Quality & Quantity (2008) 42:513–525
DOI 10.1007/s11135-006-9057-z

Using Clinical Trials to Benchmark Effects Produced in Clinical Practice

TAKUYA MINAMI[1,*], RONALD C. SERLIN[2], BRUCE E. WAMPOLD[3], JOHN C. KIRCHER[1] and G.S. (JEB) BROWN[4]
[1]*Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA;*
[2]*Department of Educational Psychology, University of Wisconsin, Madison, WI, USA;*
[3]*Department of Counseling Psychology, University of Wisconsin, Madison, WI, USA;*
[4]*Center for Clinical Informatics, Salt Lake City, UT, USA*

Abstract. This paper proposes an intuitive yet statistical advancement of the benchmarking method (e.g., Weersing and Weisz, 2002, *Journal of Consulting and Clinical Psychology* 70: 299–310) that could facilitate the assessment of pre-post treatment effectiveness of psychotherapy and other interventions delivered in clinical settings against efficacy observed in clinical trials. Primary development was in the use of the "good-enough principle" (Serlin and Lapsley, 1985 *American Psychologist* 40: 73–83, 1993, In: G. Keren & C. Lewis (eds.), A handbook for Data Analysis in Behavioral Sciences: Methodological Issues. Hillsdale, NJ: Lawrence Erlbaum Associated, pp. 199–228), which allowed for setting a clinically relevant margin between the benchmarks and the effect sizes observed in clinical settings so as to avoid attaining statistical significance with clinically trivial differences. Examples are given using clinical trials benchmarks of adult depression treatment, followed by instructions and limitations for its use.

Key words: benchmarking, efficacy and effectiveness, meta-analysis, psychotherapy, clinical practice

1. Introduction

Although psychotherapy efficacy (i.e., effect of psychotherapy in clinical trials) has been well established, psychotherapy effectiveness (i.e., effect of psychotherapy in clinical settings) has not. In the past decade, interest in evaluating effectiveness has been emphasized (e.g., Seligman, 1995; Goldfried and Wolfe, 1998; Shadish et al., 2000). However, investigations regarding effectiveness of treatment-as-usual (TAU) in these settings have

*Author for correspondence: Takuya Minami, Department of Educational Psychology, University of Utah, 1705 East Campus Center Drive RM 342, Salt Lake City, UT, 84112–9599, USA. Tel.: (1)-801-5817191; Fax: (1)-801-5815566; E-mail: takuya.minami@ed.utah.edu

not progressed, seemingly due to two difficulties: (a) few clinical settings utilize standardized outcome measures in systematic ways to assess outcomes, and (b) even if outcomes are measured, sound methodology has been lacking for the translation of the measurements into an intuitively comprehensible quantity. As the reason for the first issue appears rather administrative and political (e.g., Addis, 2002), it cannot be addressed simply from alterations of research designs or statistical methods. However, the second issue could be adequately addressed by improvements on a demonstrated strategy, namely benchmarking (e.g., Weersing and Weisz, 2002). Specifically, benchmarking allows for a direct statistical comparison of pre-post treatment outcomes between clinical trials and clinical settings. In other words, with benchmarking, effectiveness could be established by assessing whether the benefits to clients in clinical practice approach the benefits that clients experience in controlled research.

A review of the very few studies that utilized a benchmarking strategy illustrates how this method has progressed until now. In 1998, Wade et al. published a benchmarking study that used clinical trial results to evaluate the pre-post effectiveness of an empirically-supported treatment (EST), namely cognitive-behavioral therapy for panic disorder (CBT-P; Barlow and Craske, 1994), implemented in a community mental health center (CMHC). Wade et al. selected two published clinical trials (viz., Barlow et al., 1989; Telch et al., 1993) to serve as pre-post treatment efficacy benchmarks for their CMHC individual and group treatments. Wade et al. concluded that the effects of CBT-P implemented in the CMHC setting was comparable to those produced in the two clinical trials, but no statistical analyses were used to support their claim. This heuristic comparison was replicated to benchmark the effectiveness of cognitive therapy (CT) for depression implemented in a CMHC (Merrill et al., 2003). Another limitation of these two studies was that the treatment implemented was not TAU, but an established treatment (in these cases, ESTs) with many of the conditions that distinguish clinical trials from clinical practice (e.g., special training of therapists, supervision; see Westen et al., 2004).

