USING MACHINE LEARNING TO PREDICT TREATMENT OUTCOME FOR ANXIOUS YOUTH

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

By
Lesley A. Norris
August 2023

Examining Committee Members:

Philip C. Kendall, Ph.D., ABPP Advisory Chair, Department of Psychology, Temple University
Thomas M. Olino, Ph.D., Department of Psychology, Temple University
Richard Heimberg, PhD, Department of Psychology, Temple University
Lauren B. Alloy. Ph.D., Department of Psychology, Temple University
Johanna M. Jarcho, PhD, Department of Psychology, Temple University
Elizabeth Gosch, PhD, ABPP, Department of Clinical Psychology, Philadelphia College of Osteopathic Medicine

ABSTRACT

Background: Efficacious treatments for youth anxiety disorders exist, but there is considerable heterogeneity in outcome. Predictors and moderators of differential treatment response have been difficult to identify. Machine Learning (ML) is a promising approach. **Methods:** Data from nine randomized controlled trials (RCTs) of youth anxiety treatments were harmonized into a dataset (N = 1362; M_{age} = 10.59, SD_{age} = 2.47; 48.9% female; 71.9% White, 5.9% Black, 1.0% Asian; 10.8% Hispanic) and supervised ML algorithms predicted treatment outcomes. ML models were also built separately for youth who completed individual cognitive behavioral therapy (CBT), family CBT, combination of sertraline (SRT) and CBT, and SRT alone to examine predictive features. Models were then externally validated in a research clinic providing CBT for youth anxiety (N = 50; M_{age} = 12.04, SD_{age} = 3.22; 56% female; 76% Caucasian, 10% Black, 6 % Asian, 2% Other; 6% Hispanic). **Results:** Lasso Regression emerged as the best performing model ($RMSE$ = 1.40), with comparable RMSEs when the same approach was applied within an external dataset ($RMSE$ = 1.40). Predictive features of poorer outcomes were primarily indicators of increased symptom severity, particularly youth depressive symptom severity, although predictors varied within subsamples (e.g., caregiver psychopathology was an important predictor for FCBT; increased somatic symptoms were important predictors for better response to SRT). **Discussion:** ML helped identify features of anxious youth who will respond to treatments.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Philip Kendall, for consistently supporting and challenging me to grow over the past six years. I could not have asked for better mentorship, on and off the tennis court, which I know will extend beyond my graduate studies. To my dissertation committee members, Drs. Thomas M. Olino, Richard Heimberg, Lauren B. Alloy, Johanna M. Jarcho and Elizabeth Gosch, I am so appreciative for your insightful feedback on this project and others. Your guidance throughout my time at Temple has been invaluable to my professional and personal growth. To my consultant and co-sponsor team, Marija Stanojevic and Dr. Zoran Obradovic, thank you for patience and guidance as I learned a new analytic language. To the incredible undergraduate and graduate student staff at the Child and Adolescent Anxiety Disorders Clinic, thank you for your dedicated efforts to collect this data. You showed me the joy that comes from working with colleagues who play nicely in the sandbox. To my unconditionally loving parents, husband, family and friends, I am forever grateful to you for always listening and encouraging. You made the good days better and the bad days brighter. Finally, this project would not have been possible without the extraordinary effort of families who contributed their time to each study trial. For their dedication to helping advance our understanding of treatment outcomes, thank you. May our person-centered research efforts always truly center you.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Background

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 (Racine et al., 2021). These disorders are associated with significant impairment in youth social, academic and family functioning (Essau et al., 2014; Settipani & Kendall, 2013; Swan & Kendall, 2016), along with substantial healthcare costs (Marciniak et al., 2004). Left untreated, youth anxiety disorders typically follow a chronic course (Essau et al., 2014) and confer additional risk for development of multiple long-term negative sequelae, including substance use (Lopez et al., 2005), suicide attempts/ideation (Rudd et al., 2004), comorbid disorders (Cummings et al., 2014) and additional societal costs (Marciniak et al., 2004).

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 (Manassis et al., 2010). 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 (i.e., demonstration of manualized treatment efficacy in > 2 randomized trials conducted by > 2 independent investigator teams; Higa-McMillan et al., 2016). Additional "probably or possibly efficacious" treatments were 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., $I^2 > 60\%$; Wang et al., 2017). On average, approximately 40% of anxious youth are classified as "non-responders" across large randomized controlled trials (RCTs; e.g., Walkup et al., 2008). Researchers have sought to identify baseline variables that clarify for whom treatments work broadly (i.e., predictors) and which treatments work for whom and under what conditions (i.e., moderators; Baron & Kenny, 1986; Holmbeck, 1997; Kraemer et al., 2002).

The identification of predictors and moderators 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 (Kazdin & Blase, 2011; Kazdin, 2019) and low retention rates following treatment initiation (Harpaz-Rotem et al., 2004), 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 (Magnavita & Lilienfeld, 2016). For those families who do access and complete a full course of evidence-based treatment, gold-standard protocols are often lengthy (up to 16 sessions) and can be associated with considerable financial cost, particularly when providers do not accept insurance. These are considerable burdens for families to bear, especially for those who do not experience clinically meaningful improvement in youth symptoms. Beyond logistical and financial concerns, adverse events have been associated with SSRI use (Murphy et al., 2008; Murphy et al., 2021; Wang et al., 2017) and anxious youth who are less responsive to CBT show an increased risk for substance use (Taylor et al., 2012), suicide and decreased

quality of life (Bystritsky, 2006) 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 (Bystritsky, 2006). 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, predictors and moderators of differential treatment response have been difficult to identify. The first review examined peer-reviewed studies of CBT predictors across varying CBT formats and protocols (Nilsen et al., 2013). 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 (Nilsen et al., 2013). The second review included both peer-reviewed and dissertation studies of CBT efficacy and replicated this pattern of null findings; socioeconomic status and parent psychopathology were also not consistent predictors of posttreatment response across a majority of trials (Knight et al., 2014). 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 (Compton et al., 2014). Results from this study suggested that baseline symptom severity and family dysfunction were potential predictors of poorer outcome. 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 been shown to moderate outcomes across several studies, 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 (Norris & Kendall, 2021; Nilsen et al., 2013; Compton et al., 2014). However, most reviews have found that demographic variables (i.e., age, sex assigned at birth, race, ethnicity, socioeconomic status), pre-treatment youth characteristics (i.e., global anxiety severity, primary diagnosis, comorbidity) and pre-treatment parent variables (i.e., global psychopathology, anxiety) do not consistently moderate response (Norris et al., 2021). Thus, decades later, the famous question that has driven psychotherapy research ["*What treatment, by whom, is most effective for this individual with that specific problem, and under which set of circumstances?*" (Paul, 1967, p. 111; Kiesler, 1966)] remains largely unanswered with regards to prediction and moderation.

Several reasons may underlie the difficulty in identifying predictor and moderator variables. First, there may not be meaningful differences in treatment response across individuals. To adequately assess whether this is the case, several problems hampering prediction/moderation efforts must be considered. One pervasive problem is statistical power. Results from one simulation study on major depressive disorder treatments suggested that at least 300 participants per treatment arm, and ideally 500, are required to predict outcome differences (Luedtke et al., 2019). Sample sizes for most RCTs fall well below these numbers, and recruitment of such large samples is time and resource intensive. Power concerns are particularly pronounced when examining interactions (Brookes et al., 2004), and yet it is likely that complex interactions across multiple

variables, rather than simple main effects, will be most informative in determining which treatments work best for whom (e.g., Tiemens et al., 2016). These kinds of complex interactions have been difficult to account for in more traditional analytic approaches due to imposed explanatory constraints and assumptions. Another critical question is the degree to which group-level predictor and moderator analyses are generalizable to the individual (i.e., whether moderation processes are ergodic; Molenaar, 2004; Molenaar 2013). Empirical demonstrations in other contexts indicate that aggregated estimates at the group-level are inconsistent with individual-level estimates (e.g., Borkenau & Ostendorf, 1998; Fisher et al., 2018; Hamaker et al., 2007; Molenaar et al., 2003; Molenaar, 2004; Na et al., 2010). For example, Fisher and colleagues (2018) compared intra- and inter-individual variation across six datasets and found that, while mean estimates were generally consistent across levels, variance estimates were two to four times larger at the individual level. If the spirit behind the search for predictors and moderators is to help practitioners answer the question "which treatments work for which client," then concerns about ergodicity suggest that group-level approaches may not provide a complete answer.

