Journal of Consulting and Clinical Psychology
2008, Vol. 76, No. 1, 116–124

0022-006X/08/$12.00 DOI: 10.1037/0022-006X.76.1.116

Benchmarking the Effectiveness of Psychotherapy Treatment for Adult Depression in a Managed Care Environment: A Preliminary Study

Takuya Minami
University of Utah

Bruce E. Wampold and Ronald C. Serlin
University of Wisconsin–Madison

Eric G. Hamilton
PacifiCare Behavioral Health

George S. (Jeb) Brown
Center for Clinical Informatics

John C. Kircher
University of Utah

This preliminary study evaluated the effectiveness of psychotherapy treatment for adult clinical depression provided in a natural setting by benchmarking the clinical outcomes in a managed care environment against effect size estimates observed in published clinical trials. Overall results suggest that effect size estimates of effectiveness in a managed care context were comparable to effect size estimates of efficacy observed in clinical trials. Relative to the 1-tailed 95th-percentile critical effect size estimates, effectiveness of treatment provided in this setting was observed to be between 80% (patients with comorbidity and without antidepressants) and 112% (patients without comorbidity concurrently on antidepressants) as compared to the benchmarks. Because the nature of the treatments delivered in the managed care environment were unknown, it was not possible to make conclusions about treatments. However, while replications are warranted, concerns that psychotherapy delivered in a naturalistic setting is inferior to treatments delivered in clinical trials appear unjustified.

Keywords: benchmarking, effectiveness, managed care, clinical trials, depression

More than a decade has passed since estimating the effect of psychotherapy as it is delivered in natural settings was identified as a critical issue in psychotherapy research (e.g., Barlow, 1981; Cohen, 1965; Luborsky, 1972; Seligman, 1995; Strupp, 1989; Weisz, Donenberg, Han, & Weiss, 1995). Although the benefits of psychotherapy have been investigated in laboratory environments with randomized clinical trials (RCTs) and found to be substantial as early as the late 1970s (Smith & Glass, 1977; also Smith, Glass, & Miller, 1980), surprisingly little is known about the effects of psychotherapy in natural settings. The dichotomy of laboratory and natural settings was emphasized by Seligman (1995), who discriminated between *efficacy*, which is now used to denote the effects of psychotherapy in RCTs, and *effectiveness*, which is used to denote the effects of psychotherapy in clinical practice.

The few studies that have investigated effectiveness over the years have provided mixed results, attributed in part to a variety of methodologies used to investigate effectiveness because of difficulty in using a randomized control group design in natural settings. Notably, three methods have been used to estimate the effects of psychotherapy in natural settings: clinical representativeness, direct comparison, and benchmarking. Clinical representativeness studies, including some of the analyses conducted by Smith et al. (1980), statistically estimate effectiveness from efficacy studies, which are based on factors that distinguish natural

Takuya Minami and John C. Kircher, Department of Educational Psychology, University of Utah; Bruce E. Wampold, Department of Counseling Psychology, University of Wisconsin–Madison; Ronald C. Serlin, Department of Educational Psychology, University of Wisconsin–Madison; Eric G. Hamilton, PacifiCare Behavioral Health, San Francisco; George S. (Jeb) Brown, Center for Clinical Informatics, Salt Lake City, UT.

Eric G. Hamilton is currently employed by United Behavioral Health, which took over PacifiCare Behavioral Health in a merger.

Part of this article is based on a doctoral dissertation, in partial fulfillment of the requirements for a doctorate in counseling psychology from the University of Wisconsin–Madison, completed by Takuya Minami under the guidance of Bruce E. Wampold and Ronald C. Serlin. An earlier version of this article was presented at the 36th annual international meeting of the Society for Psychotherapy Research, Montreal, Quebec, Canada. Partial funding for this study was provided by the Department of Counseling Psychology, University of Wisconsin–Madison, as a doctoral research award to Takuya Minami. Bruce E. Wampold and George S. (Jeb) Brown have periodically consulted with PacifiCare Behavioral Health (PBH) and, subsequent to its merger with United Behavioral Health (UBH), with UBH. All analyses in this article were conducted independently of PBH and UBH, and no editorial oversight was exercised by either organization, as per an agreement between PBH and the authors prior to undertaking this project.

Correspondence concerning this article should be addressed to Takuya Minami, Department of Educational Psychology, University of Utah, 1705 East Campus Center Drive, Room 327, Salt Lake City, UT 84112. E-mail: takuya.minami@ed.utah.edu

settings from clinical trials (e.g., random assignment of patients, use of treatment manuals). Most comprehensive clinical representativeness studies are meta-analyses conducted by Shadish and colleagues (Shadish, Matt, Navarro, & Phillips, 2000; Shadish et al., 1997) that estimated effectiveness of psychotherapy from an original pool of approximately 1,000 independent clinical trials. Although their investigations led them to conclude that the benefits of psychotherapy provided in clinically representative environments are similar to those attained in treatments delivered in clinical trials, their conclusions must be taken tentatively because only about 5% of treatments in these meta-analyses met even minimal criteria for clinical representativeness (Shadish et al., 1997). Estimates of effectiveness were based on statistical estimations rather than actual clinical outcomes in natural settings.

The second means of estimating effectiveness involves direct comparisons of treatments tested in RCTs with treatments delivered in natural settings. In these studies, empirically supported treatments (ESTs) or other manualized treatments are transported into natural settings and their pre- and posttreatment effects are compared with treatments that are already being offered in these natural settings, which are known as treatments-as-usual (TAU). Numerous studies have investigated the feasibility of ESTs in natural settings for various disorders, including panic disorder, depression, and substance abuse (e.g., Addis et al., 2004; Merrill, Tolbert, & Wade, 2003; Morgenstern, Blanchard, Morgan, Labouvie, & Hayaki, 2001).

