Evidence-Based Mental Health Online First, published on March 16, 2016 as 10.1136/eb-2016-102341

Are all psychotherapies equally effective in the treatment of adult depression? The lack of statistical power of comparative outcome studies

Pim Cuijpers[1,2]

[1]Department of Clinical, Neuro and Developmental Psychology, VU University Amsterdam, Amsterdam, The Netherlands; [2]EMGO Institute for Health and Care Research, VU University and VU University Medical Center Amsterdam, Amsterdam, The Netherlands
Correspondence to Professor Pim Cuijpers, Department of Clinical Psychology, VU University Amsterdam, Van der Boechorststraat 1, Amsterdam 1081 BT, The Netherlands; p.cuijpers@vu.nl

ABSTRACT

More than 100 comparative outcome trials, directly comparing 2 or more psychotherapies for adult depression, have been published. We first examined whether these comparative trials had sufficient statistical power to detect clinically relevant differences between therapies of d=0.24. In order to detect such an effect size, power calculations showed that a trial would need to include 548 patients. We selected 3 recent meta-analyses of psychotherapies for adult depression (cognitive behaviour therapy (CBT), interpersonal psychotherapy and non-directive counselling) and examined the number of patients included in the trials directly comparing other psychotherapies. The largest trial comparing CBT with another therapy included 178 patients, and had enough power to detect a differential effect size of only d=0.42. None of the trials in the 3 meta-analyses had enough power to detect effect sizes smaller than d=0.34, but some came close to the threshold for detecting a clinically relevant effect size of d=0.24. Meta-analyses may be able to solve the problem of the low power of individual trials. However, many of these studies have considerable risk of bias, and if we only focused on trials with low risk of bias, there would no longer be enough studies to detect clinically relevant effects. We conclude that individual trials are heavily underpowered and do not even come close to having sufficient power for detecting clinically relevant effect sizes. Despite this large number of trials, it is still not clear whether there are clinically relevant differences between these therapies.

INTRODUCTION

Several different types of psychological treatment for adult depression have been examined in dozens of randomised controlled trials and have been found to result in significantly better outcomes than no-treatment control conditions. That is true for cognitive behaviour therapy, interpersonal psychotherapy (IPT) and behavioural activation therapy.[1] One of the problems in this field is that all types of therapy seem to be equally or about equally effective,[1] and there does not seem to be one type of therapy that is significantly more effective than others.[2]

Therefore, it is not surprising that researchers have conducted trials in which different types of therapy were directly compared with each other to examine whether a new type of therapy is more effective than an established therapy, or to examine whether one established therapy might be more effective than another for a specific target population. In a regularly updated database of randomised trials examining the effects of psychotherapies for adult depression,[3] we found more than a hundred such comparative outcome trials, directly comparing two or more therapies for adult depression. For example, between 2006 and 2010, we found 20 such comparative trials, and between 2011 and 2014, we found 34.

Unfortunately, these comparative trials pose a big problem with statistical power bringing about the conclusion that all therapies are equally effective very uncertain. Because the differential effects between psychological treatments are small or non-existent, large sample sizes are needed. What researchers sometimes do is to also include a control arm (a waiting list, care-as-usual or another type), so that the trial has three arms (two psychotherapy and one control condition). This allows them to examine whether either of the active treatments is effective compared with the control group. And this difference with the control group is also what they use as the starting point for their power calculations. But comparison between the two active treatments requires a completely different power calculation, and much larger sample sizes. So, basically, such trials can say very little about the differential effects between treatments, because they do not have the statistical power for that. That is a major problem in these comparative trials, whether or not they have included a third (control) condition.

To illustrate this problem, we took two steps. First, we examined whether individual comparative studies have sufficient statistical power to detect differential effects between psychotherapies. In a second step, we examined whether meta-analyses of comparative outcome studies have sufficient power to detect such differential effects.

THE FIRST STEP: STATISTICAL POWER OF INDIVIDUAL TRIALS

In the first step, aimed at examining whether individual comparative studies have sufficient statistical power, we calculated the number of patients that need to be included in a trial to find a differential effect size (Cohen's d) ranging from d=0.1 to d=0.6. These numbers were calculated using G*Power software,[4] assuming a statistical power of 0.8 and an α level of 5% (based on a two-sided t test indicating the difference between two independent means).[5] The blue line in figure 1 gives these numbers, ranging from 90 patients to find a differential effect size of d=0.6 to 3142 patients to find an effect size of d=0.1 (the number for the effect size of 0.1 is outside the range of the figure).