Machine Learning (ML) is an analytic tool that may address the gaps in traditional approaches to prediction and moderation (Chekroud et al., 2021; Coutanche & Hallion, 2019; Dwyer et al., 2018). Although a standard definition of ML has not been widely accepted, 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 (Samuel, 1959). ML's focus on predictive fit rather than explanatory inference represents a more flexible analytic technique to detect treatment predictors and moderators in comparison to

traditional approaches with imposed explanatory constraints. 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 represent an optimal analytic strategy for use in the recent push towards person-centered treatment approaches (Hamburg & Collins, 2010). 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 psychotherapy research (Aafjes-van Doorn et al., 2020; Lee et al., 2018), although such studies have rarely been conducted using youth samples (Aafjes-van Doorn et al., 2020) and only one has applied ML (the Personalized Advantage Index) to the identification of predictors and moderators of CBT for youth anxiety (Lebowitz et al., 2021). As ML studies have proliferated, concerns have been raised about study quality and limitations of ML more broadly (Wilkinson et al., 2020). 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; for the single study to apply a machine learning approach within a sample of youth with anxiety, the sample size was $N = 124$ (Lebowitz et al., 2021) and 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 (Siontis et al., 2015; van Bronswijk et al., 2020; Adibi et al., 2020). Third, concerns have been raised about whether ML truly outperforms traditional approaches,

although several studies have shown that ML outperformed standard regression (e.g., Kessler et al., 2016; Rosellini et al., 2018; Webb et al., 2019). 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 (Vollmer et al., 2020). Thus, although ML remains a promising technique, the application of ML to predict treatment outcomes is still in its infancy.

Current Study

The present study built ML models predicting outcome using a harmonized dataset of nine RCTs of youth anxiety treatments (N = 1362; Phase 1). Treatment conditions across studies included various CBT modalities [individual (ICBT), family (FCBT) and group (GCBT)], along with medication conditions [sertraline (SRT) and combination of SRT and ICBT (COMB)] and inclusion of additional parent components to ICBT protocols [cognitive parent training (CPT) and CBT involving parents (CBT/P)]. To facilitate a preliminary examination of moderation, models were built separately for each active treatment condition available in $\geq$ 10% of the dataset using an algorithm that allowed for examination of predictive features. In Phase 2, the models were externally validated in a separate sample of youth (N = 50) who completed ICBT in the Child and Adolescent Anxiety Disorders Clinic (CAADC).

Specific Aims

Primary Aim 1

The first aim was to train and test ML models of treatment outcome in a harmonized dataset. Although there is limited consistent domain knowledge to inform

feature selection, studies have pointed to the following categories of variables as potential predictors/moderators of outcome: (1) demographics, (2) diagnosis, (3) anxiety severity, (4) behavioral problems and (5) caregiver psychopathology (Compton et al., 2014; Knight et al., 2014; Nilsen et al., 2013; Norris et al., 2021). To reflect these categories and based on measure availability across datasets, the following variables were included as model features: (1) demographics (age, race, ethnicity), (2) Anxiety Disorders Interview Schedule (ADIS) composite severity scores for all assessed youth diagnoses, (3) Child Behavior Checklist (CBCL) subscale T-scores and (4) Anxiety Disorders Interview Schedule for DSM-IV, lifetime version, (ADIS-IV-L) composite severity scores for caregiver diagnoses available within a subset of the data. In a preliminary examination of moderation, models were then built separately for each active treatment condition available in ≥ 10% (n = 136) of the dataset (ICBT, FCBT, COMB; SRT was available in n = 133 and was also examined) using an algorithm that facilitated comparison of predictive features across conditions.

Primary Aim 2

The second aim was to test the external validity of ML models with a new sample of anxious youth (N = 50) who received ICBT in the Child and Adolescent Anxiety Disorders Clinic (CAADC), an outpatient clinic at Temple University. The accuracy of the model trained on new CAADC data was assessed using root mean square error (RMSE).

Hypotheses

Expectations were that older age, female sex assigned at birth, severe social anxiety, cooccurring externalizing disorders, and elevated caregiver psychopathology

would be model features predictive of poorer outcome. However, because ML methods take a data (not hypothesis) driven approach and have not been applied in this context, there may be more complex, non-linear patterns among features not yet identified, and some algorithms tested may be too complex for interpretation. Second, it was expected that model performance would drop when tested in the external validation sample.

CHAPTER 2

METHODS

Participants

Participants were 1362 youth with a primary anxiety disorder ages 6-17 (*M* = 10.59, *SD* = 2.47; 48.9% female; 71.9% White, 5.9% Black, 1.0% Asian; 10.8% Hispanic) and their caregivers who enrolled in one of the nine RCTs included in the harmonized dataset. 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) and their caregivers who completed in-person ICBT 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 harmonized dataset (Table 1) was selected to mirror the criteria used in the most recent Cochrane review of CBT for youth anxiety (see James et al., 2015), 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 (e.g., ADIS, KSADS), 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. These updates were made because telehealth, prevention, school-based and pilot studies (update a) and anxiety treatments for youth ages 4-19 and with other primary disorders (updates b c) were outside of the scope of the current project and to increase generalizability of the dataset (update d) and feasibility of project completion (update e).

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

Note. * indicates updates to Cochrane review criteria

Datasets from nine RCTs (Bodden et al., 2008; Kendall, 1994; Kendall, 1997; Kendall et al., 2008; Nauta et al., 2003; Silverman et al., 2009; Villabo et al., 2018;

Walkup et al., 2008; Wood et al., 2006) that met study inclusion criteria were collected from study principal investigators (PIs) and harmonized into a single harmonized dataset. Brief details of the methodology for each RCT available for use in the harmonized dataset are presented in Table 2. All trials received Institutional Review Board (IRB) approval across institutions, which included discussions of data sharing.

Table 2. Study procedures for harmonized trials

Study	Design	Ages	Sessions	*N*
Bodden et al., 2008	ICBT(*n*=64), FCBT(*n*=64), WL(*n*=19)	8-18	13	147
Kendall, 1994	ICBT(*n*=27), WL(*n*=20)	9-13	16	47
Kendall, 1997	ICBT(*n*=60), WL(*n*=34)	9-13	16	94
Kendall et al., 2008	ICBT(*n*=55), FCBT (*n*=56), FESA (*n*=50)	7-14	16	161
Nauta et al., 2003	ICBT(*n*=29), ICBT + CPT(*n*=30), WL(*n*=20)	7-18	12	79
Silverman et al., 2009	ICBT(*n*=60), CBT/P(*n*=59),	7-16	12-14	119
Villabo et al., 2018	ICBT(*n*=55), GCBT(*n*=55), WL(*n*=55)	7-13	14	165
Walkup et al., 2008	ICBT(*n*=139), SRT(*n*=133), COMB(*n*=140), PBO(*n*=76)	7-17	12	488
Wood et al., 2006	ICBT(*n*=20), FCBT(*n*=20)	6-13	12-16	40

Note. ICBT = individual cognitive behavioral therapy; FCBT = family cognitive behavioral therapy; WL = waitlist; 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 = ICBT with SRT; PBO = pill placebo.

A project codebook was generated to facilitate future use of the dataset by other investigator teams, as a secondary aim of the current study was to develop a deidentified, centralized dataset of youth anxiety treatment data (in line with the NIMH RDoC db and NDAR dataset) that can facilitate continued cross-site collaborations in the identification of predictor and moderator variables. A section was included titled "Data Decision Transparency," which detailed decisions made during the dataset harmonization process. For example, in situations where trial data included apparent errors (e.g., caregiver age listed as 5 years), the error was removed and listed as missing. To further facilitate future use of the harmonized dataset, another section was included titled "Measures Overlap" to detail all available measures across trials.

