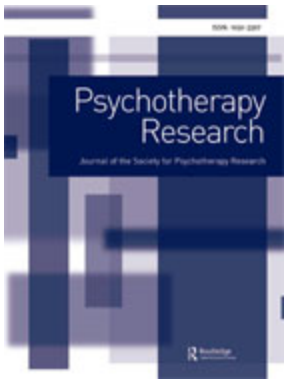
Article:
Henderson, C.E., Woodard, G.S., Simmonds-Buckley, M. et al. (2024) Prediction of adolescent psychotherapy outcomes using youth- and caregiver-reported symptoms data. Psychotherapy Research, 35 (7). pp. 1185-1197. ISSN: 1050-3307

https://doi.org/10.1080/10503307.2024.2394187



Prediction of Adolescent Psychotherapy Outcomes Using Youth- and Caregiver-Reported Symptoms Data

Journal:	*Psychotherapy Research*
Manuscript ID	TPSR-2023-0317.R2
Manuscript Type:	Research Article
Keywords:	Measurement-based care, adolescent psychotherapy
Classifications:	Child Psychotherapy, Mental Health Services Research, Outcome Research

Running head: ADOLESCENT PSYCHOTHERAPY OUTCOMES

Prediction of Adolescent Psychotherapy Outcomes Using Youth- and Caregiver-Reported Symptoms Data

Abstract

Objective: We used longitudinal youth- and caregiver-reports of adolescent psychological symptoms from three samples of youth receiving mental health services in routine treatment settings to derive expected change trajectories and identify cases at risk for treatment failure.

Method: Participants were 1906 youth (1053 caregivers) receiving treatment in community mental health settings, merged across three samples. The Symptoms and Functioning Severity Scale (SFSS) was used as an indicator of weekly clinical change. Multilevel modeling methods were used to develop expected change trajectories and identify cases at risk for treatment failure (not on track; NOT). Logistic regression was used to predict client improvement as a function of NOT status.

Results: The SFSS was a reliable indicator of therapeutic change according to youth-reported symptoms. Caregiver reports were not as robust. Whereas predictive accuracy of NOT status yielded moderately high sensitivity in detecting improvement according to youth report, caregiver reports were not as predictive.

Conclusions: The youth-reported version of the SFSS-based algorithm seems appropriate for implementation in clinical care. Future studies should search for similarly predictive measures for caregivers.

Keywords: Measurement-based care, adolescent psychotherapy, caregiver report, treatment outcomes, multilevel modeling

Clinical or Methodological Significance of This Article: Findings from this adolescent treatment study indicated that the SFSS is a reliable indicator of weekly change in treatment according to youth report. According to youth report, NOT cases were reliably detected by

approximately the 8th treatment session, and remaining on track predicted clinical improvement with moderately high sensitivity. Predictions using caregiver reports were not as robust.

Objective

Rates of youth mental health symptom improvement in community mental health settings are low; more than half of youth treated in these settings deteriorate or do not improve (Authors Masked, 2017; Nelson et al., 2013; Sale et al., 2021; Warren et al., 2010). Measurement-based care (MBC), defined as the process of regularly administering outcome measures to clients to inform clinical decision making (Lewis et al., 2019), may help to improve rates of symptom improvement (Lambert et al., 2018). Several meta-analyses have found small but significant effects (effect sizes ranging from 0.17-0.28) of MBC on youth mental health symptoms (de Jong et al., 2021; Rognstad et al., 2023; Tam & Ronan, 2017). MBC is an evidence-based practice which is widely applicable across diagnostic categories and treatment settings (Lewis et al., 2019). More than 20 professional organizations have endorsed MBC as an evidence-based practice (Coalition for the Advancement and Application of Psychological Science, 2018) and a cornerstone of youth mental health evidence-based practice (American Psychological Association Presidential Task Force on Evidence-Based Practice, 2006).

There is substantial evidence in the adult literature that MBC is particularly effective for clients who are not-on-track (NOT) to improve, operationalized as a client's level of symptomatology exceeding the expected trajectory (de Jong et al., 2021; Rognstad et al., 2023; Shimokawa et al., 2010). Some computerized measurement-feedback systems have a function that alerts clinicians when a client is NOT if their symptoms are more severe than expected by comparison to clinical norms. When MBC systems alert clinicians that their client is not improving, clinicians have a timely opportunity to identify and resolve problems that are interfering with the client's progress. This feedback-informed method of treatment adaptation has been shown to be more effective than usual psychological treatment, according to meta-

analyses of clinical trials (e.g., see Shimokawa et al., 2010). For example, client outcomes were similar for clients in the MBC and no MBC conditions until an alert that a client was NOT was given to a clinician (Probst et al., 2013). After the alert was given, clients in the MBC condition had better mental health outcomes (Probst et al., 2013). One study found that providing information about the client's progress to the clinician only, or to the clinician *and* the client together, both had better outcomes than treatment as usual without MBC (Hawkins et al., 2004). Thus, it is important to be able to identify clients who are NOT in order to reap the full benefit of MBC.

