PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Does cognitive behaviour therapy have an enduring effect that is superior to keeping patients on continuation pharmacotherapy? A meta-analysis
AUTHORS	Cuijpers, Pim; Hollon, Steven; van Straten, Annemieke; Bockting, Claudi; Berking, Matthias; Andersson, Gerhard

VERSION 1 - REVIEW

REVIEWER	Charles F. Reynolds III, M.D. UPMC Endowed Professor in Geriatric Psychiatry University of Pittsburgh School of Medicine I have published with the senior author, Professor Pim Cuijpers, in the past and am currently collaborating with his group on a new manuscript.
REVIEW RETURNED	20-Jan-2013

THE STUDY	Conformity with criteria of the PRISMA checklist is clearly delineated.
GENERAL COMMENTS	The question addressed by this meta-analysis is of first rank importance to patients with major depression and to the clinicians who care for them. This is so because depression is, more often than not, a relapsing, recurrent illness which, with each episode, risks chronicity and increasing treatment resistance. Thus, an objective of treatment must be to maintain wellness, as well as to achieve response and remission in the first place. The evidence reviewed in this paper is directly relevant to clinical decision making in the pursuit of sustaining remission and promoting recovery. As I understand the data presented here, there was no statistically or clinically significant difference between risk for relapse during the year after acute treatment between CBT (with boosters, in some studies) and continued receipt of antidepressant pharmacotherapy. In my opinion, the authors place too much emphasis on the non-significant "trend" observed for this comparison. When, however, prior CBT was compared with medication discontinuation during one year follow up (eight studies), patients who received prior CBT were less likely to relapse than patients who were withdrawn from medication treatment. This observation was statistically reliable and clinically meaningful. Several points concerning the studies used were not clear but have clinical and scientific relevance to the interpretation of the results: (1) did study participants receive further treatment outside the the study protocol during the one-year follow up? (2) how many CBT boosters did participants receive? (3) was the dose of pharmacotherapy administered during follow up the same as used during acute

	treatment to achieve initial response/remission? (4) how was relapse determined and defined? It may not be possible to answer these questions based upon information contained in the studies used; however, these are important methodological issues in continuation/maintenance treatment research that deserve mention in the Discussion. In my opinion, the authors overstate their recommendation that CBT be the treatment of choice for most depressed patients (lines 28-29, page 11). As they themselves appropriately emphasize, the number of good studies addressing the question posed is still small and the finding should be interpreted with caution pending replication

REVIEWER	Peter Bower, Centre for Primary Care, University of Manchester
REVIEW RETURNED	23-Jan-2013

THE STUDY	This is a meta analysis which derives from analysis of a large database of depression studies which has been developed by the authors and used for a large number of analyses relating to a number of important questions in psychological therapy for depression The research question being considered here is likely to be of interest to clinical readers and provides an interesting perspective on the comparative long term benefits of CBT and medication. This is not an area of particular expertise for me, but it seems as if the analysis is original (which is surprising in itself). As noted, there have been a large number of analyses related to this database, and the overall methods are well thought through and rigorous. The main issue where I think the paper could be improved is in terms of the presentation of the different designs which have been used in the included studies, which I found a little opaque, both in the text and the tables. The variability in the included study designs almost asks for 9 separate CONSORTS to clarify: but more practically I wonder if a separate DESIGN table is warranted to show this more clearly – it would also help to clarify suggestions about optimal designs for later trials, which would add impact to the paper. Those studies that only include responders to the acute phase are presumably at high risk of bias in a way that is somewhat specific to the analyses under test? I wonder if more needed to be made of that design type. The reference to quality being ‘relatively high’ does beg the question ‘relative to what’? Is this in relation to other analyses conducted from this database? This really would benefit from a referent I don’t like the label ‘prior CBT’ – I think ‘acute phase CBT’ is both clearer and more clinically meaningful and would suggest using that unless there are strong arguments to the contrary. The justification of the inclusion of studies with 5 booster sessions (as long as these were not ‘regularly planned’) will need better justification. The results of the analyses are likely to generate

