Child Psychiatry & Human Development
https://doi.org/10.1007/s10578-025-01873-9

RESEARCH

Using Machine Learning to Predict Treatment Outcome in a Concatenated Dataset of Youth Anxiety Treatments

Lesley A. Norris[1] · Marija Stanojevic[2] · Laura C. Skriner[3] · Brian C. Chu[4] · Marianne Aalberg[5] · Wendy K. Silverman[6] · Denise Bodden[7] · John C. Piacentini[8] · Zoran Obradovic[9] · Philip C. Kendall[9]

Accepted: 11 June 2025

Abstract

Machine Learning (ML) is a promising approach for predicting outcomes of youth anxiety treatments. To this end, data from nine randomized controlled trials of youth anxiety treatments were concatenated into a dataset (N=1362; M_{age}=10.59, SD_{age}=2.47; 48.9% female; 71.9% White, 5.9% Black, Other, 5.9%; 10.8% Hispanic) and ML algorithms were used to predict outcomes. Models were then applied on an external validation sample in a research clinic (N=50; M_{age}=12.04, SD_{age}=3.22; 56% female; 76% Caucasian, 10% Black, 6% Asian, 2% Other; 6% Hispanic). To examine predictive features by treatment type, Lasso Regression models were built separately for youth who completed individual cognitive behavioral therapy (CBT), family CBT (FCBT), sertraline alone (SRT), and combination of SRT and CBT (COMB). Automatic relevance determination (ARD) emerged as the best performing model in the concatenated ($RMSE$=1.84, R^2=0.28) and external validation datasets ($RMSE$=1.87, R^2=0.11). Predictive features of poorer outcomes were primarily indicators of symptom severity and trial effects, although predictors varied within treatments (e.g., caregiver psychopathology was predictive for FCBT; depressive symptoms were predictive for COMB). Implications for use of ML to predict outcomes are discussed.

Keywords Machine learning · Youth anxiety · Prediction · Harmonization

Anxiety disorders are one of the most common forms of child and adolescent (referred to hereafter as youth) psychopathology, with global prevalence rates rising to 20.5% following the COVID-19 pandemic [1]. These disorders are associated with significant disruptions to youth social, academic and family functioning [2–4]. Left untreated, youth anxiety disorders typically follow a chronic course [2] and confer additional risk for development of multiple long-term negative sequelae, including substance use [5], suicide attempts/ideation [6], and comorbid disorders [7].

Efficacious intervention for youth with anxiety disorders is consequently critical. Cognitive behavioral therapy (CBT), selective serotonin reuptake inhibitors (SSRIs) and their combination have been identified as efficacious treatments for youth anxiety disorders in a synthesis of the Cochrane Database of Systematic Reviews [8]. An evidence base update of 111 treatment outcome studies similarly found that CBT and CBT with medication, as well as exposure, CBT including parents, modeling, and education, met criteria for "well-established" treatments [9]. Additional "probably or possibly efficacious" treatments were

✉ Lesley A. Norris
Lesley.norris@brown.edu

[1] Department of Psychiatry and Human Behavior, Warren Alpert Medical School of Brown University, Providence, RI, USA

[2] Cambridge Cognition, Toronto, ON, Canada

[3] The Center for Stress, Anxiety, and Mood, LLC, Summit, NJ, USA

[4] Department of Clinical Psychology, Graduate School of Applied and Professional Psychology, Rutgers University, Piscataway, NJ, USA

[5] Akershus University Hospital, Lørenskog, Norway

[6] Yale University, New Haven, CT, USA

[7] Department of Clinical Child and Family Studies, Utrecht University, Utrecht, The Netherlands

[8] Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles School of Medicine, Los Angeles, CA, USA

[9] Department of Psychology, Temple University, Philadelphia, PA, USA

Published online: 03 July 2025

identified, including group therapy and family psychoeducation. Although efficacy for CBT, SSRIs, and their combination has been documented in aggregate, there remains heterogeneity in outcomes (i.e., $I^2>60\%$; [10]). On average, approximately 40% of anxious youth are classified as "non-responders" across large randomized controlled trials (RCTs,e.g., [11]). Researchers have sought to identify baseline variables that clarify a more personalized approach to care [12–14].

The identification of variables that predict outcomes is important for several reasons. First, indicators of non-response might provide insights into needed adaptations to current protocols. Second, given the difficulty associated with accessing care [15, 16] and low retention rates following treatment initiation [17], better classification at baseline of who responds to which treatments might help to efficiently leverage both limited resources and a small window for change, while buttressing against clinical decision-making biases [18]. For those families who do access and complete a full course of evidence-based treatment, gold-standard protocols are often lengthy (e.g., 16 sessions) and can be associated with financial cost if sessions are not reimbursed by insurance. These are considerable burdens for families to bear, especially for those who do not experience meaningful improvement in youth symptoms. Beyond logistical and financial concerns, adverse events have been associated with SSRI use [10, 19, 20] and anxious youth who are less responsive to CBT show an increased risk for substance use [21], suicide and decreased quality of life [22] in adulthood compared to responders. Youth may also experience a decrease in self-efficacy as a result of treatment non-response, and consequently, show a decreased willingness to engage in future, potentially more beneficial treatments [22]. Thus, treatment non-response is not without its negative consequences at an individual, family and systems level, and ideally could be accurately predicted before treatment initiation.

According to three reviews of youth anxiety treatment studies, indicators of differential treatment response have been difficult to identify. In the first review of CBT predictors [23] no baseline youth demographic (i.e., sex assigned at birth, age, ethnicity, intellectual functioning) or clinical factors (i.e., primary anxiety diagnosis, anxiety severity, symptom duration, general comorbidity, co-occurring externalizing or other internalizing disorders) consistently predicted differential outcome across a majority of studies [23]. The second review replicated this pattern of null findings,socioeconomic status and parent psychopathology were also not consistent predictors of posttreatment response across a majority of trials [24]. Findings from both reviews were in contrast with results from a third review of both psychotherapy and medication treatments for youth anxiety and obsessive–compulsive disorders [25]. Results from this study suggested that baseline symptom severity and family dysfunction were potential predictors of poorer outcome, which was consistent with a later review of outcomes for anxiety and depressive disorders that found parental psychopathology predicted worse CBT outcomes for anxious youth [26]. However, no predictor variables emerged consistently across the three reviews, and null findings were the norm. A similar pattern has been reported in the moderator literature. Primary diagnosis has sometimes been found to moderate outcomes, with some indication that youth with social anxiety disorder (SoP) may respond better to treatments involving medication and that social anxiety may moderate response to group versus individual CBT [23, 25, 27]. However, most reviews have found that demographic variables (i.e., age, sex assigned at birth, race, ethnicity, socioeconomic status), pretreatment youth characteristics (i.e., global anxiety severity, primary diagnosis, comorbidity) and pretreatment parent variables (i.e., global psychopathology, anxiety) do not consistently moderate response (Norris et al., 2021).

Machine Learning (ML) is an analytic tool that may address the gaps in traditional approaches to prediction [28–30]. ML is broadly defined as a branch of artificial intelligence that is able to model the system and predict outcomes from data without explicit programming [31]. ML's focus on predictive fit rather than explanatory inference represents a more flexible analytic technique to predict treatment outcomes in comparison to traditional approaches with explicit assumptions about the underlying relationship between variables. In addition, certain ML algorithms can better facilitate the identification of complex patterns of variables and their interpretation at the individual patient level, and thus may represent an optimal analytic strategy for use in the recent push towards person-centered treatment approaches [32]. These models can then be complemented by traditional analyses examining variables that emerge as important parameters in ML models, thus helping to inform development of theoretical models of treatment response.

Recent years have seen an increase in implementation of ML in treatment research [33, 34], although such studies have rarely been conducted using youth samples [33] and to our knowledge only two have applied ML to the identification of predictors CBT for youth anxiety [35, 36]. As ML studies have proliferated, concerns have been raised about study quality and limitations of ML more broadly [37]. First, although required sample sizes for ML analyses depend on both learning algorithm complexity and the unknown underlying function that relates model input to output, small sample sizes remain the norm across studies,one review found that only fourteen studies of ML in psychotherapy studies included samples > 200 (Aafjes-van Doom et al., 2020.

Second, model performance is rarely examined in additional samples, despite widespread recognition of the importance of external validation in model development [38–40]. Third, concerns have been raised about whether ML truly outperforms traditional approaches, although several studies have shown that ML outperformed standard regression (e.g., [41–43]). These and other weaknesses in ML studies have been organized into a "TREE concerns" framework (Transparency, Reproducibility, Ethics and Effectiveness), which has been used to frame updated reporting guidelines for ML studies [44]. Thus, although ML remains a promising technique, the application of ML to predict treatment outcomes is still in its infancy.

The present exploratory study built ML models to predict outcome using a concatenated dataset of nine RCTs of youth anxiety treatments (N=1362; Phase 1). Treatment conditions across studies included various CBT modalities [individual (CBT), family (FCBT) and group (GCBT)], along with medication conditions [sertraline (SRT) and combination of SRT and CBT (COMB)] and inclusion of additional individual parent components to CBT protocols [cognitive parent training (CPT) and CBT involving parents (CBT/P)]. Features utilized from the concatenated dataset included: (1) demographics (age, race, ethnicity, caregiver education), (2) ADIS composite CSRs for all assessed youth diagnoses (N=24), (3) all available CBCL subscale T-scores (N=24) and (4) ADIS-IV-L composite severity scores for caregiver diagnoses (N=16 per parent) available within a subset of the data. To facilitate a preliminary examination of prediction differences within treatment conditions, models were trained, validated, and tested separately for each active treatment condition available in a subset ≥ 10% of the dataset using an algorithm that allowed for examination of predictive features (i.e., independent variables). In Phase 2, the models were examined in a separate sample of youth (N=50) who completed CBT in the Child and Adolescent Anxiety Disorders Clinic (CAADC) to assess model external validity outside of a clinical trial context.

