Benchmarking the Effectiveness of Psychotherapy: Program Evaluation as a Component of Evidence-Based Practice

V. ROBIN WEERSING, PH.D.

The medical director of a child guidance center is starting a new treatment program. Following intense media coverage of adverse events associated with selective serotonin reuptake inhibitors, depressed teens and their families are refusing to accept pharmacotherapy. The director has paid for three social work therapists to attend a cognitive-behavioral therapy (CBT) workshop, and these clinicians will begin seeing depressed teens next week. The director, however, is worried. From what she has heard and read, CBT has a good track record in treating depression although it may work best in combination with the medications that her patients are refusing to take. In addition, she wonders how well will CBT work when delivered by her social work therapists to the population of poor, Spanish-speaking teens served by the clinic?

In this example, the medical director struggles with how to bring the principles of evidence-based practice (EBP) to bear on the problem of program evaluation. She is asking a critical, pragmatic question: Does therapy work, here, now, and with my patients? This column discusses a method for addressing this query that uses the results of published psychotherapy clinical trials as a gold-standard benchmark against which the outcomes of practice can be measured (McFall, 1996). This methodology views research evidence on the effects of psychotherapy as, literally, a base—solid ground that researchers and practitioners can use as a foundation when trying to solve problems, whether theoretical or applied.

A BRIEF HISTORY OF BENCHMARKING

In the early 1990s, a community mental health center (CMHC) in Bloomington, IN, faced a dilemma not unlike the medical director in the example—a desire to provide evidence-based services that were both acceptable to and effective in their local patient population (see McFall, 1991, for discussion). The CMHC, in partnership with researchers from Indiana University, trained staff therapists in a well-supported CBT treatment manual for adults with panic disorder (Wade et al., 1998). In randomized, controlled trials of this program, more than 80% of patients were panic-free at the end of treatment (e.g., Barlow et al., 1989), but how well would the intervention perform in the CMHC? In the CMHC setting, the treatment would be delivered by general outpatient staff, and a broader patient population would be accepted for care, including those with severe agoraphobia and some using concomitant benzodiazepines. Whereas there was a solid base of clinical trial evidence supporting the efficacy of CBT for panic, there were no data on the effectiveness of the intervention under practice conditions such as these (see Weisz et al., 1992, for discussion). Would the effects of CBT for panic generalize across these differences between research and practice, and, more important, how should one measure and interpret the magnitude of treatment effects in the CMHC?

The procedure adopted by the clinic research team forms the backbone of the benchmarking method

Accepted December 20, 2004.

Dr. Weersing is with the Child Study Center, Yale University School of Medicine, New Haven, CT.

Preparation of this column was facilitated by support from the William T. Grant Foundation, the Klingenstein Third Generation Foundation, The Robert Wood Johnson Foundation, and the National Institute of Mental Health (MH064503-01A1 and MH066371-01). The author thanks Teresa Treat her helpful comments on an earlier draft of this article and Erin Warnick and Alissa Mohammed for assistance with manuscript preparation.

Correspondence and reprint requests to Dr. Robin Weersing, Child Study Center, Yale University School of Medicine, 230 South Frontage Road, New Haven, CT 06520; e-mail: robin.weersing@yale.edu.

DOI: 10.1097/01.chi.0000172682.71384.80

(see Wade et al., 1998). To assess the generalizability of CBT effects, they compared the outcomes of CBT in their sample and setting, point by point, to the results of two published clinical trials, using the same outcome measures and definitions of improvement. If the effects of CBT in the community replicated these ideal outcomes, logic would dictate that CBT "works" in the practice conditions of the new setting. Indeed, this is exactly what Wade and colleagues found. Eighty-seven percent of CMHC patients were panic-free post-treatment, even as benzodiazepine use decreased from 60% to 20% of the patient sample. Changes in dimensional symptom measures paralleled these improvements and were similar in magnitude to the results of the two benchmarks (Barlow et al., 1989; Telch et al., 1993). Furthermore, 1 year later, CMHC patients had maintained their treatment gains (Stuart et al., 2000). In the years since the publication of this report, a number of benchmarking studies have appeared in the psychotherapy literature testing the generalizability of treatments with strong outcomes in efficacy studies, including exposure and response prevention for obsessive-compulsive disorder (Franklin et al., 2000), CBT for bulimia (Tuschen-Caffier et al., 2001), cognitive therapy for depression (Merrill et al., 2003), and CBT for social phobia (Lincoln et al., 2003). In general, these investigations have demonstrated the robustness of evidence-based treatments in a variety of patient, provider, and setting parameters.