The results of direct comparison studies have provided mixed results. For example, TAUs conducted in a community-based substance abuse treatment program showed benefits comparable to those of cognitive–behavioral therapy, which was implemented in the same setting (Morgenstern et al., 2001). However, Addis et al. (2004) reported that the delivery of an EST in a managed care environment, notably panic control therapy, attained significantly better outcomes for some variables than TAU did. The mixed results may be due to significant methodological issues. In many studies, training and supervision is provided to therapists in the EST condition, whereas therapists in the TAU condition do not receive any additional training relative to treatments being delivered or the disorder being treated (e.g., Addis et al., 2004). In some studies, there are also significant differences in the dose of treatment between ESTs and TAUs (e.g., Verheul et al., 2003). In addition, it is possible that implementation of ESTs produces stronger allegiance and expectancy effects in the EST condition as compared with the TAU condition, because therapists are delivering the EST as part of an experimental arrangement (e.g., Addis et al., 2004). In the youth literature, Weisz, Jensen-Doss, and Hawley (2006) conducted a meta-analysis of comparisons of evidence-based treatments to usual care and found that no study adequately controlled for confounding variables such as setting, therapists, training, and dose of treatment. It may well be that conducting direct comparisons in natural settings by implementing an alternative treatment produces biased estimates of TAU effectiveness.

One promising method for evaluating psychotherapy effectiveness, without altering any aspect of TAUs and obviating comparison to a treatment delivered with favorable conditions (e.g., extra therapist training and supervision), is to use benchmarks created from clinical trials. Specifically, benchmarking allows pre- and posttreatment data in natural settings to be compared with pre- and posttreatment data from clinical trials. For example, in assessing the effectiveness of TAUs for children and adolescents, Weersing and Weisz (2002) conducted a benchmarking study by comparing the symptom trajectory of depressed youths who were provided TAUs in community mental health centers with aggregated symptom trajectories derived from clinical trials. Their study significantly improved on previous benchmarking studies (e.g., Merrill et al., 2003; Wade, Treat, & Stuart, 1998) by constructing benchmarks on the basis of meta-analysis rather than by using a select number of clinical trials. Benchmarking thus allows for statistical evaluation of TAUs against rigorous standards established in clinical trials without altering any aspects of the TAUs delivered in natural settings and without delivering an established treatment with augmentations, such as supervision and training of therapists, in the established treatment condition.

The purpose of the current study was to evaluate the effectiveness of TAUs delivered in a managed health care organization (i.e., HMO) by means of a benchmarking strategy. Specifically, data on adult patients diagnosed with major depressive disorder (American Psychiatric Association, 1994) were statistically compared with benchmarks derived from clinical trials for adult major depression treatment (Minami, Serlin, Wampold, Kircher, & Brown, 2006; Minami, Wampold, Serlin, Kircher, & Brown, 2007). Benchmarking was conducted with samples that were progressively matched to the clinical population most commonly observed in clinical trials on the basis of their inclusion–exclusion criteria and antidepressant medication use. Additionally, we provided indices of relative strength of the observed treatment effect size estimates as compared with treatment efficacy observed in clinical trials.

Method

Participants

The original database for this study contained patient outcome data for 99,004 adult patients (i.e., 18 years or older) who began psychotherapy treatment with 7,593 treatment providers[1] between February 8, 1999, and December 31, 2004, under the insurance coverage of PacifiCare Behavioral Health, Inc. (PBH). Among the 99,004 patients, 12,743 patients diagnosed with major depression by their provider constituted the base dataset for this study. Available patient and provider demographics for this database are provided in Tables 1 and 2. Because of some aspects of the naturalistic setting, such as concerns for privacy, PBH did not routinely collect other patient and provider demographic information such as race–ethnicity, education, and income. Providers were licensed in their jurisdictions and had a master's degree or higher in one of the following fields: counseling or clinical psychology, marriage and family therapy, clinical social work, psychiatry, or nursing (with a specialization in psychiatry). PBH did not mandate or monitor the treatment approach used by the providers. Claims data were also

[1] Treatment providers include providers who practice individually (i.e., individual providers) and those who are in group practice (group providers). Group providers have an ID solely for their group and thus do not have ID numbers for each practicing provider within the group. Therefore, the same group ID may have two or more different credentials and/or other provider demographic information, reflecting on the unique providers participating in the group practice. Thus, the actual number of providers exceeds what is reported.

