AJP in Advance. Published September 15, 2010 (doi: 10.1176/appi.ajp.2010.08060843)

A Quality-Based Review of Randomized Controlled Trials of Psychodynamic Psychotherapy

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

James H. Kocsis, M.D.

Barbara L. Milrod, M.D.

Steven P. Roose, M.D.

Jacques P. Barber, Ph.D.

Michael E. Thase, M.D.

Patrick Perkins, Ph.D.

Andrew C. Leon, Ph.D.

Objective: The Ad Hoc Subcommittee for Evaluation of the Evidence Base for Psychodynamic Psychotherapy of the APA Committee on Research on Psychiatric Treatments developed the Randomized Controlled Trial Psychotherapy Quality Rating Scale (RCT-PQRS). The authors report results from application of the RCT-PQRS to 94 randomized controlled trials of psychodynamic psychotherapy published between 1974 and May 2010.

Method: Five psychotherapy researchers from a range of therapeutic orientations rated a single published paper from each study.

Results: The RCT-PQRS had good interrater reliability and internal consistency. The mean total quality score was 25.1 (SD=8.8). More recent studies had higher total quality scores. Sixty-three of 103 comparisons between psychodynamic psychotherapy and a nondynamic comparator were of "adequate" quality. Of 39 comparisons of a psychodynamic treatment and an "active" comparator, six showed dynamic treatment to be superior, five showed dynamic treatment to be inferior, and 28 showed no difference (few of which were powered for equivalence). Of 24 adequate comparisons of psychodynamic psychotherapy with an "inactive" comparator, 18 found dynamic treatment to be superior.

Conclusions: Existing randomized controlled trials of psychodynamic psychotherapy are promising but mostly show superiority of psychodynamic psychotherapy to an inactive comparator. This would be sufficient to make psychodynamic psychotherapy an "empirically validated" treatment (per American Psychological Association Division 12 standards) only if further randomized controlled trials of adequate quality and sample size replicated findings of existing positive trials for specific disorders. We do not yet know what will emerge when other psychotherapies are subjected to this form of quality-based review.

(Am J Psychiatry Gerber et al.; AiA:1–10)

The depth and quality of the evidence base for psychodynamic psychotherapy has long been a subject of controversy among psychodynamic and nonpsychodynamic clinicians and researchers. Some have argued that the relative absence of randomized controlled trials has doomed psychodynamic treatments to obsolescence (1). This argument is sometimes used to support the further marginalization or even elimination of training in psychodynamic techniques in psychiatry, psychology, and social work training programs, to be replaced by "evidence-based treatments" (2).

Others have argued that a body of empirical evidence for psychodynamic treatments exists and is underappreciated in the context of contemporary emphasis on short-term, manualized, and symptom-focused treatments and targeted medications (3–5). Over the past several years, meta-analyses have appeared in mainstream research outlets that argue for the efficacy of psychodynamic treatments for specific disorders (6–11). Gabbard et al. (12) describe a "hierarchy of evidence" ranging from case studies and uncontrolled trials to randomized controlled trials that support the utility, if not the efficacy, of various forms of psychodynamic psychotherapy for treatment of patients with a wide range of DSM-IV axis I and II psychiatric disorders. Although each of these meta-analyses attempts, in its own way, to make use only of studies that are considered of high enough quality to warrant inclusion in a meta-analysis, there is sharp disagreement in the field about whether the quality and number of studies included is sufficient to warrant the conclusions drawn.

Ambiguity about the state of psychodynamic empirical research presents a significant problem for training and practice in the mental health fields. The objective evaluation of the quality of randomized controlled trials of psychodynamic psychotherapy is a cogent place to begin the process of correcting this problem. Such trials are widely accepted in medicine as the gold standard for assessing treatment efficacy, and there is good conceptual agreement about what constitutes a well-conducted trial (13). The CONSORT (Consolidated Standards of Reporting Trials) Statement, which has been adopted by most major medical journals, identifies 22 elements that are

important in reporting randomized controlled trials (14). However, the CONSORT checklist and other similar measures, designed primarily to assess studies of pharmacological or medical interventions, fail to adequately assess the psychotherapy literature for several reasons. First, they do not include items that are specific and essential to psychotherapy trials, such as the length of follow-up or the extent of training or supervision of psychotherapy. Second, they focus on the quality of description in the written article, with less explicit focus on what was *done* in the actual trial, thus deemphasizing such essential details as assessment of adherence to the treatment actually delivered. Third, they pay insufficient attention to the quality and credibility of comparison treatments, as this is less of an issue for studies using pill placebo for the comparison group. And fourth, psychometric evaluation of CONSORT items has never been reported, nor have the items been used to quantify overall quality scores. An extension of the CONSORT Statement to randomized trials of nonpharmacological treatment (including, for example, surgery, technical interventions, devices, rehabilitation, psychotherapy, and behavioral intervention) published in 2008 consists of further elaboration of 11 of the checklist items, addition of one item, and modification of the flow diagram (15). Although helpful, this extension remains nonspecific to psychotherapy, focuses on description over conduct of the trials, and lacks psychometric evaluation.

As part of an effort to clarify the state of psychodynamic psychotherapy research, the Ad Hoc Subcommittee for Evaluation of the Evidence Base for Psychodynamic Psychotherapy (appointed in 2004 by the APA Committee on Research on Psychiatric Treatments) developed the Randomized Controlled Trial of Psychotherapy Quality Rating Scale (RCT-PQRS). This 25-item questionnaire was designed by experienced psychiatric and psychotherapy researchers from a variety of theoretical backgrounds as a systematic way to rate the quality of randomized controlled trials of psychotherapy. The scale is designed to be used by individuals with considerable experience in reading and executing psychotherapy trials but requires only 10 to 15 minutes to rate, in addition to the time spent reading the paper to which it is applied. The scale yields a 24-item total score that has good psychometric properties and captures the overall quality of design, implementation, and reporting of psychotherapy trials. (Item 25 is an omnibus item, more about which below.)

