BMC Psychiatry

RESEARCH

Open Access

Comparison between clinician and machine learning prediction in a randomized controlled trial for nonsuicidal self-injury

Moa Pontén[1*], Oskar Flygare[1], Martin Bellander[1], Moa Karemyr[1], Jannike Nilbrink[1], Clara Hellner[1], Olivia Ojala[1] and Johan Bjureberg[1]

Abstract

Background Nonsuicidal self-injury is a common health problem in adolescents and associated with future suicidal behavior. Predicting who will benefit from treatment is an urgent and a critical first step towards personalized treatment approaches. Machine-learning algorithms have been proposed as techniques that might outperform clinicians' judgment. The aim of this study was to explore clinician predictions of which adolescents would abstain from nonsuicidal self-injury after treatment as well as how these predictions match machine-learning algorithm predictions.

Methods Data from a recent trial evaluating an internet-delivered emotion regulation therapy for adolescents with nonsuicidal self-injury was used. Clinician predictions of which patients would abstain from nonsuicidal self-injury (measured using the youth version of Deliberate Self-harm Inventory) were compared to a random forest model trained on the same available data from baseline assessments.

Results Both clinician (accuracy = 0.63) and model-based (accuracy = 0.67) predictions achieved significantly better accuracy than a model that classified all patients as reaching NSSI remission (accuracy = 0.49 [95% CI 0.41 to 0.58]), however there was no statistically significant difference between them. Adding clinician predictions to the random forest model did not improve accuracy. Emotion dysregulation was identified as the most important predictor of nonsuicidal self-injury absence.

Conclusions Preliminary findings indicate comparable prediction accuracy between clinicians and a machine-learning algorithm in the psychological treatment of nonsuicidal self-injury in youth. As both prediction approaches achieved modest accuracy, the current results indicate the need for further research to enhance the predictive power of machine-learning algorithms. Machine learning model indicated that emotion dysregulation may be of importance in treatment planning, information that was not available from clinician predictions.

Trial Registration NCT03353961||https://www.clinicaltrials.gov/, registered 2017–11-21.

Preregistration at Open Science Framework: https://osf.io/vym96/.

Keywords Artificial intelligence, Random forest analysis, Nonsuicidal self-injury, Emotion regulation, Machine learning

*Correspondence:
Moa Pontén
moa.ponten@ki.se
Full list of author information is available at the end of the article

Introduction

Nonsuicidal self-injury (NSSI) is a common health problem in adolescents [1]. Possible adverse outcomes associated with NSSI include suicidal behavior, thus making NSSI an important treatment target [2]. Nonsuicidal self-injury disorder (NSSID) was proposed as a diagnostic category in the Fifth Edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). To meet the diagnostic criteria, the following criteria need to be fulfilled: (1) engagement in NSSI on 5 or more days in the past year (Criterion A); (2) the expectation that NSSI will solve an interpersonal problem, provide relief from unpleasant thoughts and/or emotions, or induce a positive emotional state (Criterion B); (3) the experience of one or more of the following: (a) interpersonal problems or negative thoughts or emotions immediately prior to NSSI, (b) preoccupation with NSSI that is difficult to manage, or (c) frequent thoughts about NSSI (Criterion C); (4) the NSSI is not socially sanctioned or restricted to minor self-injurious behaviors (Criterion D); (5) the presence of NSSI-related clinically significant distress or interference across different domains of functioning (e.g., work, relationships; Criterion E); and (6) the NSSI does not occur only in the context of psychosis, delirium, or substance use/withdrawal and is not better accounted for by another psychiatric disorder or medical condition (Criterion F) [3].

Treatment of NSSI is often time-consuming and difficult to access [4]. Novel treatment approaches, such as Internet-delivered Emotion Regulation Individual Therapy (IERITA [5]), are developed to make treatment more accessible for this population. Given the heterogeneity across adolescents with NSSI [1], understanding which patients benefit from different therapies is an essential objective in a personalized treatment approach. This approach may help clinicians identify patients who could benefit from more intensive or supplementary interventions if a poor treatment outcome is predicted.

Clinical work involves decision-making, ranging from micro-level choices during a patient visit to macro-level predictions necessary for treatment triangulation based on clinical guidelines and patients' characteristics. By combining clinical experience with the clinical presentation of the individual's symptoms and history, the clinician forms an intuitive "gut-feeling" whether the treatment will be effective [6]. This clinical judgement is often preferred by the clinicians themselves [6]. However, clinical prediction tends to overestimate treatment effect, failing to identify individuals at risk for worse treatment outcome [7]. Further, experience, training and/or consultation seem to only marginally improve clinical judgement [8]. With the goal to improve prediction, statistical methods have been explored as an alternative or complement to clinical judgement [9–11].

Machine learning (ML) is a set of techniques that are promising in predicting disease and treatment outcome [12] and marks a paradigm shift also in psychiatry research [13]. These statistical tools allow multiple variables to be examined simultaneously, even correlated ones, and can illustrate complex non-linear patterns. Historically, research has focused on single variables, such as clinical characteristics, genes or brain data in predicting treatment outcome. However, none of these predictors alone have shown a large effect in psychiatric research. Therefore, machine learning methods are exceptionally well suited for predicting treatment outcomes as they allow aggregating small effects. One example of clinical application is from a trial for body dysmorphic disorder where the treatment outcome was predicted with 78% accuracy, indicating potential for clinical utility for these methods [14].

Although ML increasingly has been used to predict treatment outcome in psychiatric research [13], only a few studies have compared ML to clinicians in predicting treatment outcome. In two studies on psychological treatment for alcohol dependence the authors found that the ML is comparable to clinical judgment [15] and that ML outperformed clinicians' intuition [11].

Based on data from a randomized clinical trial conducted by our group [16], this study sought to explore the accuracy of clinician prediction of treatment response following IERITA, an internet-delivered intervention for adolescents with NSSID. We also wanted to explore the difference between a machine learning (ML) algorithm compared to clinicians in predicting treatment outcomes and explore the most important predictors. As an exploratory model development study, we did not form specific a priori hypotheses regarding predictors.

