Evidence-Based Psychiatric Treatment

A Quality-Based Review of Randomized Controlled Trials of Cognitive-Behavioral Therapy for Depression: An Assessment and Metaregression

Nathan C. Thoma, Ph.D.

Dean McKay, Ph.D.

Andrew J. Gerber, M.D., Ph.D.

Barbara L. Milrod, M.D.

Anna R. Edwards, Ph.D.

James H. Kocsis, M.D.

Objective: The authors assessed the methodological quality of randomized controlled trials of cognitive-behavioral therapy (CBT) for depression using the Randomized Controlled Trial Psychotherapy Quality Rating Scale (RCT-PQRS). They then compared the quality of CBT trials with that of psychodynamic therapy trials, predicting that CBT trials would have higher quality. The authors also sought to examine the relationship between quality and outcome in the CBT trials.

Method: An independent-samples t test was used to compare CBT and psychodynamic therapy trials for average total quality score. Metaregression was used to examine the relationship between quality score and effect size in the CBT trials.

Results: A total of 120 trials of CBT for depression met inclusion criteria. Their mean total quality score on the RCT-PQRS was 25.7 (SD=8.90), which falls into the lower range of adequate quality. In contrast to our prediction, no significant difference was observed in overall quality between CBT and psychodynamic therapy trials. Lower quality was related to both larger effect sizes and greater variability of effect sizes when analyzed across all available comparisons to CBT.

Conclusions: On average, randomized controlled trials of CBT and of psychodynamic therapy did not differ significantly in quality. In CBT trials, low quality appeared to reduce the reliability and validity of trial results. These findings highlight the importance of discerning quality in individual psychotherapy trials and also point toward specific methodological standards for the future.

(Am J Psychiatry 2012; 169:22–30)

In recent years, the methodological quality of randomized controlled trials of a variety of treatments has been more closely examined, particularly in meta-analyses and systematic reviews (1). In assessing the quality of randomized controlled trials of psychotherapy, special considerations are necessary because of the unique nature of psychotherapy as an experimental condition. Unlike medication, in psychotherapy the treatment and treatment delivery are highly interwoven, taking shape in a complicated human interaction unfolding over time. Issues such as training and supervision of psychotherapists, replicability of the treatment protocol, and verification of psychotherapist adherence and competence are all important to the validity of a psychotherapy trial. For these reasons, a subcommittee of APA's Committee on Research on Psychiatric Treatments developed a rating scale for the quality of randomized controlled trials of psychotherapy, the Randomized Controlled Trial of Psychotherapy Quality Rating Scale (RCT-PQRS) (2).

Gerber et al. (3) applied this scale to all known randomized controlled trials (through May 2010) of psychodynamic therapy, for any diagnosis, in an attempt to assess the quality and depth of the research base for psychodynamic therapy. The authors found both strengths and limitations in the trials studied. Notably, only 54 of the 94 trials analyzed (57.4%) received a total quality score of 24 or higher (of a possible score of 48), which had been established a priori as the cutoff for a minimally adequate level of quality. The authors also found that several specific methodological practices had been implemented poorly in more than half of the trials: the reporting of safety and adverse events; use of intent-to-treat analyses; and statistical consideration of therapist and site effects. A question Gerber et al. posed but left unanswered was how other psychotherapy treatments, such as cognitive-behavioral therapy (CBT), would fare when subjected to a similar quality-based review—on their own and in comparison with psychodynamic therapy.

CBT, from its inception, grew out of basic and applied research (4, 5), and it remains closely tied to ongoing research. Psychodynamic therapy, in contrast, grew out of a clinical tradition, and critics argue that psychoanalysis

went on to develop a culture that eschewed the use of science (6). Whereas Gerber et al. (3) found 94 randomized controlled trials of psychodynamic therapy in the literature, there are far greater numbers of randomized controlled trials of CBT. Thus, we predicted that the quality of CBT trials would be significantly higher than that of psychodynamic therapy trials.

We also sought to examine the relationship between quality and outcome. Results of previous studies of this relationship in the psychotherapy literature have been inconsistent (7–12). However, few (if any) of these analyses incorporated the use of a psychometrically validated quality scale specifically designed to assess randomized controlled trials of psychotherapy.

Using the RCT-PQRS, Gerber et al. (3) did not find any relationship between quality and outcome in trials of psychodynamic therapy. However, that study examined trials of psychodynamic therapy for a wide variety of diagnoses. Limiting the analysis to a single treatment and a single diagnosis eliminates confounding factors related to differential response of specific problems to specific treatments. We thus chose to limit the scope of the present study to randomized controlled trials of CBT for depression, which we believed would yield the greatest number of CBT trials for any single diagnosis (13).

We predicted that lower quality would be related to larger effect sizes, reasoning that a looser internal validity could provide more room for a variety of experimenter biases to influence outcome in ways that would yield more significant results. We also predicted that lower quality would be related to greater variability of outcomes as a result of the general effects of less tightly controlled experimental conditions.

Method

Sample

In June 2010, we conducted a search using the Cochrane Central Register of Controlled Trials, a registry of randomized controlled trials that has been developed through systematic searches of MEDLINE, Embase, CINAHL, LILACS, the "gray literature" of unpublished results, and hand searches (14). (Our keyword list is included in the data supplement that accompanies the online edition of this article.) In addition to the electronic search, we examined the reference sections of more than 30 meta-analyses of psychotherapy for depression, as well as the reference sections of the articles from the electronic search that met our inclusion criteria.

