, 88.43% of participants reported a lifetime history of NSSI; 82.03% of all participants and 92.77% of those with NSSI histories specifically reported a history of self-cutting. On average, participants reported 3.61 (*SD* episodes = 8.45) NSSI episodes in the month prior to the baseline assessment and 1.71 (*SD* episodes = 3.97) self-cutting episodes in the month prior to the baseline assessment. NSSI engagement remained high at each follow-up time point, with 19.92%, 34.66%, and 41.25% endorsing NSSI 3, 14, and 28 days after baseline, respectively. Most participants reported a lifetime history of suicide plans (96.14%), and suicide attempts (61.73%).

## Measures

We sought to include a wide-range of factors previously shown to be associated with SITB risk (e.g., demographics, psychopathology, prior SITBs, cognition, affect, behavior; for reviews, see Fox et al., 2015; Franklin et al., 2017). We focused on a combination of more static, presumably distal predictors (e.g., history of prior suicidal behavior; attitudes toward suicide/self-injury stimuli) and more dynamic, presumably proximal predictors (e.g., changes in arousal and affect). Notably, there are fewer studies of NSSI prediction presently, particularly over the shorter-term (Fox et al., 2015), relative to studies predicting SITBs (Franklin et al., 2017). In light of the strong association between NSSI and suicidal thoughts and behaviors (Andover et al., 2012; Hamza et al., 2012; Ribeiro et al., 2016), we reasoned that inclusion of factors related to both NSSI and suicidal thoughts and behaviors could be useful for NSSI prediction.

**Demographic information.** Demographics were assessed at baseline using a self-report measure asking about participants' age, sex, race, sexual orientation, employment status, and education.

**Acquired Capability for Suicide Scale-Fearlessness About Death (ACSS-FAD; Ribeiro et al., 2014).** The ACSS-FAD is a seven-item self-report measure that assesses fearlessness about death. Each item is rated on a 4-point Likert scale with higher total scores indicating greater fearlessness about death. This scale has demonstrated good convergent and discriminant validity, as well as generalizability across gender (Ribeiro et al., 2014). The summed score of this scale was included a predictor. The ACSS-FAD was administered at all time points, with Cronbach's alpha ranging from .85–.89.

**Affective States Questionnaire (ASQ; Hendin, Al Jurdi, Houck, Hughes, & Turner, 2010; Hendin, Maltsberger, & Szanto, 2007).** The ASQ is a nine-item self-report measure that assesses intense negative affective states (e.g., "Have you been experiencing feelings of anxiety?"). Each item is rated either as "yes" or "no." This scale has demonstrated good validity and ability to predict suicidal behavior (Hendin et al., 2010). All nine items of the ASQ were included as predictors (i.e., abandonment, anxiety, desperation, guilt, hope, humiliation, loneliness, rage, and self-hatred). The ASQ was administered at all time points, with Cronbach's alpha ranging from .88–.93.

**Disgust With Life Scale (DWLS; Chu, Bodell, Ribeiro, & Joiner, 2015; Ribeiro, Bodell, & Joiner, 2012).** The DWLS is a 12-item self-report scale assessing disgust with the self, and disgust with others and the world. Items are rated on a Likert-type scale from 0 (*not at all true of me*) to 6 (*very much true of me*); higher scores indicate higher levels of disgust. The subscales of this measure have demonstrated strong convergent validity with related measures of disgust alongside excellent internal consistency (Ribeiro et al., 2012). Each subscale was adopted as a predictor of NSSI at follow-ups. The DWLS demonstrated strong internal consistency across timepoints, with Cronbach's alpha ranging from 0.91 to 0.92.

**Beck Scale for Suicide Ideation (BSS; Beck, Kovacs, & Weissman, 1979; Beck & Steer, 1991).** The BSS is a 21-item self-report measure that assesses suicidal ideation. Each item is rated on a 3-point Likert scale with higher scores indicating greater suicidal ideation. This scale has demonstrated strong reliability and validity (Beck, Steer, & Ranieri, 1988). The summed score of the BSS was included as a predictor. The BSS was administered at all time points, with Cronbach's alpha ranging from .87–.88.

**Brief Agitation Measure (BAM; Ribeiro, Bender, Selby, Hames, & Joiner, 2011).** The BAM is a three item self-report measure that assesses past week agitation. Each item is rated on a seven-point Likert scale. Higher scores indicate greater severity of agitation symptoms. This scale has demonstrated good validity and reliability (Ribeiro et al., 2011). The total score of the BAM was used as a predictor. The BAM was administered at all time points, with Cronbach's alpha ranging from .84–.89.

**Brief Symptoms Inventory-18 (BSI-18; Derogatis, 2001).** The BSI-18 is a brief version of the full 53-item BSI that assesses past week psychological symptoms. Each item is rated on a 5-point Likert scale, with higher scores indicating higher psychological distress. This scale has shown good reliability and validity as a measure of psychological distress (Derogatis, 2001). The total score of the BSI was adopted as a predictor. The BSI was administered at all time points, with Cronbach's alpha ranging from .92–.94.