Given the limited literature on what informs and impacts clinician prediction of outcomes we wanted to explore the following secondary aims (1) if clinician confidence was associated with accuracy in prediction; (2) if the confidence and/or accuracy of clinician predictions improved as therapist treated more patients in the trial; (3) if clinician predictions were related to the amount of time they spent on treatment and how many messages they sent to the patient; and finally (4) if clinician predictions of NSSI outcome were related to the number of treatment modules the patient completed.

Materials and methods

Design

Data was gathered from a recent randomized clinical trial ($N=166$) conducted at three sites within Child and Adolescent Mental Health Services in Sweden

(NCT03353961||https://www.clinicaltrials.gov/,registered 2017–11–21) where the intervention was superior to the treatment as usual (TAU) control condition in reducing nonsuicidal self-injury [17]. Those randomized to TAU were offered IERITA after six months. Recruitment took place between November 20, 2017, to April 9, 2020 with follow-up January 2021. The trial was approved by the Stockholm Regional Ethical Review Board (no. 2017/1807–31), and all participants provided written informed consent, with older participants providing written consent and younger participants verbal consent with parental written consent. Random allocation sequence was conducted by an independent researcher using a true random number service (Random.org) in blocks of 4 or 6 for each treatment clinic and stored in sealed, opaque envelopes. The Consolidated Standards of Reporting Trials (CONSORT) and TRIPOD reporting guidelines were followed in the reporting of this study [18].

Participants

The current study included participants ($n=138$) across groups, i.e. those who received Internet-delivered Emotion Regulation Individual Therapy (IERITA) immediately ($n=84$) and those assigned to TAU who later enrolled in IERITA ($n=54$) (see Fig. 1). Both groups received treatment for 12 weeks. For baseline demographics such as NSSI frequency see Table 1. No a priori power analysis was made since the study was based on already collected data. Inclusion criteria comprised adolescents aged 13 to 17 with NSSI disorder and experiencing ≥ 1 NSSI episode during the past month (measured using the youth version of Deliberate Self-harm Inventory), with one parent willing to join the parent program [17]. Exclusion criteria involved a Children's Global Assessment Scale (CGAS) score below 40 [19], insufficient understanding of the Swedish language, immediate suicide risk, psychotic or bipolar I disorder diagnosis, current (past month) substance use disorder, life circumstances hindering participation, or other psychiatric disorder requiring immediate treatment [17].

Interventions

IERITA is a clinician-supported 12-week acceptance-based behavioral therapy, aiming to reduce NSSI by improving emotion regulation ability. Adolescents receive 11 weekly modules, incorporating text, films, and interactive exercises to learn emotion regulation skills. A supplementary mobile app facilitates learning and skill practice. A detailed description of IERITA is found in Bjureberg et al., 2023 [17].

IERITA was offered as an addition to TAU. TAU was delivered within regular healthcare services according to needs (e.g., pharmacological treatment or supportive therapy), resulting in varying types and frequencies of treatments for the participating adolescents outside the trial. The TAU only condition was enhanced by referral to adequate treatment if necessary, the establishment of a safety-plan, self-rated assessments every week, and follow-up assessments.

During IERITA, patients had asynchronous contact via a message function in the platform with a dedicated clinician who reinforced treatment engagement and assisted with homework assignment by giving corrective feedback and psychoeducation. The clinicians ($n=15$) were either psychologists or psychotherapists and the vast majority worked within child and adolescent healthcare. The clinicians received structured training and support, emphasizing adherence to the protocol.

Measures

The outcome of interest was the proportion of patients with an absence of NSSI (yes/no) at one-month post-treatment, as reported in a youth version of Deliberate Self-harm Inventory (DSHI-Y; [20, 21]), which has shown adequate construct, convergent, and discriminant validity [20]. DSHI-Y measures the frequency of the 6 most common forms of NSSI (eg, cutting and burning in the past 30 days) without conscious suicidal intent, but resulting in injury severe enough for tissue damage (e.g., scarring) to occur. For the adolescents receiving IERITA straight after randomization, DSHI-Y was assessed by a clinician blind to allocation and independent from the scientific team. For the adolescents receiving IERITA six months later, DSHI-Y was self-reported one month post-treatment. The outcome was measured with reference to the 30 days post-intervention in both groups.

Predictors

Clinicians predicted adolescent NSSI abstinence with a yes/no response and rated confidence in percentages in steps of 10 (i.e., 0, 10 […] or 100%) after a face-to-face assessment but prior to randomization. There was one prediction made per patient. Demographic and clinical variables measured at baseline were available to clinicians and used in the ML model (Table 1). Predictor variables in the ML model were selected to match the information available to clinicians. Clinicians did not receive specific instructions on what information to use in their predictions or how to weigh predictors but were instructed to use their "clinical hunch".

Statistical analyses

Preprocessing and feature selection

For the machine learning model a random forest model was used [22]. The data was pre-processed by keeping