Weersing and Weisz's (2002) study of youth depression treatment in CMHCs advanced benchmarking methodology in several ways. First, by not modifying any aspect of the TAUs delivered in the CMHCs where they collected the data, Weersing and Weisz allowed for their results to be generalized to TAUs rather than to treatments that are different from what is typically practiced. Second, rather than choosing only a few clinical trials to compare the CMHC data against, Weersing and Weisz aggregated the clinical trials results from 13 published studies, using meta-analysis to obtain their pre-post treatment efficacy benchmark, thus attaining "a best practice benchmark from a review of the entire youth depression treatment literature" (p. 300). Third, in addition to a treatment efficacy benchmark,

they obtained a benchmark for the natural history of youth depression by aggregating the results of clinical trials reporting pre-post data on wait-list and other control conditions. This permitted Weersing and Weisz to benchmark the CMHC data against natural remission of depressive symptoms, as well. Four, they evaluated whether or not the CMHC data fell within the two-tailed 95% confidence interval of the benchmarks for each time point, rather than subjectively evaluating whether the CMHC data and the benchmarks appear similar. These improvements significantly advanced the utility of the benchmarking strategy in evaluating pre-post effect sizes of TAUs in clinical settings.

However, there were certain conceptual and statistical issues with Weersing and Weisz's (2002) benchmarking strategy that needed to be addressed. Despite conceptualizing that the benchmarks derived from their meta-analyses to be estimations of population values, Weersing and Weisz utilized these values as variable – that is, containing error – while fixing the value obtained in clinical practice (i.e., not contain error). This is in contrast to considering the benchmark to be a fixed value against which to compare effects obtained in clinical practice, which are conceptualized as being sampled from a population. The latter strategy allows making the generalization that the effects of practice of the sort delivered are meeting a benchmark established by a large number of clinical trials; otherwise, the results would be restricted to the clients and therapists of the particular clinic studied.

In addition, Weersing and Weisz (2002) did not consider whether or not the differences between the benchmarks and the value in the population were clinically significant. The reason as to why this was not addressed in their study could be because the similarity of the trajectories between the TAU and control group benchmark was evident from both visual inspection and statistical analyses. However, prior specification of a statistical criterion for clinical significance becomes important as clinical settings data become large, as differences between the population and the benchmark that could be considered clinically trivial would reach statistical significance in the sample, simply due to increased power. Therefore, to benchmark the effect of TAU obtained, a decision must be made (prior to statistical analysis) on a value that indicates the largest permissible difference between the benchmark effect size and that in the population represented by the data that would still allow for the conclusion that the two are clinically equivalent. For example, one might conclude that the benchmark and the population being within 1/5th of a standard deviation (i.e., $\Delta = 0.2$) of one another is close enough, as this value is classified as a small effect (Cohen, 1988). In this case, any observed difference between the benchmark and the population represented by the data that is within $\Delta = 0.2$ is considered clinically trivial, and therefore, the two are considered clinically equivalent. On

the other hand, if the observed difference exceeds $\Delta = 0.2$, then the two are clinically dissimilar. Consequently, statistical analyses that compare benchmarks to population data must incorporate this maximum deviation for clinical equivalence, while maintaining an overall Type I error of $\alpha = 0.05$ (Serlin and Lapsley, 1985, 1993).

In the current paper, we aimed to further advance Weersing and Weisz's (2002) benchmarking method to a more clinically and statistically sound evaluation procedure for clinical settings data. Specifically, the revised benchmarking method statistically (a) determined the benchmark as a fixed value and the clinical settings data as a variable, and (b) allowed for prior designation of a maximum statistical difference between the benchmarks and the population that would still be considered clinically trivial, while maintaining rigor in statistical analyses. In doing so, both treatment efficacy and natural history benchmarks for adult depression by Minami et al. (in press) were utilized as an example.

2. Benchmark Methodology

Overall, three steps are required in the benchmarking strategy: (a) constructing the pre-post benchmarks from clinical trials, (b) estimating the effectiveness of TAU using pre-post effect size, and (c) benchmarking the TAU effect size against the constructed benchmarks.