In the adult MBC literature, some studies have used *rationally derived methods* to identify NOT cases, for instance by assuming that a client is NOT if their symptoms have not improved by a magnitude greater than the reliable change index for the measure used to track their progress after the early phase of treatment. Other studies have used *empirically derived methods* to identify NOT cases, which involves the continuous monitoring of symptoms and their comparison to an expected trajectory of change which is modelled using data from a relevant clinical sample (Lutz et al., 2009). Studies have found that empirically derived methods are more accurate than rational methods and were able to identify 100% of clients NOT with the majority of those cases being identified by session 3 (Lambert et al., 2002). The most common empirically derived method of feedback uses *expected treatment response* (ETR) curves to identify NOT cases. The ETR curves represent confidence intervals around the expected (e.g., average) trajectory of change observed over time in a clinical sample that has a similar baseline severity as the client whose symptoms are being evaluated (Finch et al., 2001). The ETR curves can be used to create NOT alerts for clients whose symptoms are more severe than the upper (e.g., 80% or 95% in some studies) confidence interval. The first few studies to create

empirically derived ETR curves for adults were done with several proprietary systems such as the OQ system (Finch et al., 2001; Lutz et al., 2006), the ComPASS system (Lueger et al., 2001), and the freely-available Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM; Lutz et al., 2005). More recently, several studies have created empirically derived ETR curves for adults based on the Patient Health Questionnaire-9 and the Generalized Anxiety Disorder-7 (Authors Masked, 2016; Authors Masked, 2021). These algorithms have been able to explain about half of the variance in treatment outcomes (Authors Masked, 2021) and identify characteristics of clients who are at risk of making poor progress in treatment (Authors Masked, 2016) using free and publicly available progress monitoring measures. The specific measure used to collect data in the present study (see below) was developed as part of a research-practice partnership (Reimer et al., 2012) to ensure that both clinical relevance and high-quality psychometric approaches were prioritized. However, "not-on-track" metrics have not been developed for the measure as it is used in practice, and the present study represents an attempt to improve its quality as an MBC measure.

However, in the youth literature, there are few studies that identify NOT clients or that examine the outcomes of these cases, and the studies that do exist have focused on a single system, the Youth Outcome Questionnaire (YOQ; Burlingame et al., 2005). Warren et al. (2009) created a warning system based on the YOQ given to youth in outpatient community mental health settings. The warning system was able to correctly identify 71% of youth who were classified as treatment failures (i.e., whose symptoms were significantly higher at the end of treatment; Warren et al., 2009). Cannon et al. (2010) created expected response trajectories for the youth and caregiver reported YOQ using the data from over 2,000 youth in managed care and community mental health clinics. The study found moderately high sensitivity rates at

identifying at-risk cases based on the youth-reported YOQ and caregiver-reported YOQ separately, but the best sensitivity was when both youth- and caregiver-reported YOQ scores were integrated (Cannon et al., 2010). Lastly, another study found that youth who completed the YOQ more frequently had faster rates of change, but both the original and simplified warning system demonstrated moderately high sensitivity at identifying NOT youth cases (Nelson et al., 2013). Although the research summarized here suggests that warning systems and analysis of ETR can be effective in monitoring youth psychotherapy outcomes, there is a need for research to move beyond this single system and identify ETR curves for other systems.

The Symptoms and Functioning Severity Scale (SFSS) is a measure of mental health symptoms for youth between 11 and 18 years of age receiving mental health treatment, which is freely available for non-profit and educational research (Athay et al., 2012; Authors Masked, 2010). In its development, brevity and sensitivity to change were emphasized along with the typical psychometric properties of reliability and validity. The SFSS was designed to be able to be repeatedly administered to monitor symptoms, functioning, and potential for NOT over the course of treatment from the perspectives of youth, caregivers, and clinicians (Authors Masked, 2010). As such, the SFSS items are designed to capture symptoms of common mental health concerns for youth, including attention deficit hyperactivity disorder, anxiety, depression, conduct disorder, and oppositional defiant disorder (Authors Masked, 2010). The SFSS has excellent convergent validity with other well-established measures of youth functioning (Authors Masked, 2010), including the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1999), Child Behavior Checklist (CBCL; Achenbach, 1991), and the Youth Outcome Questionnaire (YOQ; (Burlingame et al., 2005). The measure was developed as part of a research-practice partnership (Riemer et al., 2012) to ensure that both clinical relevance and high-quality

psychometric approaches were prioritized, and has support from two randomized controlled trials (Authors Masked, 2016; Bickman et al., 2011).

It is ideal for clinicians to have more than just scores to inform treatment planning (e.g., comparative information). There is a range of sophistication in providing interpreted information to clinicians. Several more simple alerting mechanisms were developed early on with the SFSS to help clinicians see when treatment was not progressing as expected or a client was declining. These include trends over time (improving, remaining stable, or decreasing since start of treatment), acute change scores (improving, worsening, or staying the same since the last time the SFSS was administered), and problem alerts available in the accompanying measurement-feedback system technology to show when specific items were in the top 25th percentile range of severity and reporter differences (Authors Masked, 2015). These problem alerts were associated with clinicians addressing problematic symptoms faster and more frequently (Authors Masked, 2015). Normative trajectories using ETR curves are more complex alert systems, and potentially more useful, in helping clinicians think about their client's comparative progress across a course of treatment. Thus far, there are no normative trajectories available for the SFSS. Identifying NOT or at-risk cases on the SFSS has been theoretically driven, not based on normative data, which means some youth NOT cases are likely missed. The ability to accurately identify youth who are NOT on the SFSS early in treatment represents an important step forward in developing a warning system by which therapists may adapt their treatment approaches to make them more effective.

The current study used a data-driven approach to derive normative trajectories for the SFSS among 1,902 youth and 1,011 caregivers treated in home-based and outpatient community

mental health clinics. These trajectories can be used to identify youth who are NOT to create alerts for clinicians based on normative ETR curves, rather than theoretical cutoff scores.

Method

Data sources and sample selection

Data from three sources were merged to form the dataset for the present study. For all sources, the SFSS was intended to be administered as part of MBC corresponding to the session schedule (e.g., every week for weekly treatment, every other week for bi-weekly treatment). The SFSS was first used in a randomized control trial (RCT) of MBC in youth from 11 to 18 years of age receiving home-based mental health treatment in community settings (Bickman et al., 2011). Clinics participated in a national behavioral health network that provided in-home services to youth and their families. Clinicians reported using a variety of therapeutic approaches, including cognitive-behavioral, integrative-eclectic, behavioral, family systems, and play therapy. Randomization occurred at the site level, with MBC feedback provided weekly to clinicians at 13 sites (the experimental group) and every 90 days to clinicians at 15 sites (the control group). Sites were in 10 states with an overall sample of 340 youths, 383 caregivers, and 144 clinicians.