**Modified Suicidal Thoughts and Behaviors Interview (SITBI; Nock, Holmberg, Photos, & Michel, 2007).** The SITBI assesses suicidal thoughts, plans, and attempts over the participant's lifetime, as well as the past year, past month, and past week. The modified SITBI is a self-report measure of the original interview. This self-report measure corresponds highly with the in-person interview, and produces similar estimates of engagement and frequency of NSSI (Franklin et al., 2014). In this study, we used only modules assessing NSSI, suicide plans, and suicidal behaviors (e.g., aborted suicide attempts, interrupted suicide attempts, suicide attempts). The modified SITBI was administered at all time points.

The following predictors were included in model analyses: lifetime suicide plan history, lifetime suicide preparation history, lifetime aborted attempt history, lifetime interrupted attempt history, lifetime suicide attempt history, multiple attempter status, past year suicide attempt history, physical injury of the most recent suicide attempt, lifetime NSSI history, past month self-cutting frequency, past month self-burning frequency, past month alternative NSSI (i.e., other than self-cutting and self-burning) frequency, desire to stop NSSI, and self-reported likelihood of future NSSI. We placed an emphasis on past month frequency of NSSI for three main reasons. First, past month frequency might be less subject to recall biases and underestimation than lifetime or past year frequency (Kruijshaar et al., 2005; Moffitt et al., 2010). Second, past month frequency of NSSI might be a more discriminative predictor, particularly considering that 88.43% of the sample reported a lifetime history of NSSI. Third, past month behaviors might be more indicative for acute (i.e., short-term) prediction of NSSI than lifetime or past year behaviors.

Additionally, we intentionally included frequencies of different forms of NSSI for two main reasons. First, self-cutting and self-burning are among the most severe forms of NSSI and signal higher clinical severity compared with other forms of NSSI (Andover, Pepper, Ryabchenko, Orrico, & Gibb, 2005; Stewart et al., 2017). Second, evidence suggests that self-cutting and self-burning are among the most commonly studied forms of NSSI (Klonsky & Muehlenkamp, 2007; Nixon, Cloutier, & Jansson, 2008; Swannell et al., 2014). To avoid overly burdening our participants, we prioritized the assessment of self-cutting and self-burning. However, we also included the assessment of NSSI as a broad category and other methods of NSSI as they might provide additional important information regarding participants' risk for future NSSI.

**Modified Affect Misattribution Procedure (AMP; Franklin et al., 2014; Payne, Cheng, Govorun, & Stewart, 2005).** The AMP is a brief computer-based task assessing implicit affect toward stimuli. On each trial, participants viewed an emotionally evocative image or word followed by an ambiguous Chinese symbol, and rated whether they judged the Chinese symbol to be more or less pleasant than the average symbol. Participants were instructed to ignore the emotionally evocative stimuli during these judgments; however, previous research has shown that these judgments are influenced by the nature of the stimuli. More pleasant stimuli generate more pleasant evaluations of the subsequent Chinese symbols, and vice versa (Payne et al., 2005). Therefore, judgment of the Chinese symbol represents an implicit affective reaction toward the emotional stimuli presented at the beginning of each trial.

In the present study, the AMP included both pleasant and death/suicide-related images, assessed across 48 trials (one for each picture). Stimuli included death/suicide related images ranging from low (e.g., body bag, noose, morgue) to extreme (e.g., corpse following fatal gunshot wound to the head) intensity. The AMP was administered at all time points, with Cronbach's alphas of 0.94 for the suicide/death images and 0.80 for the pleasant images. Participants' implicit affect toward suicide/death images and positive images were included as two predictors in the models.

**Explicit affective ratings (Franklin et al., 2014; Lang, Bradley, & Cuthbert, 2005).** We assessed explicit affect toward a range of suicide and pleasant stimuli using a 10-point Likert scale. Higher scores indicated more pleasant images. To assess explicit affect toward suicide, 10 images, drawn from the set used in the AMP task described above, were used. Images were of moderate intensity and depicted a range of suicide death methods (Cronbach's alpha = .90). Five images were used to assess affect toward positive stimuli (Cronbach's alpha = .79). Participants' explicit affect toward suicide stimuli and pleasant stimuli were employed as two predictors.

## Data Analyses

**Predictors.** A total of 39 predictors were included in the models. The predictors consisted of all measures described above, as well as specific demographic variables (e.g., race, gender, employment).

**Missing data.** Participants who did not complete the follow-up survey were excluded from analyses at that time point only (i.e., NSSI at follow-ups was not imputed). Missing data of predictor variables were extremely limited (0.04%), and they were addressed via single imputation of the mean.

