Epidemiology and Psychiatric Sciences (2019), **28**, 21–30. © Cambridge University Press 2018
doi:10.1017/S2045796018000057

SPECIAL ARTICLE

Was Eysenck right after all? A reassessment of the effects of psychotherapy for adult depression

P. Cuijpers[1]*, E. Karyotaki[1], M. Reijnders[1] and D. D. Ebert[2]

[1] *Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, The Netherlands*
[2] *Clinical Psychology and Psychotherapy, Institute for Psychology, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany*

Aims. In the 1950s, Eysenck suggested that psychotherapies may not be effective at all. Twenty-five years later, the first meta-analysis of randomised controlled trials showed that the effects of psychotherapies were considerable and that Eysenck was wrong. However, since that time methods have become available to assess biases in meta-analyses.

Methods. We examined the influence of these biases on the effects of psychotherapies for adult depression, including risk of bias, publication bias and the exclusion of waiting list control groups.

Results. The unadjusted effect size of psychotherapies compared with control groups was $g = 0.70$ (limited to Western countries: $g = 0.63$), which corresponds to a number-needed-to-treat of 4.18. Only 23% of the studies could be considered as a low risk of bias. When adjusting for several sources of bias, the effect size across all types of therapies dropped to $g = 0.31$.

Conclusions. These results suggest that the effects of psychotherapy for depression are small, above the threshold that has been suggested as the minimal important difference in the treatment of depression, and Eysenck was probably wrong. However, this is still not certain because we could not adjust for all types of bias. Unadjusted meta-analyses of psychotherapies overestimate the effects considerably, and for several types of psychotherapy for adult depression, insufficient evidence is available that they are effective because too few low-risk studies were available, including problem-solving therapy, interpersonal psychotherapy and behavioural activation.

Received 11 November 2017; Accepted 20 January 2018; First published online 28 February 2018

Key words: Depression, outcome studies, psychotherapy, randomised controlled trials.

Introduction

It is now 65 years ago that Eysenck wrote an influential paper that shocked the community of psychotherapists. He suggested that psychotherapies are not effective in the treatment of mental disorders (Eysenck, 1952). Based on naturalistic studies and a small sample of outcome studies, Eysenck said that the majority of patients with mental health problems get better anyway, whether or not they are treated with psychotherapy. Since the publication of Eysenck's paper, some smaller reviews of controlled studies wanted to refute Eysenck's conclusion and tried to show that psychotherapy did have an effect of the mental health of patients (Bergin & Lambert, 1971; Luborsky *et al.* 1975). However, these were small review papers, using the voting method (in which the number of studies with positive effects is counted), and the results could not be used as strong evidence for positive effects of psychotherapy and against Eysenck's conclusions.

It took 25 years to counter the findings by Eysenck in a convincing way. In the 1970s, Gene Glass developed the methods of meta-analyses (Glass, 1977), in which the effects of individual studies could be integrated statistically into one overall estimate of the effect size of an intervention compared with a control group. He calculated the effects in terms of standard deviations (effect sizes), which is not depending on the type of outcome measure; and by weighing the studies by sample size (larger studies contribute more to the pooled outcome), they made it possible to integrate all these studies into one overall estimate of the effect size on an intervention.

In 1977, together with Mary Lee Smith, Glass wrote the first modern meta-analysis of psychotherapy outcome studies (Smith & Glass, 1977). They pooled the results of 400 controlled trials of psychotherapy, in about 50 000 patients, and found that the pooled effect size of these studies was $d = 0.68$. This indicates that the patients who received psychotherapy scored 0.68

* Address for correspondence: Pim Cuijpers, Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands. (E-mail: p.cuijpers@vu.nl)

standard deviation better than the patients in the control groups. The average patient receiving therapy was better off than 75% of the untreated controls. This settled the case. Eysenck was wrong. Although Eysenck himself did not consider the meta-analysis to be a credible method, all of health care embraced this new method with tens of thousands of meta-analyses published since then. Meta-analyses are now considered to give the best evidence of the effects of treatments.

But did this early meta-analysis convincingly show that psychotherapies are effective? It was recognised from the beginning that meta-analyses had weaknesses and that we should consider their results cautiously (Hunt, 1997). For example, the 'garbage in, garbage out' principle was recognised early (Hunt, 1997), saying that if the quality of the studies included in the meta-analysis was low, then the meta-analysis itself could never solve that and could not be better than the included studies. 'Apples and oranges' was seen as another problem, recognizing that most trials on psychotherapy differed from each other considerably, in terms of recruitment, patient characteristics, setting, delivered therapy and therapists. Was it really possible to pool the results of such different studies into one overall effect size? Another problem, called the 'file drawer problem' was that not all studies were published; and if especially negative studies were not published, then pooling the results of published studies overestimated the true effect size.

Another problem was that the control conditions in psychotherapy trials were not straightforward (Mohr *et al.* 2009; Gold *et al.* 2017). In medication trials, it was possible to blind patients and not inform them about whether they received the real medication or only a pill placebo. That was not possible in psychotherapy trials, in which patients were randomised to a waiting list control group, or to usual care. Patients knew they were getting the active treatment or not. This may generate expectations and hope for improvement that has nothing to do with the therapy in itself. Waiting list control groups may stimulate patients to do nothing about their problems because they will get a treatment after the waiting period. Recent meta-analyses suggest that waiting lists may be a nocebo, and artificially inflate the effect sizes of therapies (Furukawa *et al.* 2014). Care-as-usual is problematic since this varies considerably across settings and health care systems, making comparisons very heterogeneous.