	discussion in what is a fairly contentious area, and any indication that the analyses might benefit one therapy or another are likely to come under scrutiny. I wasn't sure of the clinical justification of this and it would be helpful to have this discussed. Similarly, the choice of the 4 specific Cochrane Risk of Bias criteria might be justified (although I have no personal problem with the ones they have chosen). The meta analytic methods chosen seem conventional, and include a range of tests of the analytic assumptions.
RESULTS & CONCLUSIONS	I struggled with the discussion on page 11, on non specific effects, and was not sure how this related to the broader point about CBT being the treatment of choice. This could be usefully reworked to clarify the points made. Similarly the points made about the effects of CBT being mediated through changes in cognitive and behavioural parameters are interesting, but given the interim nature of the core results presented here, and the complexity of such mediation analyses in the presence of confounding, the final comment might be softened somewhat to indicate these complexities. I wonder if readers (and prospective funders) might like to see a clear statement from the authors as to the type of design that they think optimal to provide a definitive answer to this question (especially given the range of designs that they found in the literature). A short section here might add further to impact The total sample size in terms of studies and total numbers of patients is pretty small given that these are two of the most used treatments in a disease of very high prevalence and impact. This might deserve more comment in the discussion.

REVIEWER	Peter Watson Statistician MRC Cognition and Brain Sciences Unit I have no competing interests with the research in this paper.
REVIEW RETURNED	31-Jan-2013

THE STUDY	Not sure which pooled estimate you used assuming between study heterogeneity (see later comments). I was also not sure what 'well' (page 6, third paragraph) corresponds to with reference to the Hamilton scale. The other thing was the Peters et al (2010) mention of inflated type I errors when using the Eggers test (pages 7 and 9) used in this study which could be added to the references.
RESULTS & CONCLUSIONS	I would like to see a little more reassurance about the low numbers of studies used and particularly the apparently high proportions of people (Goven in the Table 1 N's columns on page 17) coming from the David and Hollon studies (Figure 2 continued pooled odds ratio) and Hollon and Shea studies (Figure 3 discontinued odds ratio) suggesting the pooled odds ratios are heavily reliant on just two studies. What could be mentioned to at least partly to defend the strength of the pooled odds ratio is that there appears agreement in most of the studies in the size of their odds ratio effects in that most are above 1.00.
GENERAL COMMENTS	Does cognitive behaviour therapy have an enduring effect that is

	superior to keeping patients on continuation medication? A meta-analysis. BMJOpen-2012-002542 This study performs a meta-analysis to obtain pooled estimates of odds ratios across studies using random effects estimates (page 6) and forest plots and assesses publication bias using funnel plots and associated methods. The results could be simplified and I suggest a simple strategy for obtaining pooled odds ratios which does not rely on assuming (second paragraph in meta-analysis section on page 6) between study heterogeneity of odds ratios. The forest plots, however, in Figures 2 and 3 (pages 21 and 22) suggest agreement across the comparatively larger studies in their odds ratios although possible weaknesses from the small number of studies comprising the meta-analyses and, in some cases, large proportions (>50%) of the total study sample contained in only a couple of the studies may need to be discussed in the paper to reassure the readers of the value of the pooled estimates. Pages 6 and 7. I find it confusing that you have 'expected' heterogeneity of study odds ratios (second paragraph in Meta-analysis section on page 6) in pooling the odds ratios yet (second paragraph on page 7) tested for this heterogeneity using Cochran's Q and I-squared. I am also not clear why you would compute both Q and I-squared as indicators of across study odds ratio homogeneity when the two are very closely related. I also wonder if you need to quote a 95% CI for the I-squared (second paragraph on page 7). I assume the random effects pooled estimate, which should be mentioned, is the Der-Simonian and Laird pooled estimate? Perhaps a better more logical strategy for obtaining pooled odds ratios would, instead of assuming heterogeneity of study odds ratios, to test for this using I-squared and if the value of I-squared is indicative of moderate or high heterogeneity using the thresholds for I-squared (quoted in the second paragraph on page 7) using the Der-Simonian and Laird pooled odds ratio estimate and, if low or no heterogeneity, using the Mantel-Haenszel estimate for the pooled odds ratio. Page 7. Peters et al (2010) suggest meta-analysis methods for the assessment of publication bias for situations where there is between study heterogeneity. Peters et al (p.578) in particular report that Deeks et al (2005) have found the Egger's test for bias used in this study (fourth paragraph on page 7 and second paragraph on page 9) can give inflated type I error when used to test for publication bias and suggest using more robust tests such as those suggested by Peters et al (2006) and Higgins and Green (2008). Deeks et al also suggested the effective sample size funnel plot should be used if odds ratios are heterogeneous. Page 9. I find the description of Orwin's Fail Safe N results confusing because different values of hypothetical odds ratios are used to describe the robustness of the odds ratios obtained in this study. Could, for example, the number of additional studies needed for the null result (odds ratio of 1) be computed assuming these additional studies had an odds ratio of 1 (the zero effect size). Can one then interpret what a 'large' number of additional studies might be? For example one might, more informatively, relate these numbers of additional studies (needed to give a pooled odds ratio of one using Orwin's Fail Safe N) with the number of studies (5 and 8)