**Modeling approach.** All statistical analyses were performed in R (Development Core Team, 2013) via *glm* in base R, and *randomForest* and *pROC* packages. To test simple models, we conducted a nonregularized univariate logistic regression for each factor in predicting NSSI. Next, we conducted a nonregularized multiple logistic regression including *all* study variables. Predictors were entered without interactions for each study time point. Consistent with prior research (e.g., Walsh et al., 2017; Walsh, Ribeiro, & Franklin, 2018), bootstrapping was used in addition to the procedures detailed above to assess and guard against overfitting. Even though the current sample size is considered large in suicide research, it is extremely small compared with risk prediction studies with machine learning approaches (cf., Kessler et al., 2015: $n$ = 40,820; Walsh et al., 2018: $n$ = 33,610). Unlike other methods (e.g., cross-validation) that can serve as alternatives to prevent capitalization of noise, bootstrapping allows the algorithms to be trained on the entirety of the data, a method more appropriate given the present sample size. This method first requires the training of the model on the complete available data. A set of bootstrap replicates are then created based on the original data. New models are trained on each bootstrap replicate, and applied to the original data, producing performance estimates called the "out of bag" performance. Averaging the difference between bootstrapped performance and "out of bag" performance yields the "optimism" (i.e., measure of overfitting) of the model. The corrected model performance is derived by subtracting the degree of optimism from the original model performance. In the present study, a total of 100 bootstrap replicates were created.

To compare performance of traditional logistic regression with machine learning, the random forest algorithm was chosen given its well-established accuracy and robustness (Amalakuhan et al., 2012; Austin, Lee, Steyerberg, & Tu, 2012; Futoma, Morris, &

Lucas, 2015). This algorithm is composed of an ensemble of decision trees with multiple procedures in place to avoid overfitting (i.e., spurious inflation of performance) and to increase generalizability. Unlike traditional decision trees, random forests force each "split" to consider only a subset of the predictors; this procedure decorrelates the trees and therefore makes the resulting algorithms less variable and more reliable. In this study, the number of predictors randomly considered at each split were set according to the common practice in the field: the square root of the total number of predictors (James, Witten, Hastie, & Tibshirani, 2013). The overall fitting process is repeated multiple times, producing a forest of decision trees. For this study, this process was repeated 500 times (Breiman, 2001, 2002). The final predictive outcome for each participant is determined via majority voting from the 500 trees; this procedure decreases the likelihood of spurious prediction and increases reliability.

The random forest algorithm also provides estimates of the importance of the predictors. The computation procedures include randomly permuting the values of each predictor, obtaining new classification accuracy based on the permutated values, subtracting the new accuracy from the original accuracy based on the unpermutated data, and averaging and standardizing such differences. Similar to how estimates of variable importance (e.g., standard coefficients) are dependent upon other variables and the relationships between the variable and outcome (e.g., binomial, polynomial, interaction terms) in traditional regression models, the estimates of predictor importance of random forest models should be interpreted within the context of the model. Although the univariate relationship between a predictor and the outcome might coincide with its importance as estimated by the random forest models, it is not a one-to-one relationship. Thus, predictor importance should not be considered as the strengths of their univariate associations with the outcome.

**Model fit indices.** We used Area Under the Receiving Operating Characteristic Curve (AUC) to assess the overall prediction accuracy. However, in samples where the occurrence of event is rare (e.g., suicide attempt, NSSI), AUC can be misleading. In such cases, a model can yield a high AUC when it predicts that no one attempts suicide or engages in NSSI, thus offering little clinical utility. It is important to also consider indices that reflect the rate of true positives. In the present study, precision (i.e., true positives divided by the sum of true positives and false positives), and recall (i.e., true positives divided by the number of true positives and false negatives) were adopted. Guidelines suggest that AUCs of 0.50 to 0.59 indicate extremely poor classification, 0.60 to 0.69 poor classification, 0.70 to 0.79 fair classification, 0.80 to 0.89 good classification, and above .90 excellent classification (Franklin et al., 2017; Simundic, 2008). The same guidelines were adopted for precision and recall.

In addition to the above discrimination indices, we also considered Brier scores as a calibration index. Traditionally in clinical psychology, discrimination has received more attention than calibration (Lindhiem, Petersen, Mentch, & Youngstrom, 2018). However, it is important to consider how the predicted probability of a model matches the actual probability of an event when evaluating model performance. Given that NSSI is a rare phenomenon, especially over a short follow-up length, it is important to assess whether the model underestimates or overestimates NSSI. A model that is not calibrated to the real-world probability of NSSI might offer little clinical utility. Computationally, Brier scores were calculated to reflect whether the algorithms were calibrated to the real-world probability of NSSI engagement. Similar to how variance of the sample is used in statistics to estimate the variance of the population, the probability of the outcome in the sample is used to estimate the probability of the outcome in the population (i.e., "real-world probability") because it is challenging if not impossible to sample the entire population. Brier scores were calculated with the following formula,

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2,$$

where $N$ is the sample size of classified individuals, $p_i$ is the projected outcome for individual i, and $o_i$ is the observed outcome (Brier, 1950). Brier scores range from zero to one, with zero as the best achievable score indicating complete match between predicted probability and the real-world probability.