In this article, we report psychometric properties of the RCT-PQRS based on application of the scale to all 94 randomized controlled trials of individual and group psychodynamic psychotherapy published between 1974 and May 2010 that we were able to locate. We then describe the results of applying this measure to the psychodynamic outcome literature. Our hypotheses were 1) that the overall quality of randomized controlled trials of psychodynamic psychotherapy has improved over time, from a largely inadequate implementation in the 1970s and 1980s to more rigorous implementation in the 1990s and 2000s; 2) that some aspects of psychodynamic psychotherapy trials, including characterization of patients, reliable and valid measurement of outcome, and manualization of treatment, are now being done quite well; and 3) that some aspects of psychodynamic psychotherapy trials, including reporting of therapist training and supervision, measurement of treatment adherence, analysis of therapist or study site effects, and reporting of concurrent treatments and adverse events during treatment, remain lacking.

Method

Sample

We conducted a MEDLINE search to locate all published randomized controlled trials of psychotherapies identified by their authors as being "psychodynamic" or "psychoanalytic." We also used meta-analyses and review articles to identify trials (6, 8–11, 16–22), examined the reference lists of recovered articles, and consulted experts in the field to ensure that we obtained a comprehensive list. All studies that randomized participants to two or more treatments, one of which was identified as "psychodynamic" or "psychoanalytic" treatment, were included. Studies that identified themselves as precursors to larger and more complete studies were excluded even if randomization was performed. To capture the full range of study quality, we did not use a sample size cutoff in selecting trials. Only studies with published articles in English were included. A total of 94 trials were identified that met inclusion and exclusion criteria.

Quality Measure: The RCT-PQRS

Items for the RCT-PQRS were generated by expert consensus. The experts included the subcommittee and several outside consultants, all of whom were senior psychotherapy or pharmacotherapy researchers. Contributors included those with experience in performing studies of psychodynamic psychotherapy, CBT, pharmacotherapy, and combinations. Items were compared with those of the CONSORT standard as well as several other measures of quality of outcome trials to ensure that all major domains for assessing study adequacy were included (13, 14, 23–26).

The RCT-PQRS consists of 25 items (Figure 1) 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/criteria but not both), or 2 (well described, executed, and, where necessary, justified design element). Item 25, an "omnibus" item, was rated from 1 (exceptionally poor study) to 7 (exceptionally good study). Because of the complex scientific choices involved in determining study design, the measure was designed to be administered by individuals experienced in reading and conducting psychotherapy outcome trials.

Items were chosen so as to represent randomized controlled trial design and description elements that are crucial to the execution of a study that can be expected to yield reliable, valid, and reproducible data. No assumptions were made about whether individual design elements are already commonly implemented in such trials. In fact, several items were specifically chosen with the knowledge that they are not common practice (e.g., item 13, reporting of safety and adverse events) but with the conviction that they should be a standard part of psychotherapy randomized controlled trial methodology. No a priori assumptions were made about the relative importance of different items for the overall quality of the trial.

We calculated a total score by computing the nonweighted sum of the first 24 individual items. Although we investigated both the

omnibus item and total scores, we believe that the first 24 items are essential in guiding the thinking of the rater, and we did not intend to test the usefulness of a single item in the absence of a longer scale. The advantage of the total score is that it provides more variability and is explicitly based on 24 (relatively) independent judgments, whereas the advantage of the omnibus item is that it allows the rater to give relative weight to strengths or weaknesses of the study that he or she feels are particularly relevant (the sum arbitrarily weights all items equally). Therefore we intend to keep both metrics in our analyses. We estimated a priori that a total score of 24 would correspond to the minimum adequate total quality-of-study score, corresponding to an average value of 1 on each scale item. As this is the first published use of the RCT-PQRS, no norms have yet been established. To illustrate the properties of the measure, we asked raters to score two well-known randomized controlled trials that are not included in this study (and do not include an explicitly psychodynamic treatment group). The National Institute of Mental Health Treatment of Depression Collaborative Research Program, which compared CBT, interpersonal psychotherapy, and medication, received a total quality score of 40 and an item 25 score of 6 (27). The Treatment for Adolescents With Depression Study, which studied the effect of fluoxetine, CBT, and their combination in depressed adolescents, received a total quality score of 38 and an item 25 score of 6 (28). Both of these are believed to represent a high standard of quality for a randomized controlled trial of psychotherapy.

Method of Rating

For each of the psychodynamic trials identified, a single published paper was selected that most thoroughly represented the methods and findings of the study. Each study was then randomly assigned to one of five senior raters (J.H.K., B.L.M., S.P.R., J.P.B., and M.E.T.). Each rater read the selected paper in its entirety and then completed the scale on the basis of the paper alone, disregarding prior knowledge that he or she may have had about the study. Raters were encouraged to go back to the paper during scoring to check on study details as needed. The reliability of these ratings has been described elsewhere (29). Interrater reliability was assessed using the intraclass correlation coefficient (ICC) formula (ICC[1,1]) from Shrout and Fleiss (30) as implemented by the INTRACC.SAS procedure (31) using the STAT software package, version 9.1, of the SAS System for Windows (SAS Institute, Inc., Cary, N.C.). This formula is appropriate for a situation in which *k* (in this case, k=2) independent raters score each member of a sample but the identity of the raters changes at random from one member of the sample to the next. The internal consistency of the first 24 items was assessed using the standard formula for Cronbach's alpha (32). For the first 70 studies included in this review, the total score ICC was 0.76 and Cronbach's alpha was 0.87. This indicates high interrater reliability of the total quality and high internal consistency of the items in the scale. For the purposes of this review, the rating of only a single senior rater was used for each study, as this is how the measure was designed to be used and is most likely how it will be used by other investigators.