Next, the SFSS was part of a second RCT of MBC in youth aged 11 to 18 years receiving outpatient mental health treatment in urban and rural community settings (Authors Masked, 2016). Randomization occurred at the youth level, with MBC feedback provided to clinicians weekly in the experimental group and every six months in the control group. A total of 257 youth had 2698 clinician sessions with at least one completed SFSS present. Twenty-one clinicians provided treatment, and in addition to the data provided by the youth, 248 caregivers also provided responses to the questionnaire. Participating clinics provided outpatient mental health services for a wide range of presenting problems and diagnoses. Although specific data are

lacking, the vast majority of clinicians practiced at the master's or doctoral level. Approximately, one third of clinicians reported a cognitive-behavioral treatment orientation, which almost 50% reporting an unspecified orientation or no particular allegiance.

The third data source was obtained from Mirah, an enterprise level software-as-a-service company that provides MBC technology. Mirah retains rights to use de-identified data to facilitate development of validated measures as part of the agreement with each organization that uses the proprietary software. Mirah has a license with Vanderbilt University to make the SFSS and the larger Peabody Treatment Progress Battery (PTPB) available to mental health service organizations. Mirah provided data for the purpose of expanding the overall sample size of this project, with data on 1309 youth and 422 caregivers across 8328 sessions. Due to the de-identified nature of the data, there was no information available on sample characteristics from the Mirah sample. Detailed sample characteristics for the pooled sample are described below.

Data collection methods differed across data source. For Authors Masked (2011), youth and caregivers were asked to complete paper measures at the close of a treatment session. For Authors Masked (2016), the protocol was for measures to be completed in the last 5-10 minutes of a treatment session. For the Mirah data, youth clients and caregivers received unique links (either in a text message or email) to complete measures electronically prior to the session.

Measures

The Symptoms and Functioning Severity Scale (SFSS; Athay, Riemer, & Bickman, 2012) is a 26-item measure of youth symptoms, with parallel youth (YSFSS) and caregiver (CSFSS) versions. (A clinician-report also exists but is not a focus of the present study.) It has three forms- a full version containing all items, and two shorter versions with equivalent forms so that only 13 items need to be administered each session. As such, each item is administered

every two weeks, so respondents are asked to rate the frequency of symptoms over the past two weeks on a five-point Likert scale ranging from "never" to "very often." The measure yields a total score, as well as internalizing and externalizing scores. The total score includes 26 items (for the youth version, 27 for caregiver) designed to measure symptoms associated with the measure categories of mental illness found in the American Psychological Association Diagnostic and Statistical Manual of Mental Disorders (DSM-V-TR; American Psychiatric Association, 2022). Each item is scaled from 1 to 5 with 1 representing "Never" and 5 "Very Often". The total score is calculated by taking the summation of all item scores. In the present study SFSS total scores were linearly transformed to set the scores on the same scale on the widely-used Child Behavior Checklist (Achenbach, 1991; Authors Masked, 2010); the transformation resulted in total scores ranging between 33 and 86. A score of 42 or above indicates clinically significant distress. The clinically significant cut-off score along with the measure's reliable change index (RCI; see below for further definition) was used to define categories of clinical change, as described in the Data Analysis section. The RCI for the measure is 4.63 points. Both youth and caregiver versions have evidenced strong internal consistency (0.92 and 0.93, respectively), moderate to strong test-retest reliability (0.68 and 0.87, respectively), and a stable factor structure, providing evidence of construct validity (Authors Masked, 2010). Consistent with data provided in the test manual, internal consistency was high for both versions of the measure, ranging from 0.89 to 0.94 depending on the sample (RCT vs. data optained from Mirah.)

Data Analysis

In the present study, we used YSFSS and CSFSS scores with the equivalent forms harmonized so that all data were on the same scale. Symptom reports were collected on a

session-by-session basis to monitor treatment progress. We first classified cases' observed treatment outcome by calculating a change score representing the difference between the first and last treatment session according to the reliable change index (RCI; Jacobson & Truax, 1991) of both youth- and caregiver-reported SFSS scores (Authors Masked, 2010). The RCI measures the smallest change in scores that is distinguishable from measurement error. Reliable improvement was operationalized as pre-post change exceeding the RCI. Clinically significant change was operationalized as a case with baseline scores in the clinical range and with post-treatment scores in the sub-clinical range based on the cutoff for the SFSS. These two classification rules were combined into a binary outcome definition, where cases with both reliable and clinically significant improvement were classed as treatment responders (code = 1), while others were classed as non-responders (code = 0). This operational definition has been used in previous studies in order to examine the prediction accuracy of empirically derived feedback models (Cannon et al., 2010; Warren et al., 2009), using a stringent definition of treatment response which facilitates evaluation using conventional indices such as odds ratios and positive and negative predictive values.

We next split the youth and caregiver reports by randomly selecting 60% of the participants in each group, comprising the training cases, and the remaining 40% comprising the test cases. Independent samples t-tests revealed no statistically significant differences between intake SFSS scores comparing the training and test cases for both youth- and caregiver-reported data. The training cases served to develop the ETR models, and the test cases were used to evaluate the models' prediction accuracy. We then trimmed the training cases according to treatment duration, excluding participants who had received fewer than 4 sessions. The exclusion of cases with fewer than 4 sessions was necessary to evaluate the accuracy of prediction models,

since at least 3 sessions are required to model nonlinear trends in time-series data, and one additional session (e.g., session 4 onwards) is necessary to measure a post-treatment outcome score following the initial 3 sessions. In addition, cases with more than 30 sessions were excluded to prevent the models from being overly influenced by sparse data (e.g., extreme outliers) from cases with unusually brief or lengthy interventions. This resulted in selected groups of 974 youth and 516 caregivers comprising the training cases, and 868 youth and 380 caregivers for the test cases. Finally, we stratified the data by baseline SFSS score among the training cases, grouping participants with similar ranges, ensuring the groups had a minimum of 132 participants (following the minimum sample size calculation by Authors Masked, 2021) and that the range of scores exceeded the RCI. This resulted in four groups of intake scores for both the training and test cases, which are the Quartiles referred to below. This was done to ensure adequate power for the analyses and that any observed change would exceed that expected for sampling error. This resulted in four subgroups for both the youth and caregiver data.