# Results

## Model Performance

Nonregularized univariate logistic regression models on average produced near chance level prediction accuracy across follow-ups (see Table 1). All individual predictors yielded accuracy in the poor to fair range (see Table S1 in the online supplemental material for model fit indices for each predictor). Nonregularized multiple logistic regression models including all study variables yielded significantly better accuracy; however, overall accuracy remained in the poor to fair ranges (see Tables 1 and 2). Precision indices were consistently in the poor range across time points, suggesting that the traditional approach produced a large proportion of false positives. Recall indices fell in the fair range, indi-

Table 1
*Random Forest and Multiple Logistic Regression Model Performance*

| Model | AUC [95% CI] | Precision | Recall | Brier |
|---|---|---|---|---|
| Logistic Regression (mean of individual variables) | | | | |
| T2 | .56 [.52, .59] | .26 | .56 | .45 |
| T3 | .56 [.53, .59] | .42 | .61 | .46 |
| T4 | .56 [.53, .59] | .49 | .57 | .45 |
| Multiple Logistic Regression (all study variables) | | | | |
| T2 | .72 [.69, .76] | .43 | .71 | .25 |
| T3 | .70 [.67, .73] | .57 | .72 | .29 |
| T4 | .70 [.68, .73] | .63 | .71 | .29 |
| Random Forests | | | | |
| T2 | .87 [.84, .90] | .94 | .76 | .06 |
| T3 | .89 [.87, .92] | .91 | .83 | .09 |
| T4 | .90 [.88, .92] | .91 | .86 | .09 |

*Note.* T2 = 3 days after baseline; T3 = 14 days after baseline; T4 = 28 days after baseline; AUC = area under the curve; 95% CI = 95% confidence interval. Precision refers to the number of true positives divided by the sum of true and false positives; recall refers to the number of true positives divided by the number of true positives and false negatives; Brier scores range from 0 to 1, with o representing the best achievable score indicating complete match between predicted probability and the real-world probability.

Table 2
*Contingency Table*

| Model | True positives | False positives | False negatives | True negatives |
|---|---|---|---|---|
| Univariate Logistic Regression (average) | | | | |
| T2 | 109 | 310 | 85 | 470 |
| T3 | 201 | 278 | 128 | 343 |
| T4 | 218 | 227 | 164 | 317 |
| Multiple Logistic Regression (all study variables) | | | | |
| T2 | 138 | 186 | 56 | 594 |
| T3 | 237 | 181 | 92 | 440 |
| T4 | 273 | 159 | 109 | 385 |
| Random Forests | | | | |
| T2 | 148 | 10 | 46 | 770 |
| T3 | 274 | 29 | 55 | 592 |
| T4 | 329 | 35 | 53 | 509 |

*Note*. T2 = 3 days after baseline; T3 = 14 days after baseline; T4 = 28 days after baseline.

cating that the models produced many false negatives as well. The Brier scores showed a substantial mismatch between the predicted probability of self-injurers among the sample and the actual probability.

Random forests produced good to excellent overall predictive accuracy across time points (see Tables 1 and 2). Precision indices were in the excellent range, indicating that the algorithms were adept at identifying true positives while avoiding false positives. For T2, the random forest approach yielded a recall in the fair range, suggesting that most individuals who engaged in NSSI had been identified by the model. Of note, 24% of those engaging in NSSI at this time-point were missed by the model. The recall indices increased for T3 and T4, reaching the good range. Random forest models for all three time points produced Brier scores near zero, indicating that the probability of someone identified as engaging in NSSI was closely matched with the real-world probability. The Brier scores of the random forest models in the current study are comparable to those reported in studies using machine learning to predict suicide attempts (e.g., Walsh et al., 2017: Brier = 0.14), and health conditions in general (e.g., An et al., 2018: Brier scores = 0.10–0.11; Motwani et al., 2017: Brier score = 0.08).

## Predictor Importance

Predictor importance estimated by random forests showed that self-cutting episodes in the month prior to baseline was the most important predictor across time points (see Table S2 and Figure S1 in the online supplemental material). To test whether the prediction accuracy of the random forest models was largely driven by this predictor, we assessed the model performance of random forest algorithms after removing this predictor; results were consistent with the original models (T2: AUC = 0.85, 95% CI [0.82, 0.88]; T3: AUC = 0.88, 95% CI [0.86, 0.90]; T4: AUC = 0.88, 95% CI [0.86, 0.90]).

The other top 10 most important predictors within the risk algorithm include variables related to self-injury (e.g., self-rated likelihood of future NSSI), suicidal thoughts and behaviors, psychopathology, self-disgust, agitation, and other relevant clinical measures. However, predictor importance was estimated within the random forest models, which considered highly complex, interactive, and nonlinear relationships among the present variables. These estimates do not reflect the strength of these predictors outside of the context of the present models.

# Discussion

The present study tested whether more complex models would improve NSSI prediction. Findings supported this possibility. Univariate models produced the poorest prediction, with multiple logistic regression models generating improved but still poor prediction. Machine learning models produced much-improved prediction compared to both univariate and multiple logistic regression. Moreover, machine learning model performance did not suffer when the most important factor in the algorithm was removed, supporting NSSI as a complex rather than a factor-dependent phenomenon. Results parallel recent findings in suicide research (e.g., Franklin et al., 2017; Ribeiro et al., 2016) and indicate that a high degree of complexity underlies NSSI prediction. Below, we discuss each of these findings in more detail.