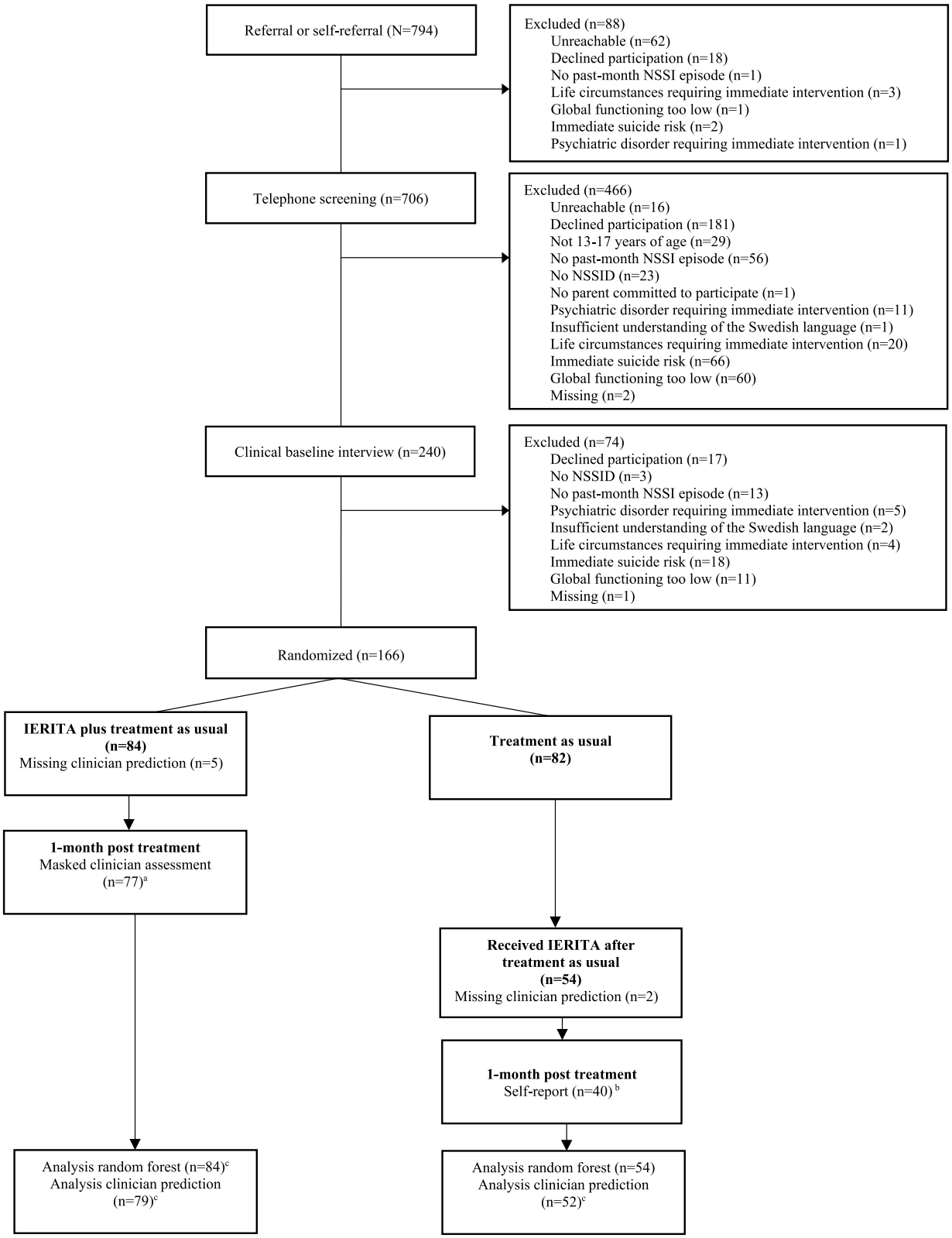

Fig. 1 Flow Diagram of Patient Enrollment and Disposition. IERITA indicates internet-delivered emotion regulation individual therapy for adolescents; NSSID, nonsuicidal self-injury disorder.[a] Masked Assessor-rated Deliberate Self-Harm Inventory–Youth Version. [b] Self-reported Deliberate Self-Harm Inventory–Youth Version. [c] All randomized participants for which a clinician prediction was made are included in the analysis

Table 1 Participant Characteristics

	No. (%)
	Total (*n* = 138)
Participant Characteristics	
Gender	
Female	126 (91)
Male	7 (5)
Non-binary	5 (4)
Age, mean (SD)	15.03 (1.24)
Sexual orientation	
Heterosexual	92 (67)
Sexual minority	43 (31)
No answer	3 (2)
Any failed grades (yes)	19 (14)
Ever been bullied (yes)	71 (51)
School absence, mean (SD)	4.39 (4.97)
Parent Characteristics[a]	
Parent living arrangement	
With children	112 (81)
With spouse/partner	101 (73)
Parent education level	
Primary school	2 (1)
Secondary school	54 (39)
College/university < 3 years	13 (9)
College/university ≥ 3 years	60 (43)
Doctorate	9 (7)
Emotion dysregulation, mean (SD)	29.15 (10.85)
Parental ability (Coping with Children's Negative Emotions Scale)	
Problem-Focused, mean (SD)	5.77 (0.83)
Emotion-Focused, mean (SD)	5.1 (0.95)
Expressive Encouragement, mean (SD)	5.24 (0.94)
Minimization, mean (SD)	2.8 (1.02)
Punitive, mean (SD)	1.49 (0.54)
Distress, mean (SD)	1.84 (0.86)
	No. (%)
	Total (*n* = 138)
Participant Clinical Characteristics	
Age NSSI onset, mean (SD),	12.67 (1.38)
Years since NSSI onset, mean (SD)	2.36 (1.31)
NSSI frequency in the past 30 days, DSHI-Y, mean (SD)	3.07 (3.54)
NSSI versatility, DSHI-Y, mean (SD)	1.30 (1.02)
Comorbidity[b]	
Major depressive disorder	77 (56)
Dysthymia[e]	6 (4)
Anxiety disorders	
Social anxiety disorder	41 (30)
Panic disorder	12 (9)
Agoraphobia	18 (13)
Specific phobia disorder	23 (17)
Generalized anxiety disorder	19 (14)

Table 1 (continued)

	No. (%)
Separation anxiety[e]	3 (2)
OCD[e]	4 (3)
BDD[e]	6 (4)
ADHD[c]	26 (19)
Autism spectrum disorder[e]	5 (4)
Anorexia[e]	1 (1)
Bulimia[e]	6 (4)
Oppositional defiant disorder[e]	4 (3)
Depression, Anxiety and Stress Scale (DASS-21)	
Depression, mean (SD)	12.09 (4.83)
Anxious, mean (SD)	8.25 (4.19)
Stress, mean (SD)	11.54 (4.36)
Insomnia symptoms, ISI, mean (SD)	11.18 (5.62)
Number of co-occurring disorders, mean (SD)	1.82 (1.57)
Number of BPD criteria[d], mean (SD)	1.93 (1.35)
Self-destructive behaviours, BSL, mean (SD)	2.8 (2.46)
Suicidality	
Low	57 (41)
Moderate	36 (26)
High	45 (33)
Life-time suicide attempt, yes	19 (14)
Ever received inpatient care, yes[e]	3 (2)
Emotion dysregulation, Ders-16, mean (SD)	58.50 (12.12)
Quality of life, Kid-Screen, mean (SD)	28.58 (4.38)
Psychological flexibility, AAQ, mean (SD)	31.68 (8.18)
Any ongoing psychopharmacological medication, mean (SD)	50 (36)
Time in ongoing counselling, mean (SD), mo	9.67 (12.36)
Ongoing counselling at inclusion	97 (70)
Global functioning, (CGAS) mean (SD)	54.27 (5.84)
Clinical severity (CGI-S) mean (SD)	
Mildly ill (3)	24 (17)
Moderately ill (4)	73 (53)
Markedly ill (5)	36 (26)
Severely ill (6)	5 (4)