Table 1
Patient Demographic Information From Base and Subset HMO Data

Variable	Base data	Clinical sample	Noncomorbid sample	Completer sample
Patients, *N* (%)	12,743 (12.87[a])	5,704 (44.76[b])	939 (7.37[b])	253 (1.99[b])
Women, *n* (%)	8,933 (70.10)	4,035 (70.74)	624 (66.45)	172 (67.98)
Age *M* ± *SD* (range)	40.13 ± 11.20 (18–91)	40.05 ± 11.02, (18–86)	38.34 ± 10.81 (18–79)	38.88 ± 10.47 (18–69)
Sessions *M* ± *SD* (*Mdn,* range)	4.16 ± 5.16 (3, 1–99)	6.66 ± 5.72 (5, 3–99)	6.15 ± 5.60 (5, 3–99)	7.94 ± 3.88 (7, 3–20)
Days *M* ± *SD* (*Mdn,* range)	42.92 ± 61.65 (23, 0–1,016)	62.08 ± 60.85 (42, 1–936)	61.35 ± 59.92 (42, 1–399)	93.07 ± 24.16 (90, 61–147)
Provider training level[c]				
Master's, *n* (%)	5,491 (43.09)	2,508 (43.97)	389 (41.43)	84 (33.20)
Doctoral, *n* (%)	2,627 (20.62)	1,174 (20.58)	183 (19.49)	53 (20.95)
Medical, *n* (%)	1,302 (10.22)	313 (5.49)	63 (6.71)	22 (8.70)
Antidepressant use				
Concurrent, *n* (%)	6,937 (54.44)	3,225 (56.54)	440 (46.86)	127 (50.20)
None, *n* (%)	4,808 (37.73)	2,080 (36.47)	435 (46.33)	112 (44.27)

[a] Percentage of initial HMO data (i.e., $N = 126{,}972$). [b] Percentage of base HMO data (i.e., $N = 12{,}743$). [c] Provider training level does not add up to 100% because of missing data.

used to determine episodes of care, treatment duration, and medication use.

Outcome Measure

The Outcome Questionnaire–30.1[2] (OQ-30; Lambert et al., 2003)—a briefer version of the Outcome Questionnaire–45.2 (OQ-45; Vermeersch, Lambert, & Burlingame, 2000; Wells, Burlingame, Lambert, & Hoag, 1996)—was used to assess outcomes of patients included in the database. Specifically designed to minimize demands on patients who complete the instrument periodically during the course of therapy, it measures patient progress in three dimensions: (a) subjective discomfort, (b) interpersonal relationships, and (c) social role performance. Lambert et al. (2001) reported high internal consistency and test–retest reliability as well as concurrent validity with other symptom measures, such as the Global Severity Index of the Symptom Checklist 90—Revised ($r = .698$; Derogatis, 1977), Inventory of Interpersonal Problems ($r = .621$, Horowitz, Rosenberg, Baer, Ureno, & Villasenor, 1988), Social Adjustment Scale ($r = .593$, Weissman & Bothwell, 1976), and Beck Depression Inventory ($r = .609$; Beck & Steer, 1987).

Procedure

Initial data collection. Patients were asked to fill out the OQ-30 before their first, third, and fifth sessions, as well as every fifth session thereafter. This assessment program was implemented systemwide at PBH as their routine clinical assessment. Consent was needed by clinicians and patients, and in the present time period, approximately 65% of eligible patients participated. Whereas clinical trials can clearly define episodes of care as the period between when the participants entered the clinical trial and when they "completed" it or "dropped out," episodes cannot be as clearly defined in natural settings. Therefore, for the current study, an episode of care was defined as the cluster of outcome assessment points or sessions, as indicated by claims data, that did not have more than a 90-day gap between two sessions. In other words, if any two sessions were more than 90 days apart, the last observation before the gap was considered the posttest score for that episode of care. Independence of observations (at the patient level) was maintained by including in the database only the first episode of care for a given patient. Similarly, as the data collection was voluntary, patients varied in terms of the number and frequency of OQ-30s that were assessed. Consequently, the lack of data points cannot be interpreted straightforwardly as attrition, as numerous reasons are possible for absence of these data points (e.g., clinician and/or patient refusal to be assessed).

Data reduction. The base dataset containing 12,743 adult patients diagnosed with major depression was sequentially reduced to match the population represented in clinical trials that investigate efficacy of psychotherapy treatments for adult depression as RCTs employ various inclusion–exclusion criteria and fix the dose of treatment. Patients were included in the first reduced sample, denoted as the clinical sample, if they met the following two criteria: (a) they had a score of 43 or above on the OQ-30, which served as the clinical cutoff score on the basis of Jacobson and Truax's (1991) formula (Lambert et al., 2003), and (b) their clinical symptoms were assessed at least two separate times. This reduction resulted in 5,704 (44.76% of base data) adult patients with major depression with at least two OQ-30 assessments who were treated by 1,859 providers. The patients in the clinical sample likely differ from clinical trial samples in that they were not excluded on the basis of comorbidity, suicidality, and other factors (see Westen, Novotny, & Thompson-Brenner, 2004), and the treatment delivered may have had a duration shorter or longer than is typical in clinical trials. Available demographic and clinical information for this sample is also provided in Tables 1 and 2.

A second sample, denoted as the noncomorbid sample, was created by incorporating the following additional exclusion criteria: (a) absence of concurrent substance abuse, (b) absence of other comorbidity including medical–physical health issues, and (c) absence of significant suicidal and homicidal ideations. These criteria were determined by OQ indicator items, PBH claims data, and provider diagnoses. Applying these criteria commonly used in clinical trials reduced the number of patients to 939 (7.37% of base data) receiving treatment from 441 providers. All available demo-

[2] Through licensing agreement, the OQ-30 is named the Life Status Questionnaire at PBH.

