



Published in final edited form as:

J Consult Clin Psychol. 2009 June ; 77(3): 439–450. doi:10.1037/a0014834.

Sources of Site Differences in the Efficacy of a Multi-site Clinical Trial: The Treatment of SSRI Resistant Depression in Adolescents

Anthony Spirito, Ph.D.

Alpert Medical School of Brown University

Kaleab Z. Abebe, M.A. and Satish Iyengar, Ph.D.

University of Pittsburgh

David Brent, M.D.

Western Psychiatric Institute and Clinic

Benedetto Vitiello, M.D.

National Institute of Mental Health

Gregory Clarke, Ph.D.

Center for Health Research Northwest Kaiser Permanente Northwest

Karen Dineen Wagner, M.D., Ph.D.

University of Texas Medical Branch, Galveston, Texas

Joan Asarnow, Ph.D.

University of California, Los Angeles

Graham Emslie, M.D.

University of Texas Southwestern Medical School

Martin Keller, M.D.

Alpert Medical School of Brown University

Abstract

Site differences in treatment outcomes are not often highlighted when the results of multisite randomized clinical trials (MRCTs) are reported. In the primary analyses of a six-site MRCT, the Treatment of SSRI-resistant Depression in Adolescents (TORDIA), there was substantial variation by site in the performance of a medication-only condition and a combined medication plus Cognitive Behavioral Therapy (CBT) condition. Two potential primary causes of site differences in MRCT outcomes are examined in this paper: sampling factors, particularly clinical characteristics of participants, and treatment protocol factors, particularly fidelity. We found that differences in the clinical characteristics of participants at baseline across-site and within-site/across conditions were the most salient explanators for site differences and differences within sites across conditions in outcome. Study findings are discussed with respect to the overall study outcomes in TORDIA as well as MRCTs in general.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at <http://www.apa.org/journals/ccp/>

Keywords

Multisite clinical trials; site differences; protocol fidelity; sampling factors; TORDIA

Multi-site randomized clinical trials (MRCTs) have been defined as studies having at least two or more clinical sites where all sites follow a standard assessment and treatment protocol and in which one site serves as the data processing and analysis center (Kraemer, 2000). Most commonly, MRCTs are conducted for low prevalence psychiatric disorders because multiple sites are needed in order to be able to accrue the necessary sample size to adequately power a study. Well-designed MRCTs have other advantages besides enabling a large sample to be accrued. For example, sites can be selected to ensure an adequate representation of important participant demographics, such as race, which may improve the generalizability of the findings (Kraemer, 2000). In addition, investigators with differing areas of expertise can be brought together on one research team to devise and implement the research, which improves the likelihood that a state-of-the-art protocol is devised and reduces the possibility of allegiance effects (Luborsky et al., 1999) influencing research implementation.

When presenting the results of MRCTs, investigators report outcome data aggregated across all the sites. However, MRCTs often find site differences in important study characteristics. Some of these, e.g. baseline differences in certain participant characteristics, can be addressed in the statistical analyses. Other differences, such as potential variability in study procedures across sites, are best prevented through careful study coordination.

Site differences in treatment outcomes are not often highlighted in the reporting of results. Some papers report site by condition interactions on some outcome variables (e.g., Kallert et al., 2007), but rarely is much more detail provided. There are exceptions. For example, Davidson et al. (2004) compared the efficacy of fluoxetine, cognitive-behavioral therapy (CBT), placebo, CBT plus fluoxetine, and CBT plus placebo for 295 adults with generalized social phobia. In this two site study, the outcome results were presented for the sample as a whole. However, the authors did report and discuss a site by treatment effect: there were more responders in the CBT plus placebo group at one site than at the other. The Davidson et al. (2004) paper is an exception. Most studies only cursorily discuss site differences, e.g., that they were investigated and found to be non-significant. However, the power to detect differences across sites is low in most MRCTs because they are powered to detect outcome differences not site differences (Kraemer & Robinson, 2005). Thus, it is likely that many more MRCTs would find site differences in study outcomes if they were powered to do so. Kraemer and Robinson (2005) note that site and site-by-treatment interaction differences can lead to misinterpretation of overall treatment response findings in MRCTs and increase Type II error. Below, we examine sampling factors and study protocol factors that may be related to differences in outcomes across sites.

Sampling Factors

Important sampling factors to examine in MRCTs include sample size per site, recruitment sources, outliers, and especially participant characteristics.

Number of Participants Per Site

Studies are typically powered to detect a main effect for the entire sample whereas an even larger sample is necessary to detect site-by-treatment interactions (Noda et al., 2006). In many MRCTs, the number of participants ultimately recruited at each site can vary substantially, and many never report on the planned recruitment per site. Although a large number of sites may be necessary to recruit the sample size necessary to adequately power a study to detect an

overall treatment effect, there is evidence that the overall effect size declines as the number of sites increases (Bridge et al., 2007). In addition, as the number of sites increases, the power to detect treatment by site interactions decreases (Kraemer & Robinson, 2005). Thus, even if a study reports a non-statistically significant difference across sites in outcomes, the possibility of site differences remains if the power to detect such a difference is low. A primary concern related to sample size differences across sites is the possibility that sites with larger sample sizes contribute unequally to the overall treatment effect size.

Recruitment Sources

The source from which patients are recruited for clinical trials can affect site outcomes. For example, in a study examining three psychosocial treatments for adolescent depression, Brent et al. (1998) found that recruitment through clinical referrals, compared to advertisements, was related to continued depression at follow-up.

Outliers

Study participants can vary widely in their response to study treatments such that a few individuals can inordinately affect the overall outcomes for a condition or a site. Many statistical analyses can handle outliers, and methods to detect outliers are also available (Tabachnik & Fidell, 2006).

Participant Characteristics

Simple randomization is the preferred method to ensure that there is equivalency across conditions in psychotherapy research (Hsu, 1989). However, when one considers a specific condition within a specific site, simple randomization is less likely to ensure equivalency due to small sample size (Shadish, Cook, & Campbell, 2002). Thus, variability in demographic and clinical characteristics can result in site differences.

Demographics—MRCTs often intentionally choose sites with diverse demographic characteristics in order to improve a study's generalizability (Meinert, 1986). This variability may lead to site differences as demographic characteristics may play a significant role in the response of patients to psychotherapy (Petty, Tennen, & Affleck, 2000). For example, higher socioeconomic status (SES) is typically associated with better response to psychotherapy (Curry et al., 2006). Therefore, MRCTs typically investigate whether demographic factors are related to outcome.

Baseline Clinical Characteristics—Researchers often use stratification techniques, such as Efron's biased coin toss (Begg & Iglewicz, 1980), to ensure that the treatment conditions of a clinical trial are equally balanced on baseline characteristics selected a priori as important to control. Such stratification techniques assign participants in a given subgroup to intervention conditions, but systematically bias the randomization in favor of balance across intervention conditions, both across and within sites (Tabachnik & Fidell, 2006).

When information is available for researchers to cogently balance treatment assignment according to a few potential predictors of treatment outcome, this approach is preferable to simple randomization (Hsu, 1989). However, not all potential variables can be controlled in the randomization process, especially within a particular site. Thus, an additional question also arises as to whether the most relevant variables were chosen for stratification. This possibility is examined post hoc by comparing sites on baseline clinical characteristics in general and clinical characteristics found to be related to outcomes in particular.

In MRCTs, differences in clinical characteristics that affect outcome must not only be examined by site, but also by condition within site to help examine possible reasons for differential effects

of condition within sites. This is important to investigate in order to determine whether factors such as clinical characteristics, or alternatively, expert delivery of the particular treatment, might account for differences in outcome across conditions.

Study Protocol Consistency across Sites

Study protocol consistency is a key factor in the execution of any MRCT. A great deal of attention must be paid to training sites in the assessment protocol to ensure consistency across sites. Treatment protocol deviations are also potential sources of outcome inconsistencies. Clinician fidelity to the treatment protocol is a very common factor examined in MRCTs (Moncher & Prinz, 1991) and may help explain potential site differences. Participant adherence to a treatment protocol, whether medication or psychotherapy, can also vary across sites and affect outcome.