Methods

Participants

Participants were 1362 youth with a primary anxiety disorder ages 6–17 years (M=10.59 years, SD=2.47; 48.9% female; 71.9% White, 10.8% Hispanic; 5.9% Black, Other, 5.9%;) and their caregivers who enrolled in one of the nine RCTs included in the concatenated dataset. Primary anxiety disorder was defined as meeting diagnostic criteria for an anxiety disorder per a semi-structured diagnostic interview. Although presence of co-occurring conditions was broadly not a rule out in included trials (Table 1; see Table 5 for severity scores on a range of co-occurring conditions), anxiety had to be the primary presenting concern to be categorized as a primary anxiety disorder. Participants in the external validation dataset were 50 youth with a primary anxiety disorder ages 7–17 (M=12.04, SD=3.22; 56% female; 76% Caucasian, 10% Black, 6% Asian, 2% Other; 6% Hispanic; note: labels were drawn directly from the clinic's demographics questionnaire) and their caregivers who completed in-person CBT treatment at the CAADC. More than half of caregivers in both datasets had completed some college training or more.

Procedure

Phase 1

Inclusion/exclusion criteria for studies included in the concatenated dataset (Table 1) was selected to mirror the criteria used in the most recent Cochrane review of CBT for youth anxiety (see [45]), with the following exceptions: (a) "types of studies" criteria were updated to specify that direct contact with the child must involve in-person (not internet-based) intervention, to exclude prevention/early intervention or school administrator-administered interventions, and to exclude preliminary/pilot investigations, (b) "participant characteristics" criteria were updated to restrict the age range between 6–18, (c) "diagnosis" and "comorbidity" criteria were updated to specify that participants must meet criteria for a primary anxiety disorder via semi-structured diagnostic assessment, not just an anxiety disorder broadly [e.g., youth presenting with autism spectrum disorders (ASD) would not be considered to present with a primary anxiety disorder], (d) "experimental intervention" criteria were updated to allow for concurrent medications for the treatment of anxiety administered naturalistically and (e) PIs agreed to provide raw data and data was provided within the timeframe for the proposed study.

Datasets from nine RCTs [46, 47], Kendall, 1997; [11, 48–52] that met study inclusion criteria were collected from study principal investigators (PIs) and the PI of an integrative data analysis of youth anxiety treatment trials [53]. Measures available across trials and in line with categories of variables examined in previous predictor studies [23–25] were concatenated into a single dataset. Brief details of the methodology for each RCT available for use in the concatenated dataset are presented in Table 2. Of note, medication use was not a consistent exclusion criterion for CBT conditions across trials, although it was often required that clients were on a "stable dose" of medication prior to enrollment and that clients maintained that dosing throughout treatment. All trials received Institutional Review Board (IRB)

Table 1 Study inclusion/exclusion criteria

Types of studies	
	RCT (including cross-over trials and cluster-randomized trials)
	Manual-based and documented modular CBT
	CBT at least 9 sessions
	Involves direct in-person* contact with the child
Types of participants	
Participant characteristics	
	Youth ages 5–18*
Diagnosis	
	Diagnostic criteria for primary anxiety disorder
	Sample does not include PTSD, SPs, SM, and OCD
Comorbidity	
	Sample does not include ASD or intellectual impairment*
Settings	
	All settings included
Intervention	
	Manual-based CBT, or modular CBT, alone or in combination with medication
	A documented, written protocol stating the specific treatment at each stage of at least nine sessions provided by trained therapists under regular supervision
	CBT had to be administered according to standard principles as a psychological model of treatment involving helping the child to (1) recognize anxious feelings and somatic reactions to anxiety, (2) clarify cognitions in anxiety-provoking situations, (3) develop coping skills that involve modification of these anxiety-provoking cognitions and (4) respond to behavioral training strategies with exposure in vivo or by imagination, usually in a gradual, hierarchical manner, and relaxation training
	CBT can be delivered individually, in a group format or with family or parental involvement. The latter spans a range of direct involvement such as (rarely) the whole family and (more usually) the parents for some conjoint or separate sessions. Family/parental CBT may include providing psycho-education for parents or even teaching parents to be co-therapists
Comparator interventions	
	Waiting list and no treatment for anxiety during that period
	Psychological treatment that did not include CBT elements, or attention only (e.g. support but with no elements of CBT)
	Treatment-As-Usual (TAU)
	Pill Placebo
Types of outcome measures	
	Primary outcome: assessed using structured interviews
	Secondary outcome: reduction in anxiety symptoms assessed with RCMAS, FSSC-R, SPAI-C, CBCL, SAS-A, STAI-C, SCARED, or SCAS

* Indicates updates to Cochrane review criteria

Table 2 Study procedures for concatenated trials

Study	Design	Ages	Sessions	*N*
[46]	CBT(n=64), FCBT(n=64), WL(n=19)	8–18	13	147
[47]	CBT(n=27), WL(n=20)	9–13	16	47
Kendall, 1997	CBT(n=60), WL(n=34)	9–13	16	94
[48]	CBT(n=55), FCBT (n=56), FESA (n=50)	7–14	16	161
[49]	CBT(n=29), CBT+CPT(n=30), WL(n=20)	7–18	12	79
[50]	CBT(n=60), CBT/P(n=59),	7–16	12–14	119
[51]	CBT(n=55), GCBT(n=55), WL(n=55)	7–13	14	165
[11]	CBT(n=139), SRT(n=133), COMB(n=140), PBO(n=76)	7–17	12	488
[52]	CBT(n=20), FCBT(n=20)	6–13	12–16	40

CBT=individual cognitive behavioral therapy; FCBT=family cognitive behavioral therapy; WL=wait-list; FESA=family-based education/support/attention (active control); CPT=cognitive parent training; CBT/P=cognitive behavioral therapy involving parents; GCBT=group cognitive behavioral therapy; SRT=sertraline; COMB=CBT with SRT; PBO=pill placebo approval across institutions, which included discussions of data sharing.

For a data quality check, an individual case for each trial was selected using a random number generator. An undergraduate volunteer checked this case against every available original trial dataset.

Phase 2

Models developed in Phase 1 were used to predict outcome in the CAADC. Archival CAADC data was used in Phase 2 so that the move to telehealth due to the COVID-19 pandemic did not confound study results. Participants who completed treatment most recently were included in the sample.

Youth and their caregivers were eligible to receive treatment at the CAADC if they (a) were between the ages of 7–17, (b) met criteria for a primary diagnosis of a DSM-5 anxiety disorder per the Anxiety Disorder Interview Schedule for DSM-5– Child and Parent Versions [54], and (c) were English-speaking and able to provide informed consent/assent. Eligibility followed multiple gating: caregivers completed a preliminary phone screen with trained study staff to determine whether youth symptoms indicated potential presence of a primary anxiety disorder and then, when caregivers endorsed elevated youth anxiety symptoms, an in-person pretreatment assessment was completed. This pretreatment assessment included (a) collection of assent/consent, (b) a semi-structured diagnostic assessment administered by reliable diagnosticians separately to caregiver and youth, and (3) completion of a battery of self-report measures (including all measures used as features in ML models). Eligible families complete sixteen sessions of CBT [*Coping Cat* [55] for children and *CAT* Project [56] for adolescents] with trained graduate student clinicians and a post-assessment (including ADIS-C/P).

Table 3 Race/ethnicity categorization across trials

Trial	Race	Ethnicity
[46]	–	–
[47]	Caucasian, Black, Asian, Hispanic, Other	–
Kendall, 1997	Caucasian, Black, Asian, Hispanic, Other	–
[48]	Caucasian, Black, Asian, Hispanic, Other	–
[49]	–	–
[50]	White, Hispanic, Black, Other	–
[51]	–	–
[11]	Black, Asian, White, Native Hawaiian/Other Pacific Islander, American Indian, Other	Non-Hispanic, Hispanic
[52]	Primary parent race assessed as African American, Asian/Pacific Islander, Caucasian, Latino, Native American, Other	–

Measures

Demographics

Youth age, sex assigned at birth, race and ethnicity and caregiver education were included as features in models. Youth age was reported in years; when more detailed child age information was available (e.g., age in months), age was rounded down to the nearest year. Youth race and ethnicity were categorized differently across trials (see Table 3) and thus had to be recoded. Datasets collected in other countries had different conventions for race/ethnicity assessment (e.g., reported country of origin, caregiver countries of origin) or did not include race/ethnicity breakdowns. Given these constraints within the concatenated dataset, race was coded sub-optimally into the following categories: White, Black, and Other. A separate category was created to indicate ethnicity based on both data provided by each team and a review of primary outcome papers. If individuals indicated "Hispanic" when asked to self-identify their race, this individual was identified as Hispanic within the ethnicity category and race was listed as missing. Missingness within the race category was not imputed, but was designated as a special class of unknown. Based on country of origin rather than imputation, ethnicity was listed as non-Hispanic for trials collected outside of the United States. When available, caregiver education was categorized as "less than high school," "high school graduate," "some college" and "graduate training."

