Baseline Depression Severity as Moderator of Depression Outcomes Between Cognitive Behavioral Therapy vs Pharmacotherapy
An Individual Patient Data Meta-analysis

Erica S. Weitz, MA; Steven D. Hollon, PhD; Jos Twisk, PhD; Annemieke van Straten, PhD; Marcus J. H. Huibers, PhD; Daniel David, PhD; Robert J. DeRubeis, PhD; Sona Dimidjian, PhD; Boadie W. Dunlop, MD, MS; Ioana A. Cristea, PhD; Mahbobeh Faramarzi, PhD; Ulrich Hegerl, MD, PhD; Robin B. Jarrett, PhD; Farzan Kheirkhah, MD; Sidney H. Kennedy, MD; Roland Mergl, PhD; Jeanne Miranda, PhD; David C. Mohr, PhD; A. John Rush, MD; Zindel V. Segal, PhD; Juned Siddique, DrPH; Anne D. Simons, PhD; Jeffrey R. Vittengl, PhD; Pim Cuijpers, PhD

IMPORTANCE Current guidelines recommend treating severe depression with pharmacotherapy. Randomized clinical trials as well as traditional meta-analyses have considerable limitations in testing for moderators of treatment outcomes.

OBJECTIVES To conduct a systematic literature search, collect primary data from trials, and analyze baseline depression severity as a moderator of treatment outcomes between cognitive behavioral therapy (CBT) and antidepressant medication (ADM).

DATA SOURCES A total of 14 902 abstracts were examined from a comprehensive literature search in PubMed, PsycINFO, EMBASE, and Cochrane Registry of Controlled Trials from 1966 to January 1, 2014.

STUDY SELECTION Randomized clinical trials in which CBT and ADM were compared in patients with a DSM-defined depressive disorder were included.

DATA EXTRACTION AND SYNTHESIS Study authors were asked to provide primary data from their trial. Primary data from 16 of 24 identified trials (67%), with 1700 outpatients (794 from the CBT condition and 906 from the ADM condition), were included. Missing data were imputed with multiple imputation methods. Mixed-effects models adjusting for study-level differences were used to examine baseline depression severity as a moderator of treatment outcomes.

MAIN OUTCOMES AND MEASURES Seventeen-item Hamilton Rating Scale for Depression (HAM-D) and Beck Depression Inventory (BDI).

RESULTS There was a main effect of ADM over CBT on the HAM-D ($\beta = -0.88; P = .03$) and a nonsignificant trend on the BDI ($\beta = -1.14; P = .08$, statistical test for trend), but no significant differences in response (odds ratio [OR], 1.24; $P = .12$) or remission (OR, 1.18; $P = .22$). Mixed-effects models using the HAM-D indicated that baseline depression severity does not moderate reductions in depressive symptoms between CBT and ADM at outcome ($\beta = 0.00; P = .96$). Similar results were seen using the BDI. Baseline depression severity also did not moderate the likelihood of response (OR, 0.99; $P = .77$) or remission (OR, 1.00; $P = .93$) between CBT and ADM.

CONCLUSIONS AND RELEVANCE Baseline depression severity did not moderate differences between CBT and ADM on the HAM-D or BDI or in response or remission. This finding cannot be extrapolated to other psychotherapies, to individual ADMs, or to inpatients. However, it offers new and substantial evidence that is of relevance to researchers, physicians and therapists, and patients.
there is no shortage of effective treatments for depression, including pharmacotherapy and psychotherapies, of which cognitive behavioral therapy (CBT) is one of the best documented. Previous meta-analyses have shown that psychotherapies are at least as effective as pharmacotherapy, used as monotherapy, in treating depression of mild and moderate symptom severity (defined by cutoff scores on depressive symptom inventories). However, less is known about the relative efficacy of psychotherapy vs pharmacotherapy in severely depressed populations.

Nonetheless, American Psychiatric Association guidelines for the treatment of depression suggest that although psychotherapy is sufficient for treating mild depression, antidepressant medications (ADMs) should be used to treat severe depression in the context of major depressive disorder. This recommendation is mainly owing to the well-known findings of the National Institute of Mental Health Treatment of Depression Collaborative Research Program, in which CBT was less effective than medications in the treatment of participants with severe depression. However, these differences were not observed in several other randomized controlled trials (RCTs) of acute-phase treatment. One limitation of RCTs is that they often include too few patients and thus lack sufficient power to detect moderation of outcomes and thoroughly examine the efficacy of these treatments in severe depression.