The issue of whether raters were assessing the quality of the study itself or of the publication was an inevitable tension in data collection. Raters were instructed, for the sake of standardization and reliability, to base their ratings only on what was reported in the single article that was given to them. Whether a similar rule would be applied for future applications of this measure would depend on the resources and goals of the researchers. For this review, to collect and read all papers on a given study would have posed methodological challenges greater than those solved by using more than one paper per study (e.g., it would be hard to determine which publications were eligible and unfair to penalize studies that had generated fewer overall publications). However, raters were instructed to base ratings of items not only on the quality and inclusivity of the publication (although this was

FIGURE 1. Items in the Randomized Controlled Trial Psychotherapy Quality Rating Scale[a]

Description of subjects
1. Diagnostic method and criteria for inclusion and exclusion
2. Documentation or demonstration of reliability of diagnostic methodology
3. Description of relevant comorbidities
4. Description of numbers of subjects screened, included, and excluded

Definition and delivery of treatment
5. Treatment(s) (including control/comparison groups) are sufficiently described or referenced to allow for replication
6. Method to demonstrate that treatment being studied is treatment being delivered (only satisfied by supervision if transcripts or tapes are explicitly reviewed)
7. Therapist training and level of experience in the treatment(s) under investigation
8. Therapist supervision while treatment is being provided
9. Description of concurrent treatments (e.g., medication) allowed and administered during course of study (if patients on medication are included, a rating of 2 requires full reporting of what medications were used; if patients on medications are excluded, this alone is sufficient for a rating of 2)

Outcome measures
10. Validated outcome measure(s) (either established or newly standardized)
11. Primary outcome measure(s) specified in advance (though does not need to be stated explicitly for a rating of 2)
12. Outcome assessment by raters blinded to treatment group and with established reliability
13. Discussion of safety and adverse events during study treatment(s)
14. Assessment of long-term post-termination outcome (should not be penalized for failure to follow comparison group if this is a wait-list or non-treatment group that is subsequently referred for active treatment)

Data analysis
15. Intent-to-treat method for data analysis involving primary outcome measure
16. Description of dropouts and withdrawals
17. Appropriate statistical tests (e.g., use of Bonferroni correction, longitudinal data analysis, adjustment only for a priori identified confounders)
18. Adequate sample size
19. Appropriate consideration of therapist and site effects

Treatment assignment
20. A priori relevant hypotheses that justify comparison group(s)
21. Comparison group(s) from same population and time frame as experimental group
22. Randomized assignment to treatment groups

Overall quality of study
23. Balance of allegiance to types of treatment by practitioners
24. Conclusions of study justified by sample, measures, and data analysis, as presented (note: useful to look at conclusions as stated in study abstract)
25. Omnibus rating: Please provide an overall rating of the quality of the study taking into account the adequacy of description, the quality of study design, data analysis, and justification of conclusions

[a] Items 1–24 are rated 0 (poor description, execution, or justification of a design element), 1 (brief description *or* either a good description or an appropriate method/criteria but not both), or 2 (well described, executed, and, where necessary, justified design element). Item 25 is rated from 1 (exceptionally poor study) to 7 (exceptionally good study).

TABLE 1. Comparison Group Categories in 94 Randomized Controlled Trials of Psychodynamic Psychotherapy[a]

Types of Comparison Groups	Number of Studies
Cognitive-behavioral therapy	18
Supportive therapy	11
Behavioral therapy	10
Family therapy	7
Group therapy	11
Medication only	6
Waiting list	13
Treatment as usual	13
No treatment	8
Miscellaneous (e.g., yoga, dietary advice, primary care, relaxation, recreation)	18

[a] Because some studies had more than one comparison group, a study may be counted in more than one category.

inevitably a significant part of what guided their ratings) but also on the quality of the study design captured in these items and the extent to which convincing justification was provided for design decisions whose rationale might not be entirely clear to someone familiar with the field of psychotherapy research. Therefore, unlike the CONSORT standard, which is more instructive about a publication than about study design (although it clearly carries implications for the latter), the RCT-PQRS was intended to capture, as directly as possible, the quality of the study design.

Study Results and Comparator Type

We classified the results of each study in terms of outcome at termination of treatment (psychodynamic psychotherapy compared with any other group in which a statistical comparison was reported). We rated whether the psychodynamic group did better, did the same, or did worse than the comparison group at the end of treatment based on statistically significant results from primary outcome measures specified by the authors (see the data supplement that accompanies the online edition of this article). When differences between groups were not statistically significant by the standards set by the study's authors, the comparison was placed in the "no difference" group. If the primary outcome measures were either insufficiently specified or provided heterogeneous results, we rated the outcome category on the basis of the authors' summary of their findings in the paper's abstract. We discussed any ambiguity about these ratings until a consensus was reached. For every paper, we felt that a clear and uncontroversial categorization could be reached.

Each comparison group was classified as "active" or "inactive," following a distinction similar to that used by other authors (33). An active comparison group is one that received a specified treatment that has been validated by any previous empirical research for the disorder being treated or is considered a specific appropriate treatment by clinical consensus. This may include various forms of psychotherapy, as well as medication alone. For the purposes of this review, the standard for an active treatment was set intentionally low in order to capture the distinction between these treatments and those that are clearly intended by the investigators to be nonspecific and/or easy targets to beat (the inactive treatments). An inactive comparison group is any group that received no treatment at all (a no-treatment or waiting list control condition), treatment as usual, or a treatment that is not thought by any experts to be effective for the disorder being treated. Although Vinnars et al. (34) showed "treatment as usual" to be efficacious for personality disorders, their research was conducted in Sweden, where strong, comprehensive universal health care with well-articulated treatment descriptions for personality disorders is the national standard; therefore similar comparator groups were deemed "active." However, most of the studies reviewed here were conducted in the United States, where "treatment as usual" often represents little or no treatment. Furthermore, because by its very nature, "treatment as usual" is nonstandardized, it most often belongs in the "inactive" category for these purposes. We discussed any ambiguous classifications until consensus was reached and all authors agreed that each classification was ultimately uncontroversial according to the standards specified above.