	actually sampled and analysed (as quoted on page 9 and in Figures 2 and 3 on pages 21-22) as an indicator of how robust a result has been obtained to get an idea of what a large number of additional studies might be. It may be, for example, that with only five studies identified as consisting of samples suitable for an analysis a fail safe N of five would be regarded as large as it would equal all the previous studies identified having comparable samples. Pages 21-22. Figures 2-3. A large percentage of the people in the CBT and pharmacotherapy groups used in the meta-analyses to obtain pooled odds ratios come from the David and Hollon (Figure 2) and Hollon and Shea (Figure 3) studies. I think, for example using the Ns in Table 1 (page 17), well over half 70% (=116/162) of the psychotherapy samples in Figure 2 come from just two (David, N=56 and Hollon, N=60) of the five studies. Is this a weakness in the pooling to obtain odds ratios that with so few studies already in the meta-analyses a proportionately large number of the people analysed were in just two of these studies so that the pooled estimates are heavily reliant on just one or two studies? I also wonder on the power of the Q statistic (significance of this is assessed on page 7) to detect between study heterogeneity when there are so few studies. Is the I-squared statistic a robustly accurate estimate of between study heterogeneity when there are so few studies? I know, for example, variances are more accurate estimates the larger the sample size they are based upon. Other comments: Page 6. I assume the positive outcome is a patient responding to treatment and remaining well for the period covered by the study? I am not familiar with the outcome measure. I think the outcome is HAM-D (which I assume is the Hamilton rating scale for depression?) but not sure what 'well' corresponds to on this scale, assuming it is a continuous scale. Are there clinical thresholds? I also wondered why the Hamilton rating scale for measuring depression was chosen in preference to the much used Beck depression inventory. Page 7. Did the authors consider including plots of the funnel plots as figures in addition to the included Figures 2 and 3 (pages 21-22) of forest plots? Page 14. The journal for reference 16 is I assume Biometrics? References Deeks, J.J., Macaskill, P. and Irwig, L. (2005.) The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin. Epidem, 58, 882-893. Higgins, J.P.T. and Green, S. (2008). Cochrane handbook for systematic reviews of interventions, Verson 5.0.1. Cochrane collaboration:Oxford. (Available at http://www.cochrane-handbook.org) Peters, J.L., Sutton, A.J., Jones, D.R., Abrams, K.R. and Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. J. Am. Med. Ass., 295, 676-680.

	Peters, J.L., Sutton, A.J., Jones, D.R. and Abrams K.R. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. Journal of the Royal Statistical Society A, 173(3), 575-591 There is an on-line copy of this paper at http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2009.00629.x/full.

VERSION 1 – AUTHOR RESPONSE

Reviewer: Charles F. Reynolds III, M.D.

I have published with the senior author, Professor Pim Cuijpers, in the past and am currently collaborating with his group on a new manuscript.

Comment: Conformity with criteria of the PRISMA checklist is clearly delineated.
Reply: No revision needed.

Comment: The question addressed by this meta-analysis is of first rank importance to patients with major depression and to the clinicians who care for them. This is so because depression is, more often than not, a relapsing, recurrent illness that, with each episode, risks chronicity and increasing treatment resistance. Thus, an objective of treatment must be to maintain wellness, as well as to achieve response and remission in the first place. The evidence reviewed in this paper is directly relevant to clinical decision making in the pursuit of sustaining remission and promoting recovery.
Reply: Thanks. No revision needed.