To be included, a study had to be a randomized controlled trial of treatment for depression that included a CBT treatment arm. CBT was defined as an active psychological treatment that either was identified as being cognitive-behavioral in nature or was described in terms of having the active use of cognitive restructuring as central to the therapy. Behavior therapies that lacked an explicit cognitive component, such as behavioral activation therapy, were excluded. The treatment could be group or individual CBT, but it had to be delivered in person by a trained therapist. Trials of bibliotherapy, teletherapy, computer therapy, or Internet therapy were excluded if they did not include a treatment arm with in-person psychotherapy. The study subjects of all trials were clinical samples of patients specifically seeking treatment for depressive symptoms. Trials of depression with psychotic features or bipolar disorder were excluded. Trials using the following patient populations were also excluded: analogue samples (e.g., undergraduates in a subject pool); persons with a medical illness that may be causal or highly related to their depression (e.g., congestive heart failure); persons with postpartum depression. Pilot studies and studies not published in English were excluded. Only one publication per trial was selected to be rated, similar to the method used in Gerber et al. (3); the publication that most thoroughly represented the methods and main posttest findings was chosen as the primary publication. Once the final list of studies was compiled, outside experts were consulted to ensure that no studies had been missed.

The sample of randomized controlled trials of psychodynamic therapy for use in the quality comparison was the same as that used by Gerber et al. (3). The sample was compiled using similar search methods, and it similarly included studies published in English through May 2010. Any randomized controlled trial that included a treatment identified as "psychodynamic" or "psychoanalytic," conducted in any modality (e.g., individual, group), for all available diagnoses, was included in that sample. As with the sample of CBT studies, pilot studies were excluded.

Quality Measure: The RCT-PQRS

The RCT-PQRS (2) is a measure of the methodological quality of randomized controlled trials of psychotherapy; it was developed by a committee of experts with divergent primary allegiances (e.g., CBT, psychodynamic therapy, pharmacology) and built on preexisting quality measures of randomized controlled trials (15–17). It was chosen for the present study because it has good psychometric properties and because it was designed specifically to assess the quality of psychotherapy trials, giving it advantages over other quality scales. The RCT-PQRS consists of 24 items corresponding to elements of study design, execution, and reporting, each rated 0 (poor description, execution, or justification of a design element), 1 (brief description or either a good description or an appropriate method or criteria set, but not both), or 2 (well described, executed, and, where necessary, justified design element). The 24 items are divided into six domains: description of subjects (e.g., diagnostic method and criteria for inclusion and exclusion); definition and delivery of treatment (e.g., method to demonstrate that the treatment being studied is the treatment being delivered); outcome measures (e.g., outcome assessment by raters blind to treatment group and with established reliability); data analysis (e.g., appropriate consideration of therapist and site effects); treatment assignment (e.g., appropriate randomization procedure performed after screening and baseline assessment); and overall quality of study (e.g., balance of allegiance to types of treatment by practitioners). Additionally, the scale offers the option of an "omnibus" item, a global rating of quality ranging from 1 (exceptionally poor study) to 7 (exceptionally good study). In the present study, an unweighted summative total of the 24 items, or quality score, was used as the primary measure.

The scale has been shown to have good internal consistency, with a Cronbach's alpha of 0.87 (2). The correlation between quality score and year of publication for 69 psychodynamic therapy randomized controlled trials was 0.51, which can be considered a measure of external validity. Criterion validity was demonstrated through ratings of two trials widely regarded as exemplars of high-quality designs for randomized controlled trials of psychotherapy—the National Institute of Mental Health Treatment of Depression Collaborative Research Program and the Treatment of Adolescents With Depression Study—with quality scores of 40 and 38, respectively (3), which can be considered relatively high scores on the scale. Interrater reliability (as measured by the in-

traclass correlation coefficient, ICC) has been reported to be 0.76 for quality score and 0.79 for the omnibus item (2). A copy of the scale is included in the online data supplement.

Method of Rating Study Quality

The first author rated all studies. To establish interrater reliability, 24 randomly selected studies were also rated by two of the scale developers (A.J.G. and B.L.M.), who each rated 12 studies. Raters came from both a CBT orientation (N.C.T.) and a psychodynamic therapy orientation (A.J.G. and B.L.M.). Each study was rated on the basis of the material represented in the publication. Raters were not blind to any details of the studies but were blind to one another's ratings. Interrater reliability was assessed using the ICC formula ICC(1,1) (18) as implemented in SAS, version 9.1 (SAS Institute, Inc., Cary, N.C.). For the 24 corated CBT trials, the ICCs were 0.88 for the quality score and 0.72 for the omnibus score, which are in the good to excellent range (19). Item-level ICCs varied, with 10 of 24 items scoring <0.60. As is common in rating scales, low ICCs at the item level can be an artifact of high agreement with low variability (16). It was expected that high ICCs would not be consistently achieved at the item level, and therefore all planned statistical analyses involved only the quality score, for which excellent reliability was achieved. Item-level descriptive analyses are presented for exploratory purposes only.

Metaregression

To examine the correlation between quality score and effect size, a random-effects metaregression was conducted using the Metareg macro (20) in Stata, version 10.0 (StataCorp, College Station, Tex.). Metaregression is a weighted regression that gives studies with larger sample sizes more weight and is recommended in a meta-analytic context (21). Effect size was used as the dependent variable, and quality score and comparison type were entered as predictors. Comparison type was used as a predictor in addition to quality score in order to control for substantial differences in effect size related to the strength of the comparator (e.g., waiting list, treatment as usual, medication) relative to CBT. This allowed the inclusion of the full breadth of CBT comparisons within a single regression model and examination of the relationship between quality and outcome across all randomized controlled trials of CBT for depression.