Fidelity to Assessment Protocol

MRCTs typically have a uniform training procedure to ensure reliability of assessment across sites, procedures in place to check forms as they are collected to discover missing data, and periodic reliability checks to catch rater drift that may occur. Reliability of assessment raters across sites can be particularly challenging to maintain in MRCTs, and can contribute to site differences in outcomes. In addition, reliability of a measure also affects power, both across and within sites. Power decays as the inverse square of reliability (Lachin, 2004). Kraemer (1991) notes that the greater the reliability of an outcome measure, the greater the power. Test-retest reliability (r) between 0.6 and 0.8 does not affect power but an $r < 0.6$ can create problems in power. Correlations below $r = 0.4$ certainly cause power problems and subsequent concerns about the stability and interpretability of the findings (Kraemer, 1991).

MRCTs typically rely on both self-report and interviewer-rated outcome measures. Regardless of assessment mode, missing data – whether it be items or entire measures – may affect outcome analyses. Missing data is particularly problematic in MRCTs if it is not missing at random (Schafer, 1997), and if this non-randomness occurs at some sites but not others. Statistical techniques are available to explore patterns of missing data, i.e. whether or not they are missing at random, and to impute missing data values in data analyses (Little & Rubin, 2002; Schafer, 1997).

Protocol Deviations

Protocol deviations refer to exceptions to the study procedures that are reviewed by the project Steering Committee to determine whether a participant remains eligible for inclusion in the clinical trial (Meinert, 1986). Protocol deviations can occur prior to or during enrollment in a study. For example, prior to enrollment, the Steering Committee might decide to allow a participant into a clinical trial who does not precisely fit an inclusion criterion, e.g., a participant slightly younger than the trial's minimum age. During a trial, a protocol deviation might include allowing a participant to remain in the trial after determining that he/she received limited additional treatment outside of the protocol.

Clinician Fidelity to Treatment Protocols

In MRCTs, a great deal of attention is paid to training both pharmacotherapists and psychotherapists in standardized treatment delivery (Moncher & Prinz, 1991). Ratings of treatment fidelity and competency of therapists are typically collected and compared across sites to ensure that any outcome differences are not related to the delivery of the protocol (Borelli et al., 2005). Rating therapist/client therapeutic alliance is also commonly used to determine whether general therapist factors might differ across sites and in turn affect outcomes (Luborsky et al., 1999).

Differences in the Use of Therapy Modules across Sites

The mandated sequence of treatment components in some manualized behavioral treatments has been criticized as interfering with therapy. More recently, however, many manuals present a number of different techniques from which only a few are prescribed as core techniques that need to be implemented with all participants. Other modules may be chosen based on both patient presentation and patient progress (Kendall, Cho, Gifford, Hays & Nauta, 1998). In addition, in many protocols, therapists can delay, eliminate, or speed up the use of certain techniques in the outlined sequence of treatment. Although this flexibility in module selection may improve a protocol's clinical application, it also results in variability across participants and sites that could theoretically be related to outcome and site differences.

Participant Differences in Protocol Adherence/Attrition

Participant adherence to treatment protocols, whether medication or psychotherapy, can have significant effects on treatment outcome. Treatment dropout is a significant issue in most MRCTs. There can be two types of attrition: one is dropout from the treatment protocol, the other is dropping out from the study altogether, i.e. not being available for the scheduled assessments of the study and therefore not contributing any more information to the analyses. Rates at which participants are retained in treatment protocols can vary significantly across sites. For example, certain baseline patient characteristics, such as severity of psychopathology, may affect retention in treatment. Alternatively, sites with less experience may have more difficulty retaining participants in the treatment protocol. Consequently, site differences in retention may affect overall study outcome. Consequently, intent-to-treat (ITT) analyses (Lachin, 2000), in which all participants enrolled in a trial are included in data analyses regardless of whether they drop out of the trial, are the standard in MRCTs (Lachin, 2000).

Current Study

Recently, a six site MRCT entitled, "Treatment of SSRI-Resistant Depression in Adolescents (TORDIA)," found overall positive outcomes of CBT plus a switch in antidepressant medication on Major Depressive Disorder (MDD) relative to a switch in medication alone. Site differences were also reported for treatment response (Brent et al., 2008). The purpose of this paper is to examine possible reasons for site differences in rates of treatment response in TORDIA. Only factors found to differ across sites, and related to treatment response rate, were examined. We discuss differences found in both the medication only and CBT plus medication conditions as determined by the major outcome variable of the original study, responder/non-responder status, at the end of the 12-week acute phase of treatment.

Method

TORDIA enrolled participants, ranging from 12 to 18 years of age, who met DSM-IV criteria for MDD as assessed by the Schedule for Affective Disorders and Schizophrenia for School-Aged Children – Present and Lifetime Versions (K-SADS-PL; Kaufman et al., 1996). All participants had to have been treated with a confirmed adequate course of an SSRI for 8 weeks and to have failed that trial in order to be eligible for the study. If eligible, they were randomized to one of four treatments: switch to another SSRI, switch to venlafaxine, switch to another SSRI plus CBT, or switch to venlafaxine plus CBT. Treatment response rates were assessed with the Children's Depression Rating Scale-Revised (CDRS-R; Poznanski, Freeman & Moknos, 1984) and the Clinical Global Improvement Scale (CGI; Guy, 1976).

Of the 334 enrolled participants, 69.2% were retained in their original treatment arm through week 12, and 85.9% were assessed through week 12. Overall findings indicated that a higher proportion of those treated with CBT plus medication were responders (i.e. a CGI ≤ 2 and an

improvement in the CDRS-R $\geq 50\%$) than those treated with medication alone, 54.8% (95% Confidence Interval: {46.9,62.5}) vs. 40.5% (33.0,48.3; Hedges' $g = 0.287$), respectively. Logistic regression indicated a main effect for CBT plus medication but not for medication alone (for details of this study, see Brent et al., 2008). Significant site differences in outcomes were also detected in TORDIA. As can be seen in Figure 1, a 2.5-fold variation in the response rate for the CBT plus medication condition was found and was statistically significant, $X^2(5, n=166) = 11.34, p < .05$. Because there were no differences in outcome between the SSRI-switch and the venlafaxine-switch, the medication conditions were collapsed for analyses in this paper. For the medication-only condition, a fourfold variation in response was also statistically significant, $X^2(1, n=168) = 16.10, p < .01$.

When the CBT plus medication condition (i.e. CBT/no CBT), site, and the interaction were included in a logistic regression, the overall fit of the model was significant [Homer-Lemeshow $X^2(11, n=334) = 35.62, p < 0.001$]. More importantly, the interaction term was significant, [$X^2(5, n=334) = 19.89; p = 0.001$], which didn't occur when the collapse was across the medication (venlafaxine/SSRI) condition, [$X^2(5, n=334) = 3.91; p = 0.5623$]. In fact, after removing the insignificant interaction terms, there were no site differences controlling for medication condition, [$X^2(5, n=334) = 6.95; p = 0.2245$].

Procedure

Site differences were examined post hoc using data collected in TORDIA. Site differences in sampling factors were investigated first, followed by potential differences in study protocol characteristics.

Sampling Factors

Four sampling factors were examined: number of participants per site, recruitment sources, outliers, and participant characteristics.

Sample Size Across Sites—TORDIA was designed with four sites expected to recruit 80 participants and two sites to recruit 40 participants each. The two sites with lower recruitment goals were chosen because of investigator expertise in CBT and pharmacotherapy trials.

Recruitment Sources—Clinical referrals, primarily from mental health professionals but also pediatric healthcare professionals, were the primary sources of patient recruitment in TORDIA. Participants were also recruited through advertisements.