Comment: As I understand the data presented here, there was no statistically or clinically significant difference between risk for relapse during the year after acute treatment between CBT (with boosters, in some studies) and continued receipt of antidepressant pharmacotherapy. In my opinion, the authors place too much emphasis on the non-significant "trend" observed for this comparison.
Reply: We have rewritten several parts of the text:
− We have rewritten most of the Discussion section, and have removed the suggestions that acute phase CBT may be superior to maintenance pharmacotherapy.
− We removed that same suggestion from the abstract.
− We have also changed the key messages of the paper completely.
− We have also added a sensitivity analysis in which we removed one potential outlier. After removal of this study, the results became significant. However, we have not referred to this explicitly in the Discussion, Abstract or key messages.

Comment: When, however, prior CBT was compared with medication discontinuation during one year follow up (eight studies), patients who received prior CBT were less likely to relapse than patients who were withdrawn from medication treatment. This observation was statistically reliable and clinically meaningful.
Reply: We have changed the Discussion section in such a way that this is now the main message of the paper.

Comment: Several points concerning the studies used were not clear but have clinical and scientific relevance to the interpretation of the results: (1) did study participants receive further treatment outside the study protocol during the one-year follow up? (2) how many CBT boosters did participants receive? (3) was the dose of pharmacotherapy administered during follow up the same as used

during acute treatment to achieve initial response/remission? (4) how was relapse determined and defined? It may not be possible to answer these questions based upon information contained in the studies used; however, these are important methodological issues in continuation/maintenance treatment research that deserve mention in the Discussion.
Reply: We have described in Table 1 which further treatment the participants receive, the number of booster sessions they received during the follow-up period, and how relapse was defined. We agree that these are important issues, but cannot be accounted for in the meta-analysis. We have added this as a limitation to the Discussion: “But there were important differences between the studies in terms of the treatment received during the continuation phase.”

Comment: In my opinion, the authors overstate their recommendation that CBT be the treatment of choice for most depressed patients (lines 28-29, page 11). As they themselves appropriately emphasize, the number of good studies addressing the question posed is still small and the finding should be interpreted with caution pending replication
Reply: See our reply to Comment 3.

Reviewer: Peter Bower, Centre for Primary Care, University of Manchester

Comment: This is a meta analysis which derives from analysis of a large database of depression studies which has been developed by the authors and used for a large number of analyses relating to a number of important questions in psychological therapy for depression
The research question being considered here is likely to be of interest to clinical readers and provides an interesting perspective on the comparative long term benefits of CBT and medication. This is not an area of particular expertise for me, but it seems as if the analysis is original (which is surprising in itself).
As noted, there have been a large number of analyses related to this database, and the overall methods are well thought through and rigorous.
Reply: Thanks. No revision needed.

Comment: The main issue where I think the paper could be improved is in terms of the presentation of the different designs which have been used in the included studies, which I found a little opaque, both in the text and the tables. The variability in the included study designs almost asks for 9 separate CONSORTS to clarify: but more practically I wonder if a separate DESIGN table is warranted to show this more clearly – it would also help to clarify suggestions about optimal designs for later trials, which would add impact to the paper.
Reply: See our reply to the second comment of the Associate Editor.

Comment: Those studies that only include responders to the acute phase are presumably at high risk of bias in a way that is somewhat specific to the analyses under test? I wonder if more needed to be made of that design type.
Reply: We have added this as one of the limitations of the study in the Discussion section: “Some studies also only included responders to the acute phase in the follow-up analyses, which may have led to bias in the overall results. If high risk patients were more likely to respond to pharmacotherapy than to CBT acute treatment could have acted as a “differential sieve” that systematically unbalanced the groups and led to differential retention being misinterpreted as an enduring effect.”

Comment: The reference to quality being ‘relatively high’ does beg the question ‘relative to what’? Is this in relation to other analyses conducted from this database? This really would benefit from a referent
Reply: We have added that the quality of the studies was relatively high, “compared with studies on psychotherapy for adult depression in general.” We have also added the references to an earlier

meta-analysis in which we examined the association between the quality of the studies in this field and the effects (Cuijpers et al., Psychol Med 2010).

Comment: I don't like the label 'prior CBT' – I think 'acute phase CBT' is both clearer and more clinically meaningful and would suggest using that unless there are strong arguments to the contrary.
Reply: We have replaced "prior CBT" with "acute phase CBT" throughout the paper.