Effect size data were extracted for all comparisons to CBT, including inactive control groups (e.g., waiting list) and active comparators (e.g., other psychotherapies or medication). Standardized mean difference effect sizes (Hedges' g) were calculated for measures of depression using the software package Comprehensive Meta-Analysis, version 2.0 (Biostat, Englewood, N.J.). Each g and the corresponding standard error were corrected for bias due to unreliability of measurement according to the methods of Hunter and Schmidt (22). When insufficient data were provided in the publication to extract an effect size, other publications related to the same trial were checked, and failing this, study authors were contacted.

Five categories were coded for comparison type: CBT versus waiting list; CBT versus treatment as usual, attention placebo, or pill placebo; CBT versus other bona fide psychotherapy; CBT versus medication; and CBT plus medication versus medication alone. Additionally, the specific criteria of Wampold et al. (23) were employed to distinguish bona fide psychotherapies from those that function as attention placebos, given empirical evidence for this distinction. Comparisons of two bona fide versions of CBT were excluded. Dummy coding was used to enter the categorical variables into the regression, with CBT versus waiting list as the reference variable.

Reliability was demonstrated for effect sizes and type of comparison by having one author familiar with meta-analytic methods (A.R.E.) extract these data from 24 randomly selected studies. The ICC was 1.00 for effect size and 1.00 for standard error. Cohen's kappa was 0.79 for comparison type. These reliabilities can be considered to be in the excellent range (19). All data used in the analysis were those extracted by the first author.

Results

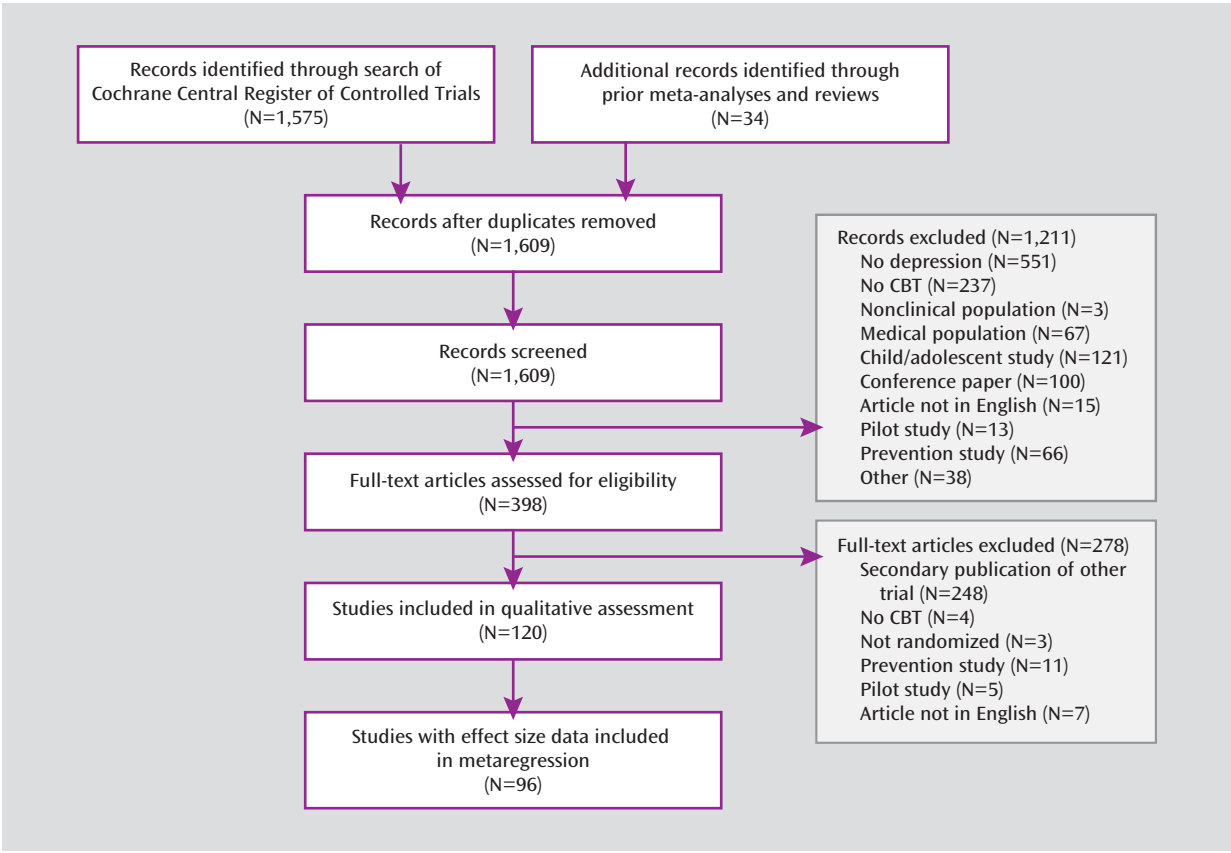
Study Characteristics

As shown in Figure 1, a total of 120 randomized controlled trials of CBT published from 1977 through May 2010 met criteria for inclusion in the quality analysis, and 96 of these provided sufficient data for effect size extraction and contained at least one non-CBT comparator. A complete table of study characteristics is included in the online data supplement; highlights are presented here. Total sample sizes ranged from 16 to 662, with a mean of 92.8 (SD=99.5, median=57). Sample sizes in treatment arms ranged from six to 226, with a mean of 32.9 (SD=35.4, median=20). Diagnostic criteria for depression varied among the 120 CBT studies. Fifty-four studies focused specifically on major depressive disorder, and 23 studies included a mix of depression diagnoses: major depressive disorder, dysthymia, and minor depression. Four studies focused specifically on minor depression or dysthymia. Thirty-two studies used a symptom measure cutoff for inclusion, and seven included subjects who were simply seeking treatment for depressive symptoms. This breadth of diagnostic methods allowed us to include and examine low-quality trials that did not use structured diagnostic interviews. All studies focused specifically on the treatment of depressive symptoms.