Anxiety Disorders Interview Schedule for Children and Parents (ADIS-C/P)

The ADIS-C/P is a semi-structured diagnostic interview used as the gold-standard measure to determine whether youth meet diagnostic criteria for a range of diagnoses [anxiety, obsessive–compulsive disorder (OCD), depression, attention-deficit/hyperactivity disorder (ADHD), etc.]. Across studies, reliable independent evaluators (IEs) administered the ADIS-C/P separately to both caregiver and youth at baseline and post-treatment and assigned a clinician severity rating (CSR) on a scale of 0 to 8 for each diagnosis. The higher of the two CSRs from caregiver and youth interviews were selected to create a composite CSR; composites were either already available within RCT datasets or were calculated in Python. A CSR of four or higher indicates that the child meets DSM-IV criteria for the diagnosis, with higher CSRs indicating a more severe impact on child functioning. Subtypes of various diagnoses (e.g., ADHD inattention/hyperactive/combined) were not available across trials; when subtypes were available, the maximum value was selected [e.g., if youth met criteria for multiple specific phobias (SPs)].

Three versions of the ADIS were used in this study. Two early trials [47, 57] used the Anxiety Disorders Interview for Children (ADIS/-C/P [58],) that provided diagnoses using DSM-III-R criteria. The ADIS-C/P has demonstrated inter-rater reliability [59], retest reliability [59–61] and sensitivity to treatment effects in samples of anxious youth (e.g., [47, 57]). The remainder of the trials included in the concatenated dataset used the ADIS-IV-C/P (Silverman, 1996) to generate DSM-IV diagnoses. The ADIS-IV-C/P has demonstrated convergent and discriminant validity [62], retest and inter-rater reliability [63] and sensitivity to treatment effects for youth anxiety disorders (e.g., [64]). Because past categorizations of overanxious disorder and avoidant disorder were eliminated in updates to the DSM, as they were viewed better classified as generalized anxiety disorder (GAD) and SoP, respectively, we employed this categorization scheme. In Phase 2, the ADIS-5-C/P was used at pre- and post-treatment, which had few and only minor changes in the anxiety disorders categorizations and therefore were compared directly. Inter-rater reliability was high (youth-reported GAD $ICC=0.82$, caregiver-reported GAD $ICC=0.89$; youth-reported SoP $ICC=0.91$, caregiver-reported SoP $ICC=0.93$; youth-reported SAD $ICC=0.94$, caregiver-reported SAD $ICC=0.93$).

Child Behavior Checklist

Child Behavior Checklist (CBCL; [65]) is a 118-item caregiver report measure that asks caregivers to report on youth behavioral and emotional problems within the past two months along a scale of 0 (not true) to 2 (very/often true). Items are used to generate the following scale scores: Competence (Activities, Social, School, Total), Syndrome (Anxious/Depressed, Withdrawn/Depressed, Somatic Complaints, Social Problems, Thought Problems, Attention Problems, Rule-Breaking Behavior, Aggressive Behavior), Internalizing Problems, Externalizing Problems, Total Problems and DSM-Oriented Scales (Depressive Problems, Anxiety Problems, Somatic Problems, Attention Deficit, Oppositional Defiant Problems, Conduct Problems) and 2007 Scale Scores (Sluggish Cognitive Tempo, Obsessive–Compulsive Problems, Stress Problems). T-scores ≥ 65 on any subscale indicate potential targets for intervention. The CBCL has demonstrated reliability, stability and validity [66].

Anxiety Disorders Interview Schedule for DSM-IV, Lifetime Version

The ADIS-IV-L [67] is a semi-structured assessment of lifetime diagnoses for the caregiver available in a subset of the concatenated dataset. Consistent with the ADIS, IEs rated the severity of caregiver diagnoses along a scale of 0 to 8, with a score ≥ 4 indicating a diagnosable disorder per ADIS-IV diagnostic criteria. Lifetime diagnoses assessed in a subset of trials included social anxiety disorder, SPs, panic disorder, agoraphobia, panic disorder with agoraphobia, generalized anxiety disorder (GAD), obsessive–compulsive disorder, post-traumatic stress disorder, dysthymia, major depressive disorder, attention deficit hyperactivity disorder, substance abuse and other. When possible, the ADIS-IV-L was administered separately to both caregivers. The ADIS-IV-L has demonstrated favorable reliability estimates [68].

Data Analytic Plan

Missingness

Although measures were selected to maximize overlap across studies, unless domain knowledge suggested otherwise (i.e., inclusion of ADIS-IV-L available within a subset of the data), missingness was identified in the concatenated dataset in both explanatory and predicted variables (ranging between 0% to 77–79% per variable; 40% in the dataset). The level of missingness could impact model precision and lead to biased outcomes (e.g., [69]). To address missingness, we separately fit models to explanatory variables (features) and predicted variables in Python Version 3.8 using a variety of imputation methods, from simple single imputation techniques to multiple imputation methods utilizing diverse imputation techniques as iterative steps. To test each imputer, 10% of non-missing values were randomly selected and masked. Each imputation algorithm was then trained and tested within this dataset. Root Mean Square Error (RMSE) values were calculated to determine the distance between imputed and actual values. This process was iterated ten times and average RMSEs were calculated across each iteration, with lower RMSE values were considered indicative of a better approach to imputation (i.e., less distance between imputed and actual values). For a comparison of the various imputation metrics, which indicated that newer techniques significantly outperformed the simpler methods, and a review of the difference in performance between simple and repeated imputations see [70]. Prediction was best using soft impute, and so this approach was used in the current project.

Data Cleaning/Feature Engineering

Features utilized from the concatenated dataset included the following measures described above: (1) demographics (age, race, ethnicity, caregiver education), (2) ADIS composite CSRs for all assessed youth diagnoses, (3) all available CBCL subscale T-scores and (4) ADIS-IV-L composite severity scores for caregiver diagnoses available within a subset of the data. Measures were selected to be in line with categories of variables that have been examined as indicators of treatment response in previous studies (Norris et al., 2020). All features were normalized. Patients from the concatenated dataset of RCTs (not the external validation clinical dataset) were randomly shuffled and split into training, validation, and test sets using a split of 80:10:10 training-validation-testing. Although 70:20:10 is a common approach in other studies minimize bias [71], 80:10:10 was used to make the training dataset larger due to high missingness and small datasets with many features. Two indicators of missingness were included in the current dataset (one to indicate a measure was not collected in the trial, and another to indicate unexpected missingness). Both values were replaced with un-known values for the purpose of imputation and prediction analyses. Study site and treatment type were replaced with nine and eight, respectively, yes/no binary features. Families who withdrew from treatment were removed into a separate dataset. Non-active treatment conditions (waitlist and pill placebo) were concatenated into a single feature.

Defining Outcome

Within supervised learning in machine learning (a problem in which the outcome is labeled), predicted outcomes can be categorical (a classification problem) or continuous (a regression problem). Definitions of treatment response varied across studies, and the only posttreatment outcome measure available across all trials in the concatenated dataset was composite posttreatment ADIS CSRs. Posttreatment assessments from waitlist controls were used so that this group had not received any treatment. Outcomes were assessed as continuous CSRs across all anxiety disorders. Continuous outcomes were selected, rather than a discrete diagnostic remission variable, to ensure grained prediction. Results are presented for the main youth anxiety disorders: separation anxiety disorder (SAD), GAD and SoP. Within the concatenated dataset, at post-treatment 15% met criteria (composite CSR $\geq$4) for a diagnosis of SAD, 18% for GAD, and 30% for SoP. Within the external validation dataset, at post-treatment 12% met criteria for SAD at post-treatment, 40% for GAD, and 46% for SoP.

Learning Algorithm Selection

A series of models were trained to predict outcome via a set of supervised learning algorithms. When selecting algorithms, emphasis was placed on (1) interpretability and identification of important features (i.e., understanding why certain predictions were made) and (2) creation of a good model when the number of features is similar to the number of participants. It was challenging to select the optimal linear methods due to feature complexity and multicollinearity, high levels of missing data, and the relatively small size of the combined dataset, so given both the exploratory nature of the current study, we opted to examine multiple linear methods. With these considerations, the following algorithms were selected: (1) Bayesian Ridge Regression, (2) Linear Regression, (3) Ridge Regression (L2), (4) Elastic Net, (5) Lasso Regression (L1), (6) Orthogonal Matching Pattern (OMP), (7) Automatic Relevance Determination (ARD) and (8) K-Nearest Neighbors (KNN). Ensemble methods combine prediction of multiple models to calculate the final prediction. The following such approaches were also implemented: (9) Decision Trees, (10) Extra Trees, (11) Gradient Boosting, (12) Random Forest, and (13) AdaBoost with Elastic Net. See Table 4 for a brief explanation of each approach.

Training, Validation and Testing

The dataset was a collection of labeled examples, with features like age, sex, and anxiety symptoms as features (input) labeled by responder status (output). Broadly, ML works by finding the function that maps input to output in data that is already labeled. The function is created using algorithms, which learn a function from the available input and output using training and validation parts of the dataset. That function is then used to predict the same label for a new sample that does not have the corresponding label and compared to the true output available in the data (testing).

Within this dataset, labeled examples (i.e., datapoints with available outcome data) were randomly shuffled and divided randomly into three sets: (1) training (80% of the sample), (2) validation (10%) and (3) testing (10%). Of note, validation here refers to the second step of ML analyses, whereas the external validation set was collected separately to determine model generalizability. Explanatory and predicted variables imputed in the training data were used to build the model. RMSE and R^2 were calculated separately within the two holdout sets (validation and testing) only on true, non-imputed predicted variables to avoid underestimation due to imputation. RMSE and R^2 were then averaged across validation and testing; this average was used

Table 4 Brief definitions of algorithms used

Algorithm	Brief definition
Bayesian Ridge Regression	Regression model incorporating Bayesian principles for parameter estimation
Linear Regression	Predicting outcomes based on a linear relationship with input variables
Ridge Regression	Linear regression with L2 regularization to prevent overfitting
Elastic Net	Combines L1 and L2 regularization for variable selection and shrinkage
Lasso Regression (L1)	Linear regression with L1 regularization for variable selection
Orthogonal Matching Pursuit (OMP)	Greedy algorithm for sparse representation
Automatic Relevance Determination (ARD)	Bayesian framework for identifying relevant features
K-Nearest Neighbors (KNN)	Classification based on proximity to training examples
Decision Trees	Tree-like model for decision-making based on feature splits
Extra Trees	Ensemble method using randomized trees for increased diversity
Gradient Boosting	Sequential ensemble method that optimizes weak learners
Random Forest	Ensemble of decision trees for improved accuracy and robustness
AdaBoost Regressor	Boosting technique that adjusts weights based on errors from prior models

as a metric of model performance (i.e., distance between predicted and actual explanatory variables).