Table 2
Provider Demographic Information From the Base and Subset HMO Data

Variable	Base HMO	Clinical sample	Noncomorbid sample	Completer sample
Providers, *N* (%)	3,225 (42.47[a])	1,859 (57.64[b])	441 (13.67[b])	139 (4.31[b])
Individual practice, *n* (%)	3,106 (96.31)	1,761 (94.73)	379 (85.94)	97 (69.78)
Women, *n* (%)	1,430 (44.34)	853 (45.88)	182 (41.27)	49 (35.25)
Men, *n* (%)	858 (26.60)	453 (24.37)	126 (28.57)	34 (24.46)
Age *M* ± *SD* (range)	57.31 ± 7.83 (34–84)	57.74 ± 7.66 (34–84)	57.81 ± 8.14 (35–82)	59.25 ± 7.25 (37–79)
Training level				
Master's, *n* (%)	1,292 (40.06)	805 (43.30)	182 (41.27)	62 (44.60)
Doctoral, *n* (%)	773 (23.97)	449 (24.15)	114 (25.85)	35 (25.18)
Medical, *n* (%)	224 (6.95)	51 (2.74)	16 (3.62)	2 (1.439)
No. of patients[c] *M* ± *SD* (range)	2.60 ± 3.26 (1–79)	2.08 ± 2.15 (1–32)	1.38 ± 1.16 (1–16)	1.14 ± 0.378 (1–4)
Years practiced *M* ± *SD* (range)	23.02 ± 8.20 (4–52)	22.71 ± 7.61 (4–51)	23.36 ± 8.36 (5–51)	24.28 ± 9.03 (8–51)

Note: As group practices do not have individual IDs for their therapists, two or more providers may report with the same IDs but with different demographic information. Percentages do not add up to 100% because of missing data.
[a] Percentage of initial HMO data (i.e., $N = 7{,}593$). [b] Percentage of initial HMO data (i.e., $N = 3{,}225$). [c] Among individual providers.

graphic and clinical information from this sample with regard to the patients and therapists is also provided in Tables 1 and 2.

The last sample, denoted as the completer sample, was constructed to include cases that approximated those who completed treatments in clinical trials, which is typically 12–20 sessions (Westen et al., 2004). This sample included only patients who were in treatment at least 60 days but less than or equal to 150 days and who did not have treatment that lasted more than 20 sessions. The completer sample included 253 (1.99%) of the initial 12,743 adult patients diagnosed with major depression.

Effect of antidepressant use. Antidepressant use was assessed with claims data so that effect size estimates could be evaluated for potential moderating effects of antidepressant use within the three subsets. Specifically, data were divided within subsets on the basis of antidepressant use, and effect size estimates were benchmarked separately.

Subset HMO Data Effect Size Calculation

Treatment effect size estimates were calculated in units of standardized pre- and post–mean change following basic meta-analytic procedures (Becker, 1988; Hedges & Olkin, 1985; Morris, 2000). The standard deviation of the intake score, rather than a pooled standard deviation, was used for standardization because it is presumably less influenced by repeated testing and/or treatment, thereby presenting a less confounded value (Becker, 1988; Morris, 2000). The estimated variance of the unbiased effect size estimate was calculated by means of an approximation given by Morris (2000; p. 19, formula 9).

Clinical Trials Benchmarks

Benchmarks for both treatment efficacy of psychotherapy for adult depression and natural history of depression were adapted from Minami, Wampold, et al. (2007). Minami, Wampold, et al. derived treatment efficacy benchmarks by meta-analytically aggregating standardized pre- and post–effect size estimates observed in published psychotherapy clinical trials for treatment of adult major depression, both for completer samples and intent-to-treat (ITT) samples. Similarly, Minami, Wampold, et al. constructed natural history benchmarks by using wait-list control groups. For the present study, ITT treatment efficacy benchmark for global symptoms self-report measures was used to make comparisons with clinical and noncomorbid subsets, as the OQ-30 was also a global well-being measure and contained all patients who completed as few as two outcome assessments. Accordingly, the ITT treatment efficacy benchmark was $d_{TE(ITT)} = 0.795$, and the natural history benchmark was $d_{NH} = 0.149$ (Minami, Wampold, et al., 2007). The mean numbers of weeks in treatment in clinical trials were approximately 16 for the efficacy benchmark and 10 for the natural history benchmark. Completer data were assessed against the completer treatment efficacy benchmark $d_{TE(C)} = 0.932$ (Minami, Wampold, et al., 2007) as well as the above natural history benchmark.

Benchmarking

Testing against the treatment efficacy benchmarks. The three samples were tested by means of the benchmarking strategy illustrated in Minami, Serlin, et al. (2006) against the respective treatment efficacy benchmarks (i.e., $d_{TE(ITT)} = 0.795$ or $d_{TE(C)} = 0.932$). This strategy tested the true effect size in the population as represented by the natural settings data against critical values derived from the benchmarks, taking into consideration a predetermined margin of effect size difference between the benchmarks and the effect size estimates observed in natural settings that could be deemed comparable while maintaining an overall Type I error rate of .05. For treatment efficacy, the margin of comparability was set at within 10% of the respective treatment efficacy benchmarks. In other words, if the natural settings effect size estimates were reliably as large as 90% of the treatment efficacy benchmarks (i.e., $d_{TE(ITT)90\%} = 0.715$ and $d_{TE(C)90\%} = 0.839$ for ITT and completers, respectively), the population effect sizes of treatments provided in natural settings as represented by the samples were considered comparable to true treatment efficacy effect sizes in clinical trials. To statistically compare the population effect size represented by the sample with the benchmark, taking into consideration the 10% margin, we adopted the "good-enough principle," which utilizes a noncentral *t* statistic (Serlin & Lapsley, 1985, 1993). This procedure allowed for hypothesis testing with a range-null rather than a point-null hypothesis while maintaining an