2.1. constructing benchmarks

2.1.1. *Selection of clinical trials*

As demonstrated by Weersing and Weisz (2002) and Minami et al. (in press), pre-post benchmarks should be constructed by meta-analytically aggregating results from well-conducted clinical trials. The criteria in selecting clinical trials for inclusion and exclusion should be predetermined by the researchers. In TAU, effects for pre-to-post change typically are based on all clients who are treated in practice as the length of treatment is indefinite. Accordingly, outcomes of TAU are assessed periodically and the last observation obtained is considered the final outcome of the treatment. This practice suggests that intent-to-treat samples yield a more appropriate benchmark than would the completer samples that contain only clients who complete the structured protocol. Consequently, only trials that report intent-to-treat samples should be included in the meta-analysis that creates the benchmark for use against clinical settings.

2.1.2. *Selection of outcome measures to aggregate*

Once the studies and samples to include in the benchmarks are selected, the researcher must select the specific outcomes to aggregate among

multiple measures often used in the studies. In doing so, two issues are of importance: (a) reactivity and specificity of outcome measures are matched both within and among the clinical trials in constructing the benchmark (Lambert et al., 1986; Minami et al., in press), and (b) only one effect size represents a study within an aggregated benchmark.

The observed magnitudes of the pre-post effect sizes in the clinical trials are significantly different based on the characteristics of the outcome measures. The reactivity of the outcome measures pertains primarily to who measured the outcomes. Specifically, outcomes measured by clinicians are known to be higher in reactivity than clients' self-report (Lambert et al., 1986; Minami et al., in press). Specificity is concerned with whether the measures assess specific, targeted symptoms of a specific diagnosis (e.g., BDI, Beck and Steer, 1987), or general, global clinical symptoms (e.g., SCL-90-R, Derogatis, 1983). The reactivity and specificity of the outcome measures used in the meta-analysis should match the reactivity and specificity of the outcome measures used in TAU.

Clinical trials typically measure their outcomes using multiple instruments. In the most ideal of cases, clinical trials considered for inclusion for a benchmark utilize identical measures, as matching in specificity and reactivity does not ensure qualitative equivalence among the measures. For example, Minami et al. (in press) constructed one benchmark by aggregating only the BDI and another by aggregating only the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960). By doing so, there was no ambiguity as to whether or not the benchmark was created by aggregating equivalent measures. In other cases, best effort of equivalence is maintained by matching the specificity and reactivity among measures. Additionally, so as to maintain independence of observation and proper weighting, only one effect size per study should be aggregated for one benchmark. In cases where researchers are interested in combining estimates from multiple measures (with the same specificity and reactivity) within a clinical trial, they are advised to consult Gleser and Olkin (1994) for methods to derive single estimates from dependent outcomes.

2.1.3. *Calculating effect size within clinical trials*

After the clinical trials and their outcomes are selected for aggregation, they are combined using standard meta-analytic procedures (e.g., Hedges and Olkin, 1985; Becker, 1988). Specifically, for each clinical trial i, the unbiased pre-post effect size estimate d_i is

$$d_i = \left(1 - \frac{3}{4n_i - 5}\right) \frac{M_{i,\text{post}} - M_{i,\text{pre}}}{SD_{i,\text{pre}}}, \tag{1}$$

where n_i is the sample size, $M_{i,\text{post}}$ the posttreatment mean of the measure, $M_{i,\text{pre}}$ the pretreatment mean of the measure, and $SD_{i,\text{pre}}$ the pretreatment standard deviation of the measure. The variance of d_i is estimated by

$$\hat{\sigma}^2_{d(i)} = \frac{2(1-r_i)}{n_i} + \frac{d_i^2}{2n_i}. \tag{2}$$

Here, r_i is the estimated correlation between the pretreatment and posttreatment scores of the outcome measure (Becker, 1988). Reasonable estimates of this value must be made. For example, Minami et al. (in press), who created benchmarks for the treatment of depression, used a value of $r = 0.5$ as the pre-post correlation of 7530 adult outpatient clients with depression in TAU (Minami et al., in press). In addition, it is noted that larger estimation of this correlation yields slightly more conservative estimates of the benchmark.

2.1.4. *Aggregating effect sizes across clinical trials*

After the effect sizes d_i are calculated for each study, they are aggregated across clinical trials to yield a single value, which would serve as the benchmark. Specifically,

$$d_B = \sum_i \frac{d_i}{\hat{\sigma}^2_{d(i)}} \Big/ \sum_i \frac{1}{\hat{\sigma}^2_{d(i)}}. \tag{3}$$

The value of d_B is considered fixed, although in traditional meta-analytic contexts, it contains a small amount of error.

The above procedure for both within and across studies is repeated when researchers are interested in constructing additional benchmarks for different types of outcome measures (e.g., different reactivity and/or specificity) or for different conditions (treatment or natural history).