For a quality assurance check of the harmonized dataset, 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 Child and Adolescent Anxiety Disorders Clinic (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 (ADIS-5-C/P; Albano & Silverman, in press), 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 ICBT [*Coping Cat* (Kendall & Hedtke, 2006) for children and *C.A.T.* Project (Kendall, Choudhury, Hudson, & Webb, 2002) for adolescents] with trained graduate student clinicians and a post-assessment (including ADIS-C/P). All procedures were approved by Temple University's IRB.

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). Datasets collected in other countries with different conventions for race/ethnicity assessment reported country of origin (Bodden et al., 2008), caregiver countries of origin (Villabo et al., 2018) or did not include race/ethnicity breakdowns (Nauta et al., 2003). Within the harmonized dataset, race was coded into the following categories: White, Black, Asian, and Other. A separate category was created to indicate

ethnicity on the basis of provided data and 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.

Table 3. Race/ethnicity categorization across trials

Trial	Race	Ethnicity
Bodden et al., 2008	-	-
Kendall, 1994	Caucasian, Black, Asian, Hispanic, Other	-
Kendall, 1997	Caucasian, Black, Asian, Hispanic, Other	-
Kendall et al., 2008	Caucasian, Black, Asian, Hispanic, Other	-
Nauta et al., 2003	-	-
Silverman et al., 2009	White, Hispanic, Black, Other	-
Villabo et al., 2018	-	-
Walkup et al., 2008	Black, Asian, White, Native Hawaiian/Other Pacific Islander, American Indian, Other	Non-Hispanic, Hispanic
Wood et al., 2006	Primary parent race assessed as African American, Asian/Pacific Islander, Caucasian, Latino, Native American, Other	-

When available, caregiver education was categorized as "less than high school," "high school graduate," "some college" and "graduate training."

Anxiety Disorder Interview Schedule

The ADIS 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 by selecting the maximum CSR across caregiver and youth reports of the same diagnosis. 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. A CSR $\geq$ 4 indicates a diagnosable disorder. Subtypes of various diagnoses (e.g., ADHD inattention/hyperactive/combined and SP type) were not available across trials; when subtypes were available, the maximum value was selected (e.g., if youth met criteria for multiple SPs).

Three versions of the ADIS were used in this study. Two early trials (Kendall, 1994; Kendall et al., 1997) used the Anxiety Disorders Interview for Children (ADIC; Silverman, 1987) that provided diagnoses using DSM-III-R criteria. The ADIC has demonstrated inter-rater reliability (Silverman & Nelles, 1988), retest reliability (κ 0.76; Silverman & Nelles, 1988) and sensitivity to treatment effects in samples of anxious youth (e.g., Kendall, 1994; Kendall et al., 1997). The remainder of the trials included in the harmonized 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 (Wood et al., 2002), retest and inter-rater reliability (κ's 0.80-0.92; Silverman et al., 2001) and sensitivity to treatment effects for youth anxiety disorders (e.g., Silverman et al., 1999). A DSM-III-R diagnosis of overanxious disorder and avoidant disorder were categorized as generalized anxiety disorder (GAD) and SoP, respectively. In Phase 2, the ADIS-5-C/P was used at pre- and post-treatment. 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

The Child Behavior Checklist (CBCL; Achenbach, 1991) 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 $\geq$ 65 on any subscale indicate potential targets for intervention. The CBCL has demonstrated reliability, stability and validity (Nakamura et al., 2009).

Anxiety Disorder Interview Schedule for DSM-IV, Lifetime Version

The ADIS-IV-L (Brown et al., 1994) is a semi-structured assessment of caregiver lifetime diagnoses available in a subset of the harmonized dataset. Consistent with the ADIS, IEs rated the severity of caregiver diagnoses along a scale of 0 to 8, with a score $\geq$ 4 indicating a diagnosable disorder per ADIS-IV diagnostic criteria. Lifetime diagnoses assessed across trials included social anxiety disorder, specific phobias (SP), 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 (DiNardo et al., 1997).

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 harmonized dataset (Table 4). The level of missingness could impact model precision and lead to biased outcomes (e.g., Ayilara et al., 2019). Missingness has traditionally been handled through deletion of missing values [i.e., listwise deletion (LD), pairwise deletion (PD)], with 97% of educational and psychological studies relying on LD or PD (Dong & Peng, 2013), despite concerns that these approaches can produce biased and inefficient estimates (Rubin, 1987; Schafer, 1997). Imputation methods provide better parameter estimates (Dong et al., 2013) and thus were prioritized to address missingness in the current study. However, no consensus exists on best practice imputation methods for intervention research.

Methods employed to date have ranged from simple mean imputation to more complex approaches like tree-based methods (e.g., Alsaber et al., 2021), K-nearest neighbors (e.g., Petrazzini et al., 2021) and other novel approaches (e.g., Emmanuel et al., 2021; Khan & Hoque, 2020).

Table 4. Missingness summary for the harmonized dataset

Variables	% missing
Child age	0%
Sex assigned at birth	0%
Race	17%
Ethnicity	0%
Caregiver education	77%
Pre diagnosis	6-67%
Post anxiety diagnoses	6-56%
Caregiver pre diagnosis	77-79%

To address missingness, these and other imputation methods were examined. The following imputation packages were implemented in Python Version 3.8: (1) Mean, (2) Median, (3) K-Nearest Neighbors, (4) Soft Impute, (5) Singular Value Decomposition and (6) Expectation-Maximization Algorithm. The following iterative imputing with the following methods were also implemented: (7) Decision Trees, (8) Extra Trees, (9) Linear Regression, (10) Bayesian Ridge, (11) Ridge Regression, (12) Elastic Net, (13) Lasso Regression, (14) Orthogonal Matching Pattern (OMP), (15) Automatic Relevance

Determination (ARD), (16) K-Nearest Neighbors, (17) Random Forest, (18) Gradient Boosting, and (19) AdaBoost Regressor.

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. Lower RMSE values were considered indicative of a better approach to imputation (i.e., less distance between imputed and actual values). For further details see (Stanojevic et al., Under Review).

For the current project, the performance of predictive algorithms was averaged across the different imputation methods with acceptable RMSE values, rather than selecting a single imputation method. This decision was made to facilitate selection of the most robust prediction method across imputers.

Data Cleaning/Feature Engineering

Features included the following: (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. All features were normalized after random shuffling into a training, validation and test set (see Training, Validation, and Testing section for further detail) so that values ranged between 0 to 1. 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 merged 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 harmonized dataset was composite ADIS CSRs. 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 focus youth anxiety disorders: separation anxiety disorder (SAD), GAD and SoP. Within the harmonized dataset, at post-treatment 15% met criteria (composite CSR ≥ 4) for 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 and (2) creation of a good model when the number of features is similar to the number of participants. With these considerations, the following algorithms were chosen: (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 several approaches to prediction within a single predictive model. The following such approaches were also selected: (9) Decision Tree, (10) Extra Trees, (11) Gradient Boosting, (12) Random Forest, and (13) AdaBoost with Elastic Net.

Training, Validation and Testing

Consistent with convention, labeled examples were randomly shuffled and divided randomly into three sets: (1) training (70% of the sample), (2) validation (15% of the sample) and (3) testing (15% of the sample). Training, cross-validation, and testing sets were tested to ensure distribution of data was consistent across all three. Of note, the validation set was different from the external validation dataset collected through the CAADC; 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 was calculated separately within the two holdout sets (validation and testing) only on true, non-imputed data to avoid underestimation due to imputation. RMSE was then averaged across validation and testing; this average was used as a metric of model performance (i.e., distance between predicted and actual explanatory variables).

Moderation

Following the same procedures outlined earlier, ML models were built separately for each active treatment condition available in $\geq$ 10% (n = 136) of the dataset (ICBT, 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.

CHAPTER 3

RESULTS

Descriptive Statistics

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

Table 5. Means and standard deviations of continuous measures

Measure	Harmonized 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)*	---
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)*

Note. --- 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.