We considered combining the youth and caregiver reports in an integrated modeling approach and in the end decided against it due to the fact that parent and youth reports frequently provide both discrepant and unique information (De Los Reyes, et al., 2022). Given this evidence base, we decided against this approach and analyzed the youth and caregiver data separately.

Analyses in the training cases examined youth- and caregiver-reported change in SFSS scores using longitudinal multilevel modeling of time-series data (MLM; Raudenbush & Bryk, 2002; Singer & Willett, 2003) as conducted in the statistical software SPSS (Version 27; IBM Corp., 2020). Longitudinal MLM estimates the initial status (intercept) and rate of change over treatment (slope) for each youth and generates average intercepts and slopes across each individual (i.e., fixed effects), along with variances representing individual variability around the

sample means (i.e., random effects). Analyses used full information maximum likelihood estimation (FIML) to accommodate and reduce potential bias due to missing data under the assumption that data were missing at random (Schafer & Graham, 2002). As recommended by Singer and Willett (2003), we compared the goodness-of-fit of a pre-specified set of unconditional (no covariates) growth models to derive the optimal covariance structure (e.g., unstructured, diagonal, autoregressive) and growth trend (e.g., linear, log-linear, quadratic, cubic) of the underlying trajectories using the Bayesian Information Criterion (BIC; Schwarz, 1978) with lower values indicating better fit to the data. Once the best-fitting models were identified, the resulting equations were used to produce growth curve models with upper and lower ETR curves. Consistent with previous research published on this topic (Authors Masked, 2021), we did not account for nesting youth reports within therapist, as accounting for this nesting could overfit the models to therapists included in the present data; that is. In addition, we were primarily interested in deriving intercept and slope parameters in the MLMs, and were secondarily concerned with their statistical significance. To this end, MLMs produce unbiased estimates for regression parameters even when not accounting for nesting, as nesting affects the standard errors of the estimates, increasing the rate of Type I errors (Bosker & Snijders, 2011).

Model parameters were then used to calculate prediction intervals representing the average SFSS score at each treatment session and the expected variation around the average trajectory. The upper bounds of the prediction intervals were calculated by multiplying the mean standard error (MSE) of the random effect by the critical value for the 90% value of the student's t distribution. Scores occurring outside the upper prediction intervals indicate more severe symptoms than expected at the relevant treatment session, indicating the case was at NOT. This statistical rule was used to classify cases as on track (OT) or NOT in both training and test cases

(using the ETR algorithm developed in the group of training cases). In total, ETR models for 4 subgroups with similar baseline scores were developed, and the classification algorithm would determine each youth's status (OT or NOT) by [1] selecting their reference subgroup and [2] comparing their symptoms to the upper boundary of the ETR curves for that reference group at a specified time-point (e.g., treatment session number).

Finally, in order to evaluate the performance of the ETR classification algorithm in the test cases, we conducted a series of logistic regression analyses using NOT status (0 = On Track, 1 = Not on Track) as a predictor of treatment response (0 = did not attain full remission of symptoms, 1 = reliable and clinically significant improvement) definitions for the categories described above. Following Authors Masked (2021), we analyzed session-by-session regression models comprising youth who remained in treatment at each session of treatment. For example, the regression model for session six included all patients who attended at least seven sessions, at session seven all participants who remained in treatment eight sessions, etc. This was done so that predictive information collected up to a given session number was not confounded with the post-treatment outcome that was measured at subsequent sessions.

Results

A summary of baseline and clinical characteristics of the participants is included in Table 1. Just over 41% of participants among the training cases (43.2%, n = 421) showed reliable improvement through treatment on the youth-report SFSS, as compared to 35.4% (n = 307) of participants in the test cases. The difference was statistically significant ($\chi^2_{[df = 2,\ N = 1842]} = 11.85$, $p < .001$). For the caregiver reports, 33.3% (n = 172) of caregivers reported reliable improvement in the training cases, and 27.1% (n = 103) of caregivers reported reliable improvement in the test cases. This difference was also statistically significant ($\chi^2_{[df = 2,\ N = 896]} = 3.99$, $p = .046$).

Table 2 displays the results of the MLM by quartile (see above for definition) for the youth-report training cases. The best-fitting model was characterized by a log-linear change trend and an unstructured covariance matrix. Participants in Quartile 1 on average began treatment with an SFSS score of 42.31 (Standard Error [SE] = 0.36) and decreased a log of -0.50 (*SE* = 0.26) per session, which was statistically significant (see Table 2). A similar pattern of statistically significant decreases was demonstrated in each Quartile (Quartile 2: Intercept = 50.44 [*SE* = 0.31], Slope = -1.46 [*SE* = 0.24]; Quartile 3: Intercept = 56.18 [*SE* = 0.37], Slope = -2.59 [*SE* = 0.32]; Quartile 4: Intercept = 63.72 [*SE* = 0.45], Slope = -3.89 [*SE* = 0.32]). Analysis of random effects by quartile indicate that a substantial amount of variance was due to between-participant differences (Quartile 1: 33%, Quartile 2: 25%, Quartile 3: 40%, Quartile 4: 34%; see Table 2). Figure 1 illustrates change trajectories for Quartile 4 (the group that showed the most improvement for both youth and caregiver reports) by participant (Panel 1) and caregiver (Panel 2). Trajectories for the other quartiles are included in supplementary material.