Among the 94 trials of psychodynamic therapy from the Gerber et al. study (3), total sample sizes ranged from 10 to 487, with a mean of 86.7 (SD=69.1, median=72). The number of comparators ranged from one (also the mode) to three. Sample sizes in each arm ranged from five to 122, with a mean of 36.6 (SD=26.0, median=30). The total number of subjects was 7,200. Seventeen studies contained group psychodynamic therapy, and 19 contained at least one nonindividual format in one of the comparator arms. The studies focused on a wide range of diagnoses, the most common of which was depression (N=19).

The total number of subjects in the 96 trials of CBT that were included in the quantitative analysis was 10,423. A total of 32 trials contained group CBT (33.3%), and one study contained family CBT (1.04%). Thirty studies (31.3%) contained a nonindividual format in one of the comparator arms. The number of comparators ranged from one to four (mean=1.85, SD=0.75). Because there were multiple treatment arms in some studies, the total number of possible comparisons to CBT (and thus effect sizes) was 153. Because including more than one effect size from a given trial would violate the assumption of independence of observations in the metaregression, an a priori hierarchy was used to select one effect size per study. The hierarchy was chosen on the basis of what we thought to be the greatest

FIGURE 1. Flow Diagram of Study Selection for a Quality Assessment of Randomized Controlled Trials of Cognitive-Behavioral Therapy (CBT)



replicability and stability of treatment in the comparator. This approach was meant to be conservative for our analyses, as we believed it would help reduce between-group effect size variability due mainly to variability in the comparator. Because the non-CBT psychotherapies and their implementations were anticipated to vary widely, psychotherapy was given low priority in the hierarchy, and treatment as usual/attention placebo was given even lower priority. Thus, the following order of priority was used in selecting the comparison arm: waiting list, medication, bona fide psychotherapy, and treatment as usual/attention placebo. To maintain balanced groups in the metaregression, the few studies comparing CBT plus medication and medication (N=4) were combined with those comparing CBT and medication. Table 1 presents data on each type of comparison.

The point estimates and confidence intervals of the aggregated effect sizes for each type of comparison are listed in Table 1, derived from a random-effects meta-analysis that was grouped by comparison type. The aggregated effect sizes are offered for descriptive purposes, as they help validate the categories of comparison type, which are in line with theoretical expectations as well as prior empirical findings (8, 23–26). Consistent with the results of Gerber et al. (3), who found that a majority of comparisons between psychodynamic therapy and another active treatment showed no significant difference, meta-analytic effect sizes for the comparison of CBT with other psychotherapies or with medication were small and nonsignificant.

Quality Scores

Quality scores for the 120 CBT studies ranged from 5 to 44 (mean=25.7, SD=8.90, median=26), indicating that the average CBT trial appears to be of minimum adequate quality, according to a cutoff of 24 on the quality score, as established by Gerber et al. (3). Quality was significantly correlated with year of study ($r=0.44$, $p<0.0001$), indicating that quality improved over time (Figure 2). Individual item scores, while limited in their potential for interpretation because of variable interrater reliability at the item level, suggest relative strengths and weaknesses of the studies in terms of methodological features. For seven of the items, at least half of the studies were scored as "good" (a rating of 2): item 1 (diagnostic method and criteria for inclusion and exclusion), item 5 (treatments are sufficiently described or referenced to allow for replication), item 10 (use of validated outcome measures), item 11 (primary outcome measure specified in advance), item 20 (a priori relevant hypotheses that justify comparison groups), item

TABLE 1. Comparisons of Cognitive-Behavioral Therapy (CBT) With Other Treatments in the Regression Analysis[a]

Comparison Type	N	Effect Size	95% CI	z	p	I^2 (%)
CBT versus waiting list	29	0.90	0.68 to 1.11	8.31	<0.001	69.1
CBT versus treatment as usual or attention placebo	18	0.40	0.17 to 0.63	3.40	<0.001	63.7
CBT versus other psychotherapy[b]	26	0.05	–0.07 to 0.18	0.83	0.41	22.6
CBT versus medication	23	0.10	–0.08 to 0.28	1.10	0.27	71.8
Total	96					

[a] One effect size was used per study in the regression, chosen according to an a priori hierarchy. The meta-analytic effect sizes are offered for descriptive purposes; they help validate the categories of comparison type in that they are in line with theoretical expectations as well as previous empirical findings. The meta-analytic effect sizes are from a random-effects meta-analysis using Hedges' g as the effect size, grouped by comparison type. The z value is the statistic used to test significance. The I^2 statistic estimates the proportion of heterogeneity within groups that is not due to sampling error.

[b] The specific types of psychotherapy were psychodynamic therapy (N=6), interpersonal therapy (N=3), behavior therapy (N=10), humanistic therapy (N=3), and other (N=4).