Results

Study Characteristics

The 94 studies included in this analysis were from articles published in English between 1974 and May 2010. Total study sample size (combining patients from all treatment groups in a given randomized controlled trial) ranged from 10 to 487 (median=73). Seventy-two studies evaluated the effects of individual psychodynamic psychotherapy lasting less than 1 year, 12 studies looked at individual psychodynamic psychotherapy lasting 1 year or longer, and 17 studies were of psychodynamic group psychotherapy. There are no existing randomized controlled trials of psychoanalysis as currently defined by the American Psychoanalytic Association (minimal session frequency four times per week).

The number of comparison groups (excluding the group treated with psychodynamic psychotherapy) ranged from one to three (mean=1.4, SD=0.71). The frequencies of various types of comparison groups are presented in Table 1. The most common comparison group was CBT (18 studies, 19%), followed by supportive therapy (11 studies, 12%). Contained within the 94 studies were 103 direct comparisons between psychodynamic psychotherapy and a nondynamic alternative treatment. Eleven studies did not report such a comparison (eight reported comparisons between two or more types of psychodynamic psychotherapy, and three failed to report comparisons of outcome at termination). Fourteen studies reported the results of separate comparisons between psychodynamic psychotherapy and two other treatments, and three studies reported comparisons with three other treatments. Of the 103 direct comparisons, 63 (61%) compared psychodynamic psychotherapy with an active treatment, and 40 (39%) compared psychodynamic psychotherapy with an inactive treatment.

The major diagnostic categories treated in the 94 studies are summarized in Table 2. The most common diagnostic categories were depressive disorders (27%), personality disorders (14%), medical illness (11%), eating disorders (11%), anxiety disorders (10%), and substance abuse (7%). Because a wide range of methods were used for assessing the diagnostic category of the patients in these studies, subdividing the results of the review by diagnosis would result in combining across heterogeneous categories; meaningful results cannot be reported in terms of diagnostic categories.

Quality Scores

The methods used to standardize the RCT-PQRS and some initial results based on 69 of the randomized

controlled trials reviewed here have been described elsewhere (29). An updated version of these findings is summarized below. The overall quality score of the 94 studies was normally distributed, with a mean of 25.1 (SD=8.8; Figure 2). More recently conducted studies had higher average total quality scores (29). The omnibus item score (item 25) showed the same trend over time (29). Individual item scores, although limited in their potential for interpretation because of the absence of high single-item interrater reliability, suggest the relative strengths and weaknesses of various study characteristics. For four of the 24 items, at least half of the studies were scored as "good" (i.e., a rating of 2): item 4 (description of numbers screened, included, and excluded), item 5 (treatment sufficiently described or referenced), item 20 (comparison groups justified by a priori hypotheses), and item 21 (comparison groups from same population and time frame). For three items, at least half of the studies were scored as "poor" (i.e., a rating of 0): item 13 (reporting of safety and adverse events), item 15 (intent-to-treat method reported in data analysis), and item 19 (consideration of therapist and study site effects). For 23 of 24 items, the mean item score was stable or increased across the time periods into which the studies were divided (1974–1988, 1989–1997, 1998–2004, 2005–2010). The mean score of item 14 (long-term post-termination outcome assessment) increased between the first two periods but then dropped and was lowest in the past 5 years, possibly reflecting greater difficulty in funding acquisition and acknowledgment of the inherent difficulties in interpreting outcome data because of artifacts of intervening nonstudy interventions that increase over time.

Rater Effects

The mean total quality scores of the five senior raters were 20.4, 24.1, 25.0, 25.4, and 27.0. For the three raters with psychodynamic orientations, the mean overall quality rating was 23.6 (SD=2.8), and for the two raters with a nondynamic orientation, the mean rating was 25.6 (SD=2.1). There was no significant relationship between mean quality rating and theoretical orientation.

Study Outcome

In the sample of 94 outcome studies, we found a total of 103 comparisons between a psychodynamic treatment and a set of either inactive or active comparators. In studies with more than one active or more than one inactive comparator, we treated the comparisons within each set together because the findings within a set of active or inactive almost always matched and because we did not want to improperly weight studies with more comparison groups in our analyses, as these studies tended to have smaller numbers of patients per group and were underpowered. Of the 103 comparisons, 63 were between psychodynamic psychotherapy and an active comparator and 40 were between psychodynamic psychotherapy and an inactive comparator. Thirty-nine (62%) of the 63 comparisons of psychodynamic psychotherapy and an active comparator came from studies with a mean total quality of at least 24. Twenty-four (60%) of the 40 comparisons of psychodynamic psychotherapy and an inactive comparator came from studies with a mean total quality of at least 24.

Of the 63 comparisons between psychodynamic psychotherapy and an active comparator, six (10%) showed greater improvement in the psychodynamic group, 10 (16%) showed greater improvement in the comparator group, and the remainder (47 studies, 75%) showed no significant difference between the groups. Of the 40 comparisons between psychodynamic psychotherapy and an inactive comparator, 27 (68%) showed greater improvement in the psychodynamic group, one (3%) showed greater improvement in the comparator group, and 12 (30%) showed no significant difference between groups.

TABLE 2. Diagnostic Categories of Patients in 94 Randomized Controlled Trials of Psychodynamic Psychotherapy

Diagnostic Category	Number of Studies
Depression (major depression, geriatric depression, postpartum depression, grief, minor/subsyndromal depression)	25
Personality disorders (borderline personality disorder, cluster C symptoms, avoidant symptoms, high utilizers of psychiatric services)	13
Complicated medical illnesses (chronic pain, irritable bowel syndrome, chronic peptic ulcer, Crohn's disease, chronic obstructive pulmonary disease, rheumatic disease, atopic dermatitis)	10
Eating disorders (anorexia, bulimia, obesity, binge eating disorder)	10
Anxiety disorders (panic disorder, posttraumatic stress disorder, social anxiety, generalized anxiety)	9
Substance abuse (opiates, cocaine, alcohol)	7
Violence or disruptive behavior	3
Hypochondriasis	1
Suicidality	1
Schizophrenia	1
Miscellaneous (childhood sexual abuse, mixed disorders)	14

Associations Between Study Outcome and Quality

Mean total quality and item 25 scores were calculated separately for studies with active and inactive comparators that found significant or no differences between groups at termination (Table 3). No relationships were observed between quality of study and the nature of the comparator (active or inactive), between quality of study and the outcome of the comparison (dynamic therapy superior or inferior, or no difference), or between quality of study and an interaction of comparator type and outcome.