For the caregiver reports, the best-fitting model was characterized by a linear change trend with a first-order heterogeneous autoregressive (AR1H) covariance structure. Participants in Quartile 1 on average began treatment with an SFSS score of 41.23 (*SE* = 0.50) and decreased a -0.13 (*SE* = 0.07) per session, a decrease that approached statistical significance (see Table 3). The decreases for other quartiles were statistically significant (Quartile 2: Intercept = 48.14 [*SE* = 0.44], Slope = -0.16 [*SE* = 0.07]; Quartile 3: Intercept = 53.66 [*SE* = 0.39], Slope = -0.30 [*SE* = 0.05]; Quartile 4: Intercept = 61.69 [*SE* = 0.50], Slope = -0.39 [*SE* = 0.08]). Analysis of random effects by quartile indicate that a substantial amount of variance was due to between-participant differences (Quartile 1: 35%, Quartile 2: 26%, Quartile 3: 30%, Quartile 4: 28%; see Table 3). See Figure 1 for a graphical depiction of change trajectories for Quartile 4 (Panel 2).

Results for logistic regression models by session number, odds ratios with corresponding 95% confidence intervals, and positive and negative predictive values (PPV and NPV) for both youth- and caregiver-report data are given in Table 4. As demonstrated by the odds ratios in the table, predictive accuracy of NOT status was modest toward the beginning of therapy for both groups of reporters, but began to show consistency by approximately session 8 with fairly robust effects, as demonstrated by the odds ratios, and remaining so through session 18, at which point data began to become much sparser due to participants ending treatment. The R-based dashboard *shinyDLRs* (Authors masked, 2022) was used to derive positive (PPV) and negative predictive values (NPV) for each of the logistic regressions. PPVs were generally larger than NPVs indicating that the models were more effective in identifying participants at risk for treatment failure (e.g., NOT cases that eventually have a poor treatment outcomes) rather than those who were likely to be responders. Caregiver results were consistent with youth reports, although there was not a consistent pattern of statistically significant results, presumably due to reduced power from smaller sample sizes. In contrast with the youth reports, the NPVs were larger than PPVs for the caregivers, indicating that this model was better at identifying responders (rather than cases at risk of treatment failure).

Discussion

The present study aimed to develop an empirically derived method to monitor treatment response and to provide data-driven feedback to youth mental health care providers. This is a rare example of data-driven feedback methods for youth and child mental health services, as most feedback systems that use expected treatment response curves have been developed in the context of university counselling centers (e.g., Finch et al. 2001) or adult psychotherapy services (e.g., Authors Masked, 2017; Lueger et al., 2001; Lutz et al., 2019). Furthermore, the majority of

these ETR models have been evaluated clinically, as a tool to support feedback-informed therapy (e.g., see meta-analysis by de Jong et al., 2021). However, there are few examples where the prediction accuracy of such models have been evaluated in statistically independent samples, using contemporary methods based on cross-validation (e.g., Authors Masked, 2021; Lutz et al., 2019). The present results indicate that growth curve equations for the SFSS measure fit the data well and predicted statistically significant changes over time. The resulting ETR curves enabled the development of a classification algorithm to identify cases at risk of treatment failure (i.e., NOT cases).

Using a cross-validation approach, we found that the classification algorithm had adequate out-of-sample accuracy, with odds ratios in the range of 0.55 – 6.80 for the youth-reported version and 0.71 – 0.80 for the caregiver-reported version. We note that the youth-reported models were more stable in showing statistically significant odds ratios at various sessions/stages of treatment. The statistical significance of these classifiers is related to the sample size and event base rate (e.g., percentage of cases who attain remission of symptoms at the end of therapy), which partly explains the variability observed in Table 4. However, the caregiver-reported model was less stable possibly due to weakened statistical power resulting from a smaller sample.

As shown in previous evaluations of ETR models (e.g., Authors Masked, 2021), the prediction accuracy of these models tends to be modest during the initial sessions of treatment, but it improves considerably around session 8 and stabilizes thereafter. This pattern is consistent with the wider literature on the dose-response effect in psychotherapy, which indicates that treatment would be expected to show reliable improvements by that stage of treatment (see review by Robinson et al., 2020). The model's statistical significance was less stable after

session 18, presumably as a function of unsuitably small sample sizes remaining in the subsequent subsamples.

The fact that caregiver-reported data was better at identifying treatment responders rather than at-risk cases raises some intriguing hypotheses for future research. From a clinical perspective, parents typically seek services for their adolescent children and often have a more targeted vision for what they would like to see happen in treatment. Consequently, they may introduce specific symptoms from the beginning sessions. Youth, on the other hand, frequently do not voluntarily present to treatment, at least initially. Therefore, the reports of youth may not be consistent toward the beginning of services and may solidify after several sessions due to increased comfort, gradually obtaining insight into their symptoms, etc. It is also possible that adolescents' tendencies to be present-focused in their orientations toward their lives may lead to more variable week-to-week reports of symptoms. These hypotheses should be tested in future research.

An additional area for future research comes from the levels of missing data being much higher with caregiver reports (e.g., Table 4). These data suggest interesting possibilities for future research. In research studies of measurement-based care (MBC) comprising caregiver reports of adolescent symptoms, parents need to be present and involved in sessions. This introduces a situation in which clinicians naturally have more access to youth. The scenario presents challenges to researchers in obtaining caregiver-reported data. Further, many clinicians are more inclined to deal individually with youth as opposed to engaging with families (Baker-Ericzén, M. J., Jenkins, M. M., & Haine-Schlagel; Walsh, 2016). As a result, parent involvement in treatment may be more sporadic for clinicians with limited comfort with a family-based treatment approach. Clinical situations such as these give rise to more caregiver missing data in a

research context. Future MBC implementation efforts should consider how best to engage caregivers in research when they are not as directly and frequently engaged in care.