Comment: The justification of the inclusion of studies with 5 booster sessions (as long as these were not 'regularly planned') will need better justification. The results of the analyses are likely to generate discussion in what is a fairly contentious area, and any indication that the analyses might benefit one therapy or another are likely to come under scrutiny. I wasn't sure of the clinical justification of this and it would be helpful to have this discussed.
Reply: We have added a sentence to explain the limit of 5 sessions: "We set the limit at 5 booster session because most psychological treatments have 6 or more treatment sessions.11"

Comment: Similarly, the choice of the 4 specific Cochrane Risk of Bias criteria might be justified (although I have no personal problem with the ones they have chosen).
Reply: We added a justification of why we did not use the other two criteria of the Cochrane risk of bias assessment tool: "The two other criteria of the 'Risk of bias' assessment tool were not used in this study, because we found no clear indication in any of the studies that these had influenced the validity of the study (suggestions of selective outcome reporting; and other problems that could put it at a high risk of bias)."

Comment: The meta analytic methods chosen seem conventional, and include a range of tests of the analytic assumptions.
Reply: No revision needed.

Comment: I struggled with the discussion on page 11, on non specific effects, and was not sure how this related to the broader point about CBT being the treatment of choice. This could be usefully reworked to clarify the points made. Similarly the points made about the effects of CBT being mediated through changes in cognitive and behavioural parameters are interesting, but given the interim nature of the core results presented here, and the complexity of such mediation analyses in the presence of confounding, the final comment might be softened somewhat to indicate these complexities.
Reply: We have removed this paragraph from the Discussion section.

Comment: I wonder if readers (and prospective funders) might like to see a clear statement from the authors as to the type of design that they think optimal to provide a definitive answer to this question (especially given the range of designs that they found in the literature). A short section here might add further to impact
Reply: We have added a paragraph to the Discussion section about this issue: "Studies on the long-term effects of treatments of depression are complicated, because subsequent treatment is difficult to control (but not impossible to influence). Another complication is that differences between treatments are typically small and therefore need large sample sizes. Furthermore, acute and maintenance treatments can be offered in several varieties and can be changed during the maintenance treatment phase. The number of possible comparisons is therefore large, but all are needed to give an adequate answer to the question which treatment is the best at the longer-term. The most important design for a future study, however, would be a sufficiently powered trial comparing acute phase CBT without subsequent continuation versus continuation pharmacotherapy (the current standard of treatment). Although some studies have used this design, none had sufficient power to find significant differences of the magnitude (modest but clinically significant) the kind of modest but clinically relevant) between the two suggested by this meta-analysis. It seems highly relevant to conduct such a trial."

Comment: The total sample size in terms of studies and total numbers of patients is pretty small given that these are two of the most used treatments in a disease of very high prevalence and impact. This might deserve more comment in the discussion.
Reply: We have added this to the Limitations subsection of the Discussion: “As noted above, the most important limitation was that the small number of studies comparing CBT with continued pharmacotherapy. Also the number of patients in these studies was relatively small…”

Reviewer: Peter Watson

I have no competing interests with the research in this paper.

Comment: Not sure which pooled estimate you used assuming between study heterogeneity (see later comments). I was also not sure what 'well' (page 6, third paragraph) corresponds to with reference to the Hamilton scale. The other thing was the Peters et al (2010) mention of inflated type I errors when using the Eggers test (pages 7 and 9) used in this study which could be added to the references.
Reply: − For the reply on the assumption on between-study heterogeneity: see later replies to the comments.
− The exact definition of how “well” was defined in each study, is reported in Table 1 (column “Outcome”). We have added this to the text: “For each study we used the number of patients who responded to treatment and remained well as outcome measure (the exact definition of the outcome in each study is reported in Table 1, column “Outcome”).”
− We have added the reference to Peters et al.

Comment: I would like to see a little more reassurance about the low numbers of studies used and particularly the apparently high proportions of people (Given in the Table 1 N's columns on page 17) coming from the David and Hollon studies (Figure 2 continued pooled odds ratio) and Hollon and Shea studies (Figure 3 discontinued odds ratio) suggesting the pooled odds ratios are heavily reliant on just two studies. What could be mentioned to at least partly to defend the strength of the pooled odds ratio is that there appears agreement in most of the studies in the size of their odds ratio effects in that most are above 1.00.
Reply: We have added a comment on this in the Results section (“As can be seen from Figure 2, the pooled odds ratios are heavily reliant on just two studies, although most of the studies pointed in the same direction. The results should, therefore, be considered with caution”; and later: “Again, these results were heavily reliant on just two studies, and the results should be considered with caution”). We have also added a sentence on this issue to the limitations subsection of the Discussion: “As noted above, the most important limitation was that the small number of studies comparing CBT with continued pharmacotherapy. Also the number of patients in these studies was relatively small, and the results of the main analyses relied heavily on just a few studies. In such a situation…”