Table 3
Effect Size Estimates of Subset HMO Data by Antidepressant Use

Sample	Antidepressants	*N*	Intake *M* (*SD*)	Last *M* (*SD*)	r_{12}	d_{HMO}	$SE_{d(HMO)}$
Clinical	All	5,704	62.01 (11.90)	53.16 (14.88)	.4994	0.7445	0.0150
	Concurrent	3,225	63.66 (12.01)	53.83 (15.30)	.4822	0.8185	0.0206
	None	2,080	59.05 (10.85)	51.72 (13.81)	.5098	0.6752	0.0241
Noncomorbid	All	939	59.29 (11.16)	49.38 (14.60)	.4745	0.8870	0.0393
	Concurrent	440	60.71 (10.87)	49.96 (14.75)	.4616	0.9876	0.0598
	None	435	57.66 (10.86)	48.60 (14.05)	.4896	0.8330	0.0563
Completer	All	253	59.59 (11.39)	46.41 (15.12)	.3684	1.1536	0.0878
	Concurrent	127	60.50 (11.11)	46.77 (14.97)	.4070	1.2286	0.1252
	None	112	58.63 (11.58)	46.30 (15.19)	.3373	1.0573	0.1314

overall Type I error rate of .05, which permits reasonable conclusions about comparability.

Testing against the natural history benchmark. For the population treatment effect size to claim any effectiveness over and above the natural symptom trajectory of depression, effect size estimates from the sample must exceed at minimum $d = 0.2$ above the natural history benchmark $d_{NH} = 0.149$ (i.e., $d_{NH} + 0.2 = 0.349$). The margin of $d = 0.2$ was selected on the basis of Cohen's (1988) suggestion that this magnitude of effect size is small. That is, for TAU to be considered superior to the natural history of depression, the obtained effect must exceed the natural history benchmark by a reasonable amount.

Relative magnitude (RM) against the benchmarks. In order to intuitively interpret the differences in magnitude of effect size estimates between the clinical trials benchmarks and natural settings data, we calculated an index to illustrate the RM between the observed effect size estimate in the HMO data and the respective benchmarks with 95% confidence.[3] The RM is easily interpreted in percentages when multiplied by 100.

Results

Effect Size Estimates

Table 3 summarizes the effect size estimates. The $N = 5{,}704$ adult patients who had clinical depression in the clinical sample had mean intake and last session scores of $M_1 = 62.01$ ($SD = 11.90$) and $M_2 = 53.16$, respectively, and a resulting effect size estimate of $d_{clinical} = 0.7445$. On the basis of the pre- and posttest correlation in this sample of $r_{12} = .4994$, the standard error was $SE_{d(clinical)} = 0.0150$. Similarly, the noncomorbid sample with $N = 939$ patients had mean intake and last scores of $M_1 = 59.29$ ($SD = 11.16$) and $M_2 = 49.38$, respectively, resulting in an effect size estimate of $d_{noncomorbid} = 0.8870$ with a standard error of $SE_{d(noncomorbid)} = 0.0393$ ($r_{12} = .4745$). The completer subset with $N = 253$ had intake and last scores of $M_1 = 59.59$ ($SD = 11.39$) and $M_2 = 46.41$, respectively, resulting in an effect size estimate of $d_{completer} = 1.1536$ with a standard error of $SE_{completer} = 0.878$. For each of the subsets, effect sizes were also estimated separately on the basis of antidepressant use (see Table 3).

Benchmarking

Benchmarking subset data against treatment efficacy. First, the clinical sample effect (viz., $d_{clinical} = 0.7445$), which included adult patients diagnosed with depression but not excluded on the basis of comorbidity or length of treatment, was tested against the ITT treatment efficacy benchmark minus 10% (i.e., $d_{TE(ITT)90\%} = 0.715$) and was statistically significant ($t = 56.23$, $\lambda = 54.01$, $p < .001$). In the range null context, the statistically significant result indicates that the magnitude of treatment effectiveness in natural settings for adult patients with depression was comparable to ITT samples in clinical trials (i.e., reliably in the range). Here, the 95th percentile one-tailed critical value was $d = 0.7398$ (Minami, Serlin, et al., 2006). RM = 0.9059 indicated that the true treatment effect as estimated by the natural settings data was reliably expected to be at or above 90.59% in magnitude of the ITT treatment efficacy benchmark $d_{TE(ITT)}$ (see Table 4).

The effect size estimate from the noncomorbid subset $d_{noncomorbid} = 0.8870$ ($N = 939$), which excluded patients with comorbidity from the clinical sample, was also tested against the ITT treatment efficacy benchmark minus 10% (i.e., $d_{TE(ITT)90\%} = 0.715$) and was statistically significant ($t = 27.18$, $\lambda = 21.91$, $p < .001$). Thus, effectiveness of treatment provided to adult patients diagnosed with major depression without comorbidity in natural settings was also comparable in magnitude to treatment efficacy observed in clinical trials. The 95th percentile one-tailed critical value that the effect size estimate observed in this subset needed to exceed was $d = 0.7766$. RM = 1.0362 indicated that the true effect is expected at or above 103.62% in magnitude of the ITT treatment efficacy benchmark $d_{TE(ITT)}$ (see Table 4).

The completer subset ($N = 253$) effect size estimate $d_{completer} = 1.1536$, which was calculated with only those patients who were in treatment between 60 and 150 days but had 20 or fewer sessions, was tested against the completer treatment efficacy benchmark $d_{TE(C)} = 0.932$, again with a 10% margin (i.e., $d_{TE(C)90\%} = 0.839$), and was statistically significant ($t = 18.35$, $\lambda = 13.35$, $p < .001$); RM = 1.0931; see Table 4).