Sixty-three of the 103 comparisons described above had a total quality score of 24 or above (the cutoff for a "reasonably well done" study established from the face validity of the measure). Because some studies contain more than one comparison, these 63 comparisons represent

FIGURE 2. Histogram of 24-Item Total Quality Scores for 94 Randomized Controlled Trials of Psychodynamic Psychotherapy

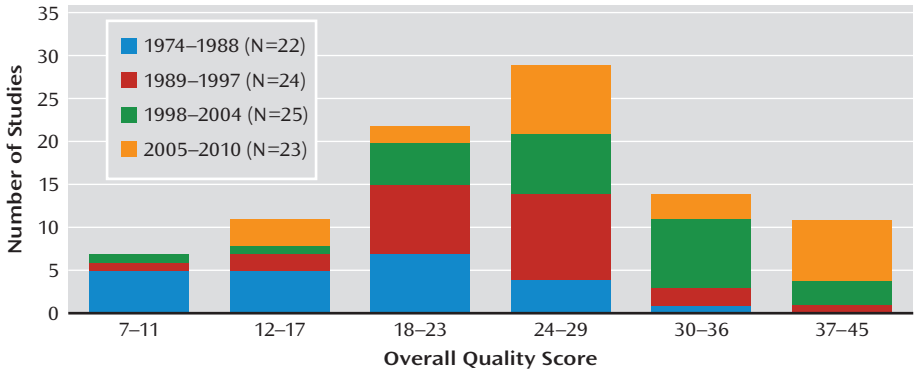

TABLE 3. Mean Total Quality Scores and Item 25 Scores, by Outcome Classification, for All Comparisons (103 Comparisons)

	Psychodynamic Treatment > Comparator					Psychodynamic Treatment = Comparator					Psychodynamic Treatment < Comparator				
	Total Quality Score			Item 25 Score		Total Quality Score			Item 25 Score		Total Quality Score			Item 25 Score	
Comparator	Mean	SD	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	N	Mean	SD
Active comparator (63 comparisons)	30.1	6.7	6	4.7	1.2	25.7	8.7	47	4.2	1.6	25.9	10.6	10	4.4	1.8
Inactive comparator (40 comparisons)	26.8	8.7	27	4.5	1.4	24.4	10.1	12	4.0	1.4	16	—	1	2	—

TABLE 4. Mean Total Quality Scores and Item 25 Scores, by Outcome Classification, for Studies With Total Quality Scores ≥ 24 (63 Comparisons)

	Psychodynamic Treatment > Comparator					Psychodynamic Treatment = Comparator					Psychodynamic Treatment < Comparator				
	Total Quality Score			Item 25 Score		Total Quality Score			Item 25 Score		Total Quality Score			Item 25 Score	
Comparator	Mean	SD	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	N	Mean	SD
Active comparator (39 comparisons)	30.1	6.7	6	4.7	1.2	31.2	5.8	28	5.1	1.1	34.2	7.9	5	5.8	1.1
Inactive comparator (24 comparisons)	31.3	5.9	18	5.2	0.86	31.7	7.3	6	5.0	1.3					

50 of the original 94 studies. Four additional studies that were excluded from the list of comparisons (because they only compared psychodynamic therapies to each other or because they did not report outcome at termination) also had total quality scores of 24 or higher. Therefore, a total of 54 of the original 94 studies reviewed were deemed "reasonably well done" by this measure.

Thirty-nine of the 63 comparisons with adequate quality compared a dynamic treatment and an active comparator: six showed dynamic treatment to be superior, five showed dynamic treatment to be inferior, and 28 showed no difference. Notably, 19 of the 24 active comparisons that were eliminated from this analysis because of a total quality below 24 came from the group that showed no difference between two treatments and the remaining five came from the group that showed dynamic therapy to be inferior to a comparator. Twenty-four of the 63 higher-quality comparisons looked at dynamic treatment against an "inactive" comparator, and 18 of these found dynamic treatment to be superior. When the total quality and item 25 scores of the higher-quality subgroup (63 studies) were analyzed with respect to outcome, no significant differences were found between groups (Table 4).

Discussion

In this evaluation, we examined a large amount of study-derived data collected to describe the utility of psychodynamic psychotherapy for psychiatric illness. On this basis, it appears that there is both good and bad news about the quality of randomized controlled trials of psychodynamic psychotherapy. The good news is that there are at least 94 randomized controlled trials published to date addressing the efficacy of psychodynamic psychotherapy, spanning a range of diagnoses, and more than half of these (57%) are of adequate quality. The 94 studies represent 103 comparisons between the outcome of psychodynamic psychotherapy and a nondynamic comparator at termination. In the

63 comparisons of dynamic treatment and an active comparator, dynamic treatment was found to be superior in almost 10% and inferior in 16%, and the analyses failed to find a difference in almost 75%. In the 40 comparisons of dynamic treatment and an inactive comparator, dynamic treatment was superior in 68%, there was no evidence for difference in 30%, and dynamic treatment was inferior in one study. We obtained similar results if only studies of good quality are included. In comparisons of good quality studies involving dynamic treatment against an active comparator, dynamic treatment was superior in 15% and inferior in 13%, and evidence was lacking in either direction in 72%. In comparisons against an inactive treatment, dynamic treatment was superior in 75% and lacked evidence in either direction in 25%.