Outliers—TORDIA data was analyzed according to whether a participant was a responder or nonresponder. Consequently, there were no outliers in the TORDIA outcome data.

Participant Characteristics—The design of TORDIA included a comprehensive baseline assessment of a wide range of adolescent symptoms and family functioning. Stratification variables included a diagnosis of chronic depression, defined as a diagnosis of MDD or Dysthymia and duration of current episode greater than or equal to 24 months, a diagnosis of comorbid anxiety disorder, and suicidality.

Study Protocol Consistency

Five aspects of study protocol consistency were emphasized in TORDIA: fidelity of the assessment procedures, limiting missing data, pharmacotherapy fidelity, CBT fidelity, therapist CBT module selection, and maintaining participants in the trial.

Assessment Reliability—In TORDIA, each site had procedures in place to check assessment forms for inaccurate or missing responses. Independent Evaluators (IEs) underwent training and certification at the Coordinating Center in the K-SADS-PL, the baseline measure used to confirm cases, and the CDRS-R, a primary outcome measure. Interviewers needed to demonstrate Kappa or Intraclass Correlations (ICC) $>.8$ on five interviews prior to certification. The certification process for site IE supervisors on the primary diagnostic and dependent variables required viewing standardized tapes of child and parent interviews and completing the CDRS-R and K-SADS-PL. Supervisor ratings of standardized K-SADS-PL and CDRS-R interviews were reviewed by the Coordinating Center. Supervisor certification on the K-SADS-PL required agreement on at least 80% of the 54 items at the summary symptoms level. For the CDRS, the Coordinating Center contacted the IE supervisor to discuss and resolve disagreements in which total score discrepancies were $\pm > 5$ points (for video recordings) or $\pm > 6$ for audio recordings. IE supervisors had to achieve CDRS-R total score ratings within the ranges specified for the standardized taped interviews in order to be certified to train IE(s) at their sites.

Missing Data—All primary statistical analyses in TORDIA (Brent et al., 2008) used intent-to-treat (ITT) approaches including last observation carried forward (LOCF) and imputation (Schafer, 1997). Results were comparable with LOCF and multiple imputation, using chained equations (ICE) in STATA (Royston, 2004). The primary outcome analyses (Brent et al., 2008) and current analyses of site differences use the LOCF data. Consequently, even though sites were not able to collect outcome data on all participants (see Results below), outcome analyses were conducted with all enrolled participants.

Pharmacotherapy Fidelity—In TORDIA, medication sessions were 30 to 60 minutes long and included assessment of vital signs, side effects, and symptomatic responses. Participants were seen weekly for 6 weeks and biweekly for weeks 7 through 12. Training in the pharmacotherapy protocol in TORDIA was designed to ensure standardized coverage of assessment and safety issues as well as to ensure that CBT techniques were not used. All pharmacotherapy sessions in TORDIA were tape recorded and rated using the 16-item Pharmacotherapy Rating Scale (PTRS) derived from the Clinical Management Scale (Hill, O'Grady, & Elkin, 1992).

CBT Protocol Fidelity—A two-day training program on the study CBT protocol was held for therapists (who had at least a Masters Degree) and supervisors, with a re-training 18 months into the trial. All CBT sessions were tape-recorded. CBT audiotapes were then rated by site supervisors on the Cognitive Therapy Rating Scale (CTRS). The CTRS (Vallis, Shaw, & Dobson, 1986) is an 11-item scale that measures therapist competence in the delivery of CBT. To be certified, a TORDIA CBT therapist had to be rated by both the site CBT supervisor and the Coordinating Center on a minimum of 12 TORDIA CBT tapes scored at an acceptable level on the CTRS, a total score of greater than or equal to 39. After a CBT therapist was officially certified, at least two CBT tapes from each participant were selected at random and reviewed from the acute phase of TORDIA by each site. All site-reviewed tapes were also rated at the data Coordinating Center to ensure both within site and across site reliability.

CBT Module Selection—In TORDIA, the therapist was required to use certain modules (e.g., psychoeducation, cognitive restructuring, problem-solving) and have at least 3 family sessions in the first 12 weeks of treatment. The therapist had flexibility in selecting the remaining modules to be used in treatment, based on on-site and Coordinating Center supervision as well as a review of 6-week case formulations in a biweekly CBT conference call.

Participant Adherence to Treatment Protocols—In TORDIA, the number of CBT sessions attended was recorded and examined across sites. Medication adherence was examined by recording pill counts at each pharmacotherapy visit.

Data Analysis

Analyses of variance with posthoc testing was used to examine site differences in continuous variables, while chi square tests were used to analyze dichotomous variables. Logistic regression analyses were conducted to examine the overall model. A recursive partitioning based on Receiver Operating Characteristics (ROC) was used to identify homogeneous subgroups where site variability was diminished (Kraemer, 1992). In this method, each of the clinical predictors that contributed to site differences and treatment outcomes was examined one at a time by comparing the diagnostic predictability for each of the variable's cutpoints. The best cutpoint of each predictor was determined by a kappa measure of agreement which combined the sensitivity and specificity of each cutpoint. The predictor whose kappa value was the highest -- as long as the corresponding Chi square value was significant -- was chosen to be the optimal cutpoint. The data were then divided into two groups based on the optimal cutpoint and the process repeated until there were non-significant associations or small sample sizes, i.e. marginal counts < 10 (for details, see Kraemer, 1992; Noda et al., 2006). We regard these analyses as exploratory because they are not based on a priori hypotheses, and consequently do not adjust for multiple testing.

Results

The results presented below are organized by potential sampling and protocol consistency factors that could result in site differences. If a site difference was found to be statistically significant, then additional analyses were conducted to determine whether the site difference was related to response to treatment. If so, further analyses differentiating the sites on the variable were conducted.

Sampling Factors

Number of participants across sites and recruitment sources, as well as demographic, stratification and clinical characteristics, are reviewed first below. Receiver operating curve analyses are then presented as a means of determining those clinical characteristics that are most strongly related to outcome.

Number of Participants By Site—By study completion, the actual number of participants recruited at the six sites was as follows: 34, 42, 45, 51, 70, and 92. The lower-than-expected sample size did not affect the ability of TORDIA to detect site by treatment interactions. Power was calculated at .94 to detect the interaction effect size observed in the TORDIA completer sample ($n = 295$).

Sensitivity analyses for clinical response were also conducted in which each site was removed one at a time. For the medication-only condition, there were only minor effects on the overall response rate (40.5%) of removing a site. Removing Site #4, the largest site, lowered the overall medication response rate the most of all sites to 36.5%. However, removing Site #1, the smallest site, had the second largest effect on lowering the overall response rate, to 37.7%. Removing Sites #2 and #3 resulted in essentially no change, and removing Sites #5 and #6 increased the response rate a few percentage points.

For the CBT plus medication condition, there were more substantial effects on the overall response rate (54.8%) by removing individual sites than medication only rates. The response rate improved a few percentage points when Site #1 (the smallest site) and Site #4 (the largest

site) were removed. Removing Site #2, the second largest site, had the most significant effect on outcomes, lowering the rate of responders to slightly above 50% (from 54.8%). Nonetheless, when taken together, the larger sites did not appear to have had an undue positive or negative effect on the overall study response rates for either condition.

Recruitment Sources—There was a significant difference in referral sources across sites, $X^2(1, n=334) = 35.19, p < 0.001$. The primary difference was that Site #1 with no advertising referrals and Site #5 with one referral (2%) from advertisements were significantly lower than the other four sites where referrals from advertisements ranged from 16.7% to 38.6%. However, referral source was not significantly related to outcome, $X^2(1, n=334) = 0.01, p = .98$; OR=0.99 (0.58, 1.70).