It is not clear what the threshold for a clinically significant effect size is for treatments in depression, but in an earlier paper, we showed that an effect size of d=0.24 could be considered as a 'minimally important difference' as seen from the patient perspective.[6] We will use this as the threshold for a clinical significant difference between treatments. According to G*Power, a trial showing that two treatments differ with this effect size should include 548 patients (274 patients in each condition).

We subsequently selected three recent meta-analyses of psychological treatments of depression that we had recently conducted and calculated the mean differential effect size of the type of psychotherapy they focused on versus other psychotherapies, the mean number of patients included in these trials and the range of the number of included patients in these trials. We selected these three meta-analyses to examine three from the seven major types of psychotherapy that have been developed for the treatment of adult depression.[2] We selected the meta-analyses that were published in the past 5 years (since 2011)

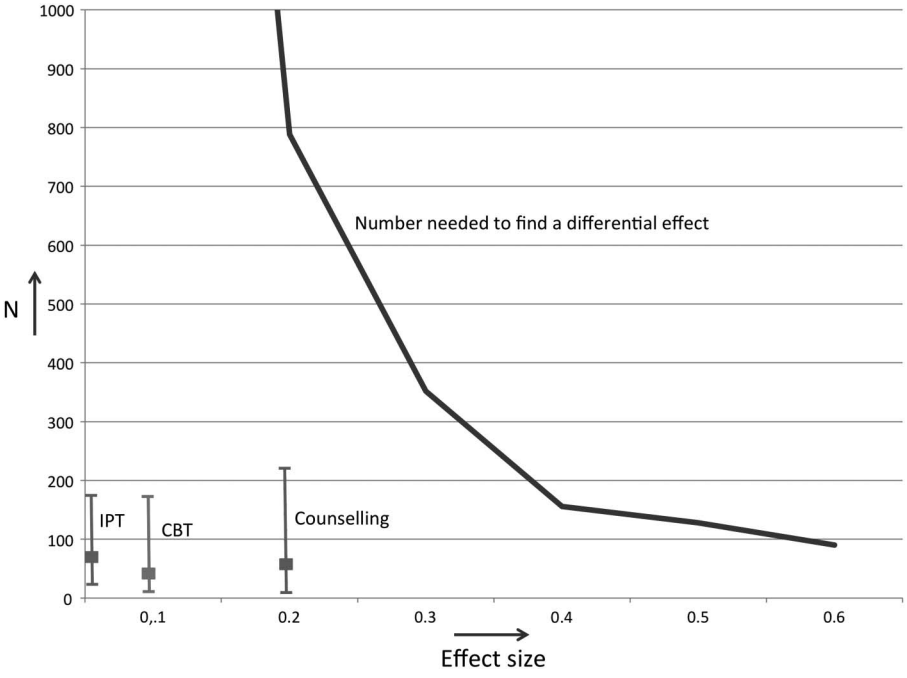

Figure 1 Number of patients needed to detect effect sizes in comparative outcome studies, and actual mean number and range of patients included in comparative trials of three psychotherapies for adult depression (CBT, interpersonal psychotherapy (IPT) and non-directive counselling).

and that used our database of randomised trials, because we had all the details of the included studies in order to conduct the power calculations.

COMPARATIVE OUTCOME TRIALS FOR COGNITIVE BEHAVIOUR THERAPY

We first examined a recent meta-analysis of cognitive behaviour therapy for adult depression.[7] This meta-analysis included 46 comparisons between cognitive behaviour therapy (CBT) and other psychotherapies, with a mean effect size of d=0.1. The mean number of patients included in comparisons between CBT and another psychotherapy was 52, and that number ranged from 13 to 178. We have given the mean and range in figure 1 to contrast them with the numbers needed to find a differential effect size. So, the mean number of included patients in these trials was 52, while in fact a total of 3142 patients would need to be included to find this differential effect size of d=0.1.