However, it is clear that there are significant quality problems in a significant percentage of randomized controlled trials of psychodynamic psychotherapy, and this may be true for trials of other psychotherapies as well. Based on the 94 studies reviewed here, reporting of safety and adverse events, intent-to-treat method in data analysis, and consideration of therapist and study site effects are lacking. Furthermore, 57% of all comparison and 54% of comparisons from studies of adequate quality failed to find any statistical differences between outcome from psychodynamic treatment and a comparator. As Milrod (35) has pointed out, if a randomized controlled trial is not powered for equivalence (that is, lacks sufficient enough power to detect differences between groups such that the absence of a significant difference can be interpreted as equivalence in efficacy or effectiveness of the treatments), it is impossible to conclude from the absence of a finding that dynamic treatment is as good as or not worse than the comparator treatment. In fact, if we use the rough approximation that a total sample size of at least 250 is required in order to be powered for equivalence (the actual number depends on the measures used), one of the comparisons reviewed here that failed to show a difference between psychodynamic treatment and an active comparator met that criterion (36).

When considered from this perspective, an important finding is that of 63 statistical comparisons based on randomized controlled trials of adequate quality, six showed psychodynamic psychotherapy to be superior to an active comparator, 18 showed it to be superior to an inactive comparator, and one showed it to be equivalent to an active comparator. The empirical support for psychodynamic psychotherapy comes down to these 25 comparisons (in boldface in online data supplement). Most of the rest of even the trials of adequate quality are uninformative (33 studies), and several suggest that psychodynamic psychotherapy is worse than an active or inactive comparator (five studies).

Therefore, it is our conclusion that although the overall results are promising, further high-quality and adequately powered randomized controlled trials of psychodynamic psychotherapy are urgently needed. The standards by which a treatment is considered "evidence-based" for the purposes of inclusion in practice standards, educational curricula, or health care reimbursement vary widely. Perhaps the most commonly cited standard, first published by Chambless and Hollon in 1996 from an American Psychological Association Division 12 (clinical psychology) task force (37), requires at least two well-conducted trials using manuals, showing superiority or equivalence for a specific disorder, and performed by separate research groups, for a treatment to be considered "well-established." By this standard, the 25 high-quality trials reviewed could have been more than enough for psychodynamic psychotherapy to be considered "empirically validated." However, these 25 trials covered a wide range of diagnoses and used different manuals or forms of dynamic therapy. In addition, the significant number of trials that failed to find differences between dynamic and active or inactive comparators, most of which were not powered for equivalence, as well as a handful of trials that found dynamic therapies to be inferior to active comparators, suggest that more work is needed. It remains necessary to identify specific dynamic treatments that are empirically validated for specific disorders.

One finding of this review is that the clearest predictor of outcome from randomized trials of psychodynamic psychotherapy is whether or not the treatment was tested against an active or an inactive comparison treatment. We suspect that this finding is common to randomized controlled trials of all psychotherapies. There is no apparent relationship in this group of studies between study quality and the outcome measured in that study.

We have taken a highly critical view of data collected in support of psychodynamic psychotherapy in this evaluation. Few domains of psychiatric intervention have yet been evaluated so critically, and we emphasize that many domains of psychotherapy outcome research (i.e., CBT, cognitive-behavioral analysis system of psychotherapy, interpersonal psychotherapy, and supportive psychotherapy, to name some) may not fare significantly better, despite the far greater number of outcome studies within these other domains to evaluate. We are currently embarking on a similar review of studies of CBT for depression. We believe that this is an appropriate stance, and we anticipate that this assessment can help move psychotherapy research in a direction toward better-designed outcome studies in the future. Thus far, proponents of other studied forms of specific psychotherapy, most notably CBT, have yet to dissect the quality of studies that support its efficacy.

For three items in the RCT-PQRS, at least half of the studies were scored as "poor": item 13 (reporting of safety and adverse events), item 15 (intent-to-treat method reported in data analysis), and item 19 (consideration of therapist and study site effects). Individual item scores must be regarded with caution because we have not yet established a high degree of interrater reliability on individual items. However, we believe that all three of these

items represent significant deficits in the psychodynamic psychotherapy research literature that are still being insufficiently addressed (1). It is well known by clinicians that all efficacious treatments, whether they be psychotherapy or psychopharmacology, carry with them some risk of adverse effects, yet the randomized controlled trial literature in psychotherapy in general does not systematically report and discuss adverse events as do, for example, most good studies of medication (2). While there has been significant improvement in the area of intent-to-treat analysis, many randomized controlled trials of psychodynamic psychotherapy still focus primarily on "completer" analyses and do not adequately employ the intent-to-treat method, which is a standard of the evidence-based medicine literature. Although intent-to-treat analyses also have certain limitations, they are the best starting point for addressing patient dropout in considering which treatments work best (38–40). Finally, a large literature supports the importance of individual therapist and study site effects in psychotherapy outcome (41–43), but the psychotherapy randomized controlled trial literature has not yet adequately responded by incorporating such considerations into a discussion of results, and even less into study design.

The RCT-PQRS focuses only on the reporting, execution, and justification of design decisions made in a given report of a randomized controlled trial, but not on how these decisions affect the broader question of when results can be applied to real-world clinical practice. In other words, the measure addresses the importance of accurately and reliably quantifying the internal validity of a randomized controlled trial, while tracking generalizability less (3). As one example of this, the diagnostic categories described in Table 2 are just one conventional way of parsing the patient samples and are by no means the only way to subdivide these studies. None of the analyses in this article relied on diagnostic categories. However, future quality-based reviews and meta-analyses will need to address this issue.