Prediction by Treatment

Following the same procedures outlined earlier, ML models were built separately for each active treatment condition available in a subset≥10% (n=136) of the dataset (CBT, FCBT, COMB). Models for SRT (n=133) were also built (n=3 participants<10%). A Lasso Regression algorithm was used so that important features in each model could be examined and compared across conditions.

Results

Descriptive Statistics

Means and standard deviations for continuous measures in the concatenated and external validation dataset are presented in Table 5.

Power

To provide insight as to whether there was any added predictive benefit to harmonizing datasets, RMSE was examined for single studies and for the fully concatenated dataset. The worst prediction model within the concatenated dataset still outperformed the best prediction model trained within single intervention studies (for further detail see [70]).

Prediction Models

Model performance (average RMSE and R^2) averaged across imputation methods is presented in Table 6. ARD yielded the smallest average RMSE value across imputation approaches ($RMSE$=1.84, R^2=0.28), indicating the most robust predictive performance; Bayesian Ridge ($RMSE$=1.85, R^2=0.27) and OMP ($RMSE$=1.85, R^2=0.27) algorithms showed similarly robust prediction. Shapley Additive exPlanations (SHAP) values explaining the contributions of features to model prediction in the concatenated dataset are presented in Fig. 1. The worst performing model was Decision Tree ($RMSE$=2.89, R^2=−0.79).

Lasso Regression by Treatment

Lasso Regression was applied separately for each treatment condition available in 10% of the dataset (CBT, FCBT and COMB). Results for SRT (n=133) are also presented. While there are constraints with explainability with Lasso Regression (L1 Regression), it was selected for the stratified analyses primarily because its ability to shed light on the importance of features. Given the high-dimensional nature of the data in the present sample, Lasso helps in identifying and retaining only the most relevant features by driving the coefficients of less important variables to zero. This not only simplifies the model but also enhances interpretability, which is crucial for an analysis of clinical predictors. In contrast, Ridge regression, while effective at managing multicollinearity, does not eliminate any features, which would require predetermined thresholds for features to report on. Unlike decision trees, which provide information regarding the most important feature, lass provides the coefficient of importance of that feature. Therefore, Lasso's capability to shrink and select features aligns better with the goals of the current stratified analyses, allowing for a focus on the key variables that drive the outcomes of interest. Cross-validation was used to tune the L1 parameter (i.e., how much L1 regularization was used). Using geometric progression,

Table 5 Means and standard deviations of continuous measures

Measure	Concatenated Data Mean *(SD)*	External Validation Mean *(SD)*
CBCL subscales		
Activities	42.39 *(13.06)*	–
Social	40.56 *(13.25)*	–
School	42.44 *(12.36)*	–
Total competence	42.11 *(10.22)*	–
Anxious/depressed	66.00 *(9.64)*	72.30 *(8.95)*
Withdrawn/depressed	62.38 *(9.90)*	64.38 *(10.49)*
Somatic complaints	63.88 *(9.49)*	62.04 *(9.33)*
Social problems	59.40 *(8.96)*	59.54 *(8.54)*
Thought problems	59.78 *(8.65)*	63.48 *(8.55)*
Attention problems	58.34 *(8.96)*	58.12 *(7.65)*
Rule-breaking behavior	53.66 *(5.94)*	54.22 *(5.44)*
Aggressive behavior	56.08 *(7.55)*	58.72 *(7.59)*
Internalizing problems	67.89 *(9.20)*	68.76 *(9.02)*
Externalizing problems	53.18 *(10.60)*	55.14 *(9.97)*
Total problems	59.79 *(13.05)*	62.52 *(8.74)*
Depressive problems	63.32 *(8.71)*	65.58 *(8.65)*
Anxiety problems	70.09 *(7.07)*	73.08 *(10.16)*
Somatic problems	63.51 *(9.72)*	61.12 *(10.24)*
Attention deficit	55.67 *(6.25)*	57.42 *(7.34)*
Oppositional defiant problems	56.69 *(6.89)*	58.18 *(7.17)*
Conduct problems	54.30 *(6.27)*	55.10 *(5.94)*
Sluggish cognitive tempo	57.30 *(7.73)*	57.36 *(6.97)*
Obsessive–compulsive problems	63.47 *(8.64)*	67.84 *(9.73)*
Stress problems	62.01 *(7.32)*	68.58 *(9.07)*
ADIS composite CSR		
SAD	2.92 *(2.68)*	1.24 *(1.95)*
SoP	3.56 *(2.70)*	3.60 *(1.95)*
GAD	3.74 *(2.56)*	4.40 *(1.36)*
SP	2.52 *(2.40)*	1.56 *(2.03)*
PD	0.21 *(1.00)*	0.20 *(1.01)*
Agoraphobia	0.13 *(0.85)*	0.30 *(1.20)*
Agoraphobia with panic	0.24 *(1.25)*	0.00 *(0.00)*
OCD	0.33 *(1.14)*	0.52 *(1.43)*
PTSD	0.13 *(0.80)*	0.08 *(0.57)*
Dysthymia	0.32 *(1.22)*	0.24 *(0.96)*
MDD	0.32 *(1.15)*	0.52 *(1.43)*
ADHD	0.82 *(1.68)*	1.46 *(2.11)*
CD	0.03 *(0.39)*	0.00 *(0.00)*
ODD	0.43 *(1.36)*	0.78 *(1.62)*
SM	0.18 *(0.88)*	0.00 *(0.00)*
Enuresis/encopresis	0.11 *(0.71)*	0.04 *(0.28)*
Sleep terrors	0.04 *(0.34)*	–
Substance abuse	0.00 *(0.00)*	0.00 *(0.00)*
Bipolar disorder	0.00 *(0.07)*	0.00 *(0.00)*
Schizophrenia	0.01 *(0.09)*	0.00 *(0.00)*
Eating disorder	0.00 *(0.00)*	0.00 *(0.00)*
MDD past	0.13 *(0.86)*	1.02 *(1.92)*
Dysthymia past	0.00 *(0.00)*	0.00 *(0.00)*
PDD	0.04 *(0.46)*	–
Tourette syndrome	0.00 (0.00)	–
ADIS-IV-L Parent*		
SoP	0.91 *(1.77)*; 0.56 *(1.33)*	–

Table 5 (continued)

Measure	Concatenated Data Mean *(SD)*	External Validation Mean *(SD)*
GAD	0.83 *(1.72)*; 0.45 *(1.35)*	–
SP	1.25 *(1.86)*; 0.59 *(1.41)*	–
PD	0.04 *(0.51)*; 0.05 *(0.59)*	–
Agoraphobia	0.14 *(0.88)*; 0.10 *(0.75)*	–
Agoraphobia with panic	0.15 *(0.81)*; 0.00 *(0.00)*	–
OCD	0.17 *(0.81)*; 0.04 *(0.36)*	–
PTSD	0.15 *(0.93)*; 0.00 *(0.00)*	–
Dysthymia	0.05 *(0.41)*; 0.15 *(0.94)*	–
MDD	0.62 *(1.73)*; 0.13 *(0.86)*	–
ADHD	0.00 *(0.00)*; 0.00 *(0.00)*	–
Substance abuse	0.02 *(0.25)*; 0.19 *(0.89)*	–
Other	0.05 *(0.59)*; 0.08 *(0.57)*	–
Posttreatment CSRs		
SAD	1.09 *(1.85)*	0.56 *(1.39)*
SoP	1.92 *(2.24)*	2.36 *(2.13)*
GAD	1.45 *(2.00)*	2.14 *(2.03)*

–- indicates measure was not collected; * caregiver 1 and 2 presented in table; CBCL = child behavior checklist; ADIS = anxiety and related disorders interview schedule; CSR = clinician severity rating; SAD = separation anxiety disorder; SoP = social anxiety disorder; SP = specific phobia; PD = panic disorder; OCD = obsessive–compulsive disorder; PTSD = post traumatic stress disorder; MDD = major depressive disorder; ADHD = attention deficit hyperactivity disorder (collapsed across subtypes); CD = conduct disorder; ODD = oppositional defiant disorder; SM = selective mutism; PDD = pervasive developmental disorders; ADIS-IV-L = anxiety disorders interview schedule for DSM-IV, lifetime version

Table 6 Model fit indices

Model	Average RMSE	SD RMSE	Average R^2	SD R^2
ARD	1.84	0.04	0.28	0.02
Bayesian Ridge	1.85	0.02	0.27	0.03
OMP	1.85	0.03	0.27	0.03
L2	1.90	0.07	0.24	0.01
AdaBoost with Elastic Net*	1.90	0.02	0.24	0.07
Bagging with Elastic Net	1.90	0.02	0.23	0.04
Elastic Net	1.90	0.02	0.23	0.03
Random Forest*	1.90	0.02	0.23	0.03
L1	1.94	0.05	0.21	0.01
Gradient Boosting*	1.95	0.03	0.19	0.03
Extra Trees*	2.07	0.07	0.09	0.12
KNN*	2.10	0.03	0.07	0.04
Lasso Lars	2.16	0.09	0.01	0.01
Decision Tree*	2.89	0.24	−0.79	0.42

* Indicates ensemble method; RMSE = root mean square error; SD = standard deviation; ARD = automatic relevance determination; Bayesian Ridge = Bayesian ridge regression; OMP = orthogonal matching pattern; L2 = ridge regression; L1 = lasso regression; KNN = k-nearest neighbors; Lars = layer-wise adaptive rate scaling

the parameter was selected between a range of 0.01–10, per convention; the optimal value for L1 that emerged was 0.30.