Abbreviations: AAQ Action and Acceptance Questionnarie, *ADHD* attention-deficit hyperactivity disorder, *BDD* body dysmorphic disorder, *BSL* Borderline Symptom List, *BPD* borderline personality disorder, *CGI-S* Clinical global impression-severity, *CGAS* Children's Global Assessment Scale, *DERS* The Difficulties in Emotion Regulation Scale, *DSHI-Y* Deliberate Self-Harm Inventory–Youth Version, *ISI* Insomnia severity index, *NSSI* nonsuicidal self-injury, *OCD* obsessive–compulsive disorder

[a] In case of two parents, one was assigned and consented to contribute to answer self-reports questions. Multiple answers were allowed

[b] Assessed by the research team using the MINI-KID International Neuropsychiatric Interview, version 6 and the Body Dysmorphic Disorder Questionnaire (administered as an interview)

[c] Includes both combined, primarily inattentive, and primarily hyperactive-impulsive subtype

[d] Assessed by the research team using the Structured Clinical Interview for DSM

[e] Not included in the final analysis due to insufficient variance

predictors that had sufficient variance and were independent from other predictor variables and had no more than 30% missing data, in line with previous studies in this field [14, 23, 24]. Missing data in all predictors were imputed using bagged trees using the *tidymodels* R-package [25, 26].

For the primary outcome variable (21 missing observations), missing data was imputed using predictive mean matching based on the weekly ratings of NSSI episodes collected during treatment [27], excluding the predictor variables to prevent leakage [28]. The random forest model was fitted without hyperparameter tuning using

the *ranger* package and with internal validation in order to reduce the risk of overfitting; i.e. always growing 500 trees, selecting the square root of the total number of predictors at each split, using a minimum node size of 10, and using tenfold cross-validation [29]. Variable importance was estimated using corrected Gini importance [30] and corresponding permutation-based *p*-values were estimated [31].

The random forest model and clinician predictions were compared using accuracy (the proportion of correct predictions), sensitivity/specificity (the proportion of patients with/without absence of NSSI correctly detected), positive predictive value/negative predictive value (the proportion of true positives/negatives among the model predictions) as well as receiver operating characteristics (ROC) curves and their corresponding area under the curve (AUC). Further, we applied McNemar's test to evaluate whether there was a statistically significant difference in classification performance between clinician and random forest predictions [32].

The association between clinician's confidence (0–100%) and the accuracy of clinician predictions was tested using a logistic mixed-effect model with random intercept, with accuracy (0/1) as the dependent variable and confidence as the independent variable ($n=130$). To investigate if clinician accuracy improved over time a logistic mixed-effects model with random intercept was employed, with prediction accuracy as the dependent variable, and the patient order, i.e. the number of patients the clinician had treated as the independent variable ($n=130$). The clinician confidence over time was tested using a mixed-effects model with random intercept, with prediction confidence as the dependent variable, and the patient order as the independent variable ($n=130$). Whether therapist predicted probabilities related to time spent treating ($n=61$) and number of messages sent ($n=63$) was tested using a mixed-effects model with random intercept for each of the two dependent variables (time spent, number of messages sent) and predicted probability as the independent variable. Patients were only included in this analysis if the therapist making the prediction were also treating the patient. The relationship between therapist predictions and number of modules completed (treatment dose) was investigated using a linear regression, with number of completed modules as the dependent variable and predicted probability and baseline NSSI as the independent variables ($n=79$).

All statistical analyses were performed using R version 4.3.1 [33]. The pre-registered statistical analysis plan, as well as scripts used to produce the results, are available on the Open Science Framework (https://osf.io/vym96/). We originally planned in secondary analyses to evaluate whether patients would improve, deteriorate, or have no change on the CGI-I (Clinical Global Impression scale–Improvement) in addition to absence of NSSI, however these analyses were not feasible as clinicians predicted that 98% of all patients would improve after treatment and CGI-I data were only available for 80 patients.

Results

The random forest model utilized data from all participants ($n=138$), however clinician predictions were unavailable for 7 patients and clinician predictions are therefore based on estimates from $n=131$ participants.

Clinicians predicted NSSI absence for 80 (58%) patients, NSSI presence for 51 (37%) patients and prediction was missing for 7 (5%) patients. During the post-treatment follow-up period 60 (44%) patients did not engage in self-harm, 57 (41%) patients did engage in self-harm, while data was missing for 21 (15%) patients. The confidence ratings by the clinicians ranged from 0 to 100 (mean = 57.3, SD = 18.1).

Clinician predictions

Area under the curve for the clinician predictions was 0.65 (95% CI 0.55 to 0.74) (see Fig. 2). The clinician predictions achieved an overall accuracy of 0.63, a sensitivity of 0.74 and specificity of 0.52. The positive predictive value was 0.61 and the negative predictive value was 0.67.

Machine learning predictions

During pre-processing, information regarding previous inpatient admissions and nine rare comorbid conditions were removed due to near-zero variance. No predictor variables were removed due to high collinearity or large proportion of missing values, and the final random forest algorithm ended up using 44 predictor variables (listed in Table 1).

Area under the curve for the random forest model was 0.7 (95% CI 0.61 to 0.78) (see Fig. 2). The model reached an accuracy of 0.67, with a sensitivity of 0.61 and specificity of 0.72. The positive predictive value was 0.69 and the negative predictive value was 0.64.

The 9 most important predictors (p-value < 0.05) in the random forest model are shown in Fig. 3, where difficulties in emotion regulation were identified as the most important predictor of nonsuicidal self-injury absence after treatment.

