

## Accuracy of a Rationally Derived Method for Identifying Treatment Failure in Children and Adolescents

Matthew J. Bishop,<sup>1</sup> Taige S. Bybee,<sup>1</sup> Michael J. Lambert, Ph.D.,<sup>2,3</sup>  
Gary M. Burlingame, Ph.D.,<sup>2</sup> M. Gawain Wells, Ph.D.,<sup>2</sup>  
and Landon E. Poppleton<sup>1</sup>

---

*Psychotherapy outcome can be enhanced by early identification of potential treatment failures before they leave treatment. In adults, compelling data are emerging that provide evidence that an early warning system that identifies potential treatment failures can be developed and applied to enhance outcome. The present study reports an analysis of early warning algorithms to identify treatment failures among child/adolescent patients (ages 3–18). The progress of 300 patients who had completed treatment was analyzed to see if algorithms could identify those children who ultimately had a negative outcome. Results indicated that the rationally derived method had a 77% success rate for identifying child/adolescent patients who were reliably worse or had deteriorated by the time that therapy was terminated.*

---

**KEY WORDS:** treatment failure; deterioration; youth outcome questionnaire; outcome monitoring; decision support tools.

Inasmuch as psychotherapy outcome research for children and adolescents has been considered as less mature than the research in adult outcome (Kazdin, 1995) this study sought to expand methodology used with adults to research in child outcome by analyzing the accuracy of a rational method which purports to identify child patients who are at risk for treatment failure. While a majority of patients who undergo a course of psychotherapy have a positive outcome, a significant

<sup>1</sup>Clinical Psychology Trainee, Department of Psychology, Brigham Young University, Provo, UT.

<sup>2</sup>Professor, Department of Psychology, Brigham Young University, Provo, UT.

<sup>3</sup>Correspondence should be addressed to Michael J. Lambert, Department of Psychology, 272 TLRB, Brigham Young University, Provo, Utah 84602; e-mail: Michael.Lambert@byu.edu.

minority of patients fail to improve or actually deteriorate while in treatment (Lambert & Bergin, 1994; Smith, Glass, & Miller, 1980). This general finding and the advent of managed care organizations have led to the development of quality management efforts aimed at improving the quality of behavioral health services. Often case-management procedures within these organizations are intended to address the failing patient and to enhance the likelihood of a positive outcome (Lambert, Huefner, & Reisinger, 2000).

Recent developments in outcome management for psychotherapy favor the continuous monitoring of patient treatment response to decide if a treatment is adequate for a particular patient (Kordy, Hannover, & Richard, 2001; Lambert, Hansen, & Finch, 2001; Lueger et al., 2001). Clinicians routinely but informally monitor treatment progress, making adjustments in their behavior in accordance with theoretical considerations, the needs and characteristics of their patients, as well as their response to treatment. Yet their ability to make accurate prognostic assessments even late in therapy has been called into question (Breslin, Sobell, Buchan, & Cunningham, 1997; Meyer & Schulte, 2002) especially with patients who show deterioration (Lambert & Bergin, 1994). Decision support tools have the possibility of assisting both the clinician and case manager as they attempt to identify patients whose progress is in doubt. With such cases a stepped-care approach in which the failing patient is stepped up to more intensive treatment may be indicated. Alternatively, the clinician may use clinical judgment without the assistance of formal methods, but in cases involving a failing patient the clinician may act more out of resignation than objective judgment (Schulte-Bahrenberg & Schulte, 1991; 1993). The case manager on the other hand, does not typically know the patient and is in need of norm-based assessments in order to support or enhance the practitioner's treatment efforts.

Quality management efforts with adults have been informed by research on patterns of change over the course of therapy and evidence that early response to therapy predicts final outcome. For example, Richard and Kordy (2002) found four distinct patterns of change in the course of treatment for bulimics, two suggesting a positive final outcome and two suggesting treatment failure. The patterns were recognizable by the fourth week of treatment. Tang and DeRubies (1999a, 1999b) found rapid improvement in depressive symptoms foretold better ultimate outcome and follow-up functioning than slow improvement. Haas, Hill, Lambert, and Morrill (2002) found the same trend in a group of patients with a variety of disorders. Positive treatment response within the first three treatment sessions occurred in the majority of clients who improved by the end of treatment and maintained gains six months to two years after termination. Wilson (1999), commenting on early response to treatment, suggested the clear practical consequence of early poor response and the possibility of modifying treatment prior to termination for these patients. Research in this area appears promising for case management and clinical practice, as patterns of treatment response can lead to early identification of particular patients who are in need of special treatment efforts.

Lambert, Whipple, Bishop, et al. (2002a) examined methods for identifying early non-responders to treatment in a large sample of adult outpatients. Examination of a rationally derived method revealed an 81% success rate for identifying patients who were reliably worse or had deteriorated by termination, and a 19% miss rate (false positives). The methodology that was used allowed for determination of possible treatment failures after only two sessions.