Therefore, several meta-analyses aggregated the results of studies examining the effects of psychotherapy and pharmacotherapy on severe depression relative to control conditions. Two rigorous meta-analyses showed that psychotherapy and pharmacotherapy were more efficacious than control treatments for severe depression. However, the findings for psychotherapy should be interpreted with caution because they were based on study-level as opposed to patient-level pretreatment severity and smaller subgroup analysis. In addition, both meta-analyses provided information only on the effectiveness of a single treatment modality and did not address the crucial issue of the relative efficacy of psychotherapy and pharmacotherapy. A conventional meta-analysis that directly compared psychotherapy and pharmacotherapy in severe depression showed no significant differences between the treatment groups. However, this finding was based on only 4 studies in the sample that reported baseline depression severity. Another meta-analysis of CBT vs pharmacotherapy using individual patient data provided substantial information, but only 4 studies were included in the analysis, thus limiting power and representativeness.

Although traditional meta-analyses are useful in aggregating evidence, they are limited in their ability to test for moderation of outcomes. In 2 of 7 of the aforementioned meta-analyses, the authors used the mean pretreatment depression scores of the full sample of the studies as an indication of severity. However, many studies have mean depression baseline scores in the moderate range even if the sample includes patients with severe depression, thus restricting the range of severity examined. The analysis is then limited to studies that include secondary comparisons of the severe sample or to studies that recruited a highly severe sample, both of which are rare.

Many of these concerns regarding conventional meta-analysis and RCTs can be addressed by using individual patient data meta-analysis (IPDMA), which includes raw data from RCTs. The use of IPDMA is a new technique in the mental health field, but it has been used successfully to examine acute and preventive treatments in medicine. Although conventional meta-analyses are appropriate for pooling outcomes, the large sample size of IPDMAs provides more power to accurately examine moderators of treatment outcomes.

Therefore, we conducted an IPDMA to provide the best estimate of the efficacy of psychotherapy relative to pharmacotherapy for the treatment of severe depression. Cognitive behavioral therapy was chosen as the comparison for specificity because it is well researched and widely available and allows for better translation to routine practice.

Methods

Identification and Selection of Studies

Study searches were conducted using several methods. First, we used a database of studies of RCTs on the psychological treatment of adult depression. This database has been described elsewhere and has been used in a series of earlier published meta-analyses (http://www.evidencebasedpsychotherapies.org). The database was developed by comprehensive literature searches (from 1966 to January 1, 2014). In these searches, 14,902 abstracts were examined from PubMed (n = 3864), PsycINFO (n = 2960), EMBASE (n = 4320), and Cochrane Registry of Controlled Trials (n = 3758).

Studies examined were on the psychological treatment of depression in general. Earlier meta-analyses were searched for confirmation that no RCTs were previously missed. From 14,902 abstracts (10,992 after the removal of duplicates), we retrieved 1613 full-text articles for possible inclusion in the database.

We included RCTs in which CBT was compared with pharmacotherapy among patients with a primary diagnosis of a depressive disorder established by a standardized diagnostic interview. No language restrictions were applied. Only studies in which the patients met diagnostic criteria for depressive disorder (major depressive disorder or dysthymia) were included. Studies used DSM-II, DSM-III, or DSM-IV diagnosis of depressive disorder. Cognitive behavioral therapies were required to be manualized and use cognitive restructuring as the main component of treatment.

Studies were excluded if they were aimed at relapse prevention or maintenance treatments or if they included adolescents or children younger than 18 years. Studies of inpatient populations were excluded because patients per definition receive more care than CBT or ADM alone and because patient characteristics likely differ between these 2 populations. Therefore, these studies were excluded to prevent high heterogeneity. Studies that included populations with comorbid general medical disorders were not excluded.

Authors of identified studies were invited via email to participate in the IPDMAs and provide original data from their trial. If the authors did not respond to the request after 1 month, a second reminder email was sent and efforts to contact coauthors were made. If no response was received, we considered the data unavailable and did not include the study in the analysis.
Quality Assessment and Data Extraction
Study quality was assessed using 4 criteria from the Cochrane Collaboration's tool for evaluating the risk of bias. This tool assesses whether there was adequate generation of randomization sequence, concealment of treatment allocation, masking of assessors, and appropriate methods for addressing missing data, which was denoted as positive when the analysis was completed in the intent-to-treat (ITT) sample, meaning that all randomized patients were included in the analysis. Only data in the published articles were used to determine the risk of bias. Two independent researchers (P.C. and an outside assessor) conducted quality assessments.