Clinician predictions vs machine learning

Both the random forest model and the clinician predictions were more accurate than a baseline model predicting that all patients would reach NSSI absence, which had an accuracy of 0.49 (95% CI 0.41 to 0.58). The random forest model and clinician predictions did not differ in terms of accuracy (0.67 and 0.63 respectively, McNemar

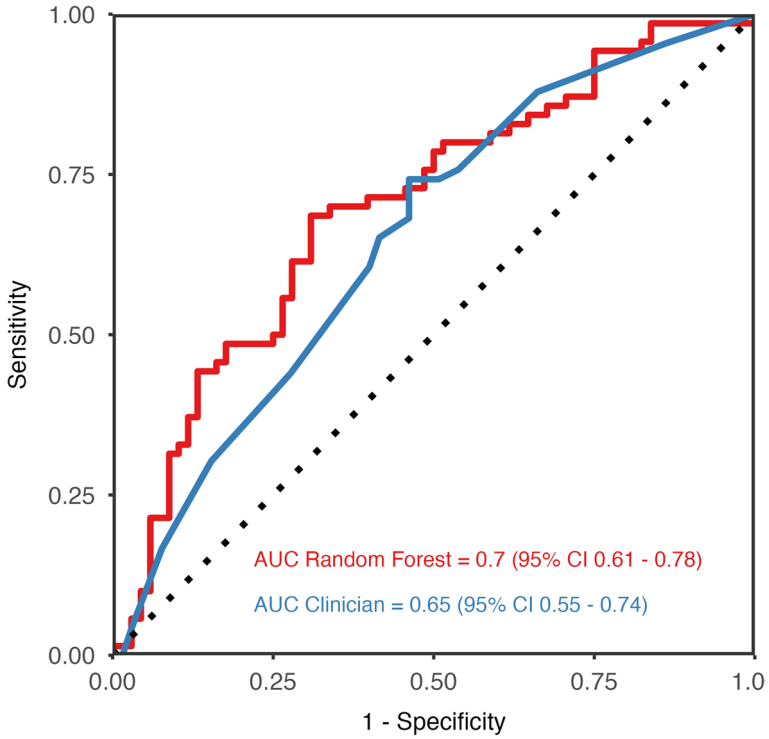

Fig. 2 Receiver operating characteristics curves for clinician and ML predictions**.** *Abbreviation:* AUC, area under the curve

test = 0.205, df = 1, $p = 0.65$), and adding the clinician prediction to the random forest model did not improve accuracy compared to the random forest model alone (Accuracy = 0.65, McNemar test = 0.308, df = 1, $p = 0.58$). Both predictions were correct for 37 (28%) patients, and there was agreement between therapists and the random forest model for 57 (44%) patients.

Secondary aims

We did not find evidence for an association between clinicians' confidence and the accuracy of their predictions (β = 0.012, SE = 0.011, $p = 0.269$). Furthermore, we failed to find evidence that clinicians' accuracy or confidence improve with number of patients they predicted and treated, since the order of patient was not associated with either clinicians' accuracy of prediction (β = 0.017, SE = 0.032, $p = 0.601$) or clinicians' confidence in their prediction (β = 0.156, SE = 0.243, $p = 0.523$). Therapist spent on average 241 min (SD = 221) treating each patient and sent on average 16.4 messages (SD = 7.22). The clinicians' prediction probabilities were not associated with the time spent on treatment (β = −11.033, SE = 93.645, $p = 0.907$) or the number of messages they sent (β = −1.913, SE = 2.532, $p = 0.453$). Lastly, the number of modules completed by patients was not associated with clinicians' prediction probabilities (β = −0.482, SE = 0.876, $p = 0.584$, $n = 79$).

Sensitivity analyses

To evaluate whether the method for assessing NSSI absence (clinician-rated or self-rated) impacted model accuracy, we compared model performance for patients receiving IERITA immediately and patients receiving IERITA after TAU. The accuracy was slightly higher in the first group (accuracy = 0.70, sensitivity = 0.67, specificity = 0.74) compared to the latter (accuracy = 0.61, sensitivity = 0.52, specificity = 0.69), however the confidence intervals of the area under the curve metric overlapped (see eFigure 1 in supplemental materials, McNemar test not possible due to differing number of participants in the groups).

Discussion

This study explored the predictive abilities of clinicians and a random forest model for treatment response following an internet-delivered emotion regulation treatment for adolescents with NSSI. Findings suggest that both clinicians and the ML algorithm performed above chance but at comparable accuracy rates (63% for clinicians and 67% for ML). However, these accuracies do not meet current thresholds for clinical application. Clinician confidence was not associated with accuracy in prediction, and neither clinician confidence nor accuracy improved over time. Therapist predictions and confidence ratings showed no association with therapist time