BENCHMARKING AS A COMPONENT OF EVIDENCE-BASED PRACTICE

The benchmarking method was developed to assess generalizability of efficacy findings, and this purpose maps directly onto the question posed in our example. The medical director wonders whether the results of CBT in clinical trials can be expected to replicate in the population of poor, Spanish-speaking teens served by her clinic. How can she answer this question?

The Answerable Question

In an EBP model, the first step is to establish the "knowns"—the existing research base relevant to the question of interest. Readers are referred to the excellent EBP handbook produced by the American Medical Association for instruction on how to access and efficiently search electronic databases for this purpose (Guyatt and Rennie, 2002). Let us imagine that our medical director uses these resources to find three of the most pertinent references available: a recent evidence-based medicine review of the effects of CBT for youths with depression (Compton et al., 2004), an "effectiveness" review of psychosocial treatments for youths (Chorpita et al., 2002), and the only CBT depression clinical trial conducted in a sample of depressed Spanish-speaking teens (Rosselló and Bernal, 1999). These references provide useful information. Broadly, CBT appears to be beneficial to depressed teens, with effect sizes on dimensional symptom measures in the medium-to-large range (Compton et al., 2004). The number needed to treat ratio for CBT is also in the acceptable range. Using estimates from Compton et al. (2004), it is necessary to treat three to six teens, on average, with CBT to produce one depression recovery (no diagnosis of major depressive disorder after treatment; see Guyatt and Rennie [2002], pp. 358–360, for discussion and definition of number needed to treat). CBT also has been successfully delivered by master's degree–level therapists (Chorpita et al., 2002), and one study did test the intervention with Spanish-speaking teens in Puerto Rico (Rosselló and Bernal, 1999). In this study, CBT was superior to waitlist but was equivalent to another well-supported psychosocial intervention for adolescent depression, interpersonal psychotherapy.

Implementing Benchmarking

By many accounts, the medical director has answered her question; there is reason to believe that CBT should produce clinically significant benefit at her clinic and in her patient population. This said, she still does not know whether CBT *does* work in her clinic. To answer this query, it is necessary to move beyond literature review to data collection. The benchmarking methodology provides one model for how this might be accomplished, within the available resources of a practice setting. Briefly, benchmarking has four logical steps:

Step 1: Define the problem, population, and treatment model
Step 2: Select or create a gold-standard outcome benchmark from the research literature
Step 3: Measure outcome in the applied setting, using comparable methods as in the benchmark
Step 4: Compare outcomes and explore reasons for any differences

In the first example, step 1 has already been accomplished: the problem is depression, the population is Spanish-speaking adolescents, and the treatment model is individual CBT. Through her literature review, the director uncovered a CBT clinical trial addressing the problem and population of interest, and this study could well fulfill step 2. The CBT program of Rosselló and Bernal (1999) consisted of 12 weekly sessions, with one third of the sessions focusing on changing negative-thinking patterns, one third on increasing pleasant activities, and one third on assertiveness and social support. For this particular study to serve as a relevant benchmark, it would be important for the social work therapists from our example clinic to be trained in a similar variant of CBT. Similarly, the sample of the clinic should match the benchmark sample as closely as is feasible; the exclusionary criteria of Rosselló and Bernal are not unusually stringent, but they did screen out youths who presented with other serious clinical issues such as significant substance abuse and intense suicidality. If the majority of depressed teens in the clinic struggle with these issues, Rosselló and Bernal still may be used as a benchmark; however, it will be difficult to pin down the cause of any gaps in outcome, if the effects of therapy in the practice setting fail to match the outcomes of the clinical trial.

Assuming that the sample and treatment of Rosselló and Bernal are the best match available in the literature, our medical director would move to step 3 and design the clinic's outcome assessments. In the benchmark clinical trial, the primary depression outcome measure was the Children's Depression Inventory (CDI; Kovacs, 1992), a 27-item youth report symptom scale. High scores on the CDI are not synonymous with a mood disorder diagnosis, nor are low scores a perfect indicator of recovery; however, the CDI is the most widely used depression symptom measure in child and adolescent therapy research (Kendall et al., 1989), with excellent psychometric characteristics (Smucker et al., 1986). Rosselló and Bernal also included measures of self-esteem, family functioning, social adjustment, and other behavioral and emotional problems. As in the Wade et al. (1998) benchmarking study, inclusion of these additional measures would provide for a more complete picture of treatment effects and the ability to test for important moderators of treatment outcomes (e.g., if teens with conflictual families had worse CBT outcomes in the clinic). To use benchmarking as a component of everyday EBP, however, regularly administering and collecting the main symptom measure of interest (e.g., the CDI) is a reasonable starting point.