21 (comparison groups from the same population and time frame as the experimental group), and item 22 (appropriate randomization to treatment groups). Three of these items (items 5, 20, and 21) overlapped with those that were rated highly in the psychodynamic therapy studies examined by Gerber et al. (3). Five items were rated as "poor" (a rating of 0) on 50% or more of the studies: item 3 (description of relevant comorbidities), item 12 (outcome assessment by raters blind to treatment group and with established reliability), item 13 (discussion of safety and adverse events during study treatments), item 15 (use of intent-to-treat method), and item 19 (appropriate statistical consideration of therapist and site effects). Three of these items (items 13, 15, and 19) were also rated poorly in the psychodynamic therapy studies (3).

Quality Comparison of CBT and Psychodynamic Therapy Trials

In the comparison of mean quality scores between the 120 randomized controlled trials of CBT for depression and the 94 psychodynamic therapy trials for a variety of diagnoses, seven studies that contained both treatments were excluded. For the remaining 113 CBT trials, the mean quality rating was 25.5 (SD=9.13). For the remaining 87 psychodynamic therapy studies, the mean quality rating was 25.1 (SD=9.04). Contrary to our prediction, an independent-samples t test indicated no significant difference in study quality between trials featuring these two psychotherapies.

The psychodynamic therapy trials focused on an array of diagnoses, whereas the CBT trials all focused on depression. Therefore, we also tested the quality scores of the CBT trials against the 13 psychodynamic therapy trials that focused on depression (mean=24.3, SD=6.47), again excluding trials that used both treatments. A Mann-Whitney U test found no significant difference.

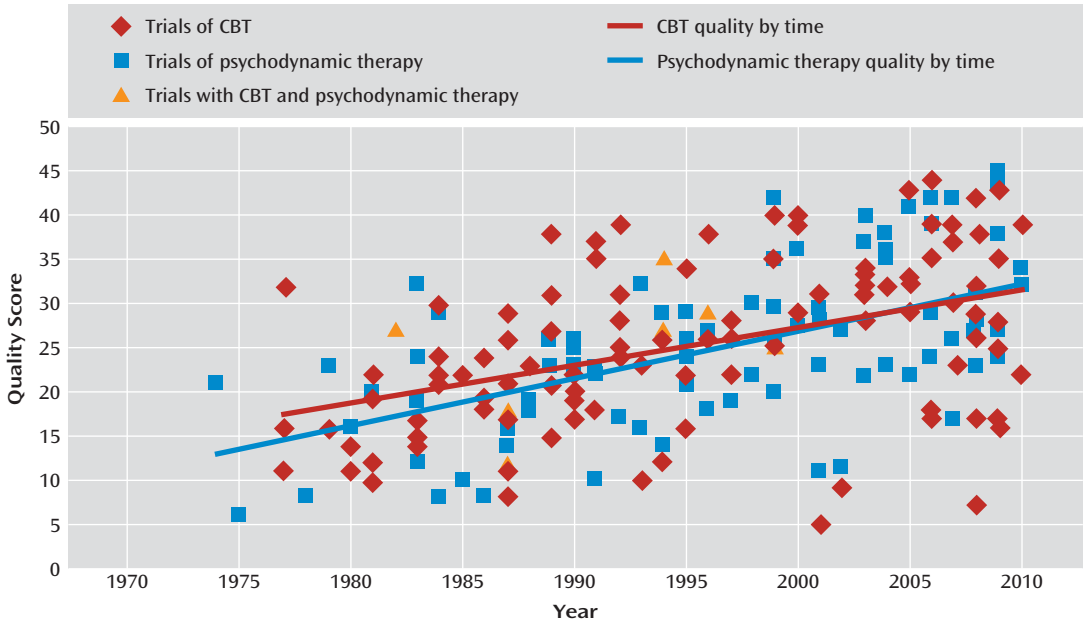
The plots of quality score and year of publication for both CBT trials and psychodynamic therapy trials are shown together in Figure 2. To further examine whether CBT and psychodynamic therapy trial quality could be statistically distinguished, we tested for differences in the degree of correlation between quality score and year for the two groups of trials, differences in the slopes of their regression lines, and differences in their intercepts. These tests would indicate whether the quality of the two groups of trials differed in their relationship to time, their rate of improvement over time, or their starting point in time, respectively. Fisher's z test found no significant differences in the degree of correlation, a t test of slopes (27) found no significant differences in slopes, and a t test of intercepts (27) found no significant differences in intercepts.

Quality and Outcome in Trials of CBT for Depression

The metaregression model that included quality score and comparator type as predictors and effect size as the dependent variable was significant ($F=12.6$, df=4, 91, $p<0.0001$; adjusted $R^2=0.525$). Residual heterogeneity was moderate ($I^2=61.0\%$). The quality score was a significant predictor of effect size ($t=–2.88$, $p=0.005$), indicating that lower quality was associated with larger effect sizes. The regression coefficient (B=–0.021, 95% CI=–0.025 to –0.008) indicated that for every 10-point increase in quality score, effect size (g) would be predicted to decrease by 0.21. Starting with a large effect of g=0.80, for example, this would represent a 25% decrease. The metaregression was rerun in a secondary analysis with all 153 possible comparisons to CBT. While using all the possible comparisons violates the assumption of independence, it is one check against the effects of the hierarchy used to select the comparators in the primary analysis. The results were significant and nearly identical. Likewise, when all observations that contained a comparison of CBT to behavior therapy were removed, the results were nearly identical and again significant.