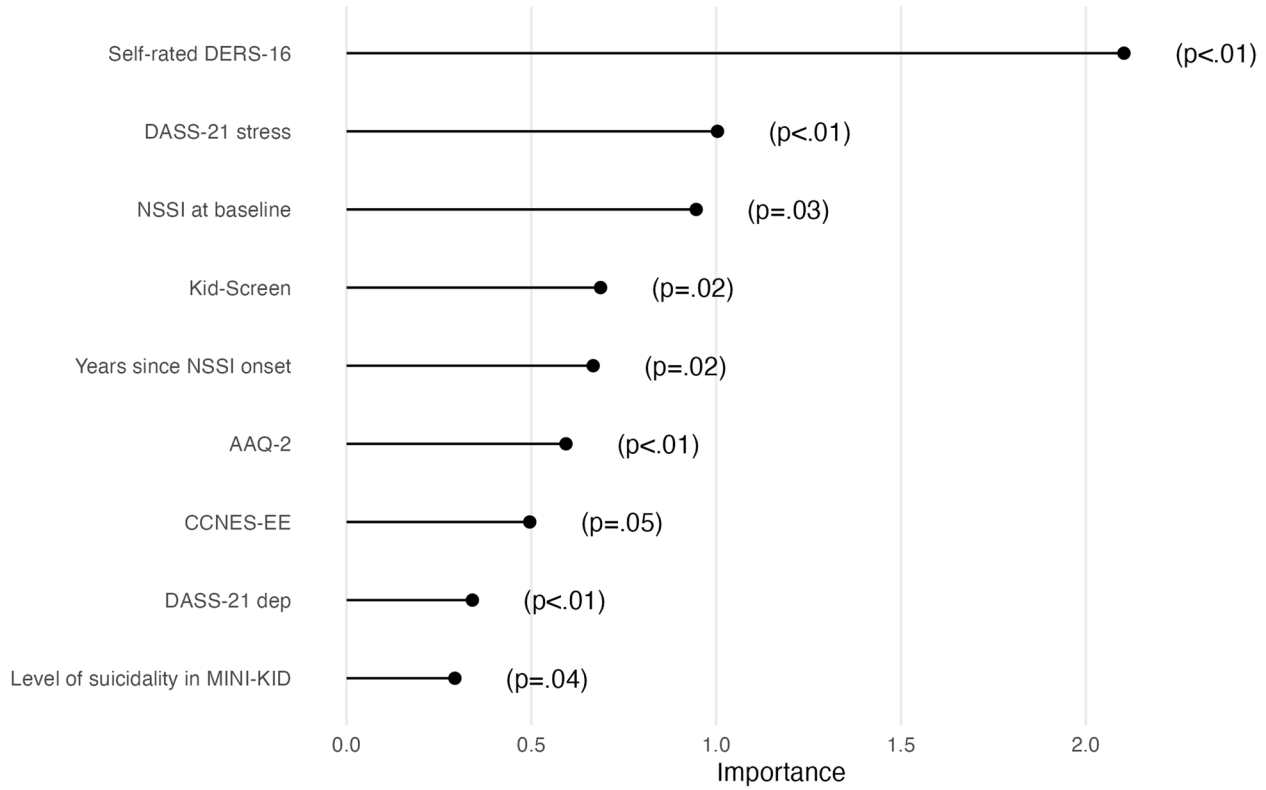

Fig. 3 Predictor importance of top 9 variables identified by the algorithm predicting self-injury post-treatment. The most important predictor was self-rated emotion dysregulation on the DERS-16 at baseline (Importance = 2.10, $p < .01$), followed by self-rated stress symptoms in DASS-21 (Importance = 1.0, $p < .01$), number of NSSI episodes on the DSHI-Y at baseline (Importance = 0.95, $p = .029$), quality of life on the Kid-Screen (Importance = 0.67, $p = .019$), number of years since NSSI onset (Importance = 0.67, $p = .019$), psychological flexibility according to AAQ-2 (Importance = 0.59, $p < .01$), parental expressive encouragement on the CCNES (Importance = 0.50, $p = .049$), self-rated depression symptoms in DASS-21 (Importance = 0.34, $p < .01$), level of suicidality in MINI-KID (Importance = 0.29, $p = .039$). Abbreviation: DERS-16, Difficulties in Emotion Regulation Scale; CCNES-EE, Coping with Children's Negative Emotions Scale, Expressive Encouragement; DASS-21, Depression, Anxiety and Stress Scale – 21; NSSI, nonsuicidal self-injury; AAQ-2, Action and Acceptance Questionnaire

spent on asynchronous contact with individual patients, including the number of messages sent and time spent per patient.

This study adds to the limited literature on mental health-clinicians vs ML predictions of post-treatment outcome. In contrast with previous findings [11], our data suggest that ML and clinician predictions are comparable. However, although better than chance, prediction accuracies of 63% (clinicians) and 67% (ML) are not useful or ready for implementation in clinical practice. There is evidence that clinicians view predictions with at least 65% accuracy as suitable for practical application [34], and recent empirical findings suggest that an accuracy of 67% should be a minimum benchmark for clinically useful predictions [35].

Importantly, the ML prediction did not improve after including the clinician prediction, suggesting that clinical intuition does not add information above and beyond the observable data available to the ML. Thus, even though clinicians interacted with patients face-to-face at baseline and had the opportunity to form clinical assumptions based on the patient's behavior during this interaction, this information did not appear to be crucial for prediction; at least in situations when rich baseline data is available. Instead, our results suggest that the information associated with clinical intuition, that was a statistically significant predictor of NSSI absence, may be observable in self-report data. Future research should delineate what information clinicians use when making clinical predictions, eg. when tailoring a treatment to an individual patient's characteristics. Additionally, this may include asking them to rank the relative importance of these factors to see similarities and differences to those factors that emerged from the random forest model.

It is noteworthy that clinician predictions were based on intuitive information from both face-to-face patient interactions and observable baseline patient data, the same information typically employed in regular care, thereby enhancing the study's ecological validity. This is in contrast to studies with clinicians only examining

baseline patient data before making their prediction [15, 36].

Further, the clinicians' accuracy in prediction did not improve with time after meeting more patients. This aligns with previous data suggesting that clinical predictions may not be useful in clinical practice across a range of disorders and treatment modalities [7, 11]. This may not be surprising given humans' limited capacity to process and weigh information based on more than four variables [37]. Our preliminary findings suggest that ML is already on par with clinicians' ability to predict outcome and future research should focus on identifying key variables that may further improve the prediction accuracy of ML.

Entering the trial with high levels of emotion dysregulation emerged as the most important predictor in the ML, information that was not available from the clinician predictions. These findings indicate that a machine learning model may add information beyond clinician prediction and that patients with more severe difficulties with emotion regulation are more likely to abstain from NSSI the month after this brief internet-delivered intervention. This aligns with prior findings demonstrating a positive association between clinical severity and treatment response in treatments targeting emotion dysregulation in patients with NSSI (e.g., [38–41]), potentially partly explained by regression to the mean, where high levels of a symptom may normalize over time. In addition, it is possible that self-harm that is not driven by emotion dysregulation may be less responsive to this particular treatment. Furthermore, it is important to note that the effect of one predictor is impacted by all other predictors in the ML model. We have previously found that emotion dysregulation and NSSI did not strongly predict treatment outcome when investigated separately in simple regression analyses [16]. Further, we have recent findings indicating that it might be particular patterns of week-to-week variability in baseline emotion dysregulation that is associated with treatment outcome [42]. Future research should delineate the predictive role of emotion dysregulation by studying how it interacts with other variables and varies over time before and after enrolling treatment.