The clinical sample effect size estimate $d_{clinical} = 0.7445$, when compared with the natural history benchmark $d_{NH} = 0.149$, was statistically significant ($t = 56.23$, $\lambda = 26.33$, $p < .001$). For

[3] Specifically, when δ_{TE} is the true population treatment efficacy benchmark (in Cohen's *d*) that one intends to compare with the natural settings data d_{HMO} of sample size *N*, then $RM = \frac{\lambda_{HMO}}{\sqrt{N}\delta_{B(TE)}}$, (1) where λ_{HMO} is the noncentrality parameter when $t_{HMO}(= \sqrt{N}d_{HMO})$ equals the noncentral *t* critical value $t_{\nu,\lambda(HMO);\alpha}$ with degrees of freedom ν (= $N - 1$) and Type I error rate α (= .05 in this study).

Table 4
Subset HMO Data Benchmarking

					vs. treatment efficacy				vs. natural history		
Sample	N	d_{HMO}	$SE_{d(HMO)}$	t_{HMO}	$t_{\nu,\lambda:\alpha}$	d_{CV}	p	RM	$t_{\nu,\lambda:\alpha}$	d_{CV}	p
Clinical	5,704	0.7445	0.0150	56.23	55.87	0.7398	.025	0.9059	28.03	0.3711	<.001
Noncomorbid	939	0.8870	0.0393	27.18	23.80	0.7766	<.001	1.0362	12.40	0.4045	<.001
Completer	253	1.1536	0.0878	18.35	15.35	0.9652	<.001	1.0931	7.28	0.4577	<.001

Note. The clinical and noncomorbid subsets are benchmarked against ITT treatment efficacy benchmark (i.e., $d_{B(TE\text{-}ITT)} = 0.795$); completer subset is benchmarked against completer treatment efficacy benchmark (i.e., $d_{B(TE\text{-}C)} = 0.932$).

reference, the 95th percentile one-tailed critical value for the natural settings effect size to exceed to claim clinical effectiveness over and above the natural trajectory of depression was $d_{CV} = 0.3711$ (Minami, Serlin, et al., 2006). Similarly, effect size estimates of the noncomorbid and completer samples were also significantly larger than the natural history benchmark (see Table 4), indicating that treatments in natural settings produce benefits beyond natural remission of depression.

Effect of antidepressants. The effect size estimates that were aggregated on the basis of whether or not antidepressants were used were separately benchmarked (see Table 5). Other than the effect size estimate from the clinical samples with no concurrent antidepressant use ($d = 0.6752$, $t = 30.79$, $p = .950$), all subset data exceeded their respective noncentral t critical values. Therefore, when adult patients diagnosed with major depression and receiving treatment in natural settings have comorbid conditions and are not on antidepressants, the magnitude of treatment effectiveness was less than treatment efficacy observed in clinical trials that excluded patients based on comorbid conditions. For this comparison, RM = 0.7993 indicated that the magnitude of the treatment effect size was estimated to be reliably at or above approximately 79.93% in magnitude against the treatment efficacy benchmark that excluded patients with comorbidity. Otherwise, treatment effectiveness in natural settings was considered comparable in magnitude to what was observed in clinical trials, including benchmarking of the most exclusive data against the most exclusive benchmark (i.e., completer no antidepressants vs. completer treatment efficacy benchmark; $d = 1.0573$, $t = 11.19$, $p = .034$; RM = 0.9235).

Discussion

There has been a dearth of studies investigating the effectiveness of TAUs delivered in natural settings. To our knowledge, the present article reports the first benchmarking study of psychotherapy TAUs for treatment of adult clinical depression. Unlike studies comparing TAUs with ESTs that were transported into natural settings, this method allowed for an assessment of effectiveness of TAUs without any changes in the natural setting.

The results of the present study suggest that psychotherapy treatment for adult depression provided in an HMO setting is effective. The providers in this study generated effect size estimates that were similar in magnitude to those observed in clinical trials. With the sample that displayed the least benefits of therapy (i.e., adult patients with major depression who were not excluded on the basis of comorbidity and were not taking antidepressants), the effect size was estimated to be approximately 80% or above in magnitude as compared to that in clinical trials that typically exclude comorbid patients. In the most favorable comparison (i.e., patients did not have comorbid conditions and were on antidepressants), the effect size was estimated to be approximately 112%, or above the treatment efficacy calculated without patients on medications. Comparability in magnitude was also observed in the most matched comparison (i.e., limited comorbidity, no antidepressants, and comparable treatment dose). In all cases, the obtained effects were reliably greater than natural history benchmarks.

There was clear evidence of an effect attributable to concurrent administration of antidepressants, regardless of the level of patient inclusion–exclusion criteria. The effect (approximately $d = 0.15$) is consistent with other studies (e.g., Thase & Jindal, 2004) and approached 20% of the estimated effect size vis-à-vis the ITT and completer treatment efficacy benchmarks. Thus, effect of medication use in natural settings warrants further investigation.