To compare CBT in practice to the benchmark (step 4), it also is important to assess symptom change on a similar schedule. In the Rosselló and Bernal (1999) trial, assessments were given at intake, 12 weeks later (post-treatment), and at 3-month follow-up, and youths continued to improve through the follow-up period. Given these data, it would behoove the medical director to obtain the personnel and financial resources necessary to conduct a follow-up assessment, even if only by phone. CDI scores from these three assessment points could then be compared point by point with the outcomes of Rosselló and Bernal. If the effects of CBT in practice match or exceed the results of the clinical trial, then the director can likely rest easy. If outcomes in the clinic do not fall within the confidence interval of the active treatment clinical trial results, then there is greater cause for concern, and a search begins (1) to understand the possible causes for the observed gaps in outcomes (e.g., sample differences, different treatment components, attrition) and (2) to evaluate alternate treatments that may be a better fit for this patient population, such as interpersonal psychotherapy or a renewed emphasis on selective serotonin reuptake inhibitors. The director also may wish to look beyond the original benchmark to the broader CBT literature, as discussed in the next section.

Benchmarking Against the Entire Evidence Base

The point-by-point method of benchmarking outlined in the previous section is a good program evaluation strategy, especially if a practice is adopting an established treatment manual with at least one published gold-standard clinical trial (e.g., Wade et al., 1998). For other applications, comparison to a single clinical trial may not be feasible or be the most desirable benchmark. For example, in a recent investigation, Weersing and Weisz (2002) constructed a benchmark from the entire CBT depression literature to assess the effects of a previously untested package of services—eclectic "treatment as usual" for youth depression in CMHCs. The investigators identified youths with diagnoses of major depression or dysthymia presenting for care at six CMHCs and observed these children and adolescents for 2 years, collecting data on symptom change and service use. To anchor the results of this descriptive study, the investigators compared the results of CMHC care to the outcomes of the entire CBT treatment literature. The authors created two composite benchmarks: one summarizing the outcomes of youths receiving CBT in all the published clinical trials to date and one summarizing the outcomes of youths randomly assigned to the control groups of these same CBT studies.

Figure 1 displays a recalculation of these benchmarks, including CBT trials not published at the time of the Weersing and Weisz article (e.g., Treatment of Adolescents With Depression Study [TADS] Team , 2004). As can be seen in the figure, between intake and post-treatment (usually 16 weeks later), youths receiving CBT typically experience a sharp drop in symptoms. These effects are generally maintained during the course of a year, although long-term follow-up data suggest that recurrence is common by 2 years post-CBT (Birmaher et al., 2000). In contrast, youths in control conditions (solid black line in Figure 1) experience gradual improvement of symptoms, a pattern that maps the course of natural remission in youth depression (Kovacs, 1996; Kovacs et al., 1997). The lines and regions in Figure 1 take into consideration variability in outcome between studies; the borders of shaded areas represent the 95% confidence interval limit on dimensional measures for that region of effects (i.e., the solid black line is the 95% upper limit on outcomes for CBT control groups).

In the Weersing and Weisz (2002) study, these two benchmarks served as hypothetical best-case (effects similar to best practices) and worse-case scenarios (effects similar to no treatment), and, unfortunately, CMHC care closely resembled the worst case. The outcome of short-term eclectic therapy for depressed youths was virtually identical to that of youths in the control conditions from the clinical trials (i.e., the solid line in Figure 1). Subgroup analyses revealed that youths receiving longer-term treatments (at least eight sessions) and white youths had somewhat better responses to therapy, with results falling between that of control and CBT (i.e., the area of intermediate effects in the figure). However, no subgroup of youths in this sample had outcomes within the range of the CBT research benchmark. (i.e., the area below the dashed line in the figure).