Finally, the sensitivity of the results in the present study did not achieve the standards set in previous research using other measures, notably research on the YOQ (Cannon et al., 2010). This is most likely due to methodological differences between the studies. For instance, Cannon et al. report results from a sample of close to 1000 more youth and used a modeling approach that integrated scores across youth and caregivers. Yet, differences in the results between Cannon et al. and the present study may also indicate may also indicate that the YOQ is, in fact, a more sensitive instrument. More research is needed prior to drawing firm conclusions. Future studies comparing different MBC tools within the same study would be very useful in elucidating the relative strengths and weaknesses of different approaches to outcome measurement.

Strengths and Limitations

The relatively large, pooled sample from multiple sources enabled a rigorous external cross-validation approach to develop and evaluate clinical prediction models. External validation of clinical prediction models is considered a hallmark of model credibility, and has been suggested to be the minimum standard of evidence before deploying such models in clinical care or prospective clinical trials (Authors Masked, 2022). We also followed methodological guidelines and evaluation procedures previously used in the field of psychotherapy (Authors Masked, 2021), to aid interpretability and comparison across studies.

Despite these strengths, the study also had some limitations. Although the pooled sample in some ways represented a strength of the present study, at the same time, it inherently injects a potential weakness, That is, the samples included in the present study may not be generalizable

to the wider population of treatment-seeking adolescents with mental health problems. However, in our estimation, the present study represents an important step toward the development of feedback models that would be clinically useful for clients treated in the current described settings, and potentially beyond them as well. The extent to which the ETR models developed in the present study generalize to populations beyond the current samples remains an area for future research.

The sample size diminished below the minimal required sample size (~n=132) in subgroups of cases with treatments that lasted more than 11 sessions, making the evaluation of prediction accuracy less certain in these subgroups. Furthermore, the base rate of reliable improvement differed between the randomly selected training and test samples. This may have resulted in more difficult-to-treat cases being represented in the test sample, given that there were no statistically significant differences in baseline severity between these samples. Nevertheless, this allowed for a more rigorous external validation of the trained classification algorithm, since we would expect that the base rate of improvement would vary in different clinical contexts/settings where such an algorithm would be implemented. Despite the possibility of more difficult-to-treat cases residing in the test sample, the test and training samples did not significantly differ in baseline symptom severity. This implies that other clinically-relevant variables are not sufficiently being captured in the initial grouping process. Future studies should incorporate larger samples and capture a richer representation of clinically-relevant background characteristics to further illuminate which characteristics are most important to clinically significant improvement as well as fine-tuning the model with respect to specific types of symptoms (e.g., internalizing and externalizing behaviors).

Conclusions

This study presents evidence that an empirically-derived outcome monitoring and feedback model based using the SFSS measure shows adequate generalizability to a statistically independent test sample, leading to accurate and stable classifications from session 8 approximately 20 or more sessions at which point data limitations prevented accurate investigation of improvement. The youth-reported version of the algorithm was more accurate and stable than the caregiver-reported version, and hence may be considered for implementation in clinical care with refinements being made in future research as suggested above. Future studies should also look to validating the SFSS and similar measures among caregivers.

References

Achenbach, T. M. (1991). *Integrative guide for the 1991 CBCL/4-18, YSR, and TRF profiles.* Department of Psychiatry, University of Vermont.

American Psychiatric Association. (2022). *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5-TR). American Psychiatric Association Publishing. https://doi.org/10.1176/appi.books.9780890425787

American Psychological Association Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285.

Athay, M. M., Riemer, M., & Bickman, L. (2012). The Symptoms and Functioning Severity Scale (SFSS): Psychometric Evaluation and Discrepancies Among Youth, Caregiver, and Clinician Ratings Over Time. *Administration and Policy in Mental Health and Mental Health Services Research*, *39*(1–2), 13–29. https://doi.org/10.1007/s10488-012-0403-2

Baker-Ericzén, M. J., Jenkins, M. M., & Haine-Schlagel, R. (2013). Therapist, parent, and youth perspectives of treatment barriers to family-focused community outpatient mental health services. *Journal of Child and Family Studies*, *22*, 854-868. https://doi.org/10.1007/s10826-012-9644-7

Bickman, L., Kelley, S. D., Breda, C., de Andrade, A. R., & Riemer, M. (2011). Effects of Routine Feedback to Clinicians on Mental Health Outcomes of Youths: Results of a Randomized Trial. *Psychiatric Services*, *62*(12), 1423–1429. https://doi.org/10.1176/appi.ps.002052011

Bosker, R., & Snijders, T. (2011). *Multilevel analysis: An Introduction to basic and advanced multilevel modeling*. Sage.

Burlingame, G. M., Cox, J. C., Wells, M. G., Lambert, M. J., Latkowski, M., & Ferre, R. (2005). The administration and scoring manual of the Youth Outcome Questionnaire. *Salt Lake City, UT: American Professional Credentialing Services*.

Cannon, J. A. N., Warren, J. S., Nelson, P. L., & Burlingame, G. M. (2010). Change trajectories for the Youth Outcome Questionnaire Self-Report: Identifying youth at risk for treatment failure. *Journal of Clinical Child and Adolescent Psychology*, *39*(3), 289–301. https://doi.org/10.1080/15374411003691727

Coalition for the Advancement and Application of Psychological Science. (2018). *Evidence-based Practice Decision-Making for Mental and Behavioral Health Care*. https://www.caaps.co/consensus-statement

de Jong, K., Conijn, J. M., Gallagher, R. A. V., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, *85*, 102002. https://doi.org/10.1016/j.cpr.2021.102002

Finch, A. E., Lambert, M. J., & Schaalje, B. G. (2001). Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology & Psychotherapy*, *8*(4), 231–242. APA PsycInfo References.