Lambert, Whipple, Smart, et al. (2001b) implemented this decision support methodology with adults using the Outcome Questionnaire-45 (OQ-45) to see if alerting clinicians to possible treatment failure would enhance patient outcome compared to treatment as usual. They reported that early identification of poorly responding patients and feedback to therapists about patient treatment improved outcome at termination. Two replications of this study supported this finding (Lambert, Whipple, Vermeersch, et al., 2002b; Whipple, et al., 2003). On average, all three studies obtained a treatment effect size of .40 and reduced deterioration for predicted failures (signal-alarm cases) from 21% in the treatment as usual control to 13% in the feedback condition. This research suggests a substantial impact of early identification of potential treatment failures based on the use of rationally derived algorithms.

Similar research with child or adolescent clients has not been reported in the literature. However, this same research group has developed algorithms for use with these younger age groups.

The current study was intended to analyze a rationally derived method, as implemented with the Youth Outcome Questionnaire (YOQ) (Burlingame et al., 1996), for identifying cases for review. The major questions addressed were: How well does the rational method predict treatment failure? Is there a tendency to over or under identify more or less disturbed patients as signal cases? These questions are seen as important because poorly responding cases, which tend to drop out of treatment early, must be identified as soon as possible if modifications to ongoing treatment are to be made. In addition, it is important to be especially careful to identify patients who are more disturbed for review because their failure to find relief coupled with their high level of disturbance places them at special risk for suicide, and severe difficulties in family relationships and school.

## METHOD

### Participants

The 300 patients in this study came from an archival repository of data regarding 4,560 youth who received therapy at both residential and non-residential settings (cf. Burlingame, Mosier, Wells, et al., 2001). Patients in the repository had received routine clinical services from a large multi-state western health care system that covers approximately 1 million lives in the intermountain west.

Specifically, all parents of children and adolescents who presented for treatment at three outpatient facilities completed the YOQ as part of their initial screening and at subsequent sessions. This sampling strategy—mall-intercept/exit poll—has a long history of use in the survey literature and is associated with producing samples with high ecological validity (Sapsford, 1999). Ecological validity is of particular interest in estimating treatment “effectiveness,” in naturalistic settings (Lyons, Howard, O’Mahoney, & Lish, 1997).

Criteria for case inclusion in the present study were: (a) available dates and scores for the repeated administration of the YOQ; (b) 14 days or less between any two consecutive administrations to ensure a valid sequential measure of distress in the children’s lives during the course of their therapy and; (c) no more than one unaccounted for therapy session between consecutive administrations of the YOQ; (d) signed an informed consent for resulting data to be used in research. These criteria yielded a sample of 300 patients representing 6.6% of the original patient pool. A comparison of this 6.6% sample was made with the larger data repository on a variety of variables with results providing no significant differences between those clients included in the present analysis and the larger sample

The sample consisted of 145 patients who received therapy in a residential setting from September 1996 to December 1998 and 155 who received therapy at a non-residential setting from February 1997 to January 2001. The near equal partitioning of the sample between residential and non-residential settings was coincidental and not pre-determined. Approximately 44% (133) of the patients were female. Patient racial profile was not recorded in the original data set. The mean age was 12.2 years, with a standard deviation of 3.8 years and a range of between 3 and 18 years.

### Measure

The YOQ (Burlingame et al., 1996) was designed to measure patient progress in therapy by repeated administration during the course of treatment and at termination. It was originally conceived as the child and adolescent equivalent of the OQ-45.2 (Lambert et al., 1996) and is a 64-item parent or guardian-report measure of functioning for children and adolescents, ranging in age from 3–18. The YOQ is specifically constructed to assess the occurrence of observed behavior change, and is not a diagnostic tool to diagnose specific psychopathology. Parents or others with reasonably extensive interaction with the client (e.g., psych-technicians and counselors in residential settings) complete the questionnaire at intake to establish a symptom severity baseline and then complete it repeatedly at regular intervals to track the child’s progress.

The YOQ is based on five domains that have been shown to be sensitive to change (Burlingame, Wells, Lambert, & Cox, 2003; Wells, Burlingame, & Lambert, 1999) in children and adolescents receiving psychosocial treatments:

(1) Interpersonal distress (ID), e.g., anxiety, depression, hopelessness and self harm; (2) Somatic distress (S), e.g., headaches, dizziness, stomach aches, nausea, bowel difficulties and pain or weakness in joints; (3) Interpersonal relationships (IR), e.g., attitude towards others, communication, interaction with friends, cooperativeness, aggressiveness, arguing and defiance; (4) Social problems (SP) items describe delinquent and aggressive behaviors (more severe aggression than that assessed on the IR scale) such as truancy, sexual problems, running away from home, destruction of property and substance abuse; (5) Behavioral dysfunction (BD), assesses ability to organize tasks, complete assignments, concentrate, and handle frustration, inattention, hyperactivity and impulsivity. A final subscale reflects critical items (CI) that are important for inpatient short term stabilization e.g., paranoia, obsessive-compulsive behaviors, hallucinations, delusions, suicide, mania, and eating disorder issues. A high score on any critical item in an outpatient setting indicates the need for immediate intervention.