2.2. ESTIMATING TAU EFFECTIVENESS

Effectiveness is also estimated by an effect size, using a procedure that is almost identical to the procedure used for a single clinical trial in preparation for aggregation across studies. Specifically, the estimated effect size of the clinical setting d_D is

$$d_D = \left(1 - \frac{3}{4N-5}\right) \frac{M_{D,\text{post}} - M_{D,\text{pre}}}{SD_{D,\text{pre}}}, \tag{4}$$

where N is the sample size, $M_{D,\text{post}}$ the posttreatment mean, $M_{D,\text{pre}}$ the pretreatment mean, $SD_{D,\text{pre}}$ the pretreatment standard deviation. The variance of this estimate is given by

$$\hat{\sigma}^2_{d(D)} = \frac{2\,(1 - r_D)}{N} + \frac{d_D^2}{2N}. \tag{5}$$

Here again, r_D is the estimated correlation between the pretreatment and posttreatment scores of the outcome measure and normally the value used in the meta-analysis would be used here as well. With sufficient data, this could be calculated by a simple Pearson r correlation coefficient between the pre- and posttreatment scores in the TAU sample.

2.3. BENCHMARKING

2.3.1. *The "good-enough" principle*

When clinical settings data have large Ns, increase in sample size leads to attaining statistical significances albeit with clinically trivial differences. Therefore, when benchmarking clinical settings data against clinical trials benchmarks, it is crucial to define a statistical criterion for clinical equivalence before conducting analyses. Following Cohen's (1988) suggestion that an effect size of $\Delta = 0.2$ is small, Minami et al. (in press) defined any difference between the benchmark and the population represented by the sample that is under $\Delta = 0.2$ to be clinically trivial. Thus, the statistical analyses employed should not reject the null hypothesis if the difference is under $\Delta = 0.2$, while maintaining an overall Type I error of $\alpha = .05$. The "good-enough principle" proposed by Serlin and Lapsley (1985, 1993) allows for such statistical analyses with a range-null hypothesis.

2.3.2. *Benchmarking against treatment efficacy benchmarks*

The implications of the $\Delta = 0.2$ margin are different depending on which – treatment efficacy or natural history – benchmark the data is compared against. When comparing the clinical settings data to a treatment efficacy benchmark, the effect size of the data must allow the conclusion that it is larger than the benchmark minus 1/5 of a standard deviation to be considered clinically equivalent with the benchmark.

When δ_D is the true effect size in the clinical setting, $\delta_{B(\mathrm{TE})}$ the true treatment efficacy benchmark from the clinical trials, and $\Delta = 0.2$ (i.e., 1/5 of a standard deviation) is the maximum difference allowed for claiming clinical equivalence, the range-null and alternative hypotheses as illustrated by Serlin and Lapsley (1985, 1993) are

$$H_0 : \delta_D \leq \delta_{B(\mathrm{TE})} - \Delta, \tag{6}$$

$$H_1 : \delta_D > \delta_{B(\mathrm{TE})} - \Delta. \tag{7}$$

With N being the sample size of the clinical settings data, the test statistic $t_{(\mathrm{TE})\nu,\lambda}$ for this hypothesis follows a noncentral t distribution with $\nu = N - 1$

degrees of freedom and a noncentrality parameter $\lambda_{\text{TE}} = \sqrt{N}\left(\delta_{B(\text{TE})} - \Delta\right)$. When $t_{(\text{TE})\nu,\lambda:.95}$ is the 95th percentile of the noncentral t distribution, then the critical value, $d_{\text{CV(TE)}}$, that the observed clinical settings effect size needs to exceed to claim clinical indifference to the clinical trials benchmark is,

$$d_{\text{CV(TE)}} = t_{(\text{TE})\nu,\lambda:.95}/\sqrt{N}. \tag{8}$$

Thus, when the observed effect size exceeds the critical value, then the null hypothesis is rejected and the clinical settings effect size is considered to lie no more than 1/5 of a standard deviation below the treatment efficacy benchmark. In this case, it is considered that the clinical settings data is clinically equivalent with the benchmark.