Goodman, R. (1999). The Extended Version of the Strengths and Difficulties Questionnaire as a Guide to Child Psychiatric Caseness and Consequent Burden. *Journal of Child Psychology and Psychiatry*, *40*(5), 791–799. https://doi.org/10.1111/1469-7610.00494

Hawkins, E. J., Lambert, M. J., Vermeersch, D. A., Slade, K. L., & Tuttle, K. C. (2004). The Therapeutic Effects of Providing Patient Progress Information to Therapists and Patients. *Psychotherapy Research*, *14*(3), 308–327. https://doi.org/10.1093/ptr/kph027

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology & Psychotherapy*, *9*(3), 149–164. https://doi.org/10.1002/cpp.333

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy*, *55*(4), 520–537. https://doi.org/10.1037/pst0000167

Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H., Hoffman, M., Scott, K., Lyon, A., Douglas, S., Simon, G., & Kroenke, K. (2019). Implementing Measurement-Based Care in Behavioral Health: A Review. *JAMA Psychiatry*, *76*(3), 324–335. https://doi.org/10.1001/jamapsychiatry.2018.3329

Lueger, R. J., Howard, K. I., Martinovich, Z., Lutz, W., Anderson, E. E., & Grissom, G. (2001). Assessing treatment progress of individual patients using expected treatment response models. *Journal of Consulting and Clinical Psychology*, *69*(2), 150–158. APA PsycInfo. https://doi.org/10.1037/0022-006X.69.2.150

Lutz, W., Lambert, M. J., Harmon, S. C., Tschitsaz, A., Schürch, E., & Stulz, N. (2006). The probability of treatment success, failure, and duration--What can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology and Psychotherapy, 13,* 223-232. https://doi.org/10.1002/cpp.496

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W.B., Evans, C., et al. (2005). Predicting change for individual psychotherapy clients based on their nearest neighbors. Journal of Consulting and Clinical Psychology, 73(5), 904–913. https://psycnet.apa.org/doi/10.1037/0022-006X.73.5.904

Lutz, W., Stulz, N., Martinovich, Z., Leon, S., & Saunders, S. M. (2009). Methodological background of decision rules and feedback tools for outcomes management in psychotherapy. *Psychotherapy Research*, *19*(4–5), 502–510. APA PsycInfo. https://doi.org/10.1080/10503300802688486

Nelson, P. L., Warren, J. S., Gleave, R. L., & Burlingame, G. M. (2013). Youth psychotherapy change trajectories and early warning system accuracy in a managed care setting. *Journal of Clinical Psychology*, *69*(9), 880–895. https://doi.org/10.1002/jclp.21963

Probst, T., Lambert, M. J., Loew, T. H., Dahlbender, R. W., Göllner, R., & Tritt, K. (2013). Feedback on patient progress and clinical support tools for therapists: Improved outcome for patients at risk of treatment failure in psychosomatic in-patient therapy under the conditions of routine practice. *Journal of Psychosomatic Research*, *75*(3), 255–261. https://doi.org/10.1016/j.jpsychores.2013.07.003

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Riemer, M., Kelley, S. D., Casey, S., & Haynes, K. T. (2012). Developing effective research-practice partnerships for creating a culture of evidence-based decision making. Administration and Policy in Mental Health and Mental Health Services Research, 39(4), 248–257. https://doi.org/10.1007/s10488-011-0368-6

Rognstad, K., Wentzel-Larsen, T., Neumer, S.-P., & Kjøbli, J. (2023). A Systematic Review and Meta-Analysis of Measurement Feedback Systems in Treatment for Common Mental Health Disorders. *Administration and Policy in Mental Health and Mental Health Services Research*, *50*(2), 269–282. https://doi.org/10.1007/s10488-022-01236-9

Sale, R., Bearman, S. K., Woo, R., & Baker, N. (2021). Introducing a Measurement Feedback System for Youth Mental Health: Predictors and Impact of Implementation in a Community Agency. *Administration and Policy in Mental Health and Mental Health Services Research*, *48*(2), 327–342. https://doi.org/10.1007/s10488-020-01076-5

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2). https://doi.org/10.1214/aos/1176344136

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *J Consult Clin Psych*, *78*(3), 298–311.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence. Oxford University Press*. USA.

Tam, H. E., & Ronan, K. (2017). The application of a feedback-informed approach in psychological service with youth: Systematic review and meta-analysis. *Clinical Psychology Review*, *55*, 41–55. https://doi.org/10.1016/j.cpr.2017.04.005

Walsh, F. (2016). Applying a family resilience framework in training, practice, and research: Mastering the art of the possible. *Family process*, *55*(4), 616-632. https://doi-org/10.1111/famp.12260

Warren, J. S., Nelson, P. L., & Burlingame, G. M. (2009). Identifying youth at risk for treatment failure in outpatient community mental health services. *Journal of Child and Family Studies*, *18*(6), 690–701. https://doi.org/10.1007/s10826-009-9275-9

Warren, J. S., Nelson, P. L., Mondragon, S. A., Baldwin, S. A., & Burlingame, G. M. (2010). Youth psychotherapy change trajectories and outcomes in usual care: Community mental health versus managed care settings. *Journal of Consulting and Clinical Psychology*, *78*(2), 144–155. https://doi.org/10.1037/a0018544

Table 1. Sample Characteristics

	Training Sample	Test Sample
Demographic Characteristics		
Age, years	13.11 (3.20)	13.37 (3.11)
Sex		
Male	377 (47.8)	328 (44.4)
Female	411 (52.1)	409 (55.4)
Other	1 (0.1)	1 (0.1)
Clinical Characteristics		
Baseline SFSS Score (Youth Report)	52.43 (9.09)	52.45 (9.20)
Baseline SFSS Score (CG Report)	51.41 (9.03)	50.60 (9.31)
Treatment Sessions	6.02 (3.61)	4.88 (4.17)

Table 2. Parameter Estimates for Multilevel Model with Youth-Reported SFSS Serving as Criterion by Quartile for Training Sample