Demographic Characteristics—Analyses revealed site differences in age, $F(5, 328) = 3.14, p < .01$; race, $X^2(5, n=334) = 18.07, p < .01$; gender, $X^2(5, n=334) = 15.86, p < .01$; and SES, $F(5, 171) = 6.48, p < .0001$. Age across sites ranged from a mean of 15.5 years to 16.2 years ($M=15.88, SD=1.57 \{15.71, 16.05\}$). The percentage of females ranged from 57.1% to 86.3% ($0.70 \{0.65, 0.75\}$), while percentage of Whites ranged from 64.3% to 91.3% ($0.83 \{0.78, 0.87\}$). The most substantial differences were on SES. Nonetheless, when examined in relation to outcome, none of the demographic variables predicted outcome.

Baseline Assessment—The baseline included a comprehensive assessment of individual and family characteristics, including some specifically selected to include in the stratification procedures, both across and within sites.

Stratification Variables: There were significant differences across sites on chronic depression and comorbid anxiety with a trend noted on suicidality (see Table 1). Rates almost doubled from the lowest to the highest sites for chronic depression and almost tripled for comorbid anxiety and suicidality.

When the relationship between the stratification variables and the main outcome variable (responder/non-responder) was examined, there were no significant relationships for chronic depression, $X^2(1, n = 334) = 0.81, p = .37$; OR=1.00 (0.53, 1.26) or comorbid anxiety, $X^2(1, n = 334) = 0.01, p = .99$; OR=1.00 (0.64, 1.56). For suicidality (BDI Item $9 \geq 2$), there was a significant finding, $X^2(1, n = 334) = 8.12, p < .005$; OR=0.35 (0.17, 0.75). Sites #1 and 6 had significantly higher levels of suicidality than the other four sites. As suicidality was both different across sites, and related to response outcomes, this variable may have contributed in part to site differences in outcomes.

Clinical Characteristics: Additional post hoc analyses identified three baseline variables (see Table 2) that differed across sites and were related to clinical response overall: duration of index MDD episode as rated on the K-SADS-PL, Hosmer-Lemeshow $X^2(1, n=326) = 4.01, p < .05$; OR=0.99 (0.98, 1.00); hopelessness, as assessed by the Beck Hopelessness Scale (Beck, 1988), Hosmer-Lemeshow $X^2(1, n=327) = 9.20, p < .01$; OR=0.94 (0.90, 0.98); and family conflict, as reported by adolescents on the Conflict Behavior Questionnaire (Robin & Weiss, 1980), Hosmer-Lemeshow $X^2(1, n=327) = 13.62, p < .01$; OR=0.93 (0.90, 0.97).

Logistic regression was conducted including these three baseline variables as covariates, along with site, treatment condition, and the interaction of the two. Stepwise backwards regression resulted in a model that included the CBT plus medication condition, and the three baseline variables, but not the medication only condition (Hosmer-Lemeshow chi-square = 48.98, $df = 20, p = .001$). Because the medication only condition was not significant in the main model, CBT plus medication, site, and the interaction were tested in another model. Again, the three baseline clinical variables were retained in the model, Hosmer-Lemeshow chi-square = 45.42,

df = 14, $p < .001$, suggesting that these baseline and interaction terms account for some of the site effects in the CBT plus medication outcomes.

Within Site, Across Condition Baseline Clinical Characteristics: There were within-site differences across conditions on the stratification variables (i.e. chronic depression, comorbid anxiety, and suicidality) found at Sites #1, #5, and #6 in TORDIA. For example, at Site #1, where the CBT plus medication response rate was much lower than the medication only response rate, the CBT plus medication group had four-fold higher rates of suicidal ideation. However, these stratification variables were not related to outcome.

The baseline clinical characteristics that were related to outcome were also examined. There were 17 within-site differences on socioeconomic status (2 sites), depression (1 site), hopelessness (2 sites), suicidality (5 sites), anxiety (1 site), substance abuse (1 site), borderline features (3 sites), and family conflict (2 sites). Three-quarters of these differences were in the direction of the more successful treatment conditions having participants with less severe clinical characteristics than the less successful condition. There was only one within site difference between conditions related to response rate: the Site #1 CBT plus medication group had almost twice the adolescent-reported family conflict than the medication-only group. Thus, it appears that the low response rate of Site #1's CBT plus medication group might have been related to the high level of family conflict in this condition relative to the medication-only condition.

Receiver Operating Curve Analyses—Given the findings regarding clinical characteristics related to outcomes, we conducted ROC analyses with all 334 participants to determine which differences in baseline clinical variables accounted best for site differences in outcome. Low family conflict (< 9) and low hopelessness (< 10) comprised the cutpoints among the predictors with the best balance of sensitivity and specificity for predicting response rate. As seen in Figure 2, the overall clinical response of 47.6% improved to 67.8% in the subgroup of participants with low family conflict and low hopelessness ($N = 87$). Conversely, for patients who had high family conflict (≥ 9) and hopelessness (≥ 10), the response rate dropped to 37.0% ($n = 119$). When sites were stratified on these characteristics, the overall site effects were no longer significant, Fisher's $X^2 = 2.10$, $p = .84$.

We also examined the percentage of participants within conditions on these characteristics. For the CBT plus medication condition, the percentage of participants who had low family conflict and low hopelessness did not differ across sites, $X^2(5, n = 166) = 10.06$, $p = .07$. For the medication only condition, there was a significant difference across sites, $X^2(5, n = 168)$, $p < .001$. Site #3 had a significantly larger percentage of patients with low family conflict and low hopelessness scores (48.9%) than sites #1(17.6%), 3(22.9%), 4(20.7%), 5(27.5%), and 6 (23.8%).

Within site, across condition differences in the percentage of participants with low family conflict and low hopelessness were also investigated. At Site #1, the medication-only group had a higher percentage of these patients compared to the CBT plus medication group, 35.3% vs. 0%; $X^2(1, n = 34) = 7.29$, $p < .01$. At Site #5, the CBT plus medication condition had a higher percentage of these patients than the medication only condition, 40% vs. 15.4%; $X^2(1, n = 51) = 3.88$, $p < .05$. At Site #6, the CBT plus medication condition had a higher percentage of these patients than the medication-only condition, 38.1% vs. 9.5%; $X^2(1, n = 42) = 4.73$, $p < .05$. Thus at these three sites, the large differences in response rates across conditions within site were related to the participant characteristics. Only at Site #3, where the medication-only condition had a higher percentage of these patients than the CBT plus medication condition, 65.2% vs. 31.8%; $X^2(1, n = 45) = 5.02$, $p < .05$, was this pattern not detected.

Protocol Consistency

Protocol factors examined included interrater reliability, missing data, protocol deviations, fidelity to the pharmacotherapy and CBT protocols, use of CBT modules, and CBT and medication condition attrition/adherence.

Interrater Reliability on the CDRS and K-SADS-PL—The Coordinating Center in TORDIA rated randomly selected tapes (N=95) of the CDRS from each site and calculated interrater reliability between the Coordinating Center and each site. Four sites had very high intraclass correlations for the CDRS total score (Site #3 = .89, Site #6 = .92, Site #1 = .93, and Site #4 = .93). The other two sites had high ICC correlations (Site #2 = .76, and Site #5 = .79). For the K-SADS-PL, Kappa coefficients were calculated for randomly selected tapes (N=85). The percent agreement was 84.7% for MDD with a Kappa of .65, while for Dysthymia, the percent agreement was 93.8% with a Kappa of .74. There were no differences across the sites. By Kraemer's (1991) criteria, these reliability ratings are good, and would not be expected to affect power or be related to site differences in outcomes.

Missing Data—Although all analyses were conducted using LOCF, we examined the percentage of missing outcome data, i.e. CDRS and CGI, across sites. Missing data ranged from 8.57% to 22.22% across sites, but this difference was not significant, $X^2(5, n=334) = 4.43$, $p = .49$.