2.3.3. *Benchmarking against treatment efficacy benchmark*

Conversely, with the natural history benchmarks, the clinical settings data had to allow the conclusion that the effect size is more than 1/5 of a standard deviation *above* the benchmark to be considered having any treatment effect at all. In other words, unless the data is at least 1/5 of a standard deviation above the natural history benchmark, we considered that the treatment has no clinical effect. Thus, for comparing the clinical settings data against the true natural history benchmark $\delta_{B(\text{NH})}$, the $\Delta = 0.2$ minimum difference for claims of clinical difference has a reverse effect as compared to the treatment efficacy benchmark. Specifically,

$$H_0 : \delta_D \leq \delta_{B(\text{NH})} + \Delta, \tag{9}$$

$$H_1 : \delta_D > \delta_{B(\text{NH})} + \Delta. \tag{10}$$

The test statistic $t_{(\text{NH})\nu,\lambda}$ for this hypothesis also follows a non-central t distribution with $\nu = N - 1$ degrees of freedom and a noncentrality parameter $\lambda_{\text{NH}} = \sqrt{N}\left(\delta_{B(\text{NH})} + \Delta\right)$. If $d_{\text{CV(NH)}}$ is the critical value that the clinical settings effect size needs to exceed to claim clinical effectiveness beyond natural remission,

$$d_{\text{CV(NH)}} = t_{(\text{NH})\nu,\lambda:.95}/\sqrt{N}. \tag{11}$$

When the observed effect size exceeds this critical value, the null hypothesis is rejected, and the effectiveness of treatment in the clinical setting is statistically (and clinically) beyond natural remission.

2.3.4. *Possible results from benchmarking*

Depending on the observed effect size of the clinical settings d_D, there are three possible cases when compared against the treatment and natural history benchmarks. First, when d_D exceeds the critical value derived from the treatment efficacy benchmark (i.e., $d_{CV(TE)}$), it is concluded that d_D is clinically equivalent to the treatment efficacy benchmark $d_{B(TE)}$. In other words, this provides evidence that psychotherapies in clinical settings are at least as clinically effective as the clinical trials. Second, in the case that d_D does not exceed $d_{CV(TE)}$ but exceeds the critical value derived from the natural history benchmark (i.e., $d_{CV(NH)}$), it is concluded that d_D is not clinically equivalent to $d_{B(TE)}$, but is larger than the natural history benchmark $d_{B(NH)}$. This would indicate that while psychotherapies in clinical settings provide clinically meaningful effect as compared to wait-list controls in clinical trials, the effectiveness is not at par with psychotherapies in clinical trials. Third, in the case that d_D does not exceed $d_{CV(NH)}$, this indicates that TAU is not clinically superior to no treatment. This would suggest that psychotherapies in clinical settings do not have clinically meaningful effect over and above wait-list controls in clinical trials.

3. Illustration

For illustration, we demonstrate the results of the above calculations with the BDI treatment efficacy and natural history benchmarks that were derived from an aggregation of clinical trials of adult depression treatment (Minami et al., in press). Table I and Figure 1 summarize the calculated critical values. Most noncentral t critical values could be calculated by statistical packages such as SPSS, SAS, and NCSS.

As an example, let us assume a clinical settings data of $N = 1,000$ that utilized the BDI for their outcome measure. From Table I, we see that the treatment efficacy critical value is $d_{CV(TE-BDI)} = 1.5843$, and the natural history critical value is $d_{CV(NH-BDI)} = 0.6282$.

Depending on the observed effectiveness of the clinical settings data, there are three possible conclusions. First, in the case that the clinical settings effect size d_D is above the treatment efficacy critical value (i.e., $d_{CV(TE-BDI)} = 1.5843$), the conclusion would be that the treatments in clinical settings are clinically equivalent the clinical trials. Second, when d_D does not exceed $d_{CV(TE-BDI)}$ but exceeds the natural history critical value (i.e., $d_{CV(NH-BDI)} = 0.6282$), it is concluded that although the treatments in clinical settings do better than wait-list controls in clinical trials, they are clinically inferior to clinical trials. Third, if d_D does not exceed $d_{CV(NH-BDI)}$, the conclusion would be that the treatments in clinical settings are no different than wait-list controls in clinical trials, and thus cannot claim clinical effectiveness at all.