Power

There is currently little information on how best to conduct power analyses for ML. However, to provide some insight as to whether there was any added predictive benefit to joining datasets, RMSE was examined for single studies and for the fully harmonized dataset. The worst prediction model within the harmonized dataset still outperformed the best prediction model trained within single intervention studies (for further detail see Stanojevic et al., Under Review).

Prediction Models

Model performance (average RMSE) averaged across the different imputation methods is presented in Table 6. Lasso Ridge regression yielded the smallest average RMSE value across imputation approaches (*RMSE* = 1.40), indicating the most robust predictive performance; Bayesian Ridge (*RMSE* = 1.43) and ARD (*RMSE* = 1.44)

algorithms showed similarly robust prediction. The worst performing model was Decision Tree (*RMSE* = 2.15).

A simple linear regression algorithm was also applied, but some values were predicted to be infinity. This was because the algorithm was calculating the inverse of a matrix to create a linear regression model, akin to dividing by zero. Therefore, simple linear regression was not able to solve the prediction problem, but regularized versions of linear regression and other more complex regression algorithms (i.e., Bayesian Ridge Regression and Lasso Regression) showed low RMSE values, indicating good results.

Table 6. Regression model root mean square errors

Model	Average RMSE across imputation methods
L1	1.40
Bayesian Ridge	1.43
ARD	1.44
OMP	1.45
Random Forest*	1.45
L2	1.46
Gradient Boosting*	1.48
Bagging with Elastic Net	1.51
Elastic Net	1.51
AdaBoost with Elastic Net*	1.51
KNN	1.51
Extra Trees*	1.51

Decision Tree*	2.15

Note. * indicates ensemble method; L1 = Lasso Regression; Bayesian Ridge = Bayesian Ridge Regression; ARD = Automatic Relevance Determination; OMP = Orthogonal Matching Pattern; L2 = Ridge Regression; KNN = K-Nearest Neighbors.

Lasso Regression: Prediction

Lasso Regression results are presented separately using the optimal performing imputer (Bagging Regression with Elastic Nets) so that important features can be examined. Cross-validation was used to tune the L1 parameter (i.e., how much L1 regularization was used). Using geometric progression, the parameter was selected between a range of 0.01-1, per convention; the optimal value for L1 that emerged was 0.10. The RMSE evaluated on imputed data using Bagging Regression with Elastic Net was 1.36.

Predictive features are presented in Table 7. 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 posttreatment 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.

Table 7. Lasso regression predictive features

Outcome	Predictive features	β
Separation anxiety	SAD CSR	0.79
	WL and PBO	0.06
	Youth sex assigned at birth	0.02
	Enuresis/encopresis CSR	0.01
	GAD CSR	0.01
	Kendall (1997)	-0.02
	COMB	-0.05
	Youth age	-0.06
Social anxiety	SoP CSR	1.02
	CBCL withdrawn/dep	0.18
	WL and PBO	0.09
	SM CSR	0.06
	Walkup et al., (2008)	0.06
	OCD CSR	0.05
	Kendall (1997)	-0.01
	Bodden et al., (2008)	-0.03
	PTSD CSR	-0.04
	GCBT	-0.05
	COMB	-0.08
Generalized anxiety	GAD CSR	0.39
	Walkup et al., (2008)	0.19

Dysthymia CSR	0.15
CBCL: depressive prob	0.08
WL and PBO	0.08
OCD CSR	0.06
MDD CSR	0.04
CBCL OC prob	0.03
CBCL anx/dep	0.03
CBCL thought prob	0.03
COMB	-0.07
Bodden et al., (2008)	-0.14

Note. CSR = Clinician Severity Rating; SAD = Separation Anxiety Disorder; WL and PBO = waitlist and pill placebo; GAD = Generalized Anxiety Disorder; COMB = combination cognitive behavioral therapy and sertraline; SoP = Social Anxiety Disorder; SM = Selective Mutism; OCD = Obsessive-Compulsive Disorder; PTSD = Post-Traumatic Stress Disorder; GCBT = Group Cognitive Behavioral Therapy; CBCL = Child Behavior Checklist for Ages 6-18; prob = Problems; MDD = Major Depressive Disorder; prob = Problems; OC = Obsessive-Compulsive; anx/dep = Anxious/Depressed.

Lasso Regression: Moderation

Lasso Regression was applied using the procedures outlined above separately for each treatment condition available in 10% of the dataset (ICBT, FCBT and COMB). Results for SRT (n = 133) are also presented.

ICBT Moderation

The RMSE evaluated on imputed data within the ICBT subset using Bagging Regression with Elastic Net was 1.36. Predictive features across the three outcomes examined are presented in Table 8.

Table 8. Lasso regression results for ICBT

Outcome	Predictive features	β
Separation anxiety	SAD CSR	0.64
	Villabo et al., 2018	0.11
	Walkup et al., (2008)	0.07
	Child sex assigned at birth	0.03
	Youth race: other	0.03
	SoP CSR	0.02
	Panic CSR	0.01
	Kendall (1997)	-0.01
	Bodden et al., (2008)	-0.08
	Youth age	-0.12
Social anxiety	SoP CSR	1.00
	Walkup et al., (2008)	0.11
	CBCL social prob	0.10
	CBCL withdrawn/dep	0.10
	Youth race: Black	0.03
	CBCL anx/dep	0.02
	Youth race: other	0.02

	Villabo et al., 2018	0.02
	Caregiver 1 other dx CSR	0.01
	CBCL sluggish cog tempo	0.00
	Kendall (1994)	-0.02
	Caregiver2 SP CSR	-0.03
	Kendall (1997)	-0.03
	GAD CSR	-0.04
	Agoraphobia CSR	-0.08
	PTSD CSR	-0.08
	SAD CSR	-0.09
	Bodden et al., (2008)	-0.19
Generalized anxiety	Walkup et al., (2008)	0.29
	Dysthymia CSR	0.12
	CBCL thought prob	0.10
	CBCL: depressive prob	0.09
	GAD CSR	0.06
	CBCL anx/dep	0.06
	Caregiver 1 SP CSR	0.05
	Youth race: Black	0.03
	ODD CSR	0.02
	CBCL OC prob	0.01
	Child sex assigned at birth	0.01
	CD CSR	0.01

	OCD CSR	-0.01
	Kendall et al., (2008)	-0.03
	Villabo et al., 2018	-0.18
	Bodden et al., (2008)	-0.28

Note. CSR = Clinician Severity Rating; SAD = Separation Anxiety Disorder; SoP = Social Anxiety Disorder; Panic = Panic Disorder; CBCL = Child Behavior Checklist for Ages 6-18; prob = Problems; withdrawn/dep = Withdrawn/Depressed; anx/dep = Anxious/Depressed; dx = Diagnosis; cog = Cognitive; SP = Specific Phobia; GAD = Generalized Anxiety Disorder; PTSD = Post-Traumatic Stress Disorder; ODD = Oppositional Defiant Disorder; OC = Obsessive-Compulsive; CD = Conduct Disorder; OCD = Obsessive-Compulsive Disorder.

FCBT Moderation

The RMSE evaluated on imputed data within the FCBT subset using Bagging Regression with Elastic Net was 1.59. Predictive features across the three outcomes examined are presented in Table 9.