Limitations of this study call for caution in interpreting the results. First, the treatment efficacy and natural history benchmarks that were used in this study, despite being the best indices

Table 5
Subset HMO Data Benchmarking by Antidepressant Use

Sample	Antidepressant	N	d_{HMO}	$SE_{d(HMO)}$	t_{HMO}	$t_{\nu,\lambda:\alpha}$	d_{CV}	p	RM
Clinical	Concurrent	3,225	0.8185	0.0206	46.48	42.48	0.7480	<.001	0.9879
	None	2,080	0.6752	0.0241	30.79	34.49	0.7561	.950	0.7993
Noncomorbid	Concurrent	440	0.9876	0.0598	20.72	19.58	0.8058	<.001	1.1219
	None	435	0.8330	0.0563	17.37	19.48	0.8064	.017	0.9326
Completer	Concurrent	127	1.2286	0.1252	13.85	11.51	1.0209	<.001	1.1081
	None	112	1.0573	0.1314	11.19	10.86	1.0259	.034	0.9235

Note. The clinical and noncomorbid subsets are benchmarked against ITT treatment efficacy benchmark (i.e., $d_{B(TE\text{-}ITT)} = 0.795$); completer subset is benchmarked against completer treatment efficacy benchmark (i.e., $d_{B(TE\text{-}C)} = 0.932$).

currently available, are a compilation of various self-report outcome measures that assess global symptoms (e.g., Symptom Checklist-90). While these benchmarks appeared to be the most representative to compare with the OQ-30 (see Minami, Wampold, et al., 2007), which is also a self-report measure of global symptoms, differences between the measures could potentially impact the results.

Second, because of the nature of naturalistic research in an HMO context, the characteristics of the psychotherapy delivered by the providers are unknown. Consequently, conclusions about the efficacy of types of treatment are precluded, as it is plausible that the TAUs practiced among the practitioners included ESTs. This speculation is reasonable given that many treatments of depression have been designated as empirically supported (Chambless et al., 1998), many providers may have been trained in programs that emphasize ESTs, and some HMOs and other payers utilize various evidence-based practice guidelines. However, it should be realized that in surveys of psychologists, about one third indicate that their theoretical orientation is eclectic or integrative and about one third indicate that they are psychoanalytically or psychodynamically oriented (e.g., Norcross, Hedges, & Castle, 2002; Norcross, Karpiak, & Santoro, 2005), suggesting that many therapists in the present study were not likely delivering ESTs as they are manualized.

Third, the managed care context of the current study would likely limit the generalizability of these findings to other natural settings (e.g., community mental health centers, university and college counseling centers). Managed care companies may engage in a variety of provider credentialing, peer review, and quality improvement activities, which not only may encourage the use of ESTs but may potentially impact outcomes in other ways that are unobservable in other natural settings. In addition, it is also feasible that the clinicians and patients in this provider network who voluntarily participated in the study may not be representative with regard to treatment outcomes in these settings. Unfortunately, there are no data that would allow for outcome assessments of the other 35% of clinicians in this provider network.

Fourth, the use of 10% (for treatment efficacy) and $d = 0.2$ (for natural history) may be objectionable as criteria for comparability. This criticism cannot be refuted unless the field reaches a consensus on an effect size that would constitute comparability (similar to the adoption of $\alpha = .05$ as the criterion Type I error rate). However, the ranges that were used in this study for hypothesis testing to determine comparability were small, and thus providers in this study produced effects that were close to the benchmarks. Indeed, in several instances, the obtained effect in the managed care context exceeded the respective clinical trial benchmarks.

Fifth, significant differences in the mean number of weeks between the benchmarks and our clinical and noncomorbid subsets were observed. Specifically, the ITT treatment efficacy benchmark averaged approximately 16 weeks in treatment, whereas the HMO subsets averaged less than 9. However, conclusions about the relative efficiency of RCT treatments and those in the current sample are tenuous, as patients in clinical trials are encouraged to continue in treatment regardless of whether they are sufficiently improved or fail to make additional progress. In light of evidence suggesting that most of the change in psychotherapy occurs early in treatment (Barkham et al., 1996; Howard, Kopta, Krause, & Orlinsky, 1986), it may well be that a 9-week effect size in clinical trials would not be much smaller than the 16-week benchmark used in this study. Nevertheless, the rate at which providers in clinical practice achieved effects (i.e., on average, less than nine sessions) is impressive.

Sixth, effects attributed to therapists have not been modeled in this study or in the clinical trials that comprise the benchmarks. In our data, intraclass correlation coefficients (ICCs) averaged approximately $\rho = .06$, indicating that about 6% of the true variance could be attributed to therapists (see also Wampold & Brown, 2005). However, ICCs as high as $\rho = .18$ were obtained when calculated with the completer sample and therapists who had two or more patients were included. Such magnitudes of therapist effects have also been reported in clinical trials (Crits-Christoph & Mintz, 1991; Kim, Wampold, & Bolt, 2006; Luborsky et al., 1986). Therefore, "treatment" effect size estimates most likely vary because of differences in therapeutic effect attributable to therapists in both natural settings and clinical trials.

Last, it is important to note that benchmarking cannot explain why clinical trials and natural settings are comparable or different. In particular, a comparable statistical effect observed between the natural settings data and clinical trials benchmark does not suggest that the "treatments" in the two settings are equivalent. Naturalistic practice settings and research environments are quite different with regard to many patient and therapist factors, such as heterogeneity among patients, funding structure, supervision and training, length of treatment, demand characteristics, patient assignment, and clinical caseload (Nathan, Stuart, & Dolan, 2000; Rounsaville, O'Malley, Foley, & Weissman, 1988; Rupert & Baird, 2004; Seligman, 1995; Wampold, 1997, 2001; Westen & Morrison, 2001; Westen et al., 2004). Thus, aggregated effect size estimates from both the clinical trials and natural settings incorporate these differences in setting as well as any differences between treatments. However, the results of the present study suggest that providers in an HMO setting are effectively and efficiently treating depression, an observation that should be comforting to patients and payers. While replications are warranted, concerns that psychotherapy treatments for adult depression practiced in natural settings is inferior to the treatments used in clinical trials appear unfounded.