Table I. BDI Critical Values for Treatment of Adult Depression

Treatment Efficacy Critical Values				Natural History Critical Values			
N	$d_{CV(TE-BDI)}$	N	$d_{CV(TE-BDI)}$	N	$d_{CV(NH-BDI)}$	N	$d_{CV(NH-BDI)}$
	$d_{B(TE-BDI)} = 1.7059$				$d_{B(NH-BDI)} = 0.3712$		
100	1.7732	1000	1.5843	100	0.7597	1000	0.6282
125	1.7420	1250	1.5758	125	0.7386	1250	0.6221
150	1.7195	1500	1.5695	150	0.7232	1500	0.6176
200	1.6886	2000	1.5608	200	0.7019	2000	0.6113
250	1.6680	2500	1.5549	250	0.6875	2500	0.6070
300	1.6530	3000	1.5505	300	0.6770	3000	0.6039
400	1.6322	4000	1.5445	400	0.6624	4000	0.5995
500	1.6182	5000	1.5403	500	0.6525	5000	0.5965
600	1.6080	6000	1.5373	600	0.6452	6000	0.5942
800	1.5939	8000	1.5330	800	0.6351	8000	0.5911
		10000	1.5301			10000	0.5890

BDI = Beck Depression Inventory (Beck & Steer, 1987).

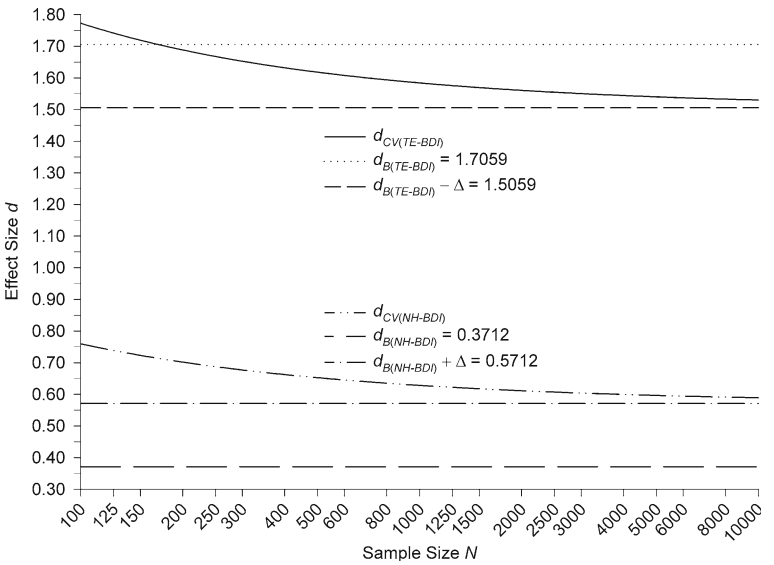

Figure 1. BDI effect size critical values by clinical settings data sample size.

4. Summary and Conclusion

The dearth of studies investigating the effectiveness of TAUs delivered in clinical settings has hampered our understanding of psychotherapy as it is practiced outside of the research environment. One probable explanation for this void is the lack of sound research designs and statistical methods that provide easily comprehensible interpretations of measured outcomes in clinical settings. Thus, this paper is an attempt to expand on a promising method, notably benchmarking, which allows for bridging clinical settings data with clinical trials data. Under the current force towards implementation of EBPs to clinical settings (e.g., Addis, 2002; Chorpita et al., 2002), it is of primary importance that the effectiveness of TAUs delivered in clinical settings be assessed before unilateral enforcement of EBPs in non-research settings.

Our advancement of the benchmarking method allows for a simple and intuitive statistical evaluation of effectiveness in clinical settings, provided that pre-post effect sizes are calculated in Cohen's *d*. After matching outcome measures by reactivity and specificity, clinicians and researchers can readily interpret how their data compares with benchmarks derived from clinical trials. Noteworthy from a clinical standpoint is the use of a range-null hypothesis testing procedure to avoid attaining statistical significance that has no clinical relevance. In addition, as more clinical trials with intent-to-treat samples are published, additional benchmarks could be established for different disorders and different outcome measures, taking into consideration their reactivity and specificity. Ideally, benchmarks would be aggregated for major outcome measures so that equivalence in measures could be sought between the clinical trials and clinical settings when benchmarking.