Table 9. Lasso regression results for FCBT

Outcome	Predictive features	β
Separation anxiety	SAD CSR	0.98
	Caregiver2 agoraphobia CSR	0.30
	Caregiver 1 SP3 CSR	0.27
	CBCL: attention prob	0.16

	Caregiver1 SP1 CSR	0.12
	Caregiver1 GAD CSR	0.12
	SP CSR	0.10
	Dysthymia CSR	0.08
	MDD past CSR	0.08
	Caregiver1 PTSD CSR	0.03
	SoP CSR	0.02
	Youth age	-0.02
	MDD CSR	-0.04
	Panic CSR	-0.05
	Caregiver1 OCD CSR	-0.06
	CBCL activities	-0.12
	Youth race: other	-0.17
	Caregiver2 dysthymia CSR	-0.20
	PDD CSR	-0.23
	Caregiver2 SP1	-0.28
Social anxiety	SoP CSR	0.97
	Caregiver1 SP CSR	0.48
	Caregiver2 agoraphobia CSR	0.35
	MDD past CSR	0.32
	Youth sex assigned at birth	0.23
	Caregiver1 GAD CSR	0.23
	OCD CSR	0.19

	Caregiver1 agoraphobia CSR	0.15
	Caregiver1 OCD CSR	0.08
	Youth age	0.05
	CBCL withdrawn/dep	0.05
	Youth race: Black	0.04
	CBCL social prob	0.01
	Kendall et al., (2008)	0.01
	GAD CSR	-0.06
	Caregiver1 dysthymia CSR	-0.11
	Caregiver2 CSR: other	-0.11
	SP CSR	-0.12
	Wood et al., (2006)	-0.12
	Caregiver1 SP1 CSR	-0.12
	Caregiver1 panic CSR	-0.13
	Caregiver1 panic and agoraphobia CSR	-0.16
	CBCL activities	-0.17
	CBCL somatic complaints	-0.21
	PTSD CSR	-0.26
	Caregiver2 CSR: dysthymia	-0.27
	CBCL thought prob	-0.36
Generalized anxiety	SoP CSR	0.28
	PTSD CSR	0.26

SAD CSR	0.19
OCD CSR	0.18
GAD CSR	0.13
SP CSR	0.11
Caregiver1 SP1 CSR	0.10
MDD past CSR	0.09
Caregiver1 PTSD CSR	0.07
Youth race: other	0.05
Dysthymia CSR	0.05
CD CSR	0.04
Caregiver1 SP3 CSR	0.03
Caregiver2 dysthymia CSR	0.01
SM CSR	-0.01
ADHD CSR	-0.01
Caregiver2 GAD CSR	-0.02
Caregiver1 SoP CSR	-0.03
CBCL somatic complaints	-0.03
Caregiver1 OCD CSR	-0.06
Bodden et al., (2008)	-0.06
Caregiver2 SP2 CSR	-0.06
CBCL social prob	-0.33

Note. CSR = Clinician Severity Rating; SAD = Separation Anxiety Disorder; SP = Specific Phobia; SP3 = third SP diagnosis; prob = Problems; SP1 = first SP diagnosis;

GAD = Generalized Anxiety Disorder; MDD = Major Depressive Disorder; PTSD = Post-Traumatic Stress Disorder; SoP = Social Anxiety Disorder; Panic = Panic Disorder; OCD = Obsessive Compulsive Disorder; PDD = Pervasive Developmental Disorders; Generalized Anxiety Disorder; withdrawn/dep = Withdrawn/Depressed; CD = Conduct Disorder; SM = Selective Mutism; ADHD = Attention-deficit Hyperactivity Disorder; SP2 = second SP diagnosis.

COMB Moderation

The RMSE evaluated on imputed data within the COMB subset using Bagging Regression with Elastic Net was 1.61. Predictive features across the three outcomes examined are presented in Table 10.

Table 10. Lasso regression results for COMB

Outcome	Predictive features	β
Separation anxiety	SAD CSR	0.44
	CBCL thought prob	0.07
	Youth race: Asian	0.06
	SZ CSR	0.06
	CBCL: depressive prob	0.06
	CBCL: anxiety prob	0.05
	Agoraphobia CSR	0.05
	CBCL stress prob	0.04
	CBCL internalizing	0.02

	Sleep terrors CSR	-0.01
	Youth age	-0.01
	Enuresis/encopresis CSR	-0.04
Social anxiety	CBCL withdrawn/dep	0.68
	SoP CSR	0.40
	SM CSR	0.26
	Sleep terrors CSR	0.16
	Youth age	0.12
	CBCL thought prob	0.11
	CBCL rule breaking	0.08
	OCD CSR	0.07
	SZ CSR	0.02
	CBCL somatic prob	0.02
	MDD CSR	0.00
	CBCL total competence	-0.02
	CBCL social	-0.03
	Enuresis/encopresis CSR	-0.06
	CBCL sluggish cog tempo	-0.13
	CBCL attention deficit	-0.13
	ODD CSR	-0.15
	CBCL school	-0.22
Generalized anxiety	MDD CSR	0.36
	Panic CSR	0.17

Sleep terrors CSR	0.15
CBCL: depressive prob	0.14
Youth race: Asian	0.10
GAD CSR	0.09
CBCL somatic prob	0.06
SoP CSR	0.03
CBCL conduct prob	0.03
Youth age	0.02
ADHD CSR	0.00
PTSD CSR	-0.08
Enuresis/encopresis CSR	-0.15

Note. CSR = Clinician Severity Rating; SAD = Separation Anxiety Disorder; CBCL = Child Behavior Checklist for Ages 6-18; prob = Problems; SZ = Schizophrenia; withdrawn/dep = Withdrawn/Depressed; SoP = Social Anxiety Disorder; SM = Selective Mutism; OCD = Obsessive Compulsive Disorder; MDD = Major Depressive Disorder; cog = Cognitive; ODD = Oppositional Defiant Disorder; Panic = Panic Disorder; GAD = Generalized Anxiety Disorder; ADHD = Attention-deficit Hyperactivity Disorder; PTSD = Post-Traumatic Stress Disorder.

SRT Moderation

The RMSE evaluated on imputed data within the SRT subset using Bagging Regression with Elastic Net was 1.37. Predictive features across the three outcomes examined are presented in Table 11.

Table 11. Lasso regression results for SRT

Outcome	Predictive features	β
Separation anxiety	SAD CSR	1.03
	PTSD CSR	0.43
	Enuresis/encopresis CSR	0.19
	Youth race: Black	0.15
	CBCL attention deficit	0.12
	SM CSR	0.03
	Youth race: other	0.01
	Sleep terrors CSR	-0.02
	SZ CSR	-0.05
	Panic CSR	-0.06
	CBCL OC prob	-0.06
	CBCL conduct prob	-0.13
	SoP CSR	-0.17
	CBCL sluggish cog tempo	-0.32
	CBCL activities	-0.49
Social anxiety	CBCL withdrawn/dep	0.68
	SoP CSR	0.34
	Agoraphobia CSR	0.29
	CBCL social prob	0.27
	SP CSR	0.23

	SM CSR	0.20
	PTSD CSR	0.14
	Enuresis/encopresis CSR	0.12
	CBCL rule breaking	0.11
	Youth race: Asian	0.03
	GAD CSR	0.01
	OCD CSR	0.01
	Youth ethnicity	0.00
	CBCL total competence	-0.02
	Sleep terror CSR	-0.05
	ODD CSR	-0.06
	Youth race: other	-0.08
	CBCL OD prob	-0.09
	CBCL externalizing	-0.13
	CBCL activities	-0.14
	SAD CSR	-0.14
	CBCL stress prob	-0.25
	CBCL somatic prob	-0.50
Generalized anxiety	GAD CSR	0.90
	PTSD CSR	0.45
	SP CSR	0.42
	SM CSR	0.32
	Youth race: Black	0.22

CBCL withdrawn/dep	0.20
CBCL social prob	0.12
Youth sex assigned at birth	0.09
CBCL OC prob	0.05
ADHD CSR	0.05
OCD CSR	0.04
CBCL school	0.03
ODD CSR	0.03
Panic CSR	-0.01
SZ CSR	-0.02
CBCL somatic prob	-0.05
CBCL somatic complaints	-0.15
CBCL activities	-0.48

Note. CSR = Clinician Severity Rating; SAD = Separation Anxiety Disorder; PTSD = Post-Traumatic Stress Disorder; CBCL = Child Behavior Checklist for Ages 6-18; SM = Selective Mutism; SZ = Schizophrenia; Panic = Panic Disorder; OC = Obsessive-Compulsive; prob = Problems; SoP = Social Anxiety Disorder; cog = Cognitive; withdrawn/dep = Withdrawn/Depressed; SP = Specific Phobia; PTSD = Post-Traumatic Stress Disorder; GAD = Generalized Anxiety Disorder; OCD = Obsessive-Compulsive Disorder; OD = Oppositional Defiant; ADHD = Attention-deficit Hyperactivity Disorder; ODD = Oppositional Defiant Disorder; Panic = Panic Disorder.