A parent or guardian rates each question on a 5-point Likert scale (0 = never, 1 = rarely, 2 = sometimes, 3 = frequently, 4 = almost always). Notably, seven questions assess positive behaviors and are accordingly given negative weights yielding a total score ranging from -16 to 240 with higher scores indicating the endorsement of more severe distress and overall pathology. Since the total score tends to be the most reliable index of change (Burlingame, Cox, Wells, & Lambert, 2003; Burlingame et al., 1996, 2001), it was used in the present study.

Burlingame et al. (2001) reported high internal consistency for the YOQ total score ( $r = .97$ ) across four samples consisting of elementary school students ( $N = 423$ ), a community normative sample ( $N = 681$ ), outpatient  $N = 342$ ) and inpatient ( $N = 174$ ) and a clinical normative sample ( $N = 490$ ). Criterion-related validity is supported by a strong relationship ( $r = .84$ ) between the total score of the Child Behavior Checklist (CBCL; Achenbach, 1991) and the YOQ total score.

Using formulas developed by Jacobson and Truax (1991) the YOQ provides two cutoff scores for assessing patient change: Reliable Change Index (RCI) and clinically significant change. Patients who change in a positive or negative direction by at least 13 points are regarded as having made reliable change; i.e., their change exceeds measurement error. The second criterion proposed by Jacobson and Truax is to identify a cut score, to separate dysfunctional and functional populations. This was estimated to be 46 which lies a little more than halfway between the community mean of 23.2 ( $N = 683$ ) and the outpatient mean of 78.6 ( $N = 342$ ). When a patient's score falls below 46 their functioning is presumed to resemble non-patient community normals (Burlingame et al., 1996).

## Procedures

Prior to initiation of the current study and independent from data used in the current study, information about the importance of early response to treatment, the

dose response relationship and clinically significant change was used by three of the investigators (G.M.B., M.J.L., and M.G.W.) to create algorithms for identifying patients who they predicted would leave treatment before receiving therapeutic benefit or who they thought to be at risk for having a negative treatment outcome. For simplicity of communication in the clinical setting, those so identified are referred to as signal-alarm cases. This is a term that has precedence in other research aimed at improving the quality of patient care (Kordy, Hannover, & Richard, 2001). The three investigators created a matrix for identification of signal cases (See Figure 1). The vertical axis was used to plot intake YOQ scores (ranging from -16 to 240) and the horizontal axis was used to plot change scores at the sessions of interest.

Inasmuch as calculation of session-by-session progress for decision-making proved to be unwieldy—necessitating 20 or more separate matrices—sessions were lumped into groups, three aggregate matrices representing change occurring between sessions two through four, five through eight, and nine and above. Presumably, insufficient change after two to four session of treatment would be less alarming as the same lack of progress occurring after five to eight or more sessions.

The three investigators then constructed specific scenarios with the intent of classifying specific patients into four classifications based on colored signals suggested by Lambert et al. (2001a). A green category signified that a patient had made significant progress while the yellow category signified insufficient

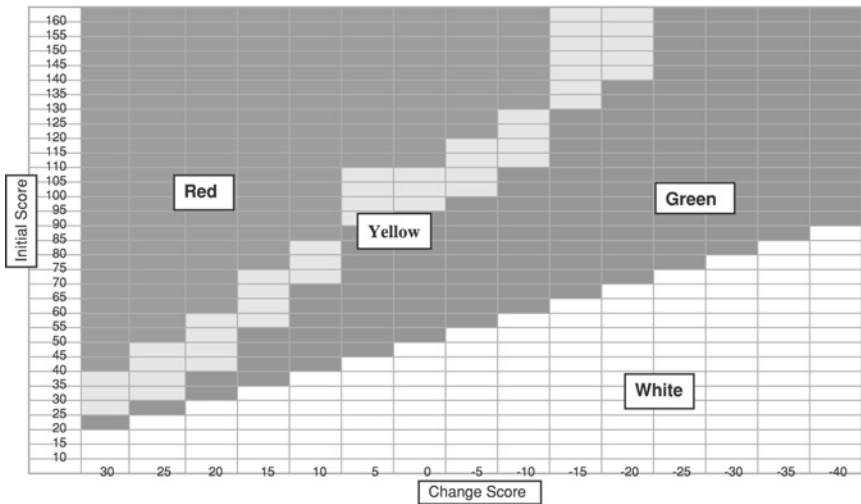


Fig. 1. Decision matrix for identifying at risk case at session 2–4.

patient progress such that the therapist was cautioned about possible treatment failure. A red category signified serious concern about a treatment failure and the need to signal the therapist to seriously reconsider the ongoing treatment regimen. Patients who get either a red or yellow signal at any time during treatment are flagged as signal alarm cases, i.e., cases that are predicted to deteriorate. Finally, a white category informed the therapist that the patient's functioning had fallen within the normal range (i.e., below 46). The authors repeated the same construction procedures for the imagined patient who had had five through eight sessions of treatment, and, again, for patients who had nine or more sessions of treatment.