Simple models—that is, univariate and multiple logistic regression models—produced poor prediction (i.e., AUCs, precisions, and recalls near 0.50, high Brier scores). We caution that this does not mean that individual factors have no relevance to NSSI. Rather, findings that many individual factors were weakly and inconsistently associated with future NSSI indicate that many factors likely have small and highly variable individual relationships with future NSSI. In other words, no one factor or small set of factors seems to play a large role in future NSSI. A wide range of factors seem to play a small role in future NSSI. This implies that some combination of these factors may improve NSSI prediction. Indeed, combining all study variables within a multiple logistic regression did improve prediction, at least in a relative sense. Nevertheless, prediction with this simple combinatory approach still produced poor prediction in an absolute sense. This finding indicated that a more complex combination of factors may be needed to improve prediction.

Consistent with this possibility, our hypotheses, and recent work in suicide research (see Walsh et al., 2017, 2018), random forests produced accurate prediction in terms of AUCs, precisions, recalls, and Brier scores. Random forests constructed slightly different optimized combinations of factors across each time point, with the top 10 most important predictors also varying across time points. Moreover, predictive accuracy did *not* suffer when the most important predictor within these algorithms was removed. These latter two findings demonstrate that there is no singular recipe for NSSI prediction. Many potential factor combinations may produce accurate NSSI risk detection. In other words, there are likely many *sufficient* algorithms for accurate NSSI prediction, but there is no evidence for a *necessary* algorithm or factor for accurate NSSI prediction thus far. This high-sufficiency, low-necessity pattern is indicative of what is called a complex adaptive system (see Miller & Page, 2009). It has been proposed that such systems underlie much of biology and psychology (e.g., Edelman & Gally, 2001), including self-injurious behaviors (e.g., Ribeiro et al., 2016b).

Of note, results highlight that random forest models can produce accurate NSSI prediction; the present results *do not* highlight a causal pathway toward NSSI engagement. On a practical level, the present findings have two key implications. First, results indicate that machine learning approaches are capable of producing accurate NSSI prediction. Such results suggest that these kinds of approaches could be adopted for accurate NSSI prediction within key clinical settings. Prior to implementing these approaches, external validation, including testing these statistical approaches across other samples, populations, time windows, and clinical variables, is needed. Pending further support, many steps must be taken to translate potential machine learning tools into practice. We emphasize that the present study represents only the first of many steps toward the potential implementation of this type of approach into clinical practice.

Second, on a conceptual level, the present findings are consistent with the idea that a high degree of complexity underlies NSSI prediction. Traditionally, researchers have taken a factor-centric approach to understanding NSSI. For example, researchers have focused on factors such as affect toward NSSI stimuli (e.g., Franklin et al., 2014) and self-criticism (e.g., Fox et al., 2019). We maintain that these and many other factors are likely important for understanding NSSI, but we reason that the present results suggest a new layer of analysis for NSSI research: understanding how NSSI can emerge from a wide range of complex factor combinations. Over the past 15 years, NSSI research has produced valuable information about the many factors that are associated with NSSI (e.g., Fox et al., 2015), but there has been little focus on investigating *how* these factors combine to produce NSSI. The present findings illustrate the importance of this *how* question: Combining factors in an additive manner produced poor prediction (see multiple logistic regression results) whereas combining these same factors in a complex manner produced accurate prediction (see machine learning results). Understanding why this occurs, and how this information can be used to treat or prevent NSSI, is an important future direction for NSSI research.

The present results should be interpreted in light of study limitations. First, NSSI risk detection using machine learning algorithms is still in a preliminary stage. Additional studies are needed to identify and overcome the logistical problems that may arise before implementing these models in real-world settings. Second, NSSI prediction was constrained by the variables included in the present study. To reduce the likelihood of overly burdening participants and increase retention rates across follow-ups, we were unable to include all predictors that might confer risk for acute NSSI, such as the number of NSSI methods previously used. However, the unique contribution of individual factors is likely small and the inclusion or omission of one particular factor is unlikely to significantly alter the present findings. Although it is *possible* that other variables may have resulted in accurate NSSI prediction via traditional statistical approaches, this possibility does not appear highly *plausible* in light of meta-analyses showing chance-level prediction of NSSI (Fox et al., 2015) and suicide and related behaviors (Franklin et al., 2017) using myriad variables.

Third and relatedly, it is tempting to interpret variables that emerged as "important" from the random forest analyses. We caution against this. The strength of these variables is highly conditional within the current algorithms and, as shown within the present univariate analyses, all univariate predictors were only poor-to-fair predictors. Fourth, machine learning algorithms from the present study were trained and tested within a single sample. Bootstrapping methods were used to avoid overfitting; however, as always, external validation in independent samples is an important step forward. Additionally, future studies are needed to test whether machine learning algorithms can be applied to accurately predict NSSI onset, frequency, and severity. Research testing this possibility using more demographically diverse samples and using both clinical and community samples where NSSI engagement may be less frequent will be particularly useful, as these samples may be more generalizable to the general population.