The development and application of our quality measure have several significant limitations. First, the accuracy of our ratings is necessarily limited by the extent of the information provided by the authors in their study descriptions. In some cases, the extent to which the authors describe a comparison treatment such as "treatment as usual" as active or inactive could have important consequences in rating the study. However, in our rating of active versus inactive comparator treatments, we observed no significant differences between raters and therefore do not believe that there was much ambiguity in these ratings for the articles reviewed. Second, we have not yet developed and published a manual for the RCT-PQRS that would allow for greater single-item reliability across raters. Lack of reliability limits our ability to interpret individual item scores. Although a manual would likely improve item reliability, it would increase the time it takes to learn and apply the measure and might not affect the overall scores. Third, we have not yet collected sufficient data from the RCT-PQRS to study other ways of aggregating or weighting individual items. We anticipate performing the appropriate analyses as we collect data on a larger number and broader range of randomized controlled trials of psychotherapies. Finally, given that raters of the studies are not blind to study outcome or, in most cases, the allegiance of the study authors, the quality scale is theoretically susceptible to bias. Blinding the studies by removing the names of treatments or the direction of findings would be difficult if not impossible, and the blinding process itself would be subject to biases. We addressed the potential for bias by making sure that our raters (and the designers of the scale) were drawn from a range of areas of clinical and research expertise, including pharmacotherapy, CBT, and psychodynamic psychotherapies. We observed no significant relationship between quality ratings and theoretical orientation of raters. In fact, the mean total quality score was slightly lower for raters with a psychodynamic orientation than for those without.

We hope that this better-operationalized evaluation of the quality of psychotherapy outcome studies will help guide investigators across all areas of psychotherapy research toward more scientifically credible, better-articulated research.

Received June 6, 2008; revisions received May 4 and Dec. 11, 2009, and July 2, 2010; accepted July 2, 2010 (doi: 10.1176/appi.ajp.2010.08060843). From Columbia College of Physicians and Surgeons and the New York State Psychiatric Institute, New York; Weill Cornell Medical College, New York; Department of Psychiatry, University of Pennsylvania, and VA Medical Center, Philadelphia; Department of Psychiatry, University of Pittsburgh, Pittsburgh. Address correspondence and reprint requests to Dr. Gerber, Columbia College of Physicians and Surgeons and New York State Psychiatric Institute, Unit 74, 1051 Riverside Dr., New York, NY 10032; gerbera@childpsych.columbia.edu (e-mail).

Dr. Gerber has received support from NIMH, 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. Kocsis has received grants and contracts from AstraZeneca, Burroughs Wellcome Trust, CNS Response, Forest, NIMH, the National Institute on Drug Abuse, Novartis, the Pritzker Consortium, Roche, and Sanofi-Aventis and serves on speakers bureaus or advisory boards for AstraZeneca, Merck, Pfizer, and Wyeth. Dr. Milrod receives research support from NIMH and through a fund in the New York Community Trust established by DeWitt Wallace. Dr. Roose serves as a consultant to Forest, Medtronic, Wyeth, and Organon and receives research support from Forest. Dr. Barber has received grants from NIMH and the National Institute on Drug Abuse, medication support from Pfizer for a randomized controlled trial on depression, and royalty income from Guilford Publications, Cambridge University Press, and Basic Books. Dr. Thase has served as a consultant or speaker for or received grant support from AstraZeneca, Bristol-Myers Squibb, Cephalon, Cyberonics, Eli Lilly, Forest Pharmaceuticals, GlaxoSmithKline, Janssen Pharmaceutica, Lundbeck, MedAvante, Neuronetics, NIMH, Novartis, Organon, Otsuka, Ortho-McNeil, PamLab, Pfizer, Sanofi-Aventis, Schering-Plough, Sepracor, Shire US, Supernus Pharmceuticals, Transept, and Wyeth-Ayerst Laboratories; he has equity holdings in MedAvante and receives royalty income from American Psychiatric Publishing, Inc., Guilford Publications, Herald House, and W.W. Norton; he has provided expert testimony

for Jones Day and Philips Lyttle, L.L.P., and Pepper Hamilton, L.L.P., and his wife is employed as the senior medical director for Embryon. Dr. Perkins receives partial salary support from studies sponsored by Novartis, AstraZeneca, Sanofi-Aventis, and Forest Laboratories. Dr. Leon has served as a consultant to NIMH, FDA, MedAvante, and Roche and as a member of data safety monitoring boards for AstraZeneca, Dainippon Sumitomo Pharma America, Organon, and Pfizer.

This work was organized by the American Psychiatric Association Committee on Research on Psychiatric Treatments (chair: Alan J. Gelenberg, M.D., 2004–2006; Jeffrey A. Lieberman, M.D., 2007–present) and the Ad Hoc Subcommittee for Evaluation of the Evidence Base for Psychodynamic Psychotherapy (James H. Kocsis [chair], Jacques Barber, Ph.D., Stephen Hollon, Ph.D., Barbara L. Milrod, M.D., Steven Roose, M.D., Michael E. Thase, M.D.). Development of the measure was done with assistance from consultants Anthony Bateman, M.D., Peter Fonagy, Ph.D., Bruce Wampold, Ph.D., Ellen Frank, M.D., John Norcross, Ph.D., and Paul Crits-Christoph, Ph.D.