It is possible that the relationship observed between quality and outcome is related to publication bias. Publication bias could contribute to the correlation between quality and outcome if small unpublished studies with null results also happen to be of low quality. To test for publication bias simultaneously across all effect sizes (and thus all comparison types), individual effect sizes were first centered on meta-analyzed effect sizes for each respective comparison type (as listed in Table 1). Subsequent application of Egger's regression test (28) and Duval

FIGURE 2. Scatterplot of Quality Score by Year of Publication for Randomized Controlled Trials of Cognitive-Behavioral Therapy and Psychodynamic Therapy[a]



[a] The quality score is the sum of the 24 items of the Randomized Controlled Trial Psychotherapy Quality Rating Scale (possible scores range from 0 to 48). The regression lines do not include the trials that contained both cognitive-behavioral therapy and psychodynamic therapy. There was no significant difference between the two groups in their mean quality scores; their degree of correlation between quality score and time; the slopes of the regression lines; or the intercepts of the regression lines.

and Tweedie's trim and fill procedure (29) both indicated a possible influence of publication bias on aggregated study outcomes. Notably, however, standard error (used as a proxy for sample size in tests of publication bias) and quality score were highly correlated ($r=-0.68$, $p<0.0001$). This indicates that the specific impacts of quality versus publication bias on outcome may not be possible to disentangle.

In addition to predicting that low quality would be related to inflated outcome, we also predicted that it would be related to greater variability of outcome. In statistical terms, this predicts heteroscedasticity related to quality score. The Breusch-Pagan test of heteroscedasticity (30) was significant ($F=8.05$, $df=1, 94$, $p=0.006$; adjusted $R^2=0.069$), indicating that lower quality was related to greater variability in outcome. This test was also significant for the model that used all 153 comparisons to CBT. It was also significant when models with fixed effects were used, which are more conservative in tests of heteroscedasticity because of their stronger weighting scheme.

Discussion

Perhaps the most surprising finding of this review was the lack of a significant difference between the mean quality scores of 113 randomized controlled trials of CBT for depression and 87 randomized controlled trials of psychodynamic therapy for a variety of diagnoses. Our hypothesis that CBT trials would be of higher quality because of the historically stronger emphasis on research within the culture of CBT than within that of psychodynamic therapy was not supported. This finding calls attention to the existence of high- and low-quality studies within both CBT and psychodynamic therapy trials. In Figure 2, the regression lines for CBT and psychodynamic therapy trial quality scores by year of publication appear to be nearly coincident. Statistical tests indicated that the strength of the correlation between quality score and year of publication was not significantly different; neither were the slopes or intercepts of the regression lines. Both sets of studies have been improving over time, and both contain randomized controlled trials with a similar range of quality.

While the psychodynamic therapy trials covered a range of diagnoses and the CBT trials centered on depression, we see no theoretical reason why diagnosis should be related to methodological quality. Furthermore, no difference was observed between the CBT trials and the subset of 13 psychodynamic therapy trials that focused on depression. Hence, we believe that the comparison of quality between the two samples of studies is a fair one.

It should be borne in mind that only the methodological quality of the trials of CBT and psychodynamic therapy was compared in this study. The question of comparing outcomes of the two treatments has been systematically addressed elsewhere (23, 31–33). Moreover, the lack of a significant difference in mean quality score between the two modalities does not indicate that the evidence base for the treatments is equivalent or even similar. The num-

ber of CBT trials far outstrips that of psychodynamic therapy trials (34), adding robustness to the overall support for CBT. An assessment of the evidence base using the specific criteria of empirically supported treatments (35) indicates "strong support" for 17 CBT treatments for a variety of DSM-IV-TR diagnoses, whereas only one psychodynamic therapy treatment—Kernberg's transference-focused psychotherapy for borderline personality disorder (36)—currently meets this level of support (37). This difference is largely due to a lack of replication with the same treatment manual by different research teams in randomized controlled trials of psychodynamic therapy (3).

The trials of CBT for depression appear to have specific areas of strength and weakness. Areas of strength appear to be in diagnostic methods for inclusion or exclusion, description of treatments, use of validated outcome measures, a priori specification of primary outcome measures, justification of comparison groups, use of same time frame for comparison groups, and appropriate randomization between groups. Areas of weakness include description of comorbidities, blinding of outcome assessment, discussion of safety and adverse events, use of intent-to-treat method, and statistical consideration of therapist and site effects. The latter three areas were also found to be deficient in the psychodynamic therapy trials (3), indicating that these are areas in particular need of attention in the design and implementation of future psychotherapy trials. Item 13, reporting of safety and adverse events, received the lowest total score of any item when individual items were summed across studies, indicating a gross lack of reporting in this area. Attention has been called to the potential for adverse effects of psychotherapy (38), and we hope that future randomized controlled trial research will bring greater transparency to this issue.

When examining the relationship between quality and outcome, our prediction of an inverse relationship between quality and effect size was supported, indicating that lower-quality CBT trials were associated with better outcome for CBT. This finding lends empirical support to the hypothesis that the manner of trial implementation may affect the validity of trial results and highlights the importance of maintaining rigorous methodological quality in psychotherapy trials. The items and item anchors of the RCT-PQRS, aimed at assessing trial features across six methodological domains, may help in defining and operationalizing standards in trial implementation. The finding of an inverse relationship between quality and effect size also indicates that the results of some previous meta-analyses may have overestimated the effects of CBT for depression, which is likely also the case for psychotherapies for depression in general (7). The moderating effect of quality on outcome speaks to the importance of incorporating validated measures of quality into meta-analyses. While we would not necessarily predict that the magnitude of the relationship between quality and outcome in the present study is universal and stable across treatments and diagnoses, the use of an instrument such as the RCT-PQRS would allow for sensitivity analyses to investigate whether such a relationship is present.