Finally, the present study was conducted entirely online, and the large majority (88.4%) of participants reported a history of NSSI at the start of the study. Although online data collection methods confer several advantages (e.g., facilitate targeted, large-scale, higher-speed data collection, increase disclosure of mental health symptoms; Casler, Bickel, & Hackett, 2013; Hauser & Schwarz, 2016), it is possible that this approach led to biases in our data. Moreover, we were unable to predict the onset of NSSI using these methodologies. Future research should consider using machine learning techniques in large samples recruited from the community and/or mental health care settings.

The present study represents a key step toward improving NSSI risk detection. Results highlighted that complex statistical models (i.e., random forest) resulted in the most accurate NSSI prediction, and this accurate prediction applied over short-term, clinically relevant (i.e., 3-day) periods. Results provide support for theories that NSSI engagement is complex and suggest that a large set of factors and sophisticated statistical modeling may be needed for accurate NSSI risk detection.

## 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,* 9915. http://dx.doi.org/10.3402/jchimp.v2i1.9915

An, S., Malhotra, K., Dilley, C., Han-Burgess, E., Valdez, J. N., Robertson, J., . . . Sun, J. (2018). Predicting drug-resistant epilepsy - A machine learning approach based on administrative claims data. *Epilepsy & Behavior, 89,* 118–125. http://dx.doi.org/10.1016/j.yebeh.2018.10.013

Andover, M. S., Morris, B. W., Wren, A., & Bruzzese, M. E. (2012). The co-occurrence of non-suicidal self-injury and attempted suicide among adolescents: Distinguishing risk factors and psychosocial correlates. *Child and Adolescent Psychiatry and Mental Health, 6,* 11. http://dx.doi.org/10.1186/1753-2000-6-11

Andover, M. S., Pepper, C. M., Ryabchenko, K. A., Orrico, E. G., & Gibb, B. E. (2005). Self-mutilation and symptoms of depression, anxiety, and borderline personality disorder. *Suicide & Life-Threatening Behavior, 35,* 581–591. http://dx.doi.org/10.1521/suli.2005.35.5.581

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 Biometrische Zeitschrift, 54,* 657–673. http://dx.doi.org/10.1002/bimj.201100251

Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: The Scale for Suicide Ideation. *Journal of Consulting and Clinical Psychology, 47,* 343–352. http://dx.doi.org/10.1037/0022-006X.47.2.343

Beck, A. T., & Steer, R. A. (1991). *Beck Scale for Suicide Ideation: Manual*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Ranieri, W. F. (1988). Scale for suicide ideation: Psychometric properties of a self-report version. *Journal of*

*Clinical Psychology, 44,* 499–505. http://dx.doi.org/10.1002/1097-4679(198807)44:4<499::AID-JCLP2270440404>3.0.CO;2-6

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L. (2002). *Manual on setting up, using, and understanding random forests V3.1*. Retrieved from https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf

Brier, G. W. (1950). Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review, 78,* 1–3. http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29,* 2156–2160. http://dx.doi.org/10.1016/j.chb.2013.05.009

Chu, C., Bodell, L. P., Ribeiro, J. D., & Joiner, T. E. (2015). Eating disorder symptoms and suicidal ideation: The moderating role of disgust. *European Eating Disorders Review, 23,* 545–552. http://dx.doi.org/10.1002/erv.2373

Cloutier, P., Martin, J., Kennedy, A., Nixon, M. K., & Muehlenkamp, J. J. (2010). Characteristics and co-occurrence of adolescent non-suicidal self-injury and suicidal behaviours in pediatric emergency crisis services. *Journal of Youth and Adolescence, 39,* 259–269. http://dx.doi.org/10.1007/s10964-009-9465-1

Derogatis, L. R. (2001). *BSI 18, Brief Symptom Inventory 18: Administration, scoring and procedures manual*. New York, NY: NCS Pearson, Incorporated.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences of the United States of America, 98,* 13763–13768. http://dx.doi.org/10.1073/pnas.231499798

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://dx.doi.org/10.1016/j.cpr.2015.09.002

Fox, K. R., Millner, A. J., & Franklin, J. C. (2016). Classifying nonsuicidal overdoses: Nonsuicidal self-injury, suicide attempts, or neither? *Psychiatry Research, 244,* 235–242.