Comment: This study performs a meta-analysis to obtain pooled estimates of odds ratios across studies using random effects estimates (page 6) and forest plots and assesses publication bias using funnel plots and associated methods.
The results could be simplified and I suggest a simple strategy for obtaining pooled odds ratios which does not rely on assuming (second paragraph in meta-analysis section on page 6) between study

heterogeneity of odds ratios. The forest plots, however, in Figures 2 and 3 (pages 21 and 22) suggest agreement across the comparatively larger studies in their odds ratios although possible weaknesses from the small number of studies comprising the meta-analyses and, in some cases, large proportions (>50%) of the total study sample contained in only a couple of the studies may need to be discussed in the paper to reassure the readers of the value of the pooled estimates.
Reply: − We assume that the reviewer suggests that we conduct the analyses according to the fixed effects model. We have done the analyses both with the fixed and the random effects model, and because we found that the results were virtually the same, we have reported only the results according to the random effects model.
− We have added the fact that the results relied on only two studies to the results and discussion section (see previous comment).
− The fact that the studies point in the same direction, does not imply that there is no heterogeneity. As Ioannidis in his BMJ 2007 paper shows the confidence intervals around the I-square often very broad, indicating that even when I-square is low, this is very uncertain and there may be considerable heterogeneity. We think therefore, that we cannot leave out the tests of heterogeneity.

Comment: Pages 6 and 7. I find it confusing that you have 'expected' heterogeneity of study odds ratios (second paragraph in Meta-analysis section on page 6) in pooling the odds ratios yet (second paragraph on page 7) tested for this heterogeneity using Cochran's Q and I-squared. I am also not clear why you would compute both Q and I-squared as indicators of across study odds ratio homogeneity when the two are very closely related. I also wonder if you need to quote a 95% CI for the I-squared (second paragraph on page 7). I assume the random effects pooled estimate, which should be mentioned, is the Der-Simonian and Laird pooled estimate?
Reply: We have removing the suggestion that we expected heterogeneity and have indicated here that we have used both the fixed and random effects model to pool the results, but report only the results according to the random effects.
We did remove the computations of Q, as they are indeed not necessary. As indicated in our reply to the previous comment, we do think that we have to calculate the 95% CI around the I-square (Ioannidis, BMJ 2007).
We have pooled the results according to the methods described by Borenstein, Hedges, Higgins, & Rothstein (2009). They do not report whether these are the Der-Simonian and Laird pooled estimates. We have added the reference to Borenstein et al. to the text.

Comment: Perhaps a better more logical strategy for obtaining pooled odds ratios would, instead of assuming heterogeneity of study odds ratios, to test for this using I-squared and if the value of I-squared is indicative of moderate or high heterogeneity using the thresholds for I-squared (quoted in the second paragraph on page 7) using the Der-Simonian and Laird pooled odds ratio estimate and, if low or no heterogeneity, using the Mantel-Haenszel estimate for the pooled odds ratio.
Reply: As indicated earlier, we now report that we have used both methods, and report only the results of the random effects model. We have also indicated already that we do not think that it is appropriate to assume there is no heterogeneity when the I-square is low (as the 95% CI is broad).

Comment: Page 7. Peters et al (2010) suggest meta-analysis methods for the assessment of publication bias for situations where there is between study heterogeneity. Peters et al (p.578) in particular report that Deeks et al (2005) have found the Egger's test for bias used in this study (fourth paragraph on page 7 and second paragraph on page 9) can give inflated type I error when used to test for publication bias and suggest using more robust tests such as those suggested by Peters et al (2006) and Higgins and Green (2008). Deeks et al also suggested the effective sample size funnel plot should be used if odds ratios are heterogeneous.
Reply: We added Begg and Muzambar's test of heterogeneity to the Methods section and have given the results in the results section (because this test is included in the CMA software). Begg and Mazumbar's test was also not significant.