More specifically, the three raters imagined a patient who started at a specific YOQ score, then imagined the patient worsening by 1 point, 2 points, 3 points, etc. At a given point, depending on the patient's intake score, a degree of worsening would become large enough that the raters reached a consensus that a yellow or red signal was warranted. Movement in a positive direction was given a white categorization if the patient's change score moved into the functional range (e.g., below a score of 46 on the YOQ).

Some examples of the procedures will help clarify the way cut-off points were devised. A patient could be imagined who began treatment with a YOQ score of 79. This score is an average score for patients entering treatment in outpatient clinics. The raters considered the patient's worsening after having two to four sessions of treatment, 1 point, 2 points, 3 points, etc., and agreed that if the patient had worsened by 11 points over two to four sessions a yellow warning was appropriate, and that getting worse by 15 points would qualify him or her for a red warning.

In contrast a patient with a score of 100 about equal to the mean inpatient score, is likely to be experiencing severe distress and a strong need for relief. Here the raters agreed that an improvement of less than 5 points after 2–4 sessions of treatment should signal a yellow caution, whereas worsening by 11 or more points should result in a red signal. A green signal categorization was given if the patient improved by 5 points or more.

Similar decisions were made to form the matrices for sessions 5 to 8, and 9 and above. The rules for classification for these matrices were based on the hypothesized meaning of changing or not changing after a higher dosage of treatment. For example, a patient who starts treatment with a score of 85 on the YOQ and does not change after two treatment sessions was not considered to be as alarming as this patient's same failure to change after nine treatment sessions. Although algorithms such as those tested in the current study may be less reliable than those based on purely statistical methods, earlier work with adults suggested high agreement between statistically derived and rationally derived methods (Lambert et al., 2002a). It is unlikely that judges who lack familiarity with both the psychometrics of the YOQ and its clinical use

would replicate the algorithms that were the subject of investigation in the current study.

The classification of patient progress based on the above rational algorithm was applied to every YOQ score for every patient ( $N = 300$ ). The YOQ was filled out for each child before their therapy session. Their data were stored electronically and was not shared with the therapist or client.

For the purpose of this study the dependent variable was reliable worsening, defined as a 13 point increase in the patient's total YOQ score from intake to termination. In the case of patients that started therapy in the functional range (i.e., below 46 on the YOQ), in addition to getting worse by 13 points, they also had to leave therapy with a final score of 46 or higher, placing them in the dysfunctional range, before they were considered to have reliably deteriorated.

## RESULTS

The average pretreatment YOQ score was 87.5, with a standard deviation of 38.4 which falls between the typical outpatient and inpatient means (Burlingame et al., 2001). As presented in Table I, 22 of the 300 patients who entered treatment (7.3%) were reliably worse or deteriorated at termination. The majority of these deteriorators, (17 or 77.3%) were correctly identified (were signal-alarm cases) using a second administration of the YOQ and the aforementioned rational method of identification. The same classification strategy correctly identified 81.7% of those clients who ultimately ended treatment with a positive outcome. Thus,

**Table I.** Success of Predicting Deterioration based on Rationally-Derived Algorithms for the Youth Outcome Questionnaire

	Predicted outcome				Total	
	Positive outcome		Negative outcome			
Actual outcome	<i>n</i>	(%)	<i>n</i>	(%)	<i>n</i>	(%)
Positive	227 <sup>a</sup>	(81.7)	51 <sup>b</sup>	(18.3)	278	(92.7)
	HIT		False-Alarm			
Negative	5	(22.7)	17	(77.3)	22	(7.3)
	False-Positive		HIT			
Total	232	(77.3)	68	(22.7)	300	(100)

*Note.* Predicted Positive Outcome includes both white and green signals given after 2nd YOQ score. Predicted Negative Outcome includes both yellow and red signals given after 2nd YOQ score. Actual Positive Outcome includes Reliably Recovered, Reliably Better, and No-Reliable-Change. Actual Negative Outcome includes both Reliably Worse and Reliably Deteriorated.

<sup>a</sup>Includes 78 patients who ended treatment with no reliable change.

<sup>b</sup>Includes 33 patients who ended treatment with no reliable change.

the overall accuracy (hit rate) for the rational method was 81.3%; i.e., the final outcome for 244 patients was correctly predicted. It should be noted that this is not a particularly impressive overall hit rate, but since the goal of the algorithms was to identify deteriorators, the 77% hit rate far surpassed chance prediction based on the base rate of treatment failure (7.2%).