Meta-analysis
We performed a meta-analysis to examine differences between the 16 studies that provided data and the 8 studies that did not. First, we calculated the effect sizes indicating the difference between CBT and ADM at posttreatment based on data reported in the published articles. The effect sizes were calculated by subtracting the average score of the CBT group from the average score of the ADM group at posttest and then dividing the results by the pooled SD. If studies included only dichotomous outcomes without reporting means and SDs, we used the effect-size calculations outlined by Hedges and Olkin.23 Comparisons between CBT and ADM outcomes in studies that provided data vs those that did not were performed as a meta-regression analysis (Stata, version 13.1; StataCorp LP). The effect size was the dependent variable, and a predictor variable indicating whether we received the data was included. Patient and study characteristics were entered as covariates. In addition, the effects of publication bias on the included studies were inspected by examining a funnel plot produced by the trim and fill procedure by Duval and Tweedie24 and by conducting a test by Egger et al25 of the intercepts in a software program (Comprehensive Meta-Analysis, version 2.2.021; http://www.meta-analysis.com).

Individual Patient Data Analysis
Continuous scores on the HAM-D and BDI at baseline were used to determine baseline severity.26-28 Fourteen of 16 studies contributed HAM-D scores (when studies provided HAM-D-21 scores; HAM-D-17 scores were calculated from individual items and used in all analyses), 4 studies contributed BDI-I scores, and 9 studies contributed BDI-II scores. Two studies were unable to contribute complete BDI scores and were removed from the BDI analysis. Ten of 16 studies contributed both HAM-D and BDI scores. The BDI-I scores were converted to BDI-II scores according to the measure's manual,28 and the aggregated BDI-I and BDI-II scores are referred to below as the BDI. Full-sample analyses were based on the ITT sample, including all randomized patients, except for 3 trials that used a modified ITT or completer sample as cited in the published trials.29,30 The details of the IPDMA are discussed further in the eMethods and eResults in the Supplement.

The studies included in our main analysis are listed in the eTable in the Supplement. We obtained data on randomized patients from 16 studies, combined the trials into 1 data set, and then imputed missing outcome data under a missing-at-random assumption, with missing data imputed using a software program (mi impute mvn in Stata, version 13.1; StataCorp LP). Using multiple imputation with a missing-at-random assumption tends to yield more unbiased results than using complete samples or mean imputation.31 Overall, 30% of HAM-D posttreatment data and 18% of BDI posttreatment data were missing. Participants’ missing outcome data were imputed 100 times using complete patient and study characteristics, such as baseline depression score, sex, length of treatment, and treatment group, as the predictor variables.2 As a robustness check, we conducted analyses only among studies with complete data. For patient-level data, we analyzed the effects of depression severity on treatment outcomes using a 1-step IPDMA approach that allows for the most sophisticated modeling of covariates.24 It has better performance in terms of power and yields less biased estimates compared with 2-step IPDMAs, in which individual patient data are used to estimate the treatment × moderator interaction within each trial, followed by a standard inverse variance meta-analysis.32-34

We used multilevel linear and logistic regression and clustered on the study level to control for unobserved heterogeneity between studies. We used the default maximum likelihood algorithm in the software program (Stata, version 13.1; StataCorp LP). A 2-level multilevel linear regression with patient-level data as level 1 and with study-level data as level 2 was used in all further analyses. The primary analysis concerned whether baseline depression severity was a moderator between CBT and ADM on depression outcomes. However, we first analyzed the effects of treatment group on depression outcomes while holding baseline severity constant. Posttreatment scores on the HAM-D or BDI were used as the outcome variable, and baseline depression score and treatment group were independent variables. To examine whether baseline depression severity was a moderator between CBT and ADM on depression outcomes, we added the interaction between baseline severity and treatment outcomes into the multilevel linear regression model. To examine the effects of patient and study variables on outcomes, we ran an adjusted model controlling for length of treatment, type of medication, demographic variables (age, sex, and marital status), and the risk of bias (sequence generation, allocation of concealment, masking, and ITT analysis). Finally, we ran the same analysis with only study completers. To examine clinically relevant outcomes, we ran the same models using response (50% reduction in scores on posttest HAM-D) and remission (score of ≤7 on posttest HAM-D) as outcomes.35 The definition of remission did not include the duration in remission, which was not reported uniformly across studies.

To test the robustness of these findings, we examined a subset of the sample that met criteria for severe depression according to a less restrictive HAM-D standard of severity (HAM-D-17 score >19),26 stricter criteria of the UK National Institute for Clinical Excellence (HAM-D-17 score >23),36 and BDI-II cutoffs (score >28).26 We ran multilevel linear regression models with posttreatment depression score as the dependent variable and intervention group as the independent variable.
In addition, several sensitivity analyses examined the effects of certain subgroups of studies on the results. This included a subset in which trials were removed that included special populations (patients with multiple sclerosis or peripartum depression), placebo-controlled trials, and lower-quality scores to examine whether the inclusion of these studies affected the results.