Predictive features are presented in Table 7–10. Features that are not included in the table can be presumed to have a coefficient of 0 and thus did not influence model output. β's can be interpreted as the slope/influence of that variable; a negative β meant that the variable influenced the post-treatment CSR down to a lower severity level. Of note, the objective of traditional approaches is to identify whether a particular β is significantly different from zero. In the context of this ML problem, the objective was to minimize the RMSE (i.e., the aggregated error of the entire model) to enhance predictive accuracy; consequently, statistical significance of β's was not provided.

CBT

The RMSE evaluated on imputed data within the CBT subset using Bagging Regression with Elastic Net was 1.91

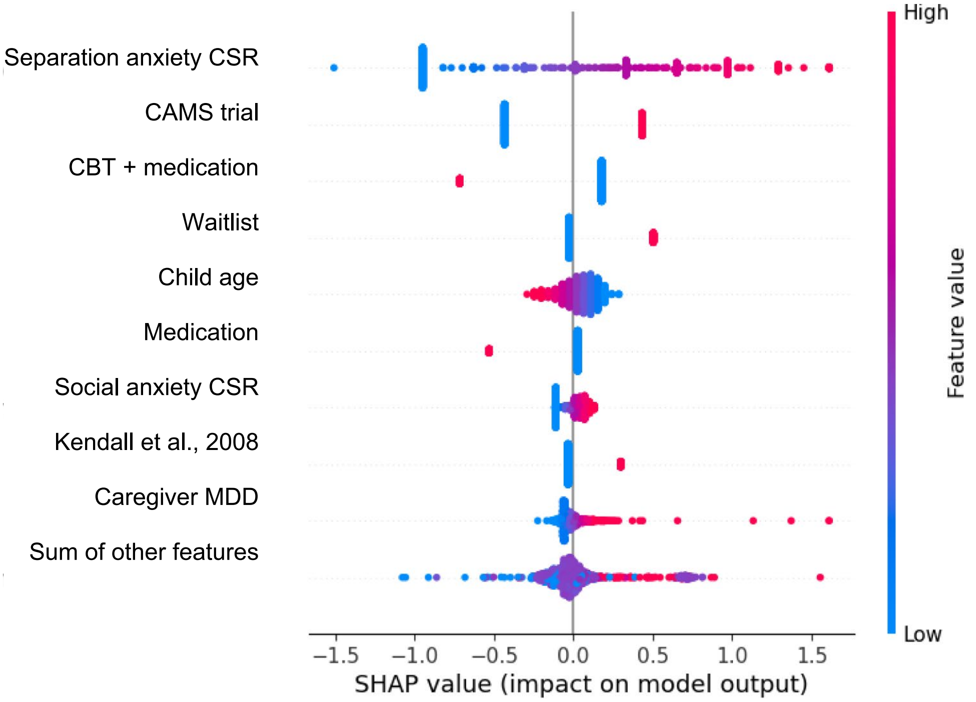


Fig. 1 Shapley additive exPlanations (SHAP) values explaining the contribution of features to model prediction in the concatenated dataset

Table 7 Lasso regression results for CBT

Outcome	Predictive features	β
Separation anxiety	SAD CSR	0.45
	[11]	0.15
	[50]	−0.06
Social anxiety	SoP CSR	0.67
	[11]	0.42
	Youth race: Black	0.01
	[46]	−0.04
	[50]	−0.33
Generalized anxiety	[11]	0.51
	Tx sessions	0.12
	GAD CSR	0.11
	[50]	−0.06

SAD = separation anxiety disorder; CSR = clinician severity rating; SoP = social anxiety disorder; Tx sessions = number of treatment sessions; GAD = generalized anxiety disorder. Citations indicate that the cited study was a predictive feature (e.g., participation in the [50] trial influenced the posttreatment CSR down to a lower severity level)

(R^2 = 0.23). Predictive features across the three outcomes examined are presented in Table 7.

FCBT

The RMSE evaluated on imputed data within the FCBT subset using the optimal performing imputer was 2.41 (R^2 = 0.05). Predictive features across the three outcomes examined are presented in Table 8.

Table 8 Lasso regression results for FCBT

Outcome	Predictive features	β
Separation anxiety	SAD CSR	0.70
	Caregiver 2 agoraphobia	0.26
	CBCL attention problems	0.06
	Panic CSR	−0.01
Social anxiety	SoP CSR	0.80
	OCD CSR	0.17
	Caregiver1 agoraphobia	0.11
	[48]	0.03
	Panic CSR	−0.00
	PTSD CSR	−0.02
	Caregiver 1 panic with agoraphobia	−0.03
	[52]	−0.07
	SP CSR	−0.12
Generalized anxiety	GAD CSR	0.32
	OCD CSR	0.14
	Caregiver 2 panic	0.11
	PTSD CSR	0.07
	[48]	0.03
	MDD CSR	0.01
	Caregiver 1 SP	0.00
	[46]	−0.04
	Caregiver 1 GAD	−0.09

SAD = separation anxiety disorder; CSR = clinician severity rating; CBCL = child behavior checklist for ages 6–18; Panic = panic disorder; SoP = social anxiety disorder; OCD = obsessive compulsive disorder; PTSD = post-traumatic stress disorder; SP = specific phobia; GAD = generalized anxiety disorder; MDD = major depressive disorder. Citations indicate that the cited study was a predictive feature

COMB

The RMSE evaluated on imputed data within the COMB subset using Bagging Regression with Elastic Net was 1.51 (R^2=−0.11). Predictive features across the three outcomes examined are presented in Table 9.

SRT

The RMSE evaluated on imputed data within the SRT subset using Bagging Regression with Elastic Net was 2.06 (R^2=0.18). Predictive features across the three outcomes examined are presented in Table 10.

External Validation

Model performance (average RMSE) averaged across the different imputation methods for the external validation set is presented in Table 11. The same three algorithms emerged as the most robust predictors: ARD (*RMSE*=1.84, R^2=0.28), Bayesian Ridge (*RMSE*=1.85, R^2=0.27) and OMP (*RMSE*=1.85, R^2=0.27), with comparable RMSE values as those observed in the concatenated dataset. Shapley Additive exPlanations (SHAP) values explaining the contributions of features to model prediction in the external validation dataset are presented in Fig. 2. The worst performing model continued to be Decision Tree (*RMSE*=2.89, R^2=−0.79).

Table 9 Lasso regression results for COMB

Outcome	Predictive features	β
Separation anxiety	SAD CSR	0.25
	Youth age	−0.04
Social anxiety	CBCL withdrawn/depressed	0.54
	SoP CSR	0.44
	SM CSR	0.13
	CBCL total problems	−0.08
Generalized anxiety	CBCL affective problems	0.15
	MDD CSR	0.08
	SoP CSR	0.06
	CBCL withdrawn/depressed	0.05
	Youth age	0.01

SAD=Separation anxiety disorder; CSR=clinician severity rating; CBCL=child behavior checklist for ages 6–18; SoP=social anxiety disorder; SM=selective mutism; MDD=major depressive disorder

Table 10 Lasso regression results for SRT

Outcome	Predictive features	β
Separation anxiety	SAD CSR	0.80
	PTSD CSR	0.08
	Youth race: other	0.03
	CBCL oppositional/defiant problems	0.01
	CBCL activities	−0.08
	CBCL sluggish cognitive tempo	−0.18
	SoP CSR	−0.21
Social anxiety	Youth age	0.19
	OCD CSR	0.16
	SoP CSR	0.16
	Youth race: Black	0.08
	CBCL somatic problems	−0.05
	CBCL externalizing	−0.05
	CBCL total	−0.31
Generalized anxiety	GAD CSR	0.52
	Youth age	0.17
	Youth race: Black	0.11
	OCD CSR	0.11
	Youth sex	0.07
	CBCL activities	−0.09
	CBCL total	−0.23

CSR=SAD=separation anxiety disorder; CSR=clinician severity rating; PTSD=post-traumatic stress disorder; CBCL=child behavior checklist for ages 6–18; SoP=social anxiety disorder; OCD=obsessive compulsive disorder; GAD=generalized anxiety disorder

Discussion

The present exploratory study applied several learning approaches to predict outcomes for anxious youth across treatment conditions (CBT, FCBT, GCBT, SRT. COMB, CPT and CBT/P) and within treatment types available within approximately 10% of the sample (CBT, FCBT, COMB and SRT). The best performing algorithms were regularized versions of linear regression and other more complex regression algorithms (i.e., ARD, Bayesian Ridge Regression). These methods help to address problems of multicollinearity and poorly distributed data, while aiding in automatic feature selection. ARD in particular emerged as the best performing algorithm, which is a type of Bayesian regression technique that models uncertainty in the weights. It provides principled estimates of uncertainty, which is critical when dealing with the uncertainty inherent in imputed data. It helps mitigate overconfidence in predictions for features with a high degree of missingness. Consistent with previous studies highlighting the utility of ML approaches in comparison to standard regression (e.g., [41–43]), Linear Regression models were not able to solve the prediction problem of the present study. Decision Trees and other ensemble methods (e.g., Extra Trees, K-Nearest Neighbors) also showed comparably lower predictive performance. The same pattern of findings was replicated in an external validation set, with comparable indicators of predictive accuracy across datasets (absolute differences in *RMSEs* ranging between 0–0.43). However, although regularized and more complex regression algorithms outperformed linear regression in the current dataset, similar important feature values emerged using ARD as those identified in previous reviews (e.g., symptom severity, demographics, family psychopathology, treatment