External Validation

Model performance (average RMSE) averaged across the different imputation methods for the external validation set is presented in Table 12. The same three algorithms emerged as the most robust predictors: L1 (*RMSE* = 1.40), Bayesian Ridge (*RMSE* = 1.23) and ARD (*RMSE* = 1.23), with comparable RMSE values as those observed in the harmonized dataset (L1 difference = 0; Bayesian Ridge difference = 0.01; ARD difference = 0.01). The worst performing model continued to be Decision Tree (*RMSE* = 2.05).

Table 12.Regression model root mean square errors external validation set

Model	External validation	Harmonized
L1	1.40	1.40
Bayesian Ridge	1.44	1.43
ARD	1.45	1.44
OMP	1.45	1.45
Random Forest*	1.46	1.45
L2	1.47	1.46
Gradient Boosting*	1.49	1.48
Extra Trees*	1.50	1.51
Elastic Net	1.51	1.51
Bagging with Elastic Net	1.51	1.51
AdaBoost with Elastic Net*	1.51	1.51
KNN	1.57	1.51

Decision Tree*	2.05	2.15

Note. * indicates ensemble method; L1 = Lasso Regression; Bayesian Ridge = Bayesian Ridge Regression; ARD = Automatic Relevance Determination; OMP = Orthogonal Matching Pattern; L2 = Ridge Regression; KNN = K-Nearest Neighbors.

CHAPTER 4

DISCUSSION

The present study applied several supervised learning approaches to predict outcomes for anxious youth across treatment conditions (ICBT, FCBT, GCBT, SRT. COMB, CPT and CBT/P) and within treatment types available within approximately 10% of the sample (ICBT, FCBT, COMB and SRT). The best performing algorithms were regularized versions of linear regression and other more complex regression algorithms (i.e., Bayesian Ridge Regression and Lasso Regression), along with ARD. These methods help to address problems of multicollinearity and poorly distributed data, while aiding in automatic feature selection. Consistent with previous studies highlighting the utility of ML approaches in comparison to standard regression (e.g., Kessler et al., 2016; Rosellini et al., 2018; Webb et al., 2019), 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, AdaBoost with Elastic Net) 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 (differences in *RMSEs* ranging between 0-0.01). Taken together, these findings suggest that regularized and more complex regression algorithms may be the best approach for future ML studies to apply in addressing questions of prediction for youth anxiety treatments.

Numerous features emerged as predictive of outcome within the harmonized dataset, unsurprisingly given the objectives of ML approaches (i.e., explanation versus prediction). Only variables with predictive influence $\geq \beta = +/- 0.05$ were reviewed for simplicity and to suggest future directions. Across SAD, SoP and GAD outcomes, (1)

increased pretreatment severity of the outcome variable and (2) randomization to non-active treatment conditions (WL and PBO) were unsurprisingly associated with less improvement, whereas randomization to COMB treatment was associated with more improvement. Findings that COMB may be the optimal treatment to address youth SAD, SoP and GAD symptoms are consistent with results from the original Walkup et al., (2008) trial. Several features emerged as predictive only for certain outcomes. For SAD outcome, youth who were younger showed more improvement. For both SoP and GAD outcomes, higher baseline depressive and obsessive-compulsive symptoms were associated with poorer treatment response. For SoP alone, less severe selective mutism (SM) and participation in GCBT were associated with more improvement. Findings that youth with SoP may respond better to group-based treatments contribute to understanding the findings across moderator studies indicating both that group formats are optimal (Liber et al., 2008) and that child-focused treatments are associated with more improvements (Manassis et al., 2002). Overall, predictive features were primarily indicators of symptom severity, consistent with previous reviews (Compton et al., 2014; Knight et al., 2014; Nilsen et al., 2013) and with measures available across datasets. However, results indicate that depressive and obsessive-compulsive symptoms may be particularly important to address for GAD and SoP outcomes. This finding suggests that inclusion of behavioral activation and response prevention treatment components could help improve GAD and SoP outcomes, and are in line with previous studies pointing to depressive symptoms as an indicator of outcome (Frank et al., 2021; O'Neil & Kendall, 2012; Silk et al., 2019)

When examining prediction within different treatments, numerous predictive features emerged (e.g., 24 predictive features with $\beta \geq 0.05$ for SoP outcomes within the FCBT subset), particularly for the FCBT, COMB and SRT subsets. Consequently, a narrative approach was taken to output interpretation, with an emphasis on predictive features that emerged across outcomes and with comparatively larger β's. Across all moderation models, increased pretreatment severity of the outcome variable continued to be associated with less improvement. For the ICBT subset, a similar pattern of predictive features emerged as what was observed within the harmonized dataset: (1) youth who were younger showed more improvement in SAD and (2) higher baseline depressive symptoms were associated with poorer SoP and GAD treatment response. New predictive features also emerged for ICBT. For SoP outcome, more severe youth agoraphobia, PTSD, and SAD were associated with more improvement, inconsistent with the general pattern of more severe symptomatology predicting lower outcome (Compton et al., 2014; Knight et al., 2014; Nilsen et al., 2013). It is possible that the flexibility of ICBT approaches allowed for simultaneous targeting of these comorbidities.

For FCBT, more severe youth depressive symptoms continued to be associated with poorer outcome, although this was the case for all outcomes within the FCBT subset. Unique to FCBT models, various forms of caregiver psychopathology emerged as key predictive features, although in varying directions across outcomes and caregiver diagnoses. For example, increased caregiver agoraphobia and GAD were separately associated with worse SAD and SoP outcomes. Conversely, severity of other caregiver diagnoses predicted better youth outcomes, including (1) caregiver OCD for youth SAD/GAD outcomes and (2) caregiver dysthymia for youth SAD/SoP outcomes.

Caregiver psychopathology has not been identified as a moderator consistently in other studies (Norris et al., 2021), although it has been shown to predict outcomes in some studies (Compton et al., 2014); 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, and require individual treatment before beginning FCBT. Other forms of caregiver psychopathology like OCD and dysthymia may be better targeted within FCBT, which in combination with youth symptom improvement may lead to a positive upward cascade across the family system.

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., depressive symptom severity associated with worse outcomes), several new predictive features emerged within COMB. For both SAD and GAD outcomes, race emerged as a predictive feature; specifically, youth who were categorized into the "Asian" race category during the data harmonization process showed lower improvement. This finding suggests a need for cultural adaptations to COMB protocols specifically for individuals who self-identify as Asian. Additional adaptations to COMB targeting sleep concerns may also be warranted, as more severe sleep terrors were associated with worse SoP and GAD outcomes. Increased severity of certain youth externalizing concerns (e.g., ODD, attention deficit symptoms) also predicted more COMB improvement in SoP symptoms. This finding is consistent with findings within the SRT subset, which found that more severe externalizing symptoms were associated

with better SoP outcomes. Race was also predictive of SRT outcomes, with youth receiving SRT and categorized into the "Black" race category during data harmonization showing poorer SAD and GAD outcomes. Physical symptoms emerged as unique predictors for SRT, with youth who reported more somatic complaints showing better 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 (Hale et al., 2018).

Study findings were considered through transparency, reproducibility, ethics and effectiveness (TREE) (Vollmer et al., 2020). The dataset was fully deidentified and the PI had no access to protected health information. The harmonized 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 harmonized sample and external validation sample were 71.9% and 76% White, respectively. Although race emerged as a predictive feature in moderation models, it is important to note that minoritized individuals were under-represented; for example, < 6% of participants within both datasets were identified as Asian and 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 harmonized dataset still showed better predictive performance than the best prediction models trained within single RCTs (Stanojevic et al., Under Review), highlighting the utility of developing larger, cross-site harmonized 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 Lasso Regression analyses, suggesting cross-site differences in outcomes. 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. 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 (Creswell et al., 2021). Finally, models were predicted of posttreatment 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.