References

1. Torrey EF: Does psychoanalysis have a future? no. Can J Psychiatry 2005; 50:743–744
2. Weissman MM, Verdeli H, Gameroff MJ, Bledsoe SE, Betts K, Mufson L, Fitterling H, Wickramaratne P: National survey of psychotherapy training in psychiatry, psychology, and social work. Arch Gen Psychiatry 2006; 63:925–934
3. Westen D, Novotny CM, Thompson-Brenner H: The empirical status of empirically supported psychotherapies: assumptions, findings, and reporting in controlled clinical trials. Psychol Bull 2004; 130:631–663
4. Blatt SJ, Zuroff DC: Empirical evaluation of the assumptions in identifying evidence based treatments in mental health. Clin Psychol Rev 2005; 25:459–486
5. Shedler J: The efficacy of psychodynamic psychotherapy. Am Psychol 2010; 65:98–109
6. Leichsenring F, Leibing E: The effectiveness of psychodynamic therapy and cognitive behavior therapy in the treatment of personality disorders: a meta-analysis. Am J Psychiatry 2003; 160:1223–1232
7. Leichsenring F, Rabung S: Effectiveness of long-term psychodynamic psychotherapy: a meta-analysis. JAMA 2008; 300:1551–1565
8. Leichsenring F, Rabung S, Leibing E: The efficacy of short-term psychodynamic psychotherapy in specific psychiatric disorders: a meta-analysis. Arch Gen Psychiatry 2004; 61:1208–1216
9. Abbass AA, Hancock JT, Henderson J, Kisely S: Short-term psychodynamic psychotherapies for common mental disorders. Cochrane Database Syst Rev 2006; 4:CD004687
10. de Maat S, Dekker J, Schoevers R, van Aalst G, Gijsbers-van Wijk C, Hendriksen M, Kool S, Peen J, Van R, de Jonghe F: Short psychodynamic supportive psychotherapy, antidepressants, and their combination in the treatment of major depression: a mega-analysis based on three randomized clinical trials. Depress Anxiety 2008; 25:565–574
11. Driessen E, Cuijpers P, de Maat SCM, Abbass AA, de Jonghe F, Dekker JJM: The efficacy of short-term psychodynamic psychotherapy for depression: a meta-analysis. Clin Psychol Rev 2010; 30:25–36
12. Gabbard GO, Gunderson JG, Fonagy P: The place of psychoanalytic treatments within psychiatry. Arch Gen Psychiatry 2002; 59:505–510
13. Jadad A: Randomised Controlled Trials. London, BMJ Books, 1998
14. Moher D, Schulz KF, Altman DG: The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001; 285:1987–1991
15. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P: Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med 2008; 148:295–309
16. Ollendick TH, King NJ: Empirically supported treatments for children and adolescents: advances toward evidence-based practice, in Handbook of Interventions That Work With Children and Adolescents: Prevention and Treatment. Edited by Barrett P, Ollendick T. New York, John Wiley & Sons, 2004, pp 3–25
17. 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
18. Churchill R, Hunot V, Corney R, Knapp M, McGuire H, Tylee A, Wessely S: A systematic review of controlled trials of the effectiveness and cost-effectiveness of brief psychological treatments for depression. Health Technol Assess 2001; 5:1–73
19. Fonagy P, Roth A, Higgitt A: The outcome of psychodynamic psychotherapy for psychological disorders. Clin Neurosci Res 2005; 4:367–377
20. Crits-Christoph P, Barber JP: Long-term psychotherapy, in Handbook of Psychological Change: Psychotherapy Processes and Practices for the 21st century. Edited by Snyder CR, Ingram RE. New York, John Wiley & Sons, 2000, pp 455–473
21. Leichsenring F: Comparative effects of short-term psychodynamic psychotherapy and cognitive-behavioral therapy in depression: a meta-analytic approach. Clin Psychol Rev 2001; 121:401–419
22. Shapiro DA, Shapiro D: Meta-analysis of comparative therapy outcome studies: a replication and refinement. Psychol Bull 1982; 92:581–604
23. Moncrieff J, Churchill R, Drummon C, 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
24. Sindhu F, Carpenter L, Seers K: Development of a tool to rate the quality assessment of randomized controlled trials using a delphi technique. J Adv Nurs 1997; 25:1262–1268
25. Lackner JM, Mesmer C, Morley S, Dowzer C, Hamilton S: Psychological treatments for irritable bowel syndrome: a systematic review and meta-analysis. J Consult Clin Psychol 2004; 72:1100–1113
26. Downs SH, Black N: The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998; 52:377–384
27. Elkin I: The NIMH Treatment of Depression Collaborative Research Program: where we began and where we are, in Handbook of Psychotherapy and Behavioral Change, 4th ed. Edited by Bergin AE, Garfield SL. New York, John Wiley & Sons, 1994, pp 114–139
28. March J, Silva S, Petrycki S, Curry J, Wells K, Fairbank J, Burns B, Domino M, McNulty S, Vitiello B, Severe J: Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents With Depression Study (TADS) randomized controlled trial. JAMA 2004; 292:807–820
29. 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
30. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86:420–428
31. Hamer RM: Compute six intraclass correlation measures. SAS Knowledge Base, 2005 (online). http://support.sas.com/kb/25/031.html

32. Cronbach LJ: Essentials of Psychological Testing. New York, HarperCollins, 1990
33. Wampold BE, Minami T, Baskin TW, Callen Tierney S: A meta-(re)analysis of the effects of cognitive therapy versus "other therapies" for depression. J Affect Disord 2002; 68:159–165
34. Vinnars B, Barber JP, Norén K, Gallop R, Weinryb RM: Manualized supportive-expressive psychotherapy versus nonmanualized community-delivered psychodynamic therapy for patients with personality disorders: bridging efficacy and effectiveness. Am J Psychiatry 2005; 162:1933–1940
35. Milrod B: Psychodynamic psychotherapy outcome for generalized anxiety disorder (editorial). Am J Psychiatry 2009; 166:841–844
36. Wiltink J, Dippel A, Szczepanski M, Thiede R, Alt C, Beutel ME: Long-term weight loss maintenance after inpatient psychotherapy of severely obese patients based on a randomized study: predictors and maintaining factors of health behavior. J Psychosom Res 2007; 62:691–698
37. Chambless DL, Hollon SD: Defining empirically supported therapies. J Consult Clin Psychol 1998; 66:7–18
38. Sheng D, Kim MY: The effects of non-compliance on intent-to-treat analysis of equivalence trials. Stat Med 2006; 25:1183–1199
39. Greenland S, Lanes S, Jara M: Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. Clin Trials 2008; 5:5–13
40. Bang H, Davis CE: On estimating treatment effects under noncompliance in randomized clinical trials: are intent-to-treat or instrumental variables analyses perfect solutions? Stat Med 2007; 26:954–964
41. Wampold BE, Serlin RC: The consequence of ignoring a nested factor on measures of effect size in analysis of variance. Psychol Methods 2000; 5:425–433
42. Crits-Christoph P, Mintz J: Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. J Consult Clin Psychol 1991; 59:20–26
43. Baldwin SA, Wampold BE, Imel ZE: Untangling the alliance-outcome correlation: exploring the relative importance of therapist and patient variability in the alliance. J Consult Clin Psychol 2007; 75:842–852