Fox, K. R., O'Sullivan, I. M., Wang, S. B., & Hooley, J. M. (2019). Self-criticism impacts emotional responses to pain. *Behavior Therapy, 50,* 410–420. http://dx.doi.org/10.1016/j.beth.2018.07.008

Franklin, J. C., Puzia, M. E., Lee, K. M., & Prinstein, M. J. (2014). Low implicit and explicit aversion toward self-cutting stimuli longitudinally predict nonsuicidal self-injury. *Journal of Abnormal, 123,* 463–469. http://dx.doi.org/10.1037/a0036436

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., . . . Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin, 143,* 187–232. http://dx.doi.org/10.1037/bul0000084

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

Garisch, J. A., & Wilson, M. S. (2015). Prevalence, correlates, and prospective predictors of non-suicidal self-injury among New Zealand adolescents: Cross-sectional and longitudinal survey data. *Child and Adolescent Psychiatry and Mental Health, 9,* 28. http://dx.doi.org/10.1186/s13034-015-0055-6

Glenn, C. R., & Klonsky, E. D. (2011). Prospective prediction of nonsuicidal self-injury: A 1-year longitudinal study in young adults. *Behavior Therapy, 42,* 751–762. http://dx.doi.org/10.1016/j.beth.2011.04.005

Gollust, S. E., Eisenberg, D., & Golberstein, E. (2008). Prevalence and correlates of self-injury among university students. *Journal of American College Health, 56,* 491–498. http://dx.doi.org/10.3200/JACH.56.5.491-498

Guan, K., Fox, K. R., & Prinstein, M. J. (2012). Nonsuicidal self-injury as a time-invariant predictor of adolescent suicide ideation and attempts in a diverse community sample. *Journal of Consulting and Clinical Psychology, 80,* 842–849. http://dx.doi.org/10.1037/a0029429

Hamza, C. A., Stewart, S. L., & Willoughby, T. (2012). Examining the link between nonsuicidal self-injury and suicidal behavior: A review of the literature and an integrated model. *Clinical Psychology Review, 32,* 482–495. http://dx.doi.org/10.1016/j.cpr.2012.05.003

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48,* 400–407. http://dx.doi.org/10.3758/s13428-015-0578-z

Hendin, H., Al Jurdi, R. K., Houck, P. R., Hughes, S., & Turner, J. B. (2010). Role of intense affects in predicting short-term risk for suicidal behavior: A prospective study. *Journal of Nervous and Mental Disease, 198,* 220–225. http://dx.doi.org/10.1097/NMD.0b013e3181d13d14

Hendin, H., Maltsberger, J. T., & Szanto, K. (2007). The role of intense affective states in signaling a suicide crisis. *Journal of Nervous and Mental Disease, 195,* 363–368.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-7138-7

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, 21,* 1366–1371. http://dx.doi.org/10.1038/mp.2015.198

Kessler, R. C., Warner, C. H., Ivany, C., Petukhova, M. V., Rose, S., Bromet, E. J., . . . Ursano, R. J. (2015). Predicting suicides after psychiatric hospitalization in U.S. Army soldiers: The Army Study To Assess Risk and rEsilience in Servicemembers (Army STARRS). *Journal of the American Medical Association Psychiatry, 72,* 49–57. http://dx.doi.org/10.1001/jamapsychiatry.2014.1754

Klonsky, E. D., & Muehlenkamp, J. J. (2007). Self-injury: A research review for the practitioner. *Journal of Clinical Psychology, 63,* 1045–1056. http://dx.doi.org/10.1002/jclp.20412

Kruijshaar, M. E., Barendregt, J., Vos, T., de Graaf, R., Spijker, J., & Andrews, G. (2005). Lifetime prevalence estimates of major depression: An indirect estimation method and a quantification of recall bias. *European Journal of Epidemiology, 20,* 103–111. http://dx.doi.org/10.1007/s10654-004-1009-0

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2005). *International Affective Picture System (IAPS)*. Gainesville, FL: Affective Ratings of Pictures and Instruction Manual.

Lindhiem, O., Petersen, I. T., Mentch, L. K., & Youngstrom, E. A. (2018). The importance of calibration in clinical psychology. *Assessment*. Advance online publication. http://dx.doi.org/10.1177/1073191117752055

Marshall, S. K., Tilton-Weaver, L. C., & Stattin, H. (2013). Non-suicidal self-injury and depressive symptoms during middle adolescence: A longitudinal analysis. *Journal of Youth and Adolescence, 42,* 1234–1242. http://dx.doi.org/10.1007/s10964-013-9919-3

Miller, J. H., & Page, S. E. (2009). *Complex adaptive systems: An introduction to computational models of social life*. Princeton, NJ: Princeton University Press.

Moffitt, T. E., Caspi, A., Taylor, A., Kokaua, J., Milne, B. J., Polanczyk, G., & Poulton, R. (2010). How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychological Medicine, 40,* 899–909. http://dx.doi.org/10.1017/S0033291709991036

Motwani, M., Dey, D., Berman, D. S., Germano, G., Achenbach, S., Al-Mallah, M. H., . . . Slomka, P. J. (2017). Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis. *European Heart Journal, 38,* 500–507.