Comment: Page 9. I find the description of Orwin's Fail Safe N results confusing because different values of hypothetical odds ratios are used to describe the robustness of the odds ratios obtained in this study. Could, for example, the number of additional studies needed for the null result (odds ratio of 1) be computed assuming these additional studies had an odds ratio of 1 (the zero effect size).
Reply: It is not possible to set the value of the OR to 1 in Orwin's fail safe N. Setting the value to zero would imply the classic fail safe N analyses. That is also possible, of course, but we choose Orwin's fail safe N to examine how many studies would be needed to find a clinically irrelevant outcome (in stead of an outcome of 1).

Comment: Can one then interpret what a 'large' number of additional studies might be? For example one might, more informatively, relate these numbers of additional studies (needed to give a pooled odds ratio of one using Orwin's Fail Safe N) with the number of studies (5 and 8) actually sampled and analysed (as quoted on page 9 and in Figures 2 and 3 on pages 21-22) as an indicator of how robust a result has been obtained to get an idea of what a large number of additional studies might be. It may be, for example, that with only five studies identified as consisting of samples suitable for an analysis a fail safe N of five would be regarded as large as it would equal all the previous studies identified having comparable samples.
Reply: We are afraid that there is no threshold for what a "large" number of studies is. This is clearly a "clinical" estimate. It would be possible to relate this to the actual number of studies, but this would be in favour of small sets of studies, which would not be reasonable.

Comment: Pages 21-22. Figures 2-3. A large percentage of the people in the CBT and pharmacotherapy groups used in the meta-analyses to obtain pooled odds ratios come from the David and Hollon (Figure 2) and Hollon and Shea (Figure 3) studies. I think, for example using the Ns in Table 1 (page 17), well over half 70% (=116/162) of the psychotherapy samples in Figure 2 come from just two (David, N=56 and Hollon, N=60) of the five studies. Is this a weakness in the pooling to obtain odds ratios that with so few studies already in the meta-analyses a proportionately large number of the people analysed were in just two of these studies so that the pooled estimates are heavily reliant on just one or two studies? I also wonder on the power of the Q statistic (significance of this is assessed on page 7) to detect between study heterogeneity when there are so few studies. Is the I-squared statistic a robustly accurate estimate of between study heterogeneity when there are so few studies? I know, for example, variances are more accurate estimates the larger the sample size they are based upon.
Reply: The fact that the results relied on only a few studies has been answered in previous comments of the reviewer. Regarding the question whether heterogeneity can be assessed with I-square when the number of studies is so small: we think that I-square can assess heterogeneity well, as long as one remembers that the 95% CI is typically broad when the number of studies is small.

Other comments:

Comment: Page 6. I assume the positive outcome is a patient responding to treatment and remaining well for the period covered by the study? I am not familiar with the outcome measure. I think the outcome is HAM-D (which I assume is the Hamilton rating scale for depression?) but not sure what 'well' corresponds to on this scale, assuming it is a continuous scale. Are there clinical thresholds? I also wondered why the Hamilton rating scale for measuring depression was chosen in preference to the much used Beck depression inventory.
Reply: We have added to the Methods section that the exact definitions of what "well" means are reported in Table 1.

Comment: Page 7. Did the authors consider including plots of the funnel plots as figures in addition to the included Figures 2 and 3 (pages 21-22) of forest plots?

Reply: Because we found no indications for publication bias, we think that the added value of these Figures is too small.

Comment: Page 14. The journal for reference 16 is I assume Biometrics?
Reply: Yes, this is Biometrics. We have given the full title now in this reference.

References

Deeks, J.J., Macaskill, P. and Irwig, L. (2005.) The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin. Epidem, 58, 882-893.
Higgins, J.P.T. and Green, S. (2008). Cochrane handbook for systematic reviews of interventions, Verson 5.0.1. Cochrane collaboration:Oxford. (Available at http://www.cochrane-handbook.org)
Peters, J.L., Sutton, A.J., Jones, D.R., Abrams, K.R. and Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. J. Am. Med. Ass., 295, 676-680.
Peters, J.L., Sutton, A.J., Jones, D.R. and Abrams K.R. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. Journal of the Royal Statistical Society A, 173(3), 575-591 There is an on-line copy of this paper at http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2009.00629.x/full.

VERSION 2 – REVIEW

REVIEWER	Peter Bower, Centre for Primary Care, University of Manchester I declare no competing interests
REVIEW RETURNED	17-Mar-2013

- The reviewer completed the checklist but made no further comments.