Of course, this is the mean effect size and it is theoretically very well possible that a particular study found a much larger differential effect size between CBT and another therapy. Therefore, we also calculated the effect size that can be found with the largest study comparing CBT with another psychotherapy. The study by Dowrick *et al*[8] compared CBT with problem-solving therapy and that comparison included 178 depressed patients. The power calculation showed that this trial had sufficient power to detect an effect size of d=0.42. The effect size that can be detected with the average trial with 52 patients was d=0.79 (as comparison: the effect size comparing CBT with untreated control groups found in this meta-analysis was d=0.71). The smallest trial with 13 patients had only sufficient power to detect an effect size of d=1.71. It is also evident that none of the trials even came close to the number needed to find a clinically relevant differential effect size of d=0.24 (N=548). The largest comparative trial had only 32% of the patients needed to find such a differential effect size.

COMPARATIVE OUTCOME TRIALS FOR NON-DIRECTIVE COUNSELLING AND IPT

We did the same calculations for non-directive counselling for depression based on another meta-analysis with 32 trials comparing counselling to other psychotherapies.[9] The mean effect size was d=−0.20 in favour of the other psychotherapies, the mean number of patients per study was 59 (range 16–221). As can be seen from figure 1, the number of patients per study did not come close to the numbers needed to find such a differential effect size (which is N=788). The largest study has sufficient power to detect a differential effect size of d=0.34.[10]

Finally, we ran these calculations for IPT,[11] using 13 comparisons with other therapies, a differential effect size of d=0.06 and 70 patients on average per trial (range 26–177). The largest trial had sufficient power to find an effect size of d=0.38.[12]

These calculations show that the trials comparing different types of psychotherapy are heavily underpowered. The largest comparative trial we found in three comprehensive meta-analyses of major types of psychotherapy included 221 patients, which is about 40% of the 548 patients needed to detect a clinically relevant effect size of d=0.24. This largest trial had enough power to detect an effect size of d=0.34. As comparison, the effect size for antidepressants versus pill placebo is d=0.31.[13] That means that none of the comparative trials even had sufficient statistical power to detect an effect similar to that of antidepressants.

THE SECOND STEP: CAN META-ANALYSES SOLVE THE PROBLEM OF STATISTICAL POWER?

In the second step, we examined whether meta-analyses of comparative outcome studies can solve the problem of low statistical power to detect differential effects. We conducted power calculations for meta-analyses according to the procedures described by Borenstein *et al*,[14] (conservatively assuming a medium level of between-study variance ($\tau 2$) and a significance level (α) of 0.05). We calculated the number of studies needed for a statistical power of 0.8 and 0.9. The studies in the three meta-analyses included on average 58 patients (29 patients per arm). In figure 2, we described how many studies would be needed (with 58 patients per study) to be able to detect a specific effect size, assuming a power of 0.8 and 0.9.

For CBT (52 patients per study), we would need 18 trials to detect a significant effect of d=0.24 with a power of 0.8, or 24 trials with a power of 0.9. The actual number of trials was 46, so this was enough to detect a clinically relevant effect. However, out of these 46 trials, only 13 had low risk of bias (defined as 3 or more positive scores on 4 items of the Cochrane risk of bias assessment tool).[15] So if we were