In addition to predicting that lower quality would be related to inflated outcome, we also predicted that lower quality would be related to greater variability of outcome as a result of less tightly controlled experimental conditions. This prediction was supported, indicating that low quality might be associated not only with bias but also with greater error and thus lower reliability of results.

Our study had several limitations. Item-level reliability was not high across all items of the RCT-PQRS, and thus item-level descriptive analyses must be considered exploratory. Furthermore, we did not conduct item-level statistical analyses aimed at uncovering which methodological features were most related to outcome or whether all items were related to outcome in the same direction. However, summation across items was able to capture an aggregation of the influence of low quality in the studies examined.

An additional limitation in this regard is that the observed correlation between quality and outcome could be caused by a third variable, such as publication bias. If trials with null results that go unpublished also happen to be low in quality, the published trials included in our analysis could represent a biased sample of the population of relevant trials and could thus contain an excess of low-quality trials with large, significant effects. The exclusion of missing (unpublished) low-quality trials with null results from the metaregression may have had a greater influence on the observed relationship between quality and outcome than differences in experimental conditions in the included trials. However, making assumptions about the quality of theoretically missing studies is inherently speculative. Furthermore, given the strong correlation observed between standard error (a proxy for sample size) and quality score, the effects of quality versus publication bias on outcome are difficult to disentangle, both conceptually and statistically. In short, we cannot say whether the observed quality-related bias in this study was caused by experimenter bias, other threats to internal validity, publication bias, or some combination of these. Although our study is observational, which limits precise causal inference, we interpret our results as evidence to support the importance of maintaining high methodological quality in randomized controlled trials, particularly in light of the additional observed relationship between low quality and greater variability of outcome.

Another limitation lies in the fact that while ratings with the RCT-PQRS are meant to assess the quality of the trials themselves, ratings are necessarily limited to the information included in the published reports. This is a limitation shared with quality rating scales in general (39). Nonetheless, it further limits our ability to discern which specific methodological features may be most related to outcome, leading us to use the aggregation of the quality of report-

ing of specific features as an approximate measure of the overall trial quality. Finally, the raters in our study were not blind to any details of the studies. While there was good interrater reliability among raters of varying orientations, and while the RCT-PQRS attempts to anchor individual item levels with enough specificity to minimize judgment and bias in item ratings (2), we must acknowledge the possibility that some ratings were affected by halo effects related to factors such as prestige level of journals or reputation of study authors. The effects, as well as feasibility, of blinding studies in the context of quality ratings may be an area for future research (15, 40).

Received March 16, 2011; revisions received June 27 and Sept. 6, 2011; accepted Sept. 15, 2011 (doi: 10.1176/appi.ajp.2011.11030433). From Weill Cornell Medical College, New York; Fordham University, Bronx, New York; and Columbia College of Physicians and Surgeons and New York State Psychiatric Institute, New York. Address correspondence to Dr. Thoma (nthoma@gmail.com).

Dr. McKay has received royalty income from Sage, American Psychological Association Press, Springer Press, Springer Science+Business, and Wiley-Blackwell, has received funding from the MINT [Mental Health Initiative] Foundation, and is co-investigator on a grant from the International Obsessive-Compulsive Disorder Foundation. Dr. Gerber has received support from an NIMH T32 grant for Research Training in Child and Adolescent Psychiatry, NARSAD, Eli Lilly (via the American Academy of Child and Adolescent Psychiatry Pilot Research Award), the American Psychoanalytic Association, the International Psychoanalytic Association, and the Neuropsychoanalysis Foundation. Dr. Milrod has received research support through a fund in the New York Community Trust established by DeWitt Wallace, the American Psychoanalytic Association Fund for Psychoanalytic Research, and NIMH grants K23 MH01849-01/05 and R01 MH 070918-01A2. Dr. Kocsis has received grants and contracts from NIMH, the National Institute on Drug Abuse, the Agency for Healthcare Research and Quality, Burroughs Wellcome Trust, the Pritzker Consortium, AstraZeneca, Forest, CNS Response, and Roche, is on the speakers bureaus of Pfizer and Merck, and serves on the advisory board of Neurosearch. Drs. Thoma and Edwards report no financial relationships with commercial interests.

The authors thank Margaret Andover, Ph.D., SeKang Kim, Ph.D., and Warren Tryon, Ph.D., for their feedback on the design and implementation of this study, and Pim Cuijpers, Ph.D., and Bruce Wampold, Ph.D., for acting as outside experts to help ensure that no studies fitting our inclusion criteria were missed.