Table 11 Model fit indices external validation dataset

	RMSE		R^2	
Model	EV	Concatenated	EV	Concatenated
ARD	1.87	1.84	0.11	0.28
Bayesian Ridge	1.85	1.85	0.13	0.27
OMP	1.61	1.85	0.34	0.27
L2	1.98	1.90	0.01	0.24
AdaBoost with Elastic Net*	1.81	1.90	0.17	0.24
Bagging with Elastic Net	–	1.90	–	0.23
Elastic Net	1.63	1.90	0.33	0.23
Random Forest*	1.61	1.90	0.34	0.23
L1	1.77	1.94	0.21	0.21
Gradient Boosting*	1.61	1.95	0.35	0.19
Extra Trees*	1.57	2.07	0.38	0.09
KNN*	1.89	2.10	0.09	0.07
Lasso Lars	1.93	2.16	0.06	0.01
Decision Tree*	2.46	2.89	−0.53	−0.79

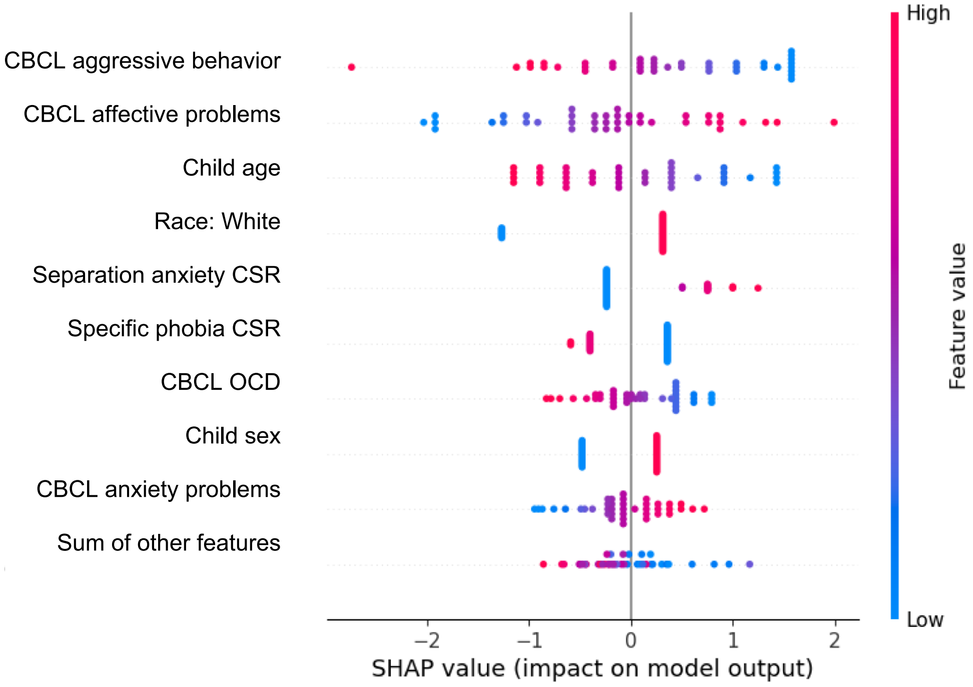
* Indicates ensemble method; RMSE = root mean square error; EV = external validation dataset.—indicates dataset too small to implement algorithm; ARD = automatic relevance determination; Bayesian Ridge = Bayesian ridge regression; OMP = orthogonal matching pattern; L2 = ridge regression; L1 = lasso regression; KNN = k-nearest neighbors; Lars = layer-wise adaptive rate scaling

condition). Thus, there may be limited added utility in application of sophisticated ML approaches towards questions of prediction. In addition, such approaches may be less intuitive and transportable to real-world clinical contexts (e.g., [72]) and have been tested less often in the clinical science literature (e.g., [73]).

When examining prediction within different treatments, numerous predictive features emerged (up to 22 predictive features), particularly for the FCBT and SRT subsets, with almost all included variables emerging as a predictive feature in one model. Consequently, a narrative approach was taken to output interpretation, with an emphasis on predictive features that emerged across outcomes and with larger absolute values of β's (i.e.,|β|≥0.05). Across all models, increased pretreatment severity of the outcome variable continued to be associated with less improvement and cross-trial differences persisted. Findings are consistent with previous reviews that have identified symptom severity as a predictor of CBT outcomes for youth with anxiety [23–25]. These findings suggest that clinicians should work to set realistic expectations around symptom change with clients presenting with increased symptom severity, normalizing that change is non-linear, validating any resulting frustration, and attending earlier and more in-depth to relapse prevention so that clients can cope effectively with ongoing symptoms at post-treatment. The therapeutic alliance may be particularly important for clients with increased symptom severity, as the alliance may motivate clients to stay connected to services even if symptoms persist after a course of outpatient care. Findings also highlight the importance of a thorough symptom assessment early in treatment, and may suggest that clients with more severe anxiety symptoms should be "stepped up" to more intensive treatments from the outset.

Specific to FCBT models, various forms of caregiver psychopathology emerged as key predictive features, although in varying directions across outcomes and caregiver

Fig. 2 SHAP values explaining the contribution of features to model prediction in the external validation dataset



diagnoses. For example, increased caregiver agoraphobia was separately associated with worse SAD and SoP outcomes, while caregiver panic disorder was associated with worse GAD outcomes. Conversely, severity of other caregiver diagnoses predicted better youth outcomes, including caregiver GAD for youth GAD outcomes. Caregiver psychopathology has not been identified as an indicator of outcomes consistently in other studies (Norris et al., 2021), although it has been shown to predict outcomes in some studies [25],findings from the current study suggest that specific forms of caregiver psychopathology, rather than psychopathology more globally, may differently influence FCBT outcomes. The mechanisms for these relationships warrant further study. For example, caregiver experience of agoraphobia may represent a barrier to attending in-person treatments, or to implementation of exposures that may parallel fears of the parent, and thus require individual treatment before beginning FCBT. Other forms of caregiver psychopathology like GAD may be better targeted within FCBT, which in combination with youth symptom improvement may lead to a positive upward cascade across the family system. Other youth comorbidities emerged as indicators of outcomes, including an association between (1) OCD and worse SoP and GAD outcomes, (2) PTSD and worse GAD outcomes and (3) SP and better SoP outcomes. Findings are inconsistent with the limited association between comorbidities and treatment outcomes found in other studies [23], but again suggest that specific comorbidities may have predictive value.

Findings from models including medication treatments (COMB and SRT) were reviewed in tandem. Although consistent patterns were observed within the COMB/SRT models as seen in other subsets (e.g., baseline symptom severity predicting worse post-treatment symptoms), several new predictive features emerged within COMB. In particular, indicators of increased youth depressive symptom severity (i.e., CBCL subscale scores and ADIS CSRs), along with SoP and selective mutism symptoms, were associated with worse response to COMB. For SoP and GAD outcomes, race emerged as a predictive feature for SRT. Specifically, youth who were categorized into the Black race category during the data harmonization process showed lower improvement, which was inconsistent with previous studies documenting minimal racial differences among anxious youth (e.g., Treadwell et al., 1995; Southam-Gerow et al., 2001; Pina, Silverman, Fuentes, et al., 2003; [25], Ginsburg et al., 2018), but consistent with findings that Black adults show poorer antidepressant response [74, 75]. This finding suggests a potential need for cultural adaptations to SRT protocols specifically for individuals who may identify as Black. Interestingly, more severe physiological symptoms and indicators of low activity predicted better response to SRT across outcomes. These findings suggest that youth who present with increased physical concerns may benefit from medication treatments specifically and are in line with previous findings suggesting that somatic symptoms may be a mechanism of change for treatments including medications [76]. Finally, inconsistent with previous reviews, youth age emerged as an important predictor of social and generalized anxiety outcomes for SRT. Specifically, older youth had lower improvement in SRT, which may indicate that older youth benefit less from a pharmacotherapy alone approach. This could be helpful information to disseminate to pediatricians, who often manage youth psychotropic medications, and may want to consider additional referrals for older youth beyond medication alone.

Study findings were considered through transparency, reproducibility, ethics and effectiveness (TREE) [44]. The dataset was fully deidentified and the PI had no access to protected health information. The concatenated dataset is available upon request, rather than through a publicly available platform (e.g., Open Science Framework),this decision was made to balance both reproducibility and ethics. A data dictionary was created to facilitate collaboration. Python code will be made available upon publication to aid in replicability efforts. Result reproducibility and external validity was examined in a research clinic setting, although it is important to note the potential for model use to exacerbate inequities given the low representation of minoritized groups within the sample.