Future studies should focus particularly on further model impact evaluation and implementation within real-world clinical settings (Vollmer et al., 2020). The model should be applied in additional external validation sets that are more diverse, both to ensure result reproducibility and to continually update model predictive power. With regards to implementation, barriers to use of ML-based tools in clinics must be assessed.

Mixed methods approaches could be a helpful tool to identify whether clinicians and patients find model output understandable, and if this is a preferred clinical decision-making aid. As ML prediction/moderation models become more fine-tuned, RCTs will be needed to ensure that use of ML decision making tools do in fact improve client outcomes (i.e., random assignment to model-based decisions versus clinician/client decisions).

REFERENCES CITED

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2020). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research,* 1-25.

Achenbach, T. M. (1991). Manual for the Child Behavior Checklist/4-18 and 1991 profile. *University of Vermont, Department of Psychiatry*.

Adibi, A., Sadatsafavi, M., & Ioannidis, J. P. (2020). Validation and utility testing of clinical prediction models: time to change the approach. *Journal of the American Medical Association, 324(3), 234-236.*

Albano, A. M., & Silverman, W. K. (in press). *The Anxiety Disorders Interview Schedule for DSM-5: Child and Parent Versions*.

Alsaber, A. R., Pan, J., & Al-Hurban, A. (2021). Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018). *International Journal of Environmental Research and Public Health*, *18*(3), 1333

Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and quality of life outcomes*, *17*(1), 1-9.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Bodden, D. H., Bogels, S. M., Nauta, M. H., 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. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 1384-1394.

Borkenau, P., & Ostendorf, F. (1998). The Big Five as states: How useful is the five-factor model to describe intraindividual variations over time?. *Journal of Research in Personality*, *32*(2), 202-221.

Brookes, S. T., Whitely, E., Egger, M., Smith, G. D., Mulheran, P. A., & Peters, T. J. (2004). Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test. *Journal of clinical epidemiology*, *57*(3), 229-236.

Brown, T. A., Barlow, D. H., & DiNardo, P. A. (1994). *Anxiety disorders interview schedule adult version: Client interview schedule*. Graywind Publications Incorporated.

Bystritsky, A. (2006). Treatment-resistant anxiety disorders. *Molecular psychiatry*, *11*(9), 805-814.

Chekroud, A. M., 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.

Compton, S. N., Peris, T. S., Almirall, D., Birmaher, B., Sherrill, J., Kendall, P. C., ... & Albano, A. M. (2014). Predictors and moderators of treatment response in childhood anxiety disorders: results from the CAMS trial. *Journal of consulting and clinical psychology*, *82*(2), 212.

Coutanche, M. N. & Hallion, L. S. (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.

Creswell, C., Nauta, M. H., Hudson, J. L., March, S., Reardon, T., Arendt, K., ... & Kendall, P. C. (2021). Research Review: Recommendations for reporting on treatment trials for child and adolescent anxiety disorders–an international consensus statement. *Journal of Child Psychology and Psychiatry*, *62*(3), 255-269.

Cummings, C. M., Caporino, N. E., & Kendall, P. C. (2014). Comorbidity of anxiety and depression in children and adolescents: 20 years after. *Psychological bulletin*, *140*(3), 816.

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.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 1-17.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*, 91-118.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O.

(2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 1-37.

Essau, C. A., Lewinsohn, P. M., Olaya, B., & Seeley, J. R. (2014). Anxiety disorders in adolescents and psychosocial outcomes at age 30. *Journal of affective disorders*, *163*, 125-132.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106-E6115.

Frank, H. E., Titone, M. K., Kagan, E. R., Alloy, L. B., & Kendall, P. C. (2021). The role of comorbid depression in youth anxiety treatment outcomes. *Child Psychiatry & Human Development*, *52*(6), 1024-1031.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. (2007). The integrated trait–state model. *Journal of Research in Personality*, *41*(2), 295-315.

Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, *363*(4), 301-304.

Hale, A. E., Ginsburg, G. S., Chan, G., Kendall, P. C., McCracken, J. T., Sakolsky, D., ... & Walkup, J. T. (2018). Mediators of treatment outcomes for anxious children and adolescents: The role of somatic symptoms. *Journal of Clinical Child & Adolescent Psychology*, *47*(1), 94-104.

Harpaz-Rotem, I., Leslie, D., & Rosenheck, R. A. (2004). Treatment retention among children entering a new episode of mental health care. *Psychiatric Services*, *55*(9), 1022-1028.

Higa-McMillan, C. K., Francis, S. E., Rith-Najarian, L., & Chorpita, B. F. (2016).

Evidence base update: 50 years of research on treatment for child and adolescent anxiety. *Journal of Clinical Child & Adolescent Psychology*, *45*(2), 91-113.

Holmbeck, G. N. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: examples from the child-clinical and pediatric psychology literatures. *Journal of consulting and clinical psychology*, *65*(4), 599.

James, A. C., James, G., Cowdrey, F. A., 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

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on psychological science*, *6*(1), 21-37.

Kazdin, A. E. (2019). Annual research review: expanding mental health services through novel models of intervention delivery. *Journal of Child Psychology and Psychiatry*, *60*(4), 455-472.

Kendall, P. C. (1994). Treating anxiety disorders in children: Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology, 62*, 100-110.

Kendall P. C., Flannery-Schroeder E., Panichelli-Mindel S. M., Southam-Gerow M., Henin, A., & Warman M. (1997). Therapy for youths with anxiety disorders: A second randomized clinical trial. *Journal of Consulting and Clinical Psychology, 65*, 366-380.

Kendall, P. C., Hudson, J. L., Gosch, E., Flannery-Schroeder, E., & Suveg, C. (2008).

Cognitive-behavioral therapy for anxiety disordered youth: A randomized clinical trial evaluating child and family modalities. *Journal of Consulting and Clinical Psychology, 76*, 282-297.

Kendall, P. C., 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.

Kendall, P. C., & Hedtke, K. A. (2006). *Cognitive-Behavioral Therapy for Anxious Children: Therapist Manual* (3rd ed.). Ardmore, PA: Workbook Publishing.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., Ebert, D. D., Hwang, I., Li, J., de Jonge, P., Nierenberg, A. A., Petukhova, M. V., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (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.

Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: an improved missing data imputation technique. *Journal of big Data*, *7*(1), 1-21.

Kiesler, D. J. (1966). Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin*, *65*(2), 110.

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.

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 1-22.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of general psychiatry*, *59*(10), 877-883.

Lebowitz, E. R., Zilcha-Mano, S., Orbach, M., Shimshoni, Y., & Silverman, W. K. (2021). Moderators of response to child-based and parent-based child anxiety treatment: a machine learning-based analysis. *Journal of Child Psychology and Psychiatry*.

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., ... & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *Journal of affective disorders*, *241*, 519-532.

Liber, J. M., Van Widenfelt, B. M., Utens, E. M., Ferdinand, R. F., Van der Leeden, A. J., Gastel, W. V., & Treffers, P. D. (2008). No differences between group versus individual treatment of childhood anxiety disorders in a randomised clinical trial. *Journal of Child Psychology and Psychiatry*, *49*(8), 886-893.

Lopez, B., Turner, R. J., & Saavedra, L. M. (2005). Anxiety and risk for substance dependence among late adolescents/young adults. *Journal of anxiety disorders*, *19*(3), 275-294.

Luedtke, A., Sadikova, E., & Kessler, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clinical Psychological Science*, *7*(3), 445-461.

Magnavita, J. J., & Lilienfeld, S. O. (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.

Manassis, K., Mendlowitz, S. L., Scapillato, D., Avery, D., Fiksenbaum, L., Freire, M., ... & Owens, M. (2002). Group and individual cognitive-behavioral therapy for childhood anxiety disorders: A randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(12), 1423-1430.

Manassis, K., Russell, K., & Newton, A. S. (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.

Marciniak, M., Lage, M. J., Landbloom, R. P., Dunayevich, E., & Bowman, L. (2004). Medical and productivity costs of anxiety disorders: case control study. *Depression and anxiety*, *19*(2), 112-120.