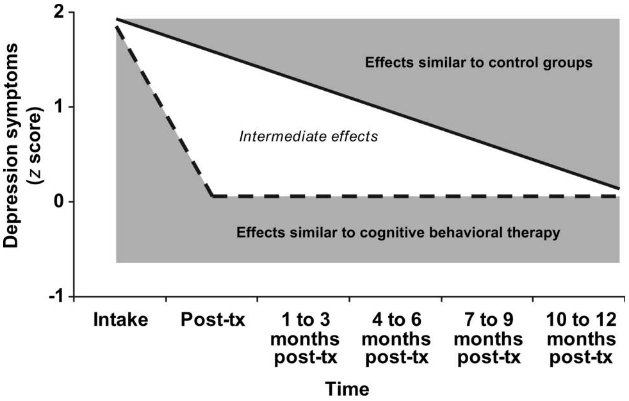

Fig. 1 Literature-wide benchmark outcomes for the cognitive-behavioral treatment of youths with depression.

Composite benchmarks such as these may be of use to our medical director, if the results of CBT in her clinic do not clearly exceed those of Rosselló and Bernal. Outcomes of therapy in her clinic would be measured with depression symptom scales, and the results plotted against literature-wide summary benchmarks, such as those shown in Figure 1. To facilitate comparison of results across different measures of depression, the example figure displays depression symptoms in a standardized, normative *z* score format. Calculating a normative *z* score is a simple arithmetic procedure that uses data from the original measure development reports to standardize scores relative to a ''normal'' community sample of youths (see Table 1 for reference). These computations take the general form $z_{nt} = (\bar{x}_t - \mu)/\sigma$,

TABLE 1
Normative Data for Common Self-report Measures of Depression for Children and Adolescents

Measure	Source Article	Mean	SD
CES-D	Roberts et al., 1991	16.98	10.65
CDI	Smucker et al., 1986	9.09	7.04
BDI	Roberts et al., 1991	7.17	7.50
MFQ-C	Kent et al., 1997	27.05	13.73
RADS	Reynolds, 1986; Reynolds and Mazza, 1998	60.18	14.29

Note: These means and SDs represent the ''best'' available data for combined samples of children and adolescents in community settings. Norms by age and sex are available for some of the most widely used measures (CDI, BDI), although these subgroup norms typically are based on small samples. For the MFQ-C, only psychometric data from samples of child and adolescent outpatients were available. These youths are likely to have higher means on the MFQ-C than would a community sample. CES-D = Center for Epidemiologic Studies Depression Scale; CDI = Children's Depression Inventory; BDI = Beck Depression Inventory; MFQ-C = Mood and Feelings Questionnaire, child version; RADS = Reynolds Adolescent Depression Scale.

where $\bar{x}_t$ is the mean score on the depression measure in the applied setting, μ is the normal population mean, and σ is the normal population SD for the depression measure (Kendall and Grove, 1988). The *z* score calculated with this formula can be interpreted as an index of depression severity; a intake *z* score of 2.0 indicates that the mean level of symptoms is 2.0 SD above the community mean for depression, a score at roughly the 98th percentile. By definition, a normative *z* score of 0 is equivalent to "normal" level of depression (the community mean), and return of symptoms to this level is one index of clinically significant change. Interested readers are referred to Weersing and Weisz (2002) for more complete discussion of how to create and interpret literature-wide benchmarks.

CAVEATS AND CONCLUSIONS

This column was intended to provide an introduction to benchmarking—a practical method for measuring treatment outcomes in applied settings. The description of benchmarking was necessarily brief; readers are referred to the published benchmarking articles referenced in this column for good examples of the method in action and more thorough discussion of the strengths and limitations of the design. More important, benchmarking is nonexperimental and suffers from the inference problems of all descriptive research (i.e., threats to internal validity, alternate explanations for observed effects). However, as a component of EBP, benchmarking may allow scientifically minded practitioners to move beyond literature review and conduct real-world program evaluation studies within the settings and samples of active clinical care.

Disclosure: The author has no financial relationships to disclose.

REFERENCES

Barlow D, Craske M, Cerny J, Klosko J (1989), Behavioral treatment of panic disorder. *Behav Ther* 20:261–282

Birmaher B, Brent DA, Kolko D et al. (2000), Clinical outcome after short-term psychotherapy for adolescents with major depressive disorder. *Arch Gen Psychiatry* 57:29–36

Chorpita B, Yim L, Donkervoet J et al. (2002), Toward large-scale implementation of empirically supported treatments for children: a review and observations by the Hawaii Empirical Basis to Services Task Force. *Clin Psychol Sci Pract* 9:165–190