Protocol Deviations—Protocol deviations occurred infrequently in TORDIA and did not differ significantly across sites. Two sites had 3 protocol deviations, two sites had 5 deviations, one site had 7 deviations, and one site had 10 deviations. Typical deviations included enrollment of participants who were slightly at variance with qualification criteria. There was no difference across sites on protocol deviations, $X^2(5, n=334) = 8.78$, $p = 0.12$.

Fidelity to the Pharmacotherapy Protocol—The Coordinating Center reviewed 153 sessions and found that 92.8% were of acceptable quality. There was a significant difference across sites on the PTRS with Site #1, having had the highest response rate of medication of all sites, reporting higher self-ratings on the PTRS than Sites #2, #3 and #5. When rated at the Coordinating Center, Site #3 had a lower PTRS score than all the other sites except Site #6. However, Site #3 had a reasonable response rate to medication, suggesting that pharmacotherapy fidelity was not related to treatment outcome. A logistic regression analysis was conducted to determine whether fidelity to the pharmacotherapy protocol played a significant role in outcome differences. Average PTRS scores did not predict response rates, whether rated by the site (Homer-Lemoshow chi square [1] = 0.16, $p = .69$) or the Coordinating Center (Homer-Lemoshow chi square [1] = 0.43, $p = .51$). Thus, pharmacotherapy fidelity did not play a major role in site outcome differences.

Fidelity to the CBT Treatment Protocol—On-site supervisors rated a total of 277 CBT tapes across all sites. Two supervisors at the Coordinating Center reviewed 351 tapes and one external consultant reviewed 49 tapes. There were no differences across sites in the mean ratings obtained on the CTRS when rated by site supervisors or the external consultant. However, on the Coordinating Center ratings, post hoc analyses revealed that one site (#3) fared less well on the CTRS than the other five sites, $F(5, 135) = 4.38$, $p < .0001$. Similar findings were obtained on the percentage of tapes rated by the Coordinating Center as acceptable (i.e., CTRS score ≥ 39). Site #3 had 64.9% of tapes rated as acceptable whereas the remaining sites had acceptable ratings ranging from 92.3% to 99.2% (Fisher's exact test = 76.52, $p < .0001$). As Site #3's response rate for CBT was higher than Site #1's, a site with a higher CTRS score, it does not appear that CBT treatment fidelity led to the site differences in the CBT plus medication response.

Logistic regression analyses of the CTRS scores on response rates did not reveal any statistically significant relationships, regardless of whether the CTRS scores were rated by the site, Hosmer-Lemeshow $X^2(1, n=103) = 0.21$; $p = .65$; OR=1.15 (0.62, 2.13); Coordinating Center, (Hosmer-Lemeshow $X^2(1, n=139) = 2.19$; $p = .14$; OR=1.58 (0.86, 2.91); or an expert consultant, Hosmer-Lemeshow $X^2(1, n=47) = 1.06$; $p = .30$; OR=1.68 (0.62, 4.57). Thus, fidelity to the CBT treatment protocol was not related to site differences in treatment outcomes.

Differences in the Use of CBT Modules across Sites—The median number of times a CBT module was used per participant was examined across sites. Only Motivational Interviewing (MI) varied by site with Site #3 using this technique more than the other sites, Kruskal-Wallis $X^2(5) = 27.46$, $p < .0001$. More frequent use of MI was also related to a lower response rate, $z = -2.16$, $p < .05$.

Site #3 had a significantly lower response rate than the response rates at Sites #2 and #5, which used motivational interviewing less frequently. Site #3 also had a lower CTRS acceptability rating than the other sites. Therefore, we ran three logistic regressions of CTRS scores (as rated by site, coordinating center, and expert consultant) on response rate controlling for use of MI. None of these regression analyses revealed statistically significant results. Thus, the interaction of MI plus CTRS ratings did not explain site differences in outcome.

Differences in Attrition—Drop-out in the first 12 weeks of the trial was primarily patient-initiated but also included protocol removal due to the occurrence of a serious event such as a report of abuse. As might be expected, patients who dropped out of the study had significantly lower response rates than those who completed treatment, $X^2(1, n=334) = 33.39$, $p = .0001$. For the CBT plus medication condition, rates of completion varied substantially, ranging from 42.2% to 81.0%, with intermediate rates of 54.6%, 63%, 74.3%, and 76%. Chi square analyses revealed a trend for significant differences across sites in drop out rates, $X^2(5, n = 166) = 10.46$, $p = .064$. The response rate for participants in the CBT plus medication condition was 23.4% higher for completers (62.7%) versus drop-outs (39.3%), $X^2(1, n=166) = 8.23$, $p < .01$; OR=2.60 (1.34, 5.04).

Dose of CBT treatment, i.e. number of sessions attended, was examined across sites. There were no statistically significant differences in number of CBT sessions attended across sites in TORDIA, $F(5, 160) = 1.43$, $p = .22$. The mean number of sessions ranged from 7.2 (SD=4.9) for Site 1 to 9.4 (SD=3.2) for Site 6. The median number of sessions ranged from 7.0 (Site 1) to 10.0 (Site 6). There was also no relation between number of sessions and response rates.

Treatment completion for the medication-only condition of TORDIA was very consistent across sites, ranging from 69.2% to 76.5%, and not significantly different. In the medication only condition, completers had much higher response rates than dropouts (49.6% vs. 17.0%), $X^2(1, n=168) = 14.90$, $p < 0.001$.

Simple regression analyses did not reveal that stratification variables or the baseline clinical characteristics were related to the number of CBT sessions attended: duration of depression, $r(162) = .01$ ($p=0.91$); CBQ, $r(162) = .10$ ($p=0.20$); or BHS, $r(163) = .03$ ($p=0.68$). For dichotomous variables, chronic depression, $F(1, 164) = 1.16$, ns; comorbid anxiety, $F(1, 164) = 1.95$, ns; and suicidality, Kruskal Wallis, $X^2(1, n=166) = 1.27$, ns. The only statistically significant difference in the relation between treatment drop-out and these clinical characteristics was parent-teen conflict as measured by the CBQ (Adolescent report), $r(327) = 0.15$ ($p=0.01$). Drop-outs had higher levels of family conflict ($M = 10.3$, S.D. = 6.7) than completers ($M = 8.3$, S.D. = 5.9) on the CBQ, $F(1, 325) = 7.80$, $p < .01$, $g=0.44$.

The difference in retention of participants across conditions within site was only statistically significant at Site #1, where there was a much higher retention in the medication only condition than CBT plus medication condition. This retention rate is consistent with the much higher response rate in the medication only condition than CBT plus medication condition at Site #1. In order to better understand drop-out, we examined the reasons for drop-out from the CBT plus medication condition at Site #1. Of the 17 assigned to this condition, two were assessed and randomized to CBT but never attended any sessions. A third participant withdrew after 2 sessions. All three of these participants said that they withdrew due to scheduling problems. One possibility for the dropout of these three participants is that the site did not significantly emphasize the importance of committing to the requirements of the research protocol during the consent process. Two patients were removed from the study when confidentiality had to be broken due to disclosure of sexual abuse, and a 6th patient became psychotic at week 5 and was withdrawn. Thus, it appears that these three patients did not receive adequate CBT treatment due to the severity of their clinical presentation. A seventh was removed at the family's request when the first Black Box warning on SSRIs was announced. Thus, it appears that the drop-out from the CBT plus medication arm of these four subjects was more likely to be related to patient and family factors than the delivery of the treatment protocol.

Medication Adherence—An ANOVA was conducted examining the average number of pills per week across site. There was a statistically significant difference across sites, $F(5,302) = 4.00, p < .01$. Post hoc tests revealed that Site #3 ($M=19.78$) had a lower rate of medication adherence than Site #5 ($M=24.21$) or Site #6 ($M=24.74$). However, medication adherence was not related to outcomes.

Discussion

In our exploration of the source of site differences in treatment response in TORDIA, we found several participant characteristics that differed across sites and were related to outcome, including suicidality, duration of the index MDD episode, hopelessness, and family conflict. A logistic regression that reported main effects of family conflict, treatment, and a treatment by hopelessness interaction fit the data well, and the addition of a treatment by site interaction did not add to the model. This can be interpreted to mean that the treatment by site interaction was in part explained by baseline differences in these clinical variables that predicted outcome.