Results

Study Selection

Figure 1 shows the inclusion process. We originally retrieved 1613 full-text articles but excluded 1589 for various reasons. Twenty-four studies met the inclusion criteria for the IPDMA. Authors from 16 of 24 identified trials (67%), with 1700 outpatients, agreed to participate and provided data from their original study. Among 8 authors who did not provide data, 4 indicated that they no longer had access to data and 4 were unreachable.

Characteristics of Included Studies

In 6 studies, patients were exclusively recruited from clinical samples, 6 studies recruited patients (in part) through the community, and 4 studies used other recruitment methods. Thirteen studies recruited adults in general while 3 studies recruited specific populations (patients with multiple sclerosis, women who earned a low income, or women with infertility). Eleven studies were conducted in the United States, 2 in Canada, and 1 each from Germany, Romania, and Iran. In 9 studies, a selective serotonin reuptake inhibitor was used for pharmacotherapy, 4 studies prescribed a tricyclic antidepressant, and 3 studies used another antidepressant or a pre-defined protocol for deciding which medication to prescribe. One study allowed for augmentation with lithium or desipramine hydrochloride and 2 studies allowed a medication switch if patients experienced adverse effects. In 14 studies, CBT was given individually, 2 studies used group sessions, and 1 study used both methods. The number of CBT sessions ranged from 8 to 28 (mean, 15.4; mode, 20), and 11 studies used 16 to 20 sessions. Eleven trials reported measuring CBT adherence or competence by rating taped sessions, and 10 trials reported that therapists received regular supervision.

The quality of the included studies based on the published reports varied (eTable in the Supplement). Seven studies reported adequate sequence generation and 6 studies reported allocation to conditions by an independent party. Eleven studies reported masking of outcome assessors, and ITT analyses were conducted in 12 studies. Five studies met all 4 quality criteria, 5 studies met 2 or 3 criteria, and the remaining 5 studies had lower-quality scores (0 or 1 of 4 criteria).

Traditional Meta-analysis Findings

Based on the results as published in all 24 articles, the difference in standardized depression scores at posttreatment between CBT and ADM was Hedges $g = -0.01$ (95% CI, -0.14 to 0.12), with low heterogeneity ($I^2 = 43\%$; 95% CI, 7%-64%). There was no significant difference ($P = .54$) in the effect size between the 16 studies that provided data for our IPDMA analyses ($g = 0.01$; 95% CI, -14.00 to 0.17 and $I^2 = 46\%$; 95% CI, 4%-70%) and the 8 studies that did not ($g = -0.08$; 95% CI, -0.33 to 0.17 and $I^2 = 40\%$; 95% CI, 0%-73%). In addition, there were no indications of publication bias in the 24 studies according to the trim and fill procedure by Duval and Tweedie (adjusted $g = -0.01$; 95% CI, -0.14 to 0.12) or the test by Egger ($P = .29$) (Figure 2).

The multivariable meta-regression analysis with the effect size as the dependent variable (Hedges $g$) and a dummy variable indicating whether a study was included in the IPDMA as the independent variable was not significant ($P = .88$) when controlling for differences in study quality and design (method of recruitment, type of medication [selective serotonin reuptake inhibitor, tricyclic antidepressant, or other]), treatment format of CBT (group or individual), and the number of intervention sessions. Again, this illustrated no differences in outcomes between studies that provided data and studies that did not.

Patient Characteristics

The sample included 1700 participants, 906 from the ADM condition and 794 from the CBT condition. Of the HAM-D out-
come sample, 793 participants (54%) met criteria for severe depression using the more lenient HAM-D criterion, and 255 participants (17%) met the more stringent National Institute for Clinical Excellence criterion. On the BDI, 509 participants (49%) met criteria for severe depression. The mean baseline scores were 19.18 on the HAM-D and 30.86 on the BDI. The mean age of the full sample was 37.38 years, 69% were female, 43% were married, and 52% were employed full-time. In total, 90% of our sample had a high school education (or 12 years of education), and 65% had a higher educational level.