	Quartile 1			
Fixed Effects	Intercept		Slope	
	Estimate	SE	Estimate	SE
ln_session	42.31***	0.35	-0.50*	0.25
	Intercept		ln_session	
Random Effects	7.33***	1.86	4.89***	0.95
	Quartile 2			
Fixed Effects	Intercept		Slope	
	Estimate	SE	Estimate	SE
ln_session	50.44***	0.31	-1.46***	0.24
	Intercept		ln_session	
Random Effects	3.76**	1.41	5.17***	0.90
	Quartile 3			
Fixed Effects	Intercept		Slope	
	Estimate	SE	Estimate	SE
ln_session	56.19***	0.37	-2.59***	0.32
	Intercept		ln_session	
Random Effects	1.99	1.63	8.23***	1.35
Fixed Effects	Quartile 4			
	Estimate	SE	Estimate	SE
ln_session	63.72***	0.45	-3.89***	0.32
	Intecept		ln_session	
Random Effects	12.36***	2.87	8.51***	1.48

*$p < .05$, **$p < .01$, ***$p < .001$

Table 3. Parameter Estimates for Multilevel Model with Caregiver-Reported SFSS Serving as Criterion by Quartile for Training Sample

	Quartile 1			
Fixed Effects	Intercept		Slope	
	Estimate	SE	Estimate	SE
Session	41.23***	0.50	-0.13*	0.07
	Intercept		Session	
Random Effects	12.48***	3.43	0.05	0.04
	Quartile 2			
Fixed Effects	Intercept		Slope	
	Estimate	SE	Estimate	SE
Session	48.14***	0.44	-0.16*	0.07
	Intercept		Session	
Random Effects	6.44*	2.84	0.14*	0.06
	Quartile 3			
Fixed Effects	Intercept		Slope	
	Estimate	SE	Estimate	SE
Session	53.66***	0.39	-0.30***	0.05
	Intercept		Session	
Random Effects	8.44***	2.50	0.04	0.03
Fixed Effects	Quartile 4			
	Estimate	SE	Estimate	SE
Session	61.69***	0.50	-0.39***	0.08
	Intecept		Session	
Random Effects	10.08**	3.52	0.17*	0.06

*$p < .05$, **$p < .01$, ***$p < .001$

Table 4. Sample Size, Odds Ratios with 95% Confidence Intervals, and Positive and Negative Predictive Values for Logistic Regressions Predicting Youth Sumptom Improvement by NOT Status by Week in Treatment.

	Youth-Reported SFSS					Caregiver-Reported SFSS				
Session	N	OR	95% CI	PPV	NPV	N	OR	95% CI	PPV	NPV
1	468	0.55	0.28 - 1.08	N/A	N/A	235	1.24	0.72 - 2.13	N/A	N/A
2	343	1.10	0.61 - 1.97	0.63	0.46	92	0.81	0.30 - 2.18	0.25	0.77
3	348	2.83***	1.59 - 5.06	0.72	0.53	101	3.72**	1.51 - 9.17	0.19	0.56
4	290	1.11	0.62 - 1.97	0.58	0.53	77	2.39	0.84 - 6.76	0.17	0.67
5	267	2.23*	1.19 - 4.18	0.67	0.58	72	0.71	0.23 - 2.18	0.29	0.79
6	239	1.84	0.98 - 3.46	0.64	0.56	67	0.73	0.25 - 2.13	0.29	0.68
7	203	1.63	0.83 - 3.18	0.56	0.6	47	1.48	0.39 - 5.63	N/A	N/A
8	188	2.36*	1.18 - 4.74	0.62	0.59	47	6.06*	1.35 - 27.29	0.11	0.57
9	181	2.22*	1.13 - 4.38	0.62	0.6	50	1.23	0.37 - 4.06	N/A	N/A
10	140	1.30	0.62 - 2.74	0.57	0.55	31	3.33	0.66 - 16.76	0.19	0.58
11	138	3.59**	1.55 - 8.31	0.72	0.6	32	6.33*	1.11 – 36.00	0.12	0.54
12	104	4.18*	1.36 - 12.79	0.76	0.58	28	3.90	0.76 - 19.95	0.21	0.57
13	101	3.80**	1.53 - 9.43	0.71	0.69	32	2.10	0.49 – 9.00	0.32	0.53
14	90	5.44**	1.84 - 16.10	0.67	0.69	26	2.29	0.44 - 11.92	0.27	0.55
15	82	2.83*	1.10 - 7.29	0.69	0.65	27	3.75	0.58 - 24.28	0.13	0.62
16	72	3.14*	1.04 - 9.48	0.65	0.68	22	0.46	0.07 - 3.14	0.36	0.78
17	61	6.80**	1.94 - 23.90	0.75	0.7	16	2.63	0.30 – 23.00	N/A	N/A
18	55	3.57*	1.11 - 11.48	0.62	0.63	23	2.72	0.48 - 15.47	0.30	0.50
19	45	2.21	0.65 - 7.54	0.61	0.65	17	3.50	0.43 - 28.45	0.27	0.60
20	40	2.62	0.61 - 11.28	0.50	0.74	13	8.00	0.46 - 139.29	0.10	0.40
21	35	1.42	0.24 - 8.26	0.57	0.62	15	2.22	0.28 - 17.63	0.43	0.44
22	37	2.25	0.49 - 10.34	0.55	0.61	14	N/A	N/A	N/A	N/A

Note. Values for weeks 23 through 30 are not included due to invalid estimates produced due to limited sample size for participants receiving more than 22 treatment sessions.
CI = Confidence Interval, OR = Odds Ratio, NOT = Not on Track, NPV = Negative Predictive Value, PPV = Positive Predictive Value

Figure 1. Expected treatment response (ETR) curves by week with 90% confidence intervals for youth- (Panel 1) and caregiver-reported SFSS scores (Panel 2) for cases in Quartile 4.