References

1. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP: Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. Health Technol Assess 1999; 3:1–98
2. Kocsis JH, Gerber AJ, Milrod B, Roose SP, Barber J, Thase ME, Perkins P, Leon AC: A new scale for assessing the quality of randomized clinical trials of psychotherapy. Compr Psychiatry 2010; 51:319–324
3. Gerber AJ, Kocsis JH, Milrod BL, Roose SP, Barber JP, Thase ME, Perkins P, Leon AC: A quality-based review of randomized controlled trials of psychodynamic psychotherapy. Am J Psychiatry 2011; 168:19–28
4. Beck AT: How an anomalous finding led to a new system of psychotherapy. Nat Med 2006; 12:1139–1141
5. McKay D, Tryon W: Behavior therapy: theoretical bases, in Encyclopedia of Psychotherapy. Edited by Hersen M, Sledge W. San Diego, Academic Press, 2002
6. Borenstein RF: The impending death of psychoanalysis. Psychoanal Psychol 2001; 18:3–20
7. Cuijpers P, van Straten A, Bohlmeijer E, Hollon SD, Andersson G: The effects of psychotherapy for adult depression are overestimated: a meta-analysis of study quality and effect size. Psychol Med 2010; 40:211–223
8. Churchill R, Hunot V, Corney R, Knapp M, McGuire H, Tylee A: A systematic review of controlled trials of the effectiveness and cost-effectiveness of brief psychological treatments for depression. Health Technol Assess 2001; 5:1–173
9. Wells-Parker E, Bangert-Drowns R, McMillen R, Williams M: Final results from a meta-analysis of remedial interventions with drink/drive offenders. Addiction 1995; 9:907–926
10. Lyons LC, Woods PJ: The efficacy of rational-emotive therapy: a quantitative review of the outcome research. Clin Psychol Rev 1991; 11:357–369
11. Shirk SR, Russell RL: A reevaluation of estimates of child therapy effectiveness. J Am Acad Child Adolesc Psychiatry 1992; 31:703–709
12. Wilson DB, Lipsey MW: The role of method in treatment effectiveness research: evidence from meta-analysis. Psychol Methods 2001; 6:413–429
13. Butler AC, Chapman JE, Forman EM, Beck AT: The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clin Psychol Rev 2006; 26:17–31
14. Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, McDonald S: Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials. Eval Health Prof 2002; 25:38–64
15. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ: Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996; 17:1–12
16. Moncrieff JA, Churchill R, Drummond DC, McGuire H: Development of a quality assessment instrument for trials of treatments for depression and neurosis. Int J Methods Psychiatr Res 2001; 10:126–133
17. Moher D, Schulz KF, Altman D; CONSORT Group: The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001; 285:1987–1991
18. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86:420–428
19. Fleiss JL, Levin B, Paik MC: Statistical Methods for Rates and Proportions, 2nd ed. New York, John Wiley & Sons, 1981
20. Harbord RM, Higgins JPT: Meta-regression in Stata. Stata J 2008; 8:493–519
21. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR: Introduction to Meta-Analysis. Chichester, UK, John Wiley & Sons, 2009
22. Hunter JE, Schmidt FL: Methods of Meta-Analysis: Correcting Error Bias in Research Findings, 2nd ed. Thousand Oaks, Calif, Sage Publications, 2004
23. Wampold BE, Minami T, Baskin TW, Tierney SC: A meta-(re) analysis of the effects of cognitive therapy versus "other therapies" for depression. J Affect Disord 2002; 68:159–165
24. Cuijpers P, van Straten A, Warmerdam L, Smits N: Characteristics of effective psychological treatments of depression: a metaregression analysis. Psychother Res 2008; 18:225–236
25. Cuijpers P, Dekker J, Hollon SD, Andersson G: Adding psychotherapy to pharmacotherapy in the treatment of depressive disorders in adults: a meta-analysis. J Clin Psychiatry 2009; 70:1219–1229
26. Pampallona S, Bollini P, Tibalbi G, Kupelnick B, Munizza C: Combined pharmacotherapy and psychological treatment for depression: a systematic review. Arch Gen Psychiatry 2004; 61:714–719
27. Kleinbaum DG, Kupper LL, Nizam A, Muller KE: Applied Regression Analysis and Other Multivariable Methods, 4th ed. Belmont, Calif, Thompson Higher Education, 2007

28. Egger M, Davey Smith G, Schneider M, Minder C: Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315:629–634
29. Duval S, Tweedie R: Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000; 56:455–463
30. Breusch TS, Pagan AR: A simple test for heteroscedasticity and random coefficient variation. J Econometric Soc 1979; 47:1287–1294
31. Tolin DF: Is cognitive-behavioral therapy more effective than other therapies? a meta-analytic review. Clin Psychol Rev 2010; 30:710–720
32. Leichsenring F: Comparative effects of short-term psychodynamic psychotherapy and cognitive-behavioral therapy in depression: a meta-analytic approach. Clin Psychol Rev 2001; 21:401–419
33. Cuijpers P, van Straten A, Andersson G, van Oppen P: Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. J Consult Clin Psychol 2008; 76:909–922
34. Thoma NC: The Quality of Randomized Controlled Trials of Cognitive Behavior Therapy for Depression: An Assessment and Meta-Regression (doctoral dissertation). New York, Fordham University, Department of Psychology, 2010
35. Chambless DL, Hollon SD: Defining empirically supported therapies. J Consult Clin Psychol 1998; 66:7–18
36. Clarkin JF, Yeomans FE, Kernberg OF: Psychotherapy for Borderline Personality: Focusing on Object Relations. Washington, DC, American Psychiatric Publishing, 2006
37. American Psychological Association Division 12: Website on Research-Supported Psychological Treatments: Transference-Focused Therapy for Borderline Personality Disorder. http://www.div12.org/PsychologicalTreatments/index.html
38. Dimidjian S, Hollon SD: How would we know if psychotherapy were harmful? Am Psychol 2010; 65:21–33
39. Jadad AR, Enkin M: Randomized Controlled Trials: Questions, Answers, and Musings. Malden, Mass, Blackwell Scientific/BMJ, 2007
40. Berlin JA: Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-Analysis Blinding Study Group. Lancet 1997; 350:185–186