ROC analyses added confidence to the logistic regression analyses and found two clinical variables that were different across sites and related to outcome. These variables were greater intensity of adolescent-reported parent-child conflict and severity of self-reported hopelessness. ROC analyses indicated that low levels of hopelessness and family conflict resulted in response rates approximately 20% higher than the remaining participants. Indeed, the site with the highest CBT plus medication response rate had a significantly higher percentage of patients with low scores on these variables than the three sites with the lowest CBT plus medication response rate. When analyses were conducted stratifying these variables, overall site differences were no longer statistically significant. We also looked at within site, across conditions differences and found statistically significant differences at four sites. At three of these four sites (Sites 1, 5 and 6), the direction of the difference was directly comparable to the response rate by condition. These findings help to explain the very discrepant response rates at these sites in which the medication only condition was superior to the CBT plus medication condition at Site #1 but the opposite finding was noted in Sites #5 and 6.

We are unaware of any studies that have examined clinical predictors of site differences. Nonetheless, TORDIA site difference predictors are very similar to those found in other studies of general predictors of treatment response in adolescent depression. The hopelessness findings in TORDIA are consistent with the predictors found in the Brent et al. (1998) treatment study

of adolescent depression as well as TADS (Curry et al., 2006). Chronic depression was also a predictor in both TADS and TORDIA. Rohde et al. (2006) found predictors comparable to TORDIA at one year follow-up of group therapy for depression: earlier MDD onset, hopelessness, and low family cohesion. However, parent-adolescent conflict, which was found to differ across sites and be related to outcome in TORDIA, was not a predictor in TADS. TADS and the Brent et al. (1998) study also found that comorbid anxiety disorder was a predictor of response, but this was not true of TORDIA.

Our examination of other potential reasons for site differences in response rates in TORDIA did not find site differences to be related to variations in fidelity to the assessment protocol, protocol deviations, the number of participants enrolled per site, recruitment sources, or fidelity to the CBT or pharmacotherapy protocol. Although there was no relationship between CBT fidelity and outcome, these findings might also have been affected by the restricted range of CTRS scores at 5 of the 6 sites (Trepka et al., 2004).¹ The use of particular CBT modules also did not differ across sites, with one exception. More frequent use of the Motivational Interviewing module was related to a lower response rate. Use of MI did not necessarily reflect a site difference in therapist effectiveness, but rather was more likely related to differences in participant characteristics across sites. That is, MI was only used if a participant reported significant substance use during the trial that a therapist believed might interfere with treatment of the patient's depression.

Attrition in the CBT plus medication condition differed by site and played a role in outcome differences. In the CBT plus medication condition, completers of the 12 week acute treatment protocol had response rates 23.4% higher than drop-outs. The response rate was 31.4% higher for patients who completed the medication only condition compared to dropouts. These findings should be interpreted cautiously because some participants were preemptively withdrawn by the research team because of a worsening clinical course. Although statistical analyses did not find a relationship between baseline clinical characteristics and number of CBT sessions attended, family conflict was related to drop-out. Thus, drop-out was related, in part, to a clinical characteristic (family conflict) which differed across sites and was related to overall response rates. Although one might speculate that CBT would be better able to address family conflict than a medication-only condition, the CBT protocol in TORDIA was primarily an individual protocol. Although there were family modules available, there was limited time to address family conflict in the first 12 weeks of treatment.

There are several limitations to this study of site differences. First, we chose to examine site differences by a dichotomous variable, clinical response to treatment, rather than a continuous variable, such as change in depressed mood as measured by the CDRS. Site differences, and the reason for site differences, might differ based on the outcome variable selected. Second, there were variables not assessed in this study which may have affected site differences. For example, site differences in attrition may have been related to the skill with which some sites were able to maintain participants in a demanding protocol. Alternatively, differences in therapist allegiance to CBT across sites may have been related to attrition. TORDIA did not systematically assess therapist beliefs about the effectiveness of the treatment they were delivering. Nonetheless, intervention staff were hired on the basis of prior CBT training and experience, which presumably is a reasonable proxy for allegiance to a given treatment approach. Congruence between patient desire to receive a particular treatment, i.e. CBT vs. no CBT, and their treatment assignment may have also affected drop-out, but this belief was not assessed in TORDIA.

¹The authors thank an anonymous reviewer for suggesting this possibility.

In conclusion, as MRCTs are conducted with more clinically complex patient populations, the inclusion of multiple sites will typically lead to a more diverse clinical sample. Broadening the range of clinical characteristics contributes to the richness of studies but may also result in site differences in outcome. Consequently, the design step of MRCTs is important because lack of attention to study procedures across sites may affect the ultimate interpretability and significance of the findings. In MRCTs it is important to ensure that study procedures, i.e. factors under the control of investigators, do not result in site differences so that any site differences can be investigated in relation to participant characteristics. In TORDIA, clinical characteristics contributed to site differences much more substantially than any inconsistencies in treatment protocol delivery across sites. This finding helps allay concerns regarding the adequacy of the implementation of the clinical trial. Our findings regarding site differences also have implications for the dissemination of evidence-based treatments. If each site in an MRCT is considered its own small study, then the variability across sites in response outcomes may reflect the range of treatment response that might be expected if this treatment were to be disseminated to the community.¹

MRCTs are rarely, if ever, powered a priori to account for site differences in outcomes. Noda et al. (2006) note that researchers should instead assume that, “power of test and precision of estimates depends not on the absence of site differences, but on the degree” (p. 932). Indeed, even though most studies do not find any, or at most only a few, site-by-treatment interactions in outcomes, the lack of statistical significance is typically due to the fact that the studies are powered to detect a main effect of treatment, not site-by-treatment interactions (Kraemer, 2000). Thus, the lack of statistically significant findings with respect to site differences does not necessarily prove the null hypothesis (Kraemer, 2000). In the future, MRCTs should ideally be powered to detect site differences given how frequently they occur.

Acknowledgments

Funding/Support: This work was supported by NIMH cooperative agreement grants MH62014 (Brown University), MH61835 (University of Pittsburgh); MH61856 (University of Texas, Galveston, TX); MH61864 (UCLA); MH61869 (Kaiser Foundation, Portland); and MH61958 (Southwestern Medical Center, Dallas, TX) and MH066371 (Center for Early Onset Mood and Anxiety Disorders, University of Pittsburgh).

References

- Beck, AT. Beck Hopelessness Scale. The Psychological Corporation; San Antonio, TX: 1988.
- Begg CB, Iglewicz B. A treatment allocation procedure for sequential clinical trials. *Biometrics* 1980;36:81–90. [PubMed: 7370375]
- Borrelli B, Sepinwall D, Ernst D, Bellg AJ, Czajkowski S, Breger R, et al. A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting and Clinical Psychology* 2005;73:852–60. [PubMed: 16287385]
- Brent DA, Emslie G, Clarke G, Wagner KD, Asarnow JR, Keller MK, et al. Switching to another SSRI or to venlafaxine with or without cognitive-behavioral therapy for adolescents with SSRI-resistant depression. The TORDIA randomized controlled trial. *Journal of the American Medical Association* 2008;299:901–13. [PubMed: 18314433]
- Brent DA, Kolko DJ, Birmaher B, Baugher M, Bridge J, Roth C, Holder D. Predictors of treatment efficacy in a clinical trial of three psychosocial treatments for adolescent depression. *Journal of the American Academy of Child and Adolescent Psychiatry* 1998;37:906–14. [PubMed: 9735610]
- Bridge JA, Iyengar S, Salary CB, Barbe R, Birmaher B, Pincus HA, et al. Clinical response and risk for reported suicidal ideation and suicide attempts in pediatric antidepressant treatment: A meta-analysis of randomized controlled trials. *Journal of the American Medical Association* 2007;297:1683–96. [PubMed: 17440145]