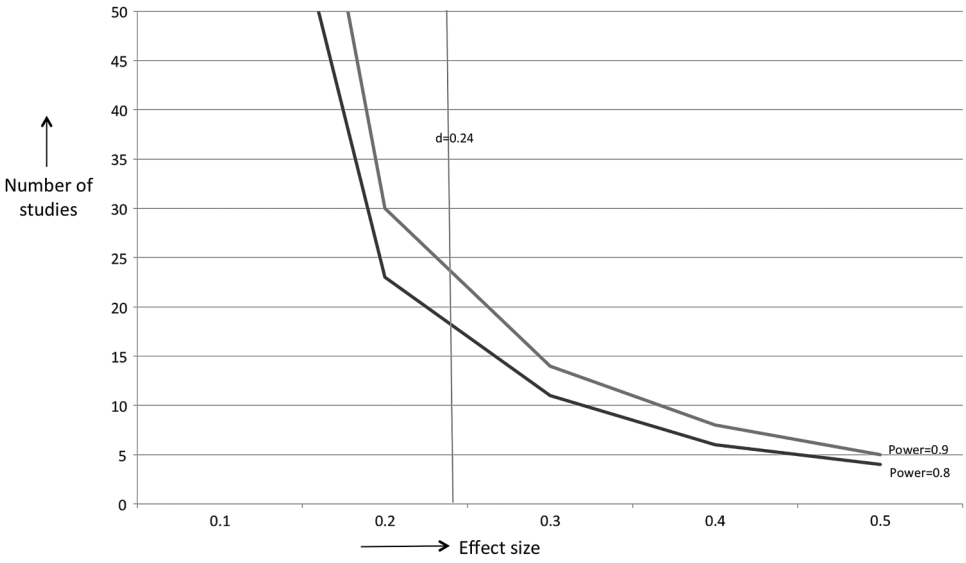

Figure 2 Number of studies needed in meta-analyses to detect differential effect sizes between therapies. IPT, interpersonal psychotherapy.

to focus only on studies with low risk of bias, there would not be enough studies to detect a clinically relevant effect.

The actual difference between CBT and other psychotherapies was d=0.1, which was not significant. In order to detect a significant effect of d=0.1, a total of 100 trials would be needed for a power of 0.8 and 133 trials for a power of 0.9.

For non-directive supportive counselling (59 patients per trial), 16 trials would be needed to detect an effect of d=0.24 with a power of 0.8 or 21 trials with a power of 0.9. The 32 trials comparing counselling with other therapies do, therefore, have enough power to detect a clinically relevant effect. However, only 14 trials had low risk of bias, so these would not be enough to detect such an effect.

For IPT (70 patients per trial), 13 trials are needed for detecting an effect size of d=0.24 for a power of 0.8 or 18 trials for a power of 0.9. The actual number of trials was 13, but only 8 had low risk of bias.

DISCUSSION

Trials comparing different types of psychotherapy for adult depression do not have sufficient power to detect clinically relevant effect sizes. In order to demonstrate a clinically significant effect size of d=0.24, a trial would need to include 548 patients, but the largest comparative trial we found in three major meta-analyses included only 221 patients. This largest trial had only enough power to detect an effect size of d=0.34, and even this trial did not have enough statistical power to detect the mean difference between antidepressant medication and placebo. The implication is that individual trials are heavily underpowered and do not even come close to having sufficient power for detecting clinically relevant effect sizes—let alone smaller effect sizes that may not be clinically significant—but are nevertheless interesting from a scientific point of view.

Meta-analyses may be able to solve this problem. By pooling the effects of multiple studies, clinically relevant effect sizes can be identified. We found that sufficient studies were available for CBT, IPT and non-directive supportive counselling. However, many of these studies have considerable risk of bias and, if we were to focus exclusively on trials with low risk of bias, there would no longer be enough studies to detect clinically relevant effects.

Thus, were we to really take the evidence seriously, we would have to conclude that it is not clear whether CBT, IPT and counselling are as effective as other therapies. Even though more than 100 comparative trials have been conducted, it still remains unclear whether one therapy is more effective than another. The evidence seems to point to no clinically relevant differences, but because of the considerable risk of bias in the majority of trials, the effectiveness of any particular therapy is still uncertain. Smaller differential effect sizes, below the threshold of clinical relevance, cannot be detected at all because many more trials are needed.

Claims that all therapies (or at least all bona fide psychotherapies) are equally effective, including therapies for adult depression, should, therefore, be considered with caution.[16–18] Individual trials do not have enough power, and meta-analyses do not include enough trials with low risk of bias. Smaller differential effect sizes of about d=0.1 or d=0.15 anyway cannot be detected with the current number of trials included in meta-analyses.

Conducting studies without sufficient statistical power to detect a realistic differential effect between treatments is not ethical. Resources as well as time and energy of patients, therapists and researchers go wasted because the trial is not suited to find the expected differences in outcome. Comparative trials of different types of psychotherapy should be expected to result in small differential effect sizes and need to include, therefore, large numbers of patients. Such trials are expensive and pose logistic challenges, but that is what power calculations point out.