Molenaar, P. C., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of interindividual and intraindividual variability: A theoretical and empirical vindication of developmental systems theory. In *Understanding human development* (pp. 339-360). Springer, Boston, MA.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201-218.

Molenaar, P. C. (2013). On the necessity to use person-specific data analysis approaches in psychology. *European Journal of Developmental Psychology*, *10*(1), 29-39.

Murphy, T. K., Segarra, A., Storch, E. A., & Goodman, W. K. (2008). SSRI adverse

events: how to monitor and manage. *International Review of Psychiatry*, *20*(2), 203-208.

Murphy, S. E., Capitão, L. P., Giles, S. L., Cowen, P. J., Stringaris, A., & Harmer, C. J. (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.

Na, J., Grossmann, I., Varnum, M. E., Kitayama, S., Gonzalez, R., & Nisbett, R. E. (2010). Cultural differences are not always reducible to individual differences. *Proceedings of the National Academy of Sciences*, *107*(14), 6192-6197.

Nakamura, B. J., Ebesutani, C., Bernstein, A., & Chorpita, B. F. (2009). A psychometric analysis of the child behavior checklist DSM-oriented scales. *Journal of Psychopathology and Behavioral Assessment*, *31*(3), 178-189.

Nauta, M. H., Scholing, A., Emmelkamp, P. M., & Minderaa, R. B. (2003). Cognitive-behavioral therapy for children with anxiety disorders in a clinical setting: No additional effect of a cognitive parent training. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*, 1270-1278.

Nilsen, T. S., Eisemann, M., & Kvernmo, S. (2013). Predictors and moderators of outcome in child and adolescent anxiety and depression: a systematic review of psychological treatment studies. *European Child & Adolescent Psychiatry*, *22*(2), 69-87.

Norris, L. A., & Kendall, P. C. (2021). Moderators of outcome for youth anxiety treatments: Current findings and future directions. *Journal of Clinical Child &*

Adolescent Psychology, *50*(4), 450-463.

O'Neil, K. A., & Kendall, P. C. (2012). Role of comorbid depression and co-occurring depressive symptoms in outcomes for anxiety-disordered youth treated with cognitive-behavioral therapy. *Child & Family Behavior Therapy*, *34*(3), 197-209.

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*(2), 109-118. https://doi.org/10.1037/h0024436

Petrazzini, B. O., Naya, H., Lopez-Bello, F., Vazquez, G., & Spangenberg, L. (2021). Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData mining*, *14*(1), 1-13.

Racine, N., McArthur, B. A., Cooke, J. E., 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 pediatrics*, *175*(11), 1142-1150.

Rosellini, A. J., Dussaillant, F., Zubizarreta, J. R., Kessler, R. C., & Rose, S. (2018). Predicting posttraumatic stress disorder following a natural disaster. *Journal of Psychiatric Research*, *96*, 15-22.

Rudd, M. D., Joiner, T. E., & Rumzek, H. (2004). Childhood diagnoses and later risk for multiple suicide attempts. *Suicide and Life-Threatening Behavior*, *34*(2), 113-125.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210-229.

Settipani, C. A., & Kendall, P. C. (2013). Social functioning in youth with anxiety disorders: Association with anxiety severity and outcomes from cognitive-behavioral therapy. *Child Psychiatry & Human Development*, *44*(1), 1-18.

Silk, J. S., Price, R. B., Rosen, D., Ryan, N. D., Forbes, E. E., Siegle, G. J., ... & Ladouceur, C. D. (2019). A longitudinal follow-up study examining adolescent depressive symptoms as a function of prior anxiety treatment. *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*(3), 359-367.

Silverman, W. (1987). *Anxiety Disorders Interview for Children (ADIC).* State University of New York at Albany: Graywind Publications.

Silverman, W. K., & Nelles, W. B. (1988). The anxiety disorders interview schedule for children. *Journal of the American Academy of Child & Adolescent Psychiatry, 27*(6), 772-778.

Silverman, W. K., & Albano, A. M. (1996). *The Anxiety Disorders Interview Schedule for DSM-IV: Child and Parent Versions.* New York, NY: Oxford University Press.

Silverman, W. K., Kurtines, W. M., Ginsburg, G. S., Weems, C. F., Rabian, B., & Serafini, L. T. (1999). Contingency management, self-control, and education support in the treatment of childhood phobic disorders: A randomized clinical trial. *Journal of Consulting and Clinical Psychology, 67,* 675-687. doi:10.1037/0022-006X.67.5.675

Silverman, W. K., Saavedra, L. M., & Pina, A. A. (2001). Test-retest reliability of anxiety symptoms and diagnoses with the Anxiety Disorders Interview Schedule for DSM-IV: Child and Parent Versions. *Journal of the American Academy of Child & Adolescent Psychiatry, 40,* 937-944. doi:10.1097/00004583-200108000-00016

Silverman, W. K., Kurtines, W. M., Jaccard, J., & Pina, A. A. (2009). Directionality of

Change in Youth Anxiety Treatment Involving Parents: An Initial Examination. *Journal of Consulting and Clinical Psychology, 77,* 474-485.

Siontis, G. C., Tzoulaki, I., Castaldi, P. J., & Ioannidis, J. P. (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of clinical epidemiology*, *68*(1), 25-34.

Skriner, L. C., Chu, B. C., Kaplan, M., Bodden, D. H., Bögels, S. M., Kendall, P. C., ... & Xie, M. G. (2019). Trajectories and predictors of response in youth anxiety CBT: Integrative data analysis. *Journal of consulting and clinical psychology*, *87*(2), 198.

Stanojevic, M., Norris, L.A., Kendall, P.C. & Obradovic, Z (under review). Can machine learning help clinical psychologists?

Swan, A. J., & Kendall, P. C. (2016). Fear and missing out: Youth anxiety and functional outcomes. *Clinical Psychology: Science and Practice*, *23*(4), 417.

Taylor, S., Abramowitz, J. S., & McKay, D. (2012). Non-adherence and non-response in the treatment of anxiety disorders. *Journal of anxiety disorders*, *26*(5), 583-589.

Tiemens, B., Bocker, K., & Kloos, M. (2016). Prediction of treatment outcome in daily generalized mental health care practice: first steps towards personalized treatment by clinical decision support. *European Journal for Person Centered Healthcare*, *4*(1), 24-32.

van Bronswijk, S. C., Bruijniks, S. J., Lorenzo-Luaces, L., Derubeis, R. J., Lemmens, L. H., Peeters, F. P., & Huibers, M. J. (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. *Psychotherapy Research*, 1-14.

Villabø, M. A., Narayanan, M., Compton, S. N., Kendall, P. C., & Neumer, S. P. (2018). Cognitive–behavioral therapy for youth anxiety: An effectiveness evaluation in community practice. *Journal of consulting and clinical psychology*, *86*(9), 751.

Vollmer, S., Mateen, B. A., Bohner, G., Kiraly, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., Granger, D., Birse, M., Branson, R., Moons, K. G. M., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *British Medical Journal*, *368*, l6927.

Walkup, J. T., Albano, A. M., Piacentini, J., Birmaher, B., Compton, S. N., Sherrill, J. T., ... & Kendall, P. C. (2008). Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. *New England Journal of Medicine*, *359*(26), 2753-2766.

Wang, Z., Whiteside, S. P., Sim, L., Farah, W., Morrow, A. S., Alsawas, M., ... & Murad, M. H. (2017). Comparative effectiveness and safety of cognitive behavioral therapy and pharmacotherapy for childhood anxiety disorders: a systematic review and meta-analysis. *JAMA pediatrics*, *171*(11), 1049-1056.

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2019). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25.

Wilkinson, J., Arnold, K. F., Murray, E. J., 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*.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of personality assessment*, *82*(1), 50-59.

Wood, J. J., Piacentini, J. C., Bergman, R. L., 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. *Journal of Clinical Child & Adolescent Psychology, 31,* 335-342. doi:10.1207/S15374424JCCP3103_05

Wood, J. J., Piacentini, J. C., Southam-Gerow, M., Chu, B. C., Sigman, M. (2006). Family cognitive behavioral therapy for child anxiety disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 45*, 314-321.