Consideration of the unsuccessful predictions produced by the rational method is also instructive. Of the misidentified cases, 51 patients (18%) were identified as signal-alarm cases or cases in need of review but who ultimately did not have a negative treatment outcome (false positives). Five of the 22 cases (22.7%) that had an actual negative outcome were also misidentified by the algorithms (false negatives). Several strategies were used to explore algorithm accuracy. The first involved examining the initial distress of patients. More specifically, the 22 patients who actually got worse were subcategorized into five groups based upon their level of disturbance at intake. The first two groupings were composed of patients whose initial intake scores fell at or below the cut score of 46 demarcating the functional and dysfunctional populations: Group A scores ranged from -16 to 23) while Group B scores ranged from one point above the non-patient sample 24 to 45. Group C included patients whose initial intake score ranged from the cut score to the mean of the normative outpatient sample (46 to 79). Group D included scores above the outpatient mean up to and including the inpatient sample mean (80-100) while group E included patients whose initial YOQ score was higher than the normative inpatient mean (above 100).

All of group A were incorrectly predicted to have positive outcomes, and all of groups D and E were correctly predicted to have deteriorated outcomes. Among the six clients in Group B, four (66.7%) were correctly predicted to have a deteriorated outcome. Among the four clients in Group C, three out of four (75%) were correctly predicted to have a deteriorated outcome. Thus, the method appeared to be reasonably accurate (75% or better) in detecting signal cases for all levels of initial disturbance except for patients that initially scored below the cut-off for being dysfunctional (groups A and B).

We also reviewed the five cases that were not identified as signal-alarm cases but ended treatment as deteriorators. One case had a dramatic negative change recorded at termination, but had demonstrated improvement up until the final YOQ administration. This patient's treatment failure could not be predicted from the algorithms and analysis of this case did not provide information about the need for modification of the decision rule cut-offs. (One could speculate from the data the occurrence of a strong, perhaps external event in the patient's life which also truncated treatment). In contrast, the other four cases began treatment in the functional range and showed reliable negative change with only one showing deterioration. All four cases missed being identified as signal-alarm cases during treatment but by just a few points. Slight modification to the existing rules (changing four green

**Table II.** Actual Outcome of 68 Cases Predicted as Treatment Failures

Actual outcome	Yellow signal		Red signal		Total	
	<i>n</i>	(%)	<i>n</i>	(%)	<i>n</i>	(%)
Reliably Recovered –13 & crossed 46	1	(25)	3	(75)	4	(5.9)
	False-Alarm		False-Alarm			
Reliably Better –13	7	(50)	7	(50)	14	(20.6)
	False-Alarm		False-Alarm			
No Reliable Change –12 to +12	17	(51.5)	16	(48.5)	33	(48.5)
	False-Alarm		False-Alarm			
Reliably Worse +13	1	(7.7)	12	(92.3)	13	(19.1)
	HIT		HIT			
Reliably Deteriorated +13 & crossed 46	3	(75)	1	(25)	4	(5.9)
	HIT		HIT			
Total Number Classified	29	(42.7)	39	(57.3)	68	(100)

messages to yellow) would have lead to more accurate predictions about these patients' final treatment status.

Table II presents data on the outcome of signal-alarm cases. As can be seen, of the 68 cases classified by the rational method as signal-alarms (red or yellow signals), only 18 (26.5%) improved or recovered at termination, whereas 33 (48.5%) showed no reliable change, and 17 (25%) actually did get worse. In contrast, of the 232 clients who were not categorized as signal-alarms (i.e., clients with only white or green signals), 149 (64%) reliably improved or recovered (see Table III). These data suggest that the outcome for patients identified as signal-alarms was not particularly good, even when deterioration was not the sole criterion. The most common outcome for signal-alarm cases is no reliable change or outright worsening (Lambert et al., 2001a).

A Chi Square analysis showed that there was a statistically significant difference in the accuracy of overall predictions according to whether the patients under consideration came from a residential or non-residential setting,  $\chi^2(1, N = 300) = 7.0112, p = .01$ . As can be seen in Table III the overall hit rate was higher among the non-residential patients (135 out of 155, or 87.1%) than among the residential patients (109 out of 145, or 75.2%). More specifically, the hit rate for predicting successful outcomes was higher among non-residential patients. Among the 144 non-residential patients with successful outcomes, 127 (88.2%) were correctly identified (true negatives) and 17 (11.8%) were misclassified as signal-alarm cases (false positives). By comparison, among the 134 residential patients with successful outcomes, 100 (74.6%) were correctly identified (true negatives) and 34 (25.4%) were misclassified as signal-alarm cases (false positives).