References

Addis, M. E., Hatgis, C., Krasnow, A. D., Jacob, K., Bourne, L., & Mansfield, A. (2004). Effectiveness of cognitive-behavioral treatment for panic disorder versus treatment as usual in a managed care setting. *Journal of Consulting and Clinical Psychology, 72,* 625–635.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barkham, M., Rees, A., Stiles, W. B., Shapiro, D. A., Hardy, G. E., & Reynolds, S. (1996). Dose-effect relations in time-limited psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 64,* 927–935.

Barlow, D. H. (1981). On the relation of clinical research to clinical practice: Current issues. *Journal of Consulting and Clinical Psychology, 49,* 147–155.

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.

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun,

K. S., Daiuto, A., et al. (1998). Update on empirically validated therapies, II. *Clinical Psychologist, 51,* 3–16.

Cohen, J. (1965). Some statistical issues in psychological research. In B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw–Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology, 59,* 20–26.

Derogatis, L. R. (1977). *The SCL-90 manual: I. Scoring, administration and procedures.* Baltimore: Johns Hopkins University School of Medicine, Clinical Psychometrics Unit.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Villasenor, V. S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology, 56,* 885–892.

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41,* 159–164.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Kim, D., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research, 16,* 161–172.

Lambert, M. J., Hatfield, D. R., Vermeersch, D. A., Burlingame, G. M., Reisinger, C. W., & Brown, G. S. (2003). *Administration and scoring manual for the OQ-30.1.* East Setauket, NY: American Professional Credentialing Services.

Luborsky, L. (1972). Research cannot yet influence clinical practice. In A. Bergin & H. Strupp (Eds.), *Changing frontiers in the science of psychotherapy* (pp. 120–127). Chicago: Aldine.

Luborsky, L., Crits-Christoph, P., McLellan, A. T., Woody, G., Piper, W., Liberman, B., et al. (1986). Do therapists vary much in their success? Findings from four outcome studies. *American Journal of Orthopsychiatry, 56,* 501–512.

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., Serlin, R. C., Wampold, B. E., Kircher, J. C. & Brown, G. S. (Jeb). (2006). Using clinical trials to benchmark effects produced in clinical practice. *Quality and Quantity* [in press]. Retrieved on December 20, 2006, at http://www.springerlink.com/content/u11t4p0740u11557/

Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (2007). Benchmarks for psychotherapy efficacy in adult major depression. *Journal of Consulting and Clinical Psychology, 75,* 232–243.

Morgenstern, J., Blanchard, K. A., Morgan, T. J., Labouvie, E., & Hayaki, J. (2001). Testing the effectiveness of cognitive-behavioral treatment for substance abuse in a community setting: Within treatment and posttreatment findings. *Journal of Consulting and Clinical Psychology, 69,* 1007–1017.

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology, 53,* 17–29.

Nathan, P. E., Stuart, S. P., & Dolan, S. L. (2000). Research on psychotherapy efficacy and effectiveness: Between Scylla and Charybdis? *Psychological Bulletin, 126,* 964–981.

Norcross, J. C., Hedges, M., & Castle, P. H. (2002). Psychologists conducting psychotherapy in 2001: A study of the Division 29 membership. *Psychotherapy: Theory, Research, Practice, and Training, 39,* 97–102.

Norcross, J. C., Karpiak, C. P., & Santoro, S. O. (2005). Clinical psychologists across the years: The Division of Clinical Psychology from 1960 to 2003. *Journal of Clinical Psychology, 61,* 1467–1483.

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* (pp. 199–228). Hillsdale, NJ: Erlbaum.

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.

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crits-Christoph, P., Hazeligg, M. D., et al. (1997). Evidence that therapy works in clinically representative conditions. *Journal of Consulting and Clinical Psychology, 65,* 355–365.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.

Strupp, H. H. (1989). Psychotherapy: Can the practitioner learn from the researcher? *American Psychologist, 44,* 717–724.

Thase, M. E., & Jindal, R. D. (2004). Combining psychotherapy and psychopharmacology for treatment of mental disorders. In M. J. Lambert (Ed.), *Handbook of psychotherapy and behavior change* (5th ed.). New York: Wiley.

Verheul, R., van den Bosch, L. M. C., Koeter, M. W. J., de Ridder, M. A. J., Stijnen, T., & van den Brink, W. (2003). Dialectical behaviour therapy for women with borderline personality disorder: 12-month, randomised clinical trial in the Netherlands. *British Journal of Psychiatry, 182,* 135–140.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74,* 242–261.

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.

Wampold, B. E. (1997). Methodological problems in identifying efficacious psychotherapies. *Psychotherapy Research, 7,* 21–43.

Wampold, B. E. (2001). *The great psychotherapy debate: Model, methods, and findings.* Mahwah, NJ: Erlbaum.

Wampold, B. E., & Brown, G. (2005). Estimating therapist variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73,* 914–923.

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.

Weissman, M. M., & Bothwell, S. (1976). Assessment of social adjustment by patient self-report. *Archives of General Psychiatry, 33,* 1111–1115.

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging

the gap between laboratory and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology, 63,* 688–701.

Weisz, J. R., Jensen-Doss, A., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. *American Psychologist, 61,* 671–689.

Wells, M. G., Burlingame, G. M., Lambert, M. J., & Hoag, M. (1996). Conceptualization and measurement of patient change during psychotherapy: Development of the Outcome Questionnaire and Youth Outcome Questionnaire. *Psychotherapy, 33,* 275–283.

Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69,* 875–899.

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

Received January 18, 2006
Revision received September 11, 2007
Accepted September 12, 2007 ■