Compton S, March J, Brent D, Albano A, Weersing V, Curry J (2004), Cognitive-behavioral psychotherapy for anxiety and depressive disorders in children and adolescents: an evidence-based medicine review. *J Am Acad Child Adolesc Psychiatry* 43:930–959

Franklin M, Abramowitz J, Kozak M, Levitt J, Foa E (2000), Effectiveness of exposure and ritual prevention for obsessive-compulsive disorder: randomized compared with nonrandomized samples. *J Consult Clin Psychol* 68:594–602

Guyatt G, Rennie D (2002), *Users' Guide to the Medical Literature: A Manual for Evidence-Based Clinical Practice.* Chicago: AMA Press

Kendall P, Grove W (1988), Normative comparisons in therapy outcome. *Behav Assess* 10:147–158

Kendall PC, Cantwell DP, Kazdin AE (1989), Depression in children and adolescents: assessment issues and recommendations. *Cognit Ther Res* 13:109–146

Kent L, Vostanis P, Feehan C (1997), Detection of major and minor depression in children and adolescents: evaluation of the Mood and Feelings Questionnaire. *J Child Psychol Psychiatry* 38:565–573

Kovacs M (1992), *Children's Depression Inventory Manual.* North Tonawanda, NY: Multi-Health Systems

Kovacs M (1996), Presentation and course of major depressive disorder during childhood and later years of the life span. *J Am Acad Child Adolesc Psychiatry* 35:705–715

Kovacs M, Obrosky D, Gatsonis C, Richards C (1997), First-episode major depressive and dysthymic disorder in childhood: clinical and sociodemographic factors in recovery. *J Am Acad Child Adolesc Psychiatry* 36:777–784

Lincoln T, Rief W, Hahlweg K et al. (2003), Effectiveness of an empirically supported treatment for social phobia in the field. *Behav Res Ther* 41:1251–1269

McFall R (1991), Manifesto for a science of clinical psychology. *Clin Psychol* 44:75–88

McFall R (1996), Consumer satisfaction as a way of evaluating psychotherapy: ecological validity and all that versus the good old randomized trial. Panel discussion at the Sixth Annual Convention of the American Association of Applied and Preventative Psychology, San Francisco

Merrill K, Tolbert V, Wade W (2003), Effectiveness of cognitive therapy for depression in a community mental health center: a benchmarking study. *J Consult Clin Psychol* 71:404–409

Reynolds W (1986), *Reynolds Adolescent Depression Scale.* Odessa, FL: Psychological Assessment Resources

Reynolds W, Mazza J (1998), Reliability and validity of the Reynolds Adolescent Depression Scale with young adolescents. *J Sch Psychol* 36:295–312

Roberts R, Lewinsohn P, Seeley J (1991), Screening for adolescent depression: a comparison of depression scales. *J Am Acad Child Adolesc Psychiatry* 30:58–66

Rosselló J, Bernal G (1999), The efficacy of cognitive-behavioral and interpersonal treatments for depression in Puerto Rican adolescents. *J Consult Clin Psychol* 67:734–745

Smucker M, Craighead W, Craighead L, Green B (1986), Normative and reliability data for the Children's Depression Inventory. *J Abnorm Child Psychol* 14:25–39

Stuart G, Treat T, Wade W (2000), Effectiveness of empirically based treatment for panic disorder delivered in a service clinic setting: 1-year follow-up. *J Consult Clin Psychol* 68:506–512

Telch M, Lucas J, Schmidt N, Hanna H, Jaimez T, Lucas R (1993), Group cognitive-behavioral treatment of panic disorder. *Behav Res Ther* 31:279–287

Treatment of Adolescents With Depression Study (TADS) Team (2004), Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for adolescents with depression study (TADS) randomized controlled trial. *JAMA* 292:807–820

Tuschen-Caffier B, Pook M, Frank M (2001), Evaluation of manual-based cognitive-behavioral therapy for bulimia nervosa in a service setting. *Behav Res Ther* 39:299–308

Wade W, Treat T, Stuart G (1998), Transporting an empirically supported treatment for panic disorder to a service clinic setting: a benchmarking strategy. *J Consult Clin Psychol* 66:231–239

Weersing V, Weisz J (2002), Community clinic treatment of depressed youth: benchmarking usual care against CBT clinical trials. *J Consult Clin Psychol* 70:299–310

Weisz J, Weiss B, Donenberg G (1992), The lab versus the clinic: effects of child and adolescent psychotherapy. *Am Psychol* 47:1578–1585