**Table III.** Comparison of Predicted vs. Actual Outcome for Residential and Non-Residential Patients

Actual outcome	White signal	Green signal	Yellow signal	Red signal
	<i>N</i> total <i>res</i> non-res	<i>N</i> total <i>res</i> non-res	<i>N</i> total <i>res</i> non-res	<i>N</i> total <i>res</i> non-res
Reliably Recovered –13 & crossed 46	22	26	1	3
	12	15	0	3
	10	11	1	0
	HIT	HIT	False-Alarm	False-Alarm
Reliably Better –13	20	81	7	7
	2	41	5	5
	18	40	2	2
	HIT	HIT	False-Alarm	False-Alarm
No Reliable Change –12 to +12	27	51	17	16
	10	20	11	10
	17	31	6	6
	HIT	HIT	False-Alarm	False-Alarm
Reliably Worse +13	0	0	1	12
	0	0	0	8
	0	0	1	4
	False-positive	False-positive	HIT	HIT
Reliably Deteriorated +13 & crossed 46	2	3	3	1
	1	1	1	0
	1	2	2	1
	False-positive	False-positive	HIT	HIT
Total Number Classified	71	161	29	39
	25	77	17	26
	46	84	12	13

However, of greater concern is the prediction of treatment failure. The hit rate for predicting treatment failure was higher among residential patients. In addition, of the 11 deteriorated patients in residential treatment, 9 (81.8%) were correctly identified (true positives), and 2 (18.2%) were misclassified based on receiving white or green signals (false negatives). Of the 11 deteriorated patients in the non-residential pool, 8 (72.7%) were correctly identified (true positives), and 3 (27.3 %) were misclassified with white or green signals (false negatives).

Finally, an analysis of all signal-alarm cases (hits *and* misses) was undertaken. Of the 68 cases that were identified with a yellow or red signal, 17 (25%) actually had a deteriorated outcome. All signal cases were divided into the aforementioned groupings (A-E) and differences in the frequency of signal alarms were examined by level of initial disturbance. The rational method tended to provide relatively fewer signal-alarms at lower levels of disturbance and suggested more signal-alarms at high disturbance levels. However, it is also at the highest level of initial disturbance (100+) that the algorithms generated the most misses using proportionate comparisons. Fifty-one of the 68 did not deteriorate as predicted (of these 51 patients, 33 had no reliable change, and 18 actually got better).

## DISCUSSION

Analysis of a rational method for identifying treatment failures found that the method was reasonably accurate. This method identified 77.3% of all cases that eventually had a negative outcome, falling slightly below the results (81%) reported by Lambert et al. (2002a) who used similar methods with adult clients. Despite this modest success, it is important to note that three-fourths of the cases that deteriorated were accurately predicted and that the level of precision far surpassed the base rate of failure (7.3%) in this sample of child patients. Although no data were collected in the present study to test clinicians' ability to predict treatment failure, past research has suggested, in the main, clinicians do not provide accurate judgments in this area (Meyer & Schulte, 2002; Schulte-Bahrenberg, & Schulte, 1993). The results are therefore viewed as promising, especially in light of the fact that the purpose of the algorithms used in the current study was aimed at identifying treatment failures rather than predicting the outcome of psychotherapy more generally. Thus, while higher overall hit rates could have been successfully achieved simply by predicting that all patients would improve, this accuracy would be obtained at the cost of losing considerable predictive power for patients who eventually deteriorated.

The small number of treatment failures who were not identified ( $n = 5$ ) made it difficult to analyze any trends that may have been related to misidentification (e.g., patient age, sex). The algorithms were more likely to fail when the YOQ scores at intake were in the functional range. A slight adjustment to the algorithms resulted in four of five patients being correctly identified as treatment failures. This adjustment, if replicated by future research, would raise the hit rate for treatment failure identification in the present sample up to 21 out of 22, or 95.5%. While a modification of decision cut-off points in the lower score range should be considered, replication of the effects of such a modification is critical to cross validate the present findings and avoid capitalizing on sample error.

Errors in predictive accuracy are not equivalent and the current algorithms were calibrated to avoid more serious predictive failures. For instance, the failure to predict deterioration in highly disturbed patients (misses) is a serious problem that needs further research. However, over identification of patients as likely treatment failures (false alarm) is not as much of a concern. Unlike some branches of medicine, where false alarms on diagnostic tests may lead to invasive and unnecessary procedures such as surgery, in psychotherapy the consequences are less intrusive. In the present sample 18.3% of cases with actual positive outcomes were misidentified (false alarms) which is similar to that found with adults (Lambert et al., 2002a). Nevertheless, like adult patients, two thirds of the child and adolescent patients identified as false alarms had a relatively poor outcome.