Finally, previous research has shown that clinicians often overestimate patients improvement and ML algorithms are better at predicting those who deteriorate [11]. Predictions based on ML can therefore complement clinician predictions, as it is clinically more relevant to identify those in need of more resources. The PPV, or the likelihood that a positive prediction indicates a true need for intervention, was 69% for ML and 61% for clinician prediction. The NPV, or the likelihood that a negative prediction accurately indicates a lack of need for intervention, was 64% for ML and 67% for clinician prediction. Given that IERITA is brief and low resource intensive [17] there is little room for reducing intervention dose based on a strong prognosis. Future research might therefore consider weighing PPV higher than NPV in the algorithm, as it may be more critical to identify those truly in need of additional support. This approach has for example been shown to lead to improved outcomes in the treatment of insomnia in adults [35].

Limitations

First, the relatively limited sample size did not allow us to separate some of the data for testing, a practice preferably undertaken in ML [25, 43] and it is currently unknown if the observed results will generalize to nhe refore come of the defnot

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12888-024-06391-x.

Supplementary Material 1.

Acknowledgements
Not applicable

CRediT authorship contribution statement
Concept and design: Pontén, Flygare, Bellander, Bjureberg. *Acquisition, analysis, or interpretation of data*: Pontén, Flygare, Ojala, Bellander, Bjureberg. *Drafting of the manuscript*: Pontén, Flygare, Ojala, Bellander, Karemyr, Nilbrink, Bjureberg. *Critical revision of the manuscript for important intellectual content*: All authors.*Statistical analysis*: Flygare, Bellander. *Obtained funding*: Bjureberg. *Administrative, technical, or material support*: Hellner, Ojala, Bjureberg. *Supervision*: Bjureberg.

Authors' contributions
Concept and design: Pontén, Flygare, Bellander, Bjureberg. Acquisition, analysis, or interpretation of data: Pontén, Flygare, Ojala, Bellander, Bjureberg. Drafting of the manuscript: Pontén, Flygare, Ojala, Bellander, Karemyr, Nilbrink, Bjureberg. Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: Flygare, Bellander. Obtained funding: Bjureberg. Administrative, technical, or material support: Hellner, Ojala, Bjureberg. Supervision: Bjureberg.

Funding
Open access funding provided by Karolinska Institute. This work was funded by the Swedish Research Council (grant Nos. 2014.1008; 2017–01506), Marcus and Amelia Wallenberg Foundation (grant No.MAW2014.0021), Fredrik and Ingrid Thuring's Foundation, Clas Groschinsky's Foundation (grant No. SF18121), Sven Jerring Foundation, Kempe-Carlgrenska Foundation, and Bror Gadelius Foundation. Johan Bjureberg was supported by the Knut and Alice Wallenberg's Foundation (grant No. 2018.0426) and The Royal Swedish Academy of Letters, History and Antiquities, and Stiftelsen Natur & Kultur. This study was supported by the National Self Injury Project in Sweden. Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data availability
The data used for analyses contain sensitive personal identifying information and are not publicly available as data sharing was not part of the written informed consent. Data are available from the corresponding author on reasonable request. Statistical code used for the analyses is publicly available from the Open Science Framework repository: https://osf.io/vym96/).

Declarations

Ethics approval and consent to participate
The trial was approved by the Stockholm Regional Ethical Review Board (no. 2017/1807–31). All participants provided informed consent, with older participants providing written consent and younger participants verbal consent with parental written consent.

Consent for publication
Not applicable.

Competing interests
Dr Flygare has received speaking fees from the Swedish OCD Association, Insight Events AB, WeMind AB, and Kry International AB, as well as reimbursement for writing articles for Inside Practice Psychiatry, all outside the submitted work. Dr Bjureberg receives royalties from Natur & Kultur, outside the submitted work. The other authors declare no potential conflicts of interest.

Author details
[1]Centre for Psychiatry Research, Department of Clinical Neuroscience, Stockholm, Karolinska Institutet, Sweden & Stockholm Health Care Services, Region Stockholm, Norra Stationsgatan 69, 113 64 Stockholm, Sweden.

Received: 17 September 2024 Accepted: 8 December 2024
Published online: 18 December 2024

References

1. Hawton K, Saunders KE, O'Connor RC. Self-harm and suicide in adolescents. The Lancet. 2012;379(9834):2373–82.
2. Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, et al. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. Psychol Med. 2016;46(2):225–36.
3. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders: Dsm-5. Arlington, VA: Amer Psychiatric Pub Incorporated; 2013.
4. Glenn CR, Esposito EC, Porter AC, Robinson DJ. Evidence Base Update of Psychosocial Treatments for Self-Injurious Thoughts and Behaviors in Youth. J Clin Child Adolesc Psychol. 2019;48(3):357–92.
5. Bjureberg J, Ojala O, Hesser H, Häbel H, Sahlin H, Gratz KL, Tull MT, Claesdotter Knutsson E, Hedman-Lagerlöf E, Ljótsson B, Hellner C. Effect of Internet-Delivered Emotion Regulation Individual Therapy for Adolescents With Nonsuicidal Self-Injury Disorder: A Randomized Clinical Trial. JAMA Network Open. 2023;6(7):e2322069.
6. Ægisdóttir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, et al. The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. Couns Psychol. 2006;34(3):341–82.
7. Hannan C, Lambert MJ, Harmon C, Nielsen SL, Smart DW, Shimokawa K, et al. A lab test and algorithms for identifying clients at risk for treatment failure. J Clin Psychol. 2005Feb 1;61(2):155–63.
8. Spengler PM, Pilipis LA. A comprehensive meta-reanalysis of the robustness of the experience-accuracy effect in clinical judgment. J Couns Psychol. 2015;62(3):360–78.
9. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: A meta-analysis. Psychol Assess. 2000;12(1):19–30.
10. Meehl PE. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press; 1954. Available from: http://content.apa.org/books/11281-000. Cited 2023 Oct 20.
11. Symons M, Feeney GFX, Gallagher MR, Young RMcD, Connor JP. Predicting alcohol dependence treatment outcomes: a prospective comparative study of clinical psychologists versus 'trained' machine learning models. Addiction. 2020;115(11):2164–75.
12. Taubitz FS, Büdenbender B, Alpers GW. What the future holds: Machine learning to predict success in psychotherapy. Behav Res Ther. 2022;156:104116.
13. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry. 2021;20(2):154–70.
14. Flygare O, Enander J, Andersson E, Ljótsson B, Ivanov VZ, Mataix-Cols D, et al. Predictors of remission from body dysmorphic disorder after internet-delivered cognitive behavior therapy: a machine learning approach. BMC Psychiatry. 2020;20(1):247.
15. Symons M, Feeney GFX, Gallagher MR, Young RMcD, Connor JP. Machine learning vs addiction therapists: A pilot study predicting alcohol dependence treatment outcome from patient data in behavior therapy with adjunctive medication. Journal of Substance Abuse Treatment. 2019;99:156–62.
16. Ojala O, Hesser H, Gratz KL, Tull MT, Hedman-Lagerlöf E, Sahlin H, et al. Moderators and predictors of treatment outcome following adjunctive internet-delivered emotion regulation therapy relative to treatment as usual alone for adolescents with nonsuicidal self-injury disorder: Randomized controlled trial. JCPP Advances. 2024;4(3):e12243.
17. Bjureberg J, Ojala O, Hesser H, Häbel H, Sahlin H, Gratz KL, et al. Effect of Internet-Delivered Emotion Regulation Individual Therapy for Adolescents With Nonsuicidal Self-Injury Disorder: A Randomized Clinical Trial. JAMA Netw Open. 2023;6(7): e2322069.
18. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015;162(1):W1-73.