- Curry J, Rohde P, Simons A, Silva S, Vitiello B, Kratochvil C, et al. Predictors and moderators of acute outcome in the Treatment for Adolescents with Depression Study (TADS). *Journal of the American Academy of Child and Adolescent Psychiatry* 2006;45:1427–39. [PubMed: 17135988]
- Davidson JRT, Foa EB, Huppert JD, Keefe FJ, Franklin ME, Compton JS, et al. Fluoxetine, comprehensive cognitive behavioral therapy, and placebo in generalized social phobia. *Archives of General Psychiatry* 2004;61:1005–1013. [PubMed: 15466674]
- Guy, W.; Guy, W. ECDEU Assessment Manual of Psychopharmacology. National Institute of Mental Health; Rockville, MD: 1976. Clinical Global Impairment Scale.
- Hill CE, O'Grady KE, Elkin I. Applying the collaborative study psychotherapy rating scale to rate therapist adherence in cognitive-behavior therapy, interpersonal therapy, and clinical management. *Journal of Consulting and Clinical Psychology* 1992;60:73–79. [PubMed: 1556289]
- Hsu L. Random sampling, randomization, and equivalence of contrasted groups in psychotherapy research. *Journal of Consulting and Clinical Psychology* 1989;57:131–137. [PubMed: 2647799]
- Kallert TW, Priebe S, McCabe R, Kiejna A, Rymaszewska J, Nawka P, et al. Are day hospitals effective for acutely ill psychiatric patients? A European multicenter randomized controlled trial. *Journal of Clinical Psychiatry* 2007;68:278–287. [PubMed: 17335327]
- Kaufman J, Birmaher B, Brent D, Ryan N, Flynn C, Moreci P, et al. The Revised Schedule for Affective Disorders and Schizophrenia for School-Aged Children: Present and lifetime version (K-SADS-PL); Preliminary reliability and validity data. *Journal of the American Academy of Child and Adolescent Psychiatry* 1996;36:980–88. [PubMed: 9204677]
- Kendall P, Chu B, Gifford A, Hays C, Nauta M. Breathing life into a manual: Flexibility and creativity with manual-based treatments. *Cognitive & Behavioral Practice* 1998;5:177–198.
- Kraemer HC. To increase power in randomized clinical trials without increasing sample size. *Psychopharmacology Bulletin* 1991;27:217–24. [PubMed: 1775591]
- Kraemer, HC. Evaluating medical tests: Objective and quantitative guidelines. Sage; Newbury Park, CA: 1992.
- Kraemer HC. Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin* 2000;26:533–41. [PubMed: 10993394]
- Kraemer HC, Robinson TN. Are certain multicenter randomized clinical trial structures misleading clinical and policy decisions? *Contemporary Clinical Trials* 2005;26:518–29. [PubMed: 15993653]
- Lachin JM. Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials* 2000;21:167–89. [PubMed: 10822117]
- Lachin JM. The role of measurement reliability in clinical trials. *Clinical Trials* 2004;1:553. [PubMed: 16279296]
- Little, RJ.; Rubin, DB. Statistical analysis with missing data. John Wiley; New York: 1987.
- Luborsky L, Diguier L, Seligman DA, Rosenthal R, Krause ED, Johnson S, et al. The researcher's own therapy allegiances: A “wild card” in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice* 1999;6:95–106.
- Meinert, C. Clinical trials: Design, conduct and analyses. Oxford; New York: 1986.
- Moncher F, Prinz F. Treatment fidelity in outcome studies. *Clinical Psychology Review* 1991;11:247–266.
- Noda A, Kraemer HC, Taylor JL, Schneider B, Ashford JW, Yesavage JA. Strategies to reduce site differences in multisite studies: A case study of Alzheimer disease progression. *American Journal of Geriatric Psychiatry* 2006;14:931–38. [PubMed: 17068315]
- Petry, NM.; Tennen, H.; Affleck, G.; Snyder, CR.; Ingram, RE. Handbook of psychological change: Psychotherapy processes & practices for the 21st century. Wiley; Hoboken, NJ: 2000. Stalking the elusive client variable in psychotherapy research; p. 88-108.
- Poznanski EO, Freeman LN, Mokros HB. Children's Depression Rating Scale – Revised. *Psychopharmacology Bulletin* 1984;21:979–89.
- Robin AL, Weiss JG. Criterion-related validity of behavior and self-report measures of problem solving communication skills in distressed and non distressed parent adolescent dyads. *Behavioral Assessment* 1980;2:339–52.

- Rohde P, Seeley JR, Kaufman NK, Clarke GN, Stice E. Predicting time to recovery among depressed adolescents treated in two psychosocial group interventions. *Journal of Consulting and Clinical Psychology* 2006;74:80–88. [PubMed: 16551145]
- Royston P. Multiple imputation of missing values. *Stata Journal* 2004;4:227–241.
- Schafer, JL. *Analysis of incomplete mutivariate data*. CRC Press; 1997.
- Shadish, W.; Cook, T.; Campbell, D. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin; New York: 2002.
- Tabachnik, B.; Fidell, L. *Using multivariate statistics*. Vol. 5th ed.. Harper Collins; New York: 2006.
- Trepka R, Rees A, Shapiro D, Hardy G, Barkham M. Therapist competence and outcome of cognitive therapy for depression. *Cognitive Therapy and Research* 2004;28:143–157.
- Vallis T, Shaw B, Dobson K. The Cognitive Therapy Scale: Psychometric properties. *Journal of Consulting and Clinical Psychology* 1986;54:381–85. [PubMed: 3722567]