The calculations we used in this article have several limitations. The power calculations are based on a t test examining the difference between two groups. Including more measurements at different follow-ups can increase statistical power without increasing the number of participants. We also looked only at a selection of meta-analyses that were focused on depression, so our conclusion may not be generalised to other therapies or disorders.

CONCLUSION

We can conclude that comparative outcome trials in the field of psychotherapy for depression are heavily underpowered and the trials that were carried out do not come close to the statistical power that is needed to examine whether one therapy is more effective than another. Comparative outcome studies that include too few patients to detect significant differences are not ethical and should not be conducted. Meta-analyses may be helpful for finding clinically relevant differences between therapies, but the number of trials with low risk of bias is too small to draw definite conclusions about the comparative effects of psychotherapies.

Although more than 100 trials have compared the outcomes of psychotherapies for adult depression, none of these trials has enough power to detect a clinically relevant difference, and the central research question of these trials (are some therapies more effective than others?), remains unanswered.

Competing interests None declared.

Provenance and peer review Not commissioned; internally peer reviewed.

doi:10.1136/eb-2016-102341

Received 11 February 2016; Revised 17 February 2016; Accepted 26 February 2016

REFERENCES

1. **Barth J**, Munder T, Gerger H, *et al*. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLoS Med* 2013;**10**:e1001454.
2. **Cuijpers P**, van Straten A, Andersson G, *et al*. Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. *J Consult Clin Psychol* 2008;**76**:909–22.
3. **Cuijpers P**, van Straten A, Warmerdam L, *et al*. Psychological treatment of depression: a meta-analytic database of randomized studies. *BMC Psychiatry* 2008;**8**:36.
4. **Faul F**, Erdfelder E, Lang A-G, *et al*. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 2007;**39**:175–91.
5. **Flint J**, Cuijpers P, Horder J, *et al*. Is there an excess of significant findings in published studies of psychotherapy for depression? *Psychol Med* 2015;**45**:439–46.
6. **Cuijpers P**, Turner EH, Koole SL, *et al*. What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depress Anxiety* 2014;**31**:374–8.
7. **Cuijpers P**, Berking M, Andersson G, *et al*. A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Can J Psychiatry Rev Can Psychiatr* 2013;**58**:376–85.
8. **Dowrick C**, Dunn G, Ayuso-Mateos JL, *et al*. Problem solving treatment and group psychoeducation for depression: multicentre randomised controlled trial Outcomes of Depression International Network [ODIN] Group. *BMJ* 2000;**321**:1450–4.
9. **Cuijpers P**, Driessen E, Hollon SD, *et al*. The efficacy of non-directive supportive therapy for adult depression: a meta-analysis. *Clin Psychol Rev* 2012;**32**:280–91.
10. **Areán PA**, Raue P, Mackin RS, *et al*. Problem-solving therapy and supportive therapy in older adults with major depression and executive dysfunction. *Am J Psychiatry* 2010;**167**:1391–8.
11. **Cuijpers P**, Geraedts AS, van Oppen P, *et al*. Interpersonal psychotherapy for depression: a meta-analysis. *Am J Psychiatry* 2011;**168**:581–92.
12. **Luty SE**, Carter JD, McKenzie JM, *et al*. Randomised controlled trial of interpersonal psychotherapy and cognitive-behavioural therapy for depression. *Br J Psychiatry* 2007;**190**:496–502.
13. **Turner EH**, Matthews AM, Linardatos E, *et al*. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;**358**:252–60.
14. **Borenstein M**, Hedges LV, Higgins JPT, *et al*. *Introduction to meta-analysis*. Chichester, UK: Wiley, 2009.
15. **Higgins JPT**, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;**343**:d5928.
16. **Baardseth TP**, Goldberg SB, Pace BT, *et al*. Cognitive-behavioral therapy versus other therapies: redux. *Clin Psychol Rev* 2013;**33**:395–405.
17. **Miller S**, Wampold B, Varhely K. Direct comparisons of treatment modalities for youth disorders: a meta-analysis. *Psychother Res* 2008;**18**:5–14.
18. **Wampold BE**, Minami T, Baskin TW, *et al*. A meta-(re)analysis of the effects of cognitive therapy versus 'other therapies' for depression. *J Affect Disord* 2002;**68**:159–65.