19. Shaffer D, Gould MS, Brasic J, Ambrosini P, Fisher P, Bird H, et al. A Children's Global Assessment Scale (CGAS). Arch Gen Psychiatry. 1983;40(11):1228–31.
20. Gratz KL. Measurement of Deliberate Self-Harm: Preliminary Data on the Deliberate Self-Harm Inventory. J Psychopathol Behav Assess. 2001;23(4):253–63.
21. Gratz KL, Latzman RD, Young J, Heiden LJ, Damon J, Hight T, et al. Deliberate self-harm among underserved adolescents: The moderating roles of gender, race, and school-level and association with borderline personality features. Personal Disord Theory Res Treat. 2012Jan;3(1):39–54.
22. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32.
23. Van Breda W, Bremer V, Becker D, Hoogendoorn M, Funk B, Ruwaard J, et al. Predicting therapy success for treatment as usual and blended treatment in the domain of depression. Internet Interv. 2018;12:100–4.
24. Wallert J, Boberg J, Kaldo V, Mataix-Cols D, Flygare O, Crowley JJ, et al. Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data. Transl Psychiatry. 2022;12(1):357.
25. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013. p. 600.
26. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. 2020. Available from: https://www.tidymodels.org.
27. Kleinke K. Multiple Imputation Under Violated Distributional Assumptions: A Systematic Evaluation of the Assumed Robustness of Predictive Mean Matching. Journal of Educational and Behavioral Statistics. 2017;42(4):371–404.
28. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804.
29. Wright MN, Ziegler A. ranger : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. J Stat Soft. 2017;77(1). Available from: http://www.jstatsoft.org/v77/i01/. Cited 2023 Jul 19.
30. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34(21):3711–8.
31. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340–7.
32. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput. 1998Oct 1;10(7):1895–923.
33. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/.
34. Eisenberg JM, Hershey JC. Derived Thresholds: Determining the Diagnostic Probabilities at Which Clinicians Initiate Testing and Treatment. Med Decis Making. 1983;3(2):155–68.
35. Forsell E, Jernelöv S, Blom K, Kraepelien M, Svanborg C, Andersson G, et al. Proof of Concept for an Adaptive Treatment Strategy to Prevent Failures in Internet-Delivered CBT: A Single-Blind Randomized Clinical Trial With Insomnia Patients. Am J Psychiatry. 2019;appiajp201818060699.
36. Sandell R. Our Varying Ability to Predict the Outcomes of Psychotherapy. Psychother Psychosom. 1988;50(3):134–40.
37. Halford GS, Baker R, McCredden JE, Bain JD. How Many Variables Can Humans Process? Psychol Sci. 2005;16(1):70–6.
38. Adrian M, McCauley E, Berk MS, Asarnow JR, Korslund K, Avina C, et al. Predictors and moderators of recurring self-harm in adolescents participating in a comparative treatment trial of psychological interventions. J Child Psychol Psychiatry. 2019;60(10):1123–32.
39. Biskin RS, Paris J, Zelkowitz P, Mills D, Laporte L, Heath N. Nonsuicidal Self-Injury in Early Adolescence as a Predictor of Borderline Personality Disorder in Early Adulthood. J Pers Disord. 2021;35(5):764–75.
40. Gratz KL, Dixon-Gordon KL, Tull MT. Predictors of treatment response to an adjunctive emotion regulation group therapy for deliberate self-harm among women with borderline personality disorder. Personal Disord. 2014;5(1):97–107.
41. Sahlin H, Bjureberg J, Gratz KL, Tull MT, Hedman-Lagerlöf E, Bjärehed J, et al. Predictors of improvement in an open-trial multisite evaluation of emotion regulation group therapy. Cogn Behav Ther. 2019;48(4):322–36.
42. Flygare O, Ojala O, Pontén M, Klintwall L, Karemyr M, Sjöblom K, et al. Subgroups of emotion dysregulation in youth with nonsuicidal self-injury: latent profile analysis of a randomized controlled trial. Cogn Behav Ther. 2024;25:1–15.
43. Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, Van Calster B, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. BMJ. 2024;22:e074821.
44. Chandler C, Foltz PW, Cohen AS, Holmlund TB, Cheng J, Bernstein JC, et al. Machine learning for ambulatory applications of neuropsychological testing. Intelligence-Based Medicine. 2020;1(1–2):100006.
45. Koenig J, Thayer JF, Kaess M. A meta-analysis on pain sensitivity in self-injury. Psychol Med. 2016;46(8):1597–612.
46. Jacobucci R, Grimm KJ. Machine Learning and Psychological Research: The Unexplored Effect of Measurement. Perspect Psychol Sci. 2020;15(3):809–16.

Publisher's Note