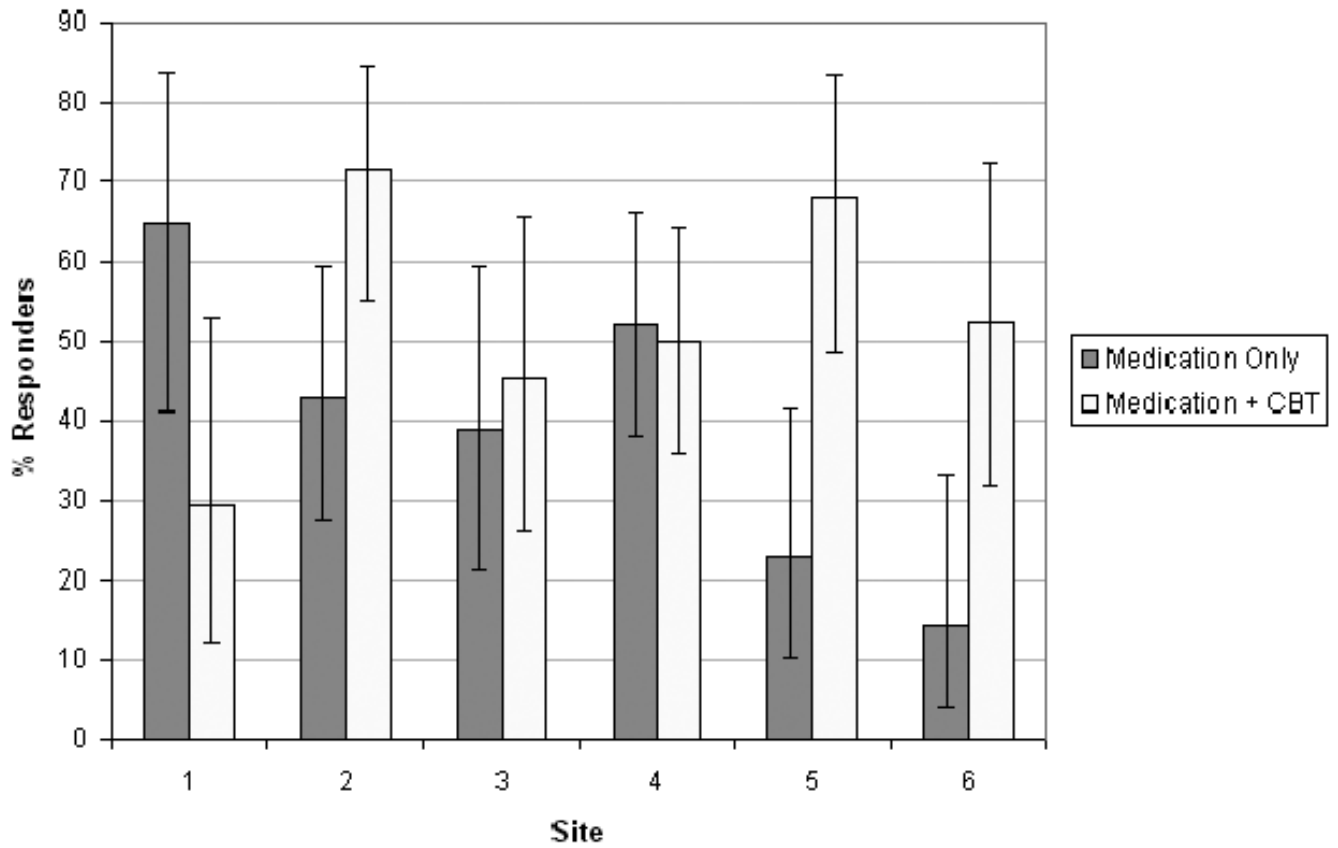


Figure 1.
Clinical Response Broken Down by Site and by Condition, Medication Only vs. Medication Plus CBT.

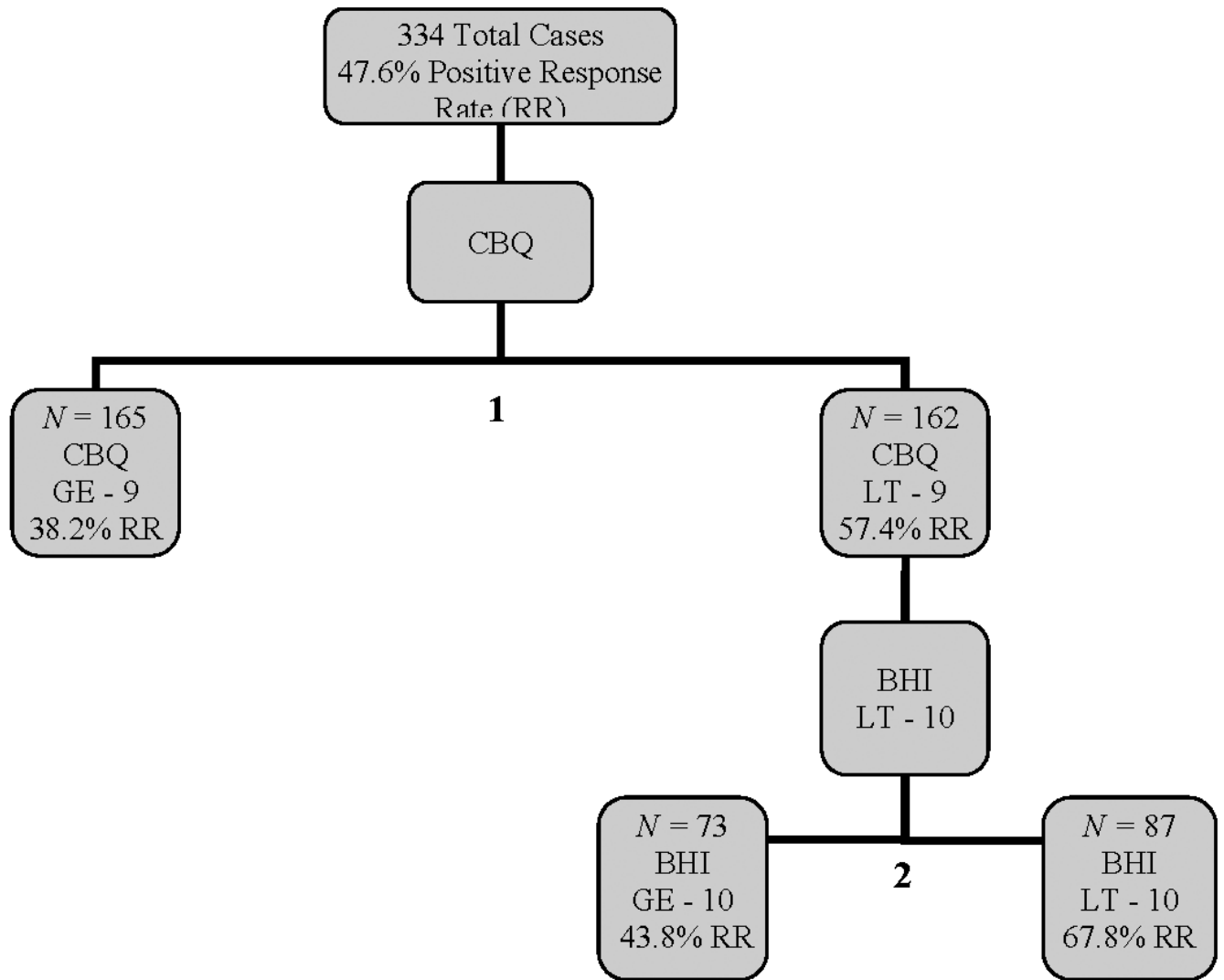


Figure 2. ROC Analyses Examining Increase in Response Rates When Controlling for Baseline Clinical Characteristics of Sample.
Note. BHS=Beck Hopelessness Scale; CBQ=Conflict Behavior Questionnaire. The tree diagram displayed here is a result of the recursive partitioning method described in the Results section.

Table 1

Stratification Variables by Site.

	Site #1 N (%)	Site #2 N (%)	Site #3 N (%)	Site #4 N (%)	Site #5 N (%)	Site #6 N (%)	Total N (%)	χ^2
Chronic depression	22 (64.7)	26 (37.1)	28 (62.2)	67 (72.8)	24 (47.1)	17 (40.5)	184 (55.1)	26.60***
Comorbid anxiety	11 (32.4)	22 (31.4)	7 (15.6)	42 (45.7)	27 (52.9)	13 (31.0)	122 (36.5)	19.37***
BDI suicidality item ≥ 2	8 (23.5)	5 (7.1)	4 (8.9)	8 (8.7)	4 (7.8)	9 (21.4)	38 (11.4)	12.00 ^f

Note. BDI = Beck Depression Inventory.

Chronic depression: Site#1>2,6, 3>6, 4>5,6; Comorbid anxiety: Site#5>2,3,6, 4>3; BDI suicidality: Sites#1,6>2,3,4,5.

^fFisher exact test = .052.

**
p < .05.

p < .0001

Table 2
Differences by Site in Baseline Clinical Characteristics Related to Response Status

	Site #1		Site #2		Site #3		Site #4		Site #5		Site #6		F
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	
Duration of Depression^a	25.2	18.5	14.7	20.1	32.1	28.6	22.4	16.8	24.6	18.4	19.4	16.8	4.35***
Hopelessness													
BHS	11.1	5.8	9.1	5.5	8.2	5.0	11.3	5.8	11.4	5.3	12.3	5.3	4.17**
Family Conflict													
Conflict Behavior Questionnaire (Adolescent)	8.9	6.3	8.4	5.8	7.7	6.5	9.7	6.3	7.7	6.0	10.7	6.2	1.82 ^b

Note. BHS = Beck Hopelessness Scale; Duration of depression: 3>6; Hopelessness: 4,5,6>3; Family conflict: 1,6>2,3,4,5.

^aFrom K-SADS.

^bSite differences on the CBQ were slightly above a trend, $p = .10$, but the CBQ was included because it was related to response rate in the overall study outcomes (Brent et al., 2008).

*** $p < .001$.

*** $p < .0001$