Lambert and colleagues (Lambert et al., 2001b, 2002b; Whipple et al., 2003) have shown that patient outcome improves when signal alarm feedback is given

to therapists even if a similarly large portion of the clients are false alarms. This suggests that the algorithms in this study, along with the estimated rate of false alarms can be applied in practice. Moreover, their use may reduce deterioration as well as increase rates of reliable and clinically significant change, probably due to increased therapist attention to signal-alarm cases.

The rational method may yield more accurate predictions of treatment failure among *residential* patients. This greater predictive accuracy in residential settings may be a function of the closer scrutiny by staff. It may also be that in residential settings the children/adolescents behavior is more closely monitored and judgments about the state of their current behavior error on the side of conservatism. It may be that the judges in residential settings are more accurate in filling out YOQ questionnaires because of their clinical experience or that children treated in residential settings are more disturbed and more vulnerable to deteriorate. Further exploration of the effects of setting variables and the person completing the measure are needed in order to clarify differences between predictions in such settings.

Several design issues limit the current study. Although the archival data base allowed for a relatively large sample drawn from routine clinical practice, it also resulted in much missing data. Although similarities were found between the study sample and the total sample on variables such as age and gender that were in the data base, it is not known if the sample analyzed is representative of the total sample that received treatment. It should be noted that the rates of deterioration (7.6% residential, 7.1% non-residential) are similar to the 5-10% estimates of negative treatment outcome based on reviews of the literature (Lambert & Bergin, 1994; Mohr, 1995). The actual number of patients that will be identified as treatment failures and who receive signal alarms will likely vary from setting to setting.

A second important limitation of this study is the definition of outcome that was employed. The YOQ was used for both signal-alarm alerts and for classification of final status. Some research (e.g., Durham et al., 2002) suggests that a test retest artifact with improving YOQ scores over administration exists, but also that this artifact is much smaller than the Reliable Change Index that was used as a marker of clinically significant change in the present study. Further research is needed to examine the accuracy of predicting outcome based on comprehensive measures of outcome drawn from a variety of reporting sources such as school-based performance, teacher ratings, peer ratings, societal records, and the like. Classification of outcome based on a variety of diverse procedures would be likely to result in lower hit rates. The problem with undertaking such research is the burden it places on patients to undergo more extensive repeated testing while keeping such research within the bounds of routine clinical practice. Obviously, studies of treatment failure require very large numbers of clients and extensive outcome measurement is problematic.

A third limitation to the present research is the definition of individual change based on clinical significance. Although the Jacobson method used in this study is the most frequently used method for estimating reliable and clinically significant change, and it seems to have some validity, further research is needed to verify that deterioration as defined here is in fact clinically meaningful (Lunnen & Ogles, 1998). The current data set employed parent ratings and staff ratings of patient status. It is not known if the cut-offs employed in the current study will work with patient YOQ self report.

Future research should be directed towards further examination of the ability of the rational method to identify cases that are not having a positive response to treatment. This research would ideally broaden the definition of final outcome, perhaps including patients who show no reliable change along with those who actually get worse. In addition, this research should study the costs and benefits of over- and under-identification of the failing patient. Future research should consider examining the effects of providing child therapists with YOQ signal's regarding their patient's progress or lack thereof. As previously mentioned, research with the OQ (Lambert et al., 2001b, 2002b) has indicated that early identification of poorly responding patients and feedback to therapists about patient treatment improved outcome at termination for adults, it is not yet known if similar positive results will be manifest with children and adolescents.

## REFERENCES

- Achenbach, T. M. (1991). *Manual for the child behavior checklist and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.
- Breslin, F., Sobell, L. C., Buchan, G., & Cunningham, J. (1997). Toward a stepped-care approach to treating problem drinkers: The predictive validity of within-treatment variables and therapist prognostic ratings. *Addiction*, *92*, 1479–1489.
- Brown, G. S., & Lambert, M. J. (1998). *Tracking patient progress: Decision making for cases that are not benefiting from therapy*. Paper presented at the 29th Annual Meeting of the Society for Psychotherapy Research, Snowbird, UT.
- Burlingame, G. M., Wells, M. G., Hoag, M., Hope, C., Nebeker, S., Konkell, K., McCollam, P., Peterson, G., Lambert, M. J., Latkowski, M., Ferre, R., & Reisinger, C. (1996). *Administration and scoring manual for the Y-OQ.2*. Setauket, NJ: American Professional Credentialing Services.
- Burlingame, G. M., Mosier, J. I., Wells, M. G., Atkin, Q. G., Lambert, M. J., Whoolery, M., & Latkowski, M. (2001). Tracking the influence of mental health treatment: The development of the Youth Outcome Questionnaire. *Clinical Psychology and Psychotherapy*, *8*, 361–379.
- Burlingame, G., Cox, J., Wells, A., & Lambert, M. (2003). Youth Outcome Questionnaire: Updated psychometric properties. In M. Maruish (Ed.), *The 2003 Behavioral outcomes and guidelines source book*. New York: Faulkner and Gray.
- Conners, C. K. (1990). *Conners' rating scales manual*. North Towanda, NY: Multi-Health Systems.
- Durham, C. J., McGrath, L. D., Burlingame, G. M., Schaalje, G. B., Lambert, M. J., & Davies, D. R. (2002). The effects of repeated administrations on self-report and parent-report scales. *Journal of Psychoeducational Assessment*, *20*, 240–257.
- Haas, E., Hill, R., Lambert, M. J., & Morrill, B. (2002). Unraveling the early response mystery: Do early responders maintain treatment gains? *Journal of Clinical Psychology*, *58*, 1157–1172.