Muehlenkamp, J. J., Kerr, P. L., Bradley, A. R., & Adams Larsen, M. (2010). Abuse subtypes and nonsuicidal self-injury: Preliminary evidence of complex emotion regulation patterns. *Journal of Nervous and Mental Disease, 198,* 258–263. http://dx.doi.org/10.1097/NMD.0b013e3181d612ab

Nixon, M. K., Cloutier, P., & Jansson, S. M. (2008). Nonsuicidal self-harm in youth: A population-based survey. *Canadian Medical Association Journal, 178,* 306–312. http://dx.doi.org/10.1503/cmaj.061693

Nock, M. K. (2010). Self-injury. *Annual Review of Clinical Psychology, 6,* 339–363. http://dx.doi.org/10.1146/annurev.clinpsy.121208.131258

Nock, M. K., Holmberg, E. B., Photos, V. I., & Michel, B. D. (2007). Self-Injurious Thoughts and Behaviors Interview: Development, reliability, and validity in an adolescent sample. *Psychological Assessment, 19,* 309–317. http://dx.doi.org/10.1037/1040-3590.19.3.309

Nock, M. K., Joiner, T. E., Jr., Gordon, K. H., Lloyd-Richardson, E., & Prinstein, M. J. (2006). Non-suicidal self-injury among adolescents: Diagnostic correlates and relation to suicide attempts. *Psychiatry Research, 144,* 65–72. http://dx.doi.org/10.1016/j.psychres.2006.05.010

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89,* 277–293. http://dx.doi.org/10.1037/0022-3514.89.3.277

R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ribeiro, J. D., Bender, T. W., Selby, E. A., Hames, J. L., & Joiner, T. E. (2011). Development and validation of a brief self-report measure of agitation: The Brief Agitation Measure. *Journal of Personality Assessment, 93,* 597–604. http://dx.doi.org/10.1080/00223891.2011.608758

Ribeiro, J. D., Bodell, L., & Joiner, T. E. (2012, November). *Disgust with self, others, and world in suicidality*. Poster Presented at the 46th Annual Meeting of the Association for Behavioral and Cognitive Therapies, National Harbor, MD.

Ribeiro, J. D., Franklin, J. C., Fox, K. R., Bentley, K. H., Kleiman, E. M., Chang, B. P., & Nock, M. K. (2016a). Letter to the Ed.: Suicide as a complex classification problem: Machine learning and related techniques can advance suicide prediction—A reply to Roaldset (2016). *Psychological Medicine, 46,* 1–2. http://dx.doi.org/10.1017/S0033291716000611

Ribeiro, J. D., Franklin, J. C., Fox, K. R., Bentley, K. H., Kleiman, E. M., Chang, B. P., & Nock, M. K. (2016b). 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://dx.doi.org/10.1017/S0033291715001804

Ribeiro, J. D., Witte, T. K., Van Orden, K. A., Selby, E. A., Gordon, K. H., Bender, T. W., & Joiner, T. E. (2014). Fearlessness about death: The psychometric properties and construct validity of the revision to the Acquired Capability for Suicide Scale. *Psychological Assessment, 26,* 115–126 12p. http://dx.doi.org/10.1037/a0034858

Selby, E. A., Bender, T. W., Gordon, K. H., Nock, M. K., & Joiner, T. E., Jr. (2012). Non-suicidal self-injury (NSSI) disorder: A preliminary study. *Personality Disorders, 3,* 167–175. http://dx.doi.org/10.1037/a0024405

Selby, E. A., Franklin, J., Carson-Wong, A., & Rizvi, S. L. (2013). Emotional cascades and self-injury: Investigating instability of rumination and negative emotion. *Journal of Clinical Psychology, 69,* 1213–1227. http://dx.doi.org/10.1002/jclp.21966

Simundic, A. M. (2008). Measures of diagnostic accuracy: Basic definitions. *Medical and Biological Sciences, 22,* 61–65.

Stewart, J. G., Esposito, E. C., Glenn, C. R., Gilman, S. E., Pridgen, B., Gold, J., & Auerbach, R. P. (2017). Adolescent self-injurers: Comparing non-ideators, suicide ideators, and suicide attempters. *Journal of Psychiatric Research, 84,* 105–112. http://dx.doi.org/10.1016/j.jpsychires.2016.09.031

Swannell, S. V., Martin, G. E., Page, A., Hasking, P., & St John, N. J. (2014). Prevalence of nonsuicidal self-injury in nonclinical samples: Systematic review, meta-analysis and meta-regression. *Suicide & Life-Threatening Behavior, 44,* 273–303. http://dx.doi.org/10.1111/sltb.12070

Taylor, P. J., Jomar, K., Dhingra, K., Forrester, R., Shahmalak, U., & Dickson, J. M. (2018). A meta-analysis of the prevalence of different functions of non-suicidal self-injury. *Journal of Affective Disorders, 227,* 759–769.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science, 5,* 457–469. http://dx.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,* 1261–1270. http://dx.doi.org/10.1111/jcpp.12916

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12,* 1100–1122. http://dx.doi.org/10.1177/1745691617693393



## Appendix

## Data Transparency Statement

The data reported in this article were collected as part of a larger study. Findings from the data collection on distinct topics have been reported in separate articles. Article 1 (in press) focused on the prediction of suicide ideation and attempt. A selective subset of data from the baseline survey were included as one of five samples included in Article 2 (under review), which examined the cross-sectional differences between suicide ideators and attempters. The models and relationships examined in the present article have not been examined in any other articles that were submitted for review.

Received November 30, 2018
Revision received April 9, 2019
Accepted April 30, 2019 ■