**IPDMA Findings**

Table 1 lists the mean scores at posttreatment categorized by baseline depression severity. There was a significant main effect of ADM over CBT on the HAM-D (β = −0.88; P = .03) and a nonsignificant trend on the BDI (β = −1.14; P = .08, statistical test for trend), but no significant differences between ADM and CBT on clinically relevant outcomes of response (odds ratio [OR], 1.24; P = .12) or remission (OR, 1.18; P = .22). In total, 63% of patients in the ADM condition and 58% of patients in the CBT condition responded to treatment, and 51% of patients in the ADM condition and 47% of patients in the CBT condition met criteria for remission.

Table 2 summarizes the results of the primary analysis examining whether baseline depression severity is a moderator between treatments. When including the interaction effect in the model, the treatment effect (ADM vs CBT) did not differ as a function of severity (β = 0.00; P = .96 for interaction effect).
In this IPDMA, we found no evidence that baseline severity of depression, whether patient or clinician rated, moderated the effect of treatment on outcomes. That is, patients with more severe depression were no more likely to require medications to improve than patients with less severe depression, and these findings were robust in sensitivity analyses. There was a modest (β = −0.73; P = .48), and the BDI (β = −1.53; P = .14).

Sensitivity analyses found no evidence of moderation of outcomes as a function of baseline depression severity in several models. These findings are reported in the Supplement.

**Discussion**

In this IPDMA, we found no evidence that baseline severity of depression, whether patient or clinician rated, moderated the effect of treatment on outcomes. That is, patients with more severe depression were no more likely to require medications than patients with less severe depression, and these findings were robust in sensitivity analyses. There was a modest effect (<1 point on the HAM-D) main effect of ADM over CBT on the continuous outcomes (HAM-D and BDI) but no evidence of any interaction, which provides new and important information for the debate about treatments for severe depression. Although guidelines, suggest that patients with severe depression require pharmacotherapy, we found no evidence that differences between ADM and CBT are moderated by baseline depression severity. Furthermore, robustness analysis on the severe sample alone showed no differential treatment response between CBT and ADM. Therefore, CBT may also be an effective first-line treatment for these patients.

However, there are some limitations to consider. The BDI and HAM-D outcome measures have been criticized. The BDI emphasizes cognitive aspects of depression and as a self-report measure may be prone to bias while the HAM-D contains some psychometric flaws and emphasizes anxiety and somatic symptoms. Moreover, neither specifically addresses functional impairment. Nonetheless, these 2 depression measures are widely used in research and clinical practice. As such, they provide an understanding of treatment outcomes for depressive symptoms.

In addition, not all studies identified as meeting the inclusion criteria for the meta-analysis contributed data. Although we tested for and did not detect bias, it is possible that the included studies were not completely representative. Some studies had quality scores that were suboptimal. Determining quality from the published articles allowed for a consistent and conservative study approach. However, quality may be higher than reported. Sensitivity analyses were performed by removing lower-quality studies, which did not affect our findings.
Samples of RCTs for depression may also not be representative of patients with depression treated in primary and psychiatric care clinics, which may be because of patients’ willingness to accept randomization, because of their previous treatment experiences, or because study criteria may exclude patients with certain comorbid disorders. In addition, the studies included did not incorporate inpatient populations; therefore, these findings cannot be extrapolated to patients having severe depression with imminent suicidality or psychosis. Outcomes comparing CBT and ADM could have varied depending on the expertise and supervision of the therapists and psychiatrists and the adherence to treatment regimens; however, it was not possible to examine the contribution of the quality of treatment in this analysis.

Furthermore, these findings might not generalize to other psychotherapies or ADMs that were not represented in the included studies. They also may not pertain to combination treatments and may reflect data only from studies of acute outcomes. Prior exposure to CBT has been found to reduce rates of relapse relative to prior exposure to medication after treatment termination. It would be important to determine whether that finding holds across the full range of initial depression severity.

Conclusions

While there are some study limitations, the defining strength of our meta-analysis is that it is the first investigation to date, to our knowledge, with sufficient power to examine baseline depression severity as a moderator of treatment outcomes between 2 active treatments. We found no evidence of any such interaction. While this IPDMA shows that pharmacotherapy provides minor improvement in the treatment of depression relative to CBT in terms of the continuous measures, there is no indication that differences between the modalities were moderated by the degree of baseline depression severity. Therefore, the data are insufficient to recommend ADM over CBT in outpatients based on baseline severity alone. More research is needed to examine whether other demographic and clinical characteristics moderate the differential response between CBT and ADM.
Baseline Depression Severity and CBT vs Pharmacotherapy Outcomes

Original Investigation Research


32. Stewart LA, Parmar MK. Meta-analysis of the literature of or of individual patient data: is there a difference? Lancet. 1993;341(8842):418-422.