- Jacobson, N. S., & Traux, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19.
- Kazdin, A. E. (1995). Scope of child and adolescent psychotherapy research: Limited sampling of dysfunctions, treatments, and client characteristics. *Journal of Clinical Child Psychology, 24*, 125–140.
- Kordy, H., Hannöver, W., & Richard, M. (2001). Computer assisted feedback driven active quality management for psychotherapy provision: The Stuttgart-Heidelberg model. *Journal of Consulting and Clinical Psychology, 69*, 173–183.
- Lambert, M. J. (1983). Introduction to assessment of psychotherapy outcome: Historical perspective and current issues. In M. J. Lambert, E. R. Christensen, & S. S. DeJulio (Eds.), *The assessment of psychotherapy outcome* (pp. 3–32). New York, NY.
- Lambert, M. J., & Bergin, A. E. (1994). The effectiveness of psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 143–189). New York: John Wiley and Sons.
- Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001a). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*, 159–172.
- Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, et al. (1996). *Administration and scoring manual for the outcome questionnaire (OQ 45.2)*. Wilmington, DE: American Professional Credentialing Services.
- Lambert, M. J., Huefner, J. C., & Reisinger, C. W. (2000). Quality improvement: Current research in outcome management. In G. Stricker, W. G. Troy, & S. A. Shueman (Eds.), *Handbook of quality management in behavioral health* (pp. 95–110). New York, Kluwer Academic/Plenum Publishers.
- Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002a). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy, 9*, 149–164.
- Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., et al. (2001b). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11*, 49–68.
- Lambert, M. J., Whipple, J. L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., Nielsen, S. L., et al. (2002b). Providing therapists with feedback on patient progress as a method of enhancing psychotherapy outcomes: A replication. *Clinical Psychology and Psychotherapy, 9*, 91–103.
- Lueger, R. J., Howard, K. I., Martinovich, Z., Lutz, W., Anderson, E. E., & Grissom, G. (2001). Assessing treatment progress with individual patients using expected treatment response models. *Journal of Consulting and Clinical Psychology, 69*, 150–158.
- Lunnen, K. M., & Ogles, B. M. (1998). A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66*, 400–410.
- Lyons, J. S., Howard, K. I., O'Mahoney, M. T., & Lish, J. D. (1997). *The measurement and management of clinical outcomes in mental health*. New York: John Wiley & Sons.
- Meyer, F., & Schulte, D. (2002). Zur Validität der Beurteilung des Therapieerfolgs durch Therapeuten (The validity of therapist's ratings of therapy outcome). *Zeitschrift für Klinische Psychologie und Psychotherapie*.
- Mohr, D. C. (1995). Negative outcome in psychotherapy. *Clinical Psychology, 2*, 1–27.
- Richard, M., & Kordy, H. (2002). *Early treatment response: Conceptualization, predictive validity and application in quality management*. Stuttgart: Center for Psychotherapy Research.
- Schulte-Bahrenberg, T., & Schulte, D. (1991). Therapiezieleveränderungen bei Therapeuten (Change in psychotherapy goals in therapists). In D. Schulte (Ed.), *Therapeutische entscheidungen (Therapeutic decisions)* (pp. 43–56). Göttingen: Hogrefe.
- Schulte-Bahrenberg, T., & Schulte, D. (1993). Change of psychotherapy goals as a process of resignation. *Psychotherapy Research, 3*, 153–165.
- Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore: John Hopkins University Press.
- Tang, T. Z., & DeRubies, R. J. (1999a). Reconsidering rapid early response in cognitive-behavioral therapy for depression. *Clinical Psychology: Science and Practice, 6*, 283–288.
- Tang, T. Z., & DeRubies, R. J. (1999b). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology, 67*, 894–904.

- Wells, M. G., Burlingame, G. M., & Lambert, M. J. (1999). Youth Outcome Questionnaire. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (2nd ed., pp. 497–453). Mahwah, NJ: Lawrence Erlbaum Associates.
- Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment failure and problem solving strategies in routine practice. *Journal of Counseling Psychology, 50*, 59–68.
- Wilson, G. T. (1999). Rapid response to cognitive behavior therapy. *Clinical Psychology: Science and Practice, 6*, 289–292.

Copyright of Journal of Child & Family Studies is the property of Springer Science & Business Media B.V. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.