There are limitations to consider. First, algorithm bias is an important concern. The concatenated sample and external validation sample were 71.9% and 76% White, respectively. Although race emerged as a predictive feature in some models, it is important to note that minoritized individuals were under-represented; for example, often Native American identity was not assessed entirely. Thus, although race emerged as an important feature in some models, study findings should not be considered generalizable across different racial and ethnic groups given low representation within each dataset. Second, the worst performing model within the concatenated dataset still showed better predictive performance than the best prediction models trained within single RCTs [70], highlighting the utility of developing larger, cross-site concatenated datasets,however, there are still concerns associated with merging datasets collected across different sites (e.g., Simpson's Paradox). Indeed, site emerged as a predictive feature within analyses, suggesting cross-site differences in outcomes. This suggests that there may be implementation differences in treatment approaches across trials (e.g., use of a waitlist versus active control comparator condition), comparisons of which have been described as "apples to oranges" (Freeman et al., 2018). Future studies could benefit from a more thorough coding of treatment manuals to better clarify variability within all conditions used in each trial. Third, there was some overlap in measures used across studies (i.e., demographics, ADIS composites, CBCL), but measure overlap was low, including definitions of treatment response/non-response used across trials, and missingness was high. Thus, features and output variables used in models were limited to primarily symptom and demographic measures. To make the best use out of RCT data, future efforts should adopt common cross-site batteries [77] and concatenate common factors measures (e.g., therapeutic alliance, expectations) and other variables (e.g., family accommodation) that have been shown to predict treatment response (e.g., [78]. Fourth, sample sizes within treatment modality subsets were still low,future studies could consider coding for active treatment components and merging trials with similar "treatment dosages" together. For example, treatment arms examining parent involvement in CBT could be concatenated with FCBT trials in future studies. Fifth, models were predicted of post-treatment CSRs, rather than clinical significance assessment or more broad-based measure of functioning/response. This situation was due to low overlap in post-treatment measures, and lack of clarity surrounding which diagnosis (or diagnoses were primary treatment targets. Sixth, we applied Lasso Regression to understand the importance of features within different treatments, although Bayesian variable selection approaches have outperformed lasso variable selections in other contexts [79]. Finally, the study was exploratory and atheoretical in nature, which can result in Type I errors, spurious discoveries, and reduced external validity [80]. Future work should develop theoretical justifications for testing potential prescriptive predictors in youth anxiety treatment trials (e.g., [81, 82]), which would serve to enhance finding clarity [30, 83]. Future studies should focus on further model impact evaluation and implementation within real-world clinical settings [44]

Author Contribution Author 1 wrote the main manuscript with supervision from author 10. Author 2 led study analyses with supervision from author 9. Authors 3–8 provided datasets (author 3 provided a harmonized dataset of several RCTs and authors 4–8 provided original trial data). All authors reviewed the manuscript.

Funding This work was supported by the National Institutes of Health [F31MH123038].

Data Availability No datasets were generated or analysed during the current study.

Declaration

Conflict of interest Philip C. Kendall receives royalties, and his spouse has employment, related to publications associated with the treatment of anxiety in youth. Other authors had no disclosures to report.

References

1. Racine N, McArthur BA, Cooke JE, Eirich R, Zhu J, Madigan S (2021) Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: a meta-analysis. JAMA Pediatr 175(11):1142–1150
2. Essau CA, Lewinsohn PM, Olaya B, Seeley JR (2014) Anxiety disorders in adolescents and psychosocial outcomes at age 30. J Affect Disord 163:125–132
3. Settipani CA, Kendall PC (2013) Social functioning in youth with anxiety disorders: Association with anxiety severity and outcomes from cognitive-behavioral therapy. Child Psychiatry & Hu Development 44(1):1–18
4. IEEE. Swan AJ, Kendall PC (2016) Fear and missing out: Youth anxiety and functional outcomes. Clinical Psychology: Science and Practice, 23(4): 417.
5. Lopez B, Turner RJ, Saavedra LM (2005) Anxiety and risk for substance dependence among late adolescents/young adults. J Anxiety Disord 19(3):275–294
6. Rudd MD, Joiner TE, Rumzek H (2004) Childhood diagnoses and later risk for multiple suicide attempts. Suicide and Life-Threatening Behavior 34(2):113–125
7. Cummings CM, Caporino NE, Kendall PC (2014) Comorbidity of anxiety and depression in children and adolescents: 20 years after. Psychol Bull 140(3):816
8. Manassis K, Russell K, Newton AS (2010) The Cochrane Library and the treatment of childhood and adolescent anxiety disorders: an overview of reviews. Evidence-Based Child Health: A Cochrane Review Journal 5(2):541–554
9. Higa-McMillan CK, Francis SE, Rith-Najarian L, Chorpita BF (2016) Evidence base update: 50 years of research on treatment for child and adolescent anxiety. J Clin Child Adolesc Psychol 45(2):91–113
10. Wang Z, Whiteside SP, Sim L, Farah W, Morrow AS, Alsawas M, Murad MH (2017) Comparative effectiveness and safety of cognitive behavioral therapy and pharmacotherapy for childhood anxiety disorders: a systematic review and meta-analysis. JAMA Pediatr 171(11):1049–1056
11. Walkup JT, Albano AM, Piacentini J, Birmaher B, Compton SN, Sherrill JT, Kendall PC (2008) Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. N Engl J Med 359(26):2753–2766
12. Baron RM, Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51(6):1173
13. Holmbeck GN (1997) Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: examples from the child-clinical and pediatric psychology literatures. J Consult Clin Psychol 65(4):599
14. Kraemer HC, Wilson GT, Fairburn CG, Agras WS (2002) Mediators and moderators of treatment effects in randomized clinical trials. Arch Gen Psychiatry 59(10):877–883
15. Kazdin AE, Blase SL (2011) Rebooting psychotherapy research and practice to reduce the burden of mental illness. Perspect Psychol Sci 6(1):21–37
16. Kazdin AE (2019) Annual research review: expanding mental health services through novel models of intervention delivery. J Child Psychol Psychiatry 60(4):455–472
17. Harpaz-Rotem I, Leslie D, Rosenheck RA (2004) Treatment retention among children entering a new episode of mental health care. Psychiatr Serv 55(9):1022–1028
18. Magnavita JJ, Lilienfeld SO (2016) Clinical expertise and decision making: An overview of bias in clinical practice. In Clinical decision making in mental health practice. (pp. 23–60). American Psychological Association.

19. Murphy TK, Segarra A, Storch EA, Goodman WK (2008) SSRI adverse events: how to monitor and manage. Int Rev Psychiatry 20(2):203–208
20. Murphy SE, Capitão LP, Giles SL, Cowen PJ, Stringaris A, Harmer CJ (2021) The knowns and unknowns of SSRI treatment in young people with depression and anxiety: efficacy, predictors, and mechanisms of action. The Lancet Psychiatry 8(9):824–835
21. Taylor S, Abramowitz JS, McKay D (2012) Non-adherence and non-response in the treatment of anxiety disorders. J Anxiety Disord 26(5):583–589
22. Bystritsky A (2006) Treatment-resistant anxiety disorders. Mol Psychiatry 11(9):805–814
23. Nilsen TS, Eisemann M, Kvernmo S (2013) Predictors and moderators of outcome in child and adolescent anxiety and depression: a systematic review of psychological treatment studies. Eur Child Adolesc Psychiatry 22(2):69–87
24. Knight A, McLellan L, Jones M, Hudson J (2014) Pre-treatment predictors of outcome in childhood anxiety disorders: a systematic review. Psychopathology Review 1(1):77–129
25. Compton SN, Peris TS, Almirall D, Birmaher B, Sherrill J, Kendall PC, Albano AM (2014) Predictors and moderators of treatment response in childhood anxiety disorders: results from the CAMS trial. J Consult Clin Psychol 82(2):212
26. Kunas SL, Lautenbacher LM, Lueken U, Hilbert K (2021) Psychological predictors of cognitive-behavioral therapy outcomes for anxiety and depressive disorders in children and adolescents: a systematic review and meta-analysis. J Affect Disord 278:614–626
27. Norris LA, Kendall PC (2021) Moderators of outcome for youth anxiety treatments: Current findings and future directions. J Clin Child Adolesc Psychol 50(4):450–463
28. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, Choi K (2021) The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 20(2):154–170
29. Coutanche MN Hallion LS (In press) Machine learning for clinical psychology and clinical neuroscience. In A. G. C. Wright and M. N. Hallquist (Eds.), The Cambridge Handbook of Research Methods in Clinical Psychology. Cambridge, UK: Cambridge University Press.
30. Dwyer DB, Falkai P, Koutsouleris N (2018) Machine Learning Approaches for Clinical Psychology and Psychiatry. Annu Rev Clin Psychol 14:91–118
31. Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM J Res Dev 3(3):210–229
32. Hamburg MA, Collins FS (2010) The path to personalized medicine. N Engl J Med 363(4):301–304
33. Aafjes-van Doorn K, Kamsteeg C, Bate J, Aafjes M (2020) A scoping review of machine learning in psychotherapy research. Psychother. Res. 1–25.
34. Lee Y, Ragguett RM, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, McIntyre RS (2018) Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord 241:519–532
35. Bertie LA, Quiroz JC, Berkovsky S, Arendt K, Bögels S, Coleman JR, Hudson JL (2024) Predicting remission following CBT for childhood anxiety disorders: a machine learning approach. Psychol. Med. 1–11
36. Lebowitz ER, Zilcha-Mano S, Orbach M, Shimshoni Y, Silverman WK (2021) Moderators of response to child-based and parent-based child anxiety treatment: a machine learning-based analysis. Journal of Child Psychology and Psychiatry.
37. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, de Kamps M, Beam A, Konigorski S, Lippert C (2020) Time to reality check the promises of machine learning-powered precision medicine. The Lancet Digital Health.
38. Adibi A, Sadatsafavi M, Ioannidis JP (2020) Validation and utility testing of clinical prediction models: time to change the approach. J Am Med Assoc 324(3):234–236
39. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP (2015) External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 68(1):25–34
40. van Bronswijk SC, Bruijniks SJ, Lorenzo-Luaces L, Derubeis RJ, Lemmens LH, Peeters FP, Huibers MJ (2020) Cross-trial prediction in psychotherapy: External validation of the Personalized Advantage Index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression. Psychother. Res. 1–14.
41. Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Cai T, Ebert DD, Hwang I, Li J, de Jonge P, Nierenberg AA, Petukhova MV, Rosellini AJ, Sampson NA, Schoevers RA, Wilcox MA, Zaslavsky AM (2016) Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol Psychiatry 21(10):1366–1371
42. Rosellini AJ, Dussaillant F, Zubizarreta JR, Kessler RC, Rose S (2018) Predicting posttraumatic stress disorder following a natural disaster. J Psychiatr Res 96:15–22
43. Webb CA, Cohen ZD, Beard C, Forgeard M, Peckham AD, Björgvinsson T (2019) Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. J Consult Clin Psychol 88(1):25
44. Vollmer S, Mateen BA, Bohner G, Kiraly FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KSL, Myles P, Granger D, Birse M, Branson R, Moons KGM, Collins GS, Ioannidis JPA, Holmes C, Hemingway H (2020) Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 368:l6927
45. James AC, James G, Cowdrey FA, Soler A, Choke A (2015) Cognitive behavioural therapy for anxiety disorders in children and adolescents. The Cochrane database of systematic reviews, 2015(2), CD004690. https://doi.org/10.1002/14651858.CD004690.pub4
46. Bodden DH, Bogels SM, Nauta MH, De Haan E, Ringrose J, Appelboom C, Appelboom-Geerts K (2008) Child versus family cognitive-behavioral therapy in clinically anxious youth: An efficacy and partial effectiveness study. J Am Acad Child Adolesc Psychiatry 47:1384–1394
47. Kendall PC (1994) Treating anxiety disorders in children: Results of a randomized clinical trial. J Consult Clin Psychol 62:100–110
48. Kendall PC, Hudson JL, Gosch E, Flannery-Schroeder E, Suveg C (2008) Cognitive-behavioral therapy for anxiety disordered youth: A randomized clinical trial evaluating child and family modalities. J Consult Clin Psychol 76:282–297
49. Nauta MH, Scholing A, Emmelkamp PM, Minderaa RB (2003) Cognitive-behavioral therapy for children with anxiety disorders in a clinical setting: No additional effect of a cognitive parent training. J Am Acad Child Adolesc Psychiatry 42:1270–1278
50. Silverman WK, Kurtines WM, Jaccard J, Pina AA (2009) Directionality of Change in Youth Anxiety Treatment Involving Parents: An Initial Examination. J Consult Clin Psychol 77:474–485
51. Villabø MA, Narayanan M, Compton SN, Kendall PC, Neumer SP (2018) Cognitive–behavioral therapy for youth anxiety: An effectiveness evaluation in community practice. J Consult Clin Psychol 86(9):751
52. Wood JJ, Piacentini JC, Southam-Gerow M, Chu BC, Sigman M (2006) Family cognitive behavioral therapy for child anxiety disorders. J Am Acad Child Adolesc Psychiatry 45:314–321
53. Skriner LC, Chu BC, Kaplan M, Bodden DH, Bögels SM, Kendall PC, Xie MG (2019) Trajectories and predictors of response