Even under the conditions that the outcome measures between the benchmarks and the clinical setting data are identical, however, there are several issues that warrant caution when interpreting a benchmarking study. First and foremost, the extent of similarity between the participants in clinical trials and clients in clinical settings needs to be taken into account on multiple dimensions, including clinical (e.g., multiple diagnoses, different treatment processes) and demographic (e.g., age, gender, race/ethnicity). Second, the differences in the environment between clinical trials and clinical settings need be considered, such as the costs and/or benefits for the client to be in treatment (e.g., time, energy, and resources) and the clients' perceptions on the use of standardized assessments in treatment. Third, there is great variability among clinical settings, and therefore, results obtained from benchmarking one type of clinical settings (e.g., managed care) may not be generalizable to other clinical settings (e.g., university and college counseling centers). Last but not least, the variability

among therapists in based on differences in their work condition, including selection, training, workload and stress, and supervision need to be taken into account (e.g., Rounsaville et al., 1988; Rupert and Baird, 2004). Therefore, the possible qualitative differences between the effect sizes obtained from the clinical settings data and the clinical trials benchmarks negate a simple conclusion based solely on numerical differences in effect sizes; unless complete qualitative equivalence regarding the above factors could be established between the clinical settings and clinical trials, it would be a gross misuse of the benchmarking strategy if definitive conclusions were drawn regarding differences in effectiveness and efficacy by its use. Benchmarking in no way determines the sources of differences; this can only be addressed by detailed investigations of differences in therapists, clients, and the treatment delivery processes between clinical trials and clinical settings.

References

Addis, M. E. (2002). Methods for disseminating research products and increasing evidence-based practice: Promises, obstacles, and future directions. *Clinical Psychology: Science and Practice* 9: 367–378.

Barlow, D. H. & Craske, M. G. (1994). *Mastery of Your Anxiety and Panic II.* Albany, NY: Graywind.

Barlow, D. H., Craske, M. G., Cerny, J. A. & Klosko, J. S. (1989). Behavioral treatment of panic disorder. *Behavior Therapy* 20: 261–282.

Beck, A. T. & Steer, R. A. (1987). *Beck Depression Inventory Manual.* San Antonio, TX: Harcourt Brace Jovanovich.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology* 41: 257–278.

Chorpita, B. F., Yim, L. M., Donkervoet, J. C., Arensdorf, A., Amundsen, M. J. & McGee, C. 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. *Clinical Psychology: Science and Practice* 9: 165–190.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Erlbaum.

Derogatis, L.R. (1983). *SCL-90-R: Administration, scoring, and procedural manual-II.* Baltimore, MD: Clinical Psychometric Research.

Gleser, L. J. & Olkin, I. (1994). Stochastically dependent effect sizes. In: H. Cooper & L. V. Hedges (eds.), *The Handbook of Research Synthesis*. New York: Russel Sage Foundation, pp. 339–355.

Goldfried, M. & Wolfe, B. (1998). Toward a clinically valid approach to therapy research. *Journal of Consulting and Clinical Psychology* 66: 143–150.

Hamilton, M. A. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry* 23: 56–62.

Hedges, L. V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* San Diego, CA: Academic Press.

Lambert, M. J., Hatch, D. R., Kingston, M. D. & Edwards, B. C. (1986). Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: A meta-analytic comparison. *Journal of Consulting and Clinical Psychology* 54: 54–59.

Merrill, K. A., Tolbert, V. E. & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting and Clinical Psychology* 71: 404–409.

Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C. & Brown, G. S. (in press). Benchmarks for psychotherapy efficacy in adult major depression. *Journal of Consulting and Clinical Psychology.*

Rounsaville, B. J., O'Malley, S., Foley, S. & Weissman, M. M. (1988). Role of manual-guided training in the conduct and efficacy of interpersonal psychotherapy for depression. *Journal of Consulting and Clinical Psychology* 56: 681–688.

Rupert, P. A. & Baird, K. A. (2004). Managed care and the independent practice of psychology. *Professional Psychology: Research and Practice* 35: 185–193.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The Consumer Reports Study. *American Psychologist* 50: 965–974.

Serlin, R. C. & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist* 40: 73–83.

Serlin, R. C. & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In: G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 199–228.

Shadish, W. R., Matt, G. E., Navarro, A. M. & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin* 126: 512–529.

Telch, M. J., Lucas, J. A., Schmidt, N. B., Hanna, H. H., Jaimez, T. L. & Lucas, R. A. (1993). Group cognitive-behavioral treatment of panic disorder. *Behaviour Research and Therapy* 31: 279–287.

Wade, W. A., Treat, T. A. & Stuart, G. L. (1998). Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy. *Journal of Consulting and Clinical Psychology* 66: 231–239.

Weersing, V. R. & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology* 70: 299–310.

Westen, D., Novotny, C. M. & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controllled clinical trials. *Psychological Bulletin* 130: 631–663.