in youth anxiety CBT: Integrative data analysis. J Consult Clin Psychol 87(2):198

54. Albano AM, Silverman WK (in press) The Anxiety Disorders Interview Schedule for DSM-5: Child and Parent Versions.
55. Kendall PC, Hedtke KA (2006) Cognitive-Behavioral Therapy for Anxious Children: Therapist Manual, 3rd edn. Workbook Publishing, Ardmore, PA
56. Kendall PC, Choudhury M, Hudson J, Webb A (2002) The C.A.T. Project Manual for the Cognitive-Behavioral Treatment of Anxious Adolescents. Ardmore, PA: Workbook Publishing.
57. Kendall PC, Flannery-Schroeder E, Panichelli-Mindel SM, Southam-Gerow M, Henin A, Warman M (1997) Therapy for youths with anxiety disorders: A second randomized clinical trial. J Consult Clin Psychol 65:366–380
58. Silverman W (1987) Anxiety Disorders Interview for Children (ADIC). State University of New York at Albany: Graywind Publications.
59. Silverman WK, Nelles WB (1988) The anxiety disorders interview schedule for children. J Am Acad Child Adolesc Psychiatry 27(6):772–778
60. Silverman WK, Eisen AR (1992) Age differences in the reliability of parent and child reports of child anxious symptomatology using a structured interview. J Am Acad Child Adolesc Psychiatry 31(1):117–124. https://doi.org/10.1097/00004583-199201000-00018
61. Silverman WK, Rabian B (1995) Test-retest reliability of the DSM-III-R childhood anxiety disorders symptoms using the Anxiety Disorders Interview Schedule for Children. J Anxiety Disord 9(2):139–150. https://doi.org/10.1016/0887-6185(94)00032-8
62. Wood JJ, Piacentini JC, Bergman RL, McCracken J, Barrios V (2002) Concurrent validity of the anxiety disorders section of the Anxiety Disorders Interview Schedule for DSM-IV: Child and Parent Versions. J Clin Child Adolesc Psychol 31:335–342. https://doi.org/10.1207/S15374424JCCP3103_05
63. Silverman WK, Saavedra LM, Pina AA (2001) Test-retest reliability of anxiety symptoms and diagnoses with the Anxiety Disorders Interview Schedule for DSM-IV: Child and Parent Versions. J Am Acad Child Adolesc Psychiatry 40:937–944. https://doi.org/10.1097/00004583-200108000-00016
64. Silverman WK, Kurtines WM, Ginsburg GS, Weems CF, Rabian B, Serafini LT (1999) Contingency management, self-control, and education support in the treatment of childhood phobic disorders: A randomized clinical trial. J Consult Clin Psychol 67:675–687. https://doi.org/10.1037/0022-006X.67.5.675
65. Achenbach TM (1991) Manual for the Child Behavior Checklist/4–18 and 1991 profile. University of Vermont, Department of Psychiatry.
66. Nakamura BJ, Ebesutani C, Bernstein A, Chorpita BF (2009) A psychometric analysis of the child behavior checklist DSM-oriented scales. J Psychopathol Behav Assess 31(3):178–189
67. Brown TA, Barlow DH, DiNardo PA (1994) Anxiety disorders interview schedule adult version: Client interview schedule. Graywind Publications Incorporated.
68. DiNardo P, Brown T, Lawton J, Barlow D (1997) The Anxiety Disorders Interview Schedule for DSM–IV Lifetime version: Description and initial reliability. Paper presented at the Association for Advancement of Behavior Therapy convention, Washington, DC
69. Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E, Lix LM (2019) Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health Qual Life Outcomes 17(1):1–9
70. Stanojevic M, Norris LA, Kendall PC, Obradovic Z (2022, December) Predicting anxiety treatment outcomes with machine learning. In 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 957–962).
71. Hastie T, Tibshirani R, Friedman J, Hastie T, Tibshirani R, Friedman J (2009) Unsupervised learning. The elements of statistical learning: Data mining, inference, and prediction 485–585.
72. Webb CA, Cohen ZD, Beard C, Forgeard M, Peckham AD, Björgvinsson T (2020) Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. J Consult Clin Psychol 88(1):25
73. Zainal NH, Newman MG (2024) Which client with generalized anxiety disorder benefits from a mindfulness ecological momentary intervention versus a self-monitoring app? Developing a multivariable machine learning predictive model. J Anxiety Disord 102:102825
74. Lesser IM, Myers HF, Lin KM, Bingham Mira C, Joseph NT, Olmos NT, Poland RE (2010) Ethnic differences in antidepressant response: a prospective multi-site clinical trial. Depress Anxiety 27(1):56–62
75. Murphy E, Hou L, Maher BS, Woldehawariat G, Kassem L, Akula N, McMahon FJ (2013) Race, genetic ancestry and response to antidepressant treatment for major depression. Neuropsychopharmacology 38(13):2598–2606
76. Hale AE, Ginsburg GS, Chan G, Kendall PC, McCracken JT, Sakolsky D, Walkup JT (2018) Mediators of treatment outcomes for anxious children and adolescents: The role of somatic symptoms. J Clin Child Adolesc Psychol 47(1):94–104
77. Creswell C, Nauta MH, Hudson JL, March S, Reardon T, Arendt K, Kendall PC (2021) Research Review: Recommendations for reporting on treatment trials for child and adolescent anxiety disorders–an international consensus statement. J Child Psychol Psychiatry 62(3):255–269
78. Cuijpers P, Reijnders M, Huibers MJ (2019) The role of common factors in psychotherapy outcomes. Annu Rev Clin Psychol 15(1):207–231
79. Bainter SA, McCauley TG, Fahmy MM et al (2023) Comparing Bayesian Variable Selection to Lasso Approaches for Applications in Psychology. Psychometrika 88:1032–1055. https://doi.org/10.1007/s11336-023-09914-9
80. Elhai JD, Montag C (2020) The compatibility of theoretical frameworks with machine learning analyses in psychological research. Curr Opin Psychol 36:83–88
81. Cheavens JS, Strunk DR, Lazarus SA, Goldstein LA (2012) The compensation and capitalization models: A test of two approaches to individualizing the treatment of depression. Behav Res Ther 50(11):699–706
82. Cohen ZD, DeRubeis RJ (2018) Treatment selection in depression. Annu Rev Clin Psychol 14(1):209–236
83. Dwyer D, Krishnadas R (2022) Five points to consider when reading a translational machine-learning paper. Br J Psychiatry 220(4):169–171

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.