In the early days of meta-analyses, these problems were known and acknowledged, but the enthusiasm about the new method was too high, and the methods to assess the influence of these problems on the actual effects of psychotherapy were not yet available. So the conclusion that psychotherapy was effective was kept as the main message from this early meta-analytic work. However, since the early days of meta-analyses, these methodological problems have been examined in much more detail; and in more recent times, methods have become available to assess the influence of most of these factors on the estimates of the effects of psychotherapies. This has made it possible to re-assess the estimates of the effects of psychotherapies found in the 1970s and examine how large the effects are in reality, after adjusting for these methodological problems. In this paper, we will try to do exactly that.

Psychotherapies for adult depression

In this paper, we will focus on psychotherapies for adult depression for several reasons. In the early meta-analyses, all trials on psychotherapy were pooled in one big meta-analysis, and that included trials on depression, anxiety, psychotic patients, delinquency and several other mental health problems (Smith & Glass, 1977; Smith *et al.* 1980). Currently, these categories are considered to be too heterogeneous to be pooled, because psychotherapies for each of these target groups are too different from each other, as are the outcome measures and many other characteristics of the participants, interventions and design of the studies. By focusing on one disorder, this major source of heterogeneity can be avoided. It narrows the scope, but this is still broad enough to explore what happens if we adjust for the methodological problems in psychotherapy research.

Another reason to focus on depression is that in the field of psychotherapy research, there is no mental disorder that has been examined as much as depression. Therapies for other mental disorders have also been examined in several dozens of trials, but depression has been examined in more than 250 trials and there is no other disorder that comes even close to this number. Furthermore, the overall (unadjusted) effects of psychotherapies are in the same range of the effects that were found by Smith and Glass (Cuijpers *et al.* 2016*a*).

Depression has also the advantage that several different types of psychotherapy have been examined. In other fields of psychotherapy, research has been heavily dominated by cognitive–behaviour therapy (CBT) (Cuijpers, 2016*a*). In depression research, several other types have been tested, such as interpersonal psychotherapy (Cuijpers *et al.* 2016*b*), counselling (Cuijpers *et al.* 2012) and brief psychodynamic therapies (Driessen *et al.* 2015*a*). This makes it possible to examine the effects of therapy across different types of therapy.

In a previous meta-analysis, we explored the influence of the use of waiting list control groups, risk

of bias and publication bias on the overall effects of CBT for major depressive and anxiety disorders (Cuijpers *et al.* 2016*a*). It was found that the effects of CBT were considerably smaller after adjustment for these biases. This study was, however, based on a relatively small number of trials, was only aimed at CBT, did not include trials using patients with increased levels of depressive symptoms, nor studies with other control groups than waiting lists and care-as-usual. In the current meta-analysis, we wanted to explore the impact of these sources of bias in all controlled trials on psychotherapy that are available.

Methods

We used a database of randomised trials of psychotherapies for adult depression that has been described in a methods paper (Cuijpers *et al.* 2008*a*). The general methods we have used in this paper have been described in a manual that is freely available (Cuijpers, 2016*b*). In brief, the database was developed through a comprehensive literature search (from 1966 until 1 January 2017), and is updated every year. We searched major bibliographical databases (PsycINFO, PubMed, Embase, Cochrane Central Register of Controlled Trials). The full search string for PubMed is given in Appendix A.

In this database, we included all randomised trials in which at least one arm was a psychological treatment for adults (>18 years) with a depressive disorder according to a diagnostic interview or an elevated level of depressive symptomatology (as indicated by a score above a cut-off score on a validated self-report depression scale). In the current study, we only used trials that compared a psychotherapy for adult depression with a control group (waiting list, care-as-usual, placebo or other; this last category included control conditions that could not be categorised into one of the three categories, such as participation in online discussion forums, in workshops on other subjects, routine care in general medical care or an information booklet).

We calculated effect sizes (Hedges' *g*) for each comparison between a psychotherapy and a comparison group. We only included outcome measures that assess depressive symptoms. If there is more than one measure of depression, these are pooled within the study, before the overall effects are pooled across studies. Because effect sizes are difficult to interpret for patients and clinicians, we also calculate the numbers-needed-to-treat (NNT; Laupacis *et al.* 1988) indicating how many patients should receive the treatment to have one more positive outcome than the comparison group (using the methods of Furukawa, 1999). In all meta-analyses, we used the random-effects model. Subgroup and metaregression analyses were conducted according to the procedures described by Borenstein *et al.* (2009).

We conducted multivariate meta-regression analyses with key characteristics of (a) patients: recruitment (community, clinical, other), target group (adults, older adults, college students, women with postpartum depression, patients with general medical disorders, other), definition of depression (diagnostic interview *v.* self-report measure); (b) psychotherapies: number of sessions, format (individual, group, guided self-help, other/mixed); and (c) trials: control group (waiting list, care-as-usual, other), year of publication, Western *v.* non-Western country and overall risk of bias. In these multivariate meta-regression analysis, we first entered all predictors simultaneously (full model). In order to avoid overfitting of the meta-regression models, we repeated the analysis, with a (manual) stepwise backward elimination of the least significant predictor until only significant predictors remained in the model (parsimonious model).

Adjusting the overall pooled effect size

We examined in the analyses the following elements that may affect the overall effect size:

Heterogeneity

This examines the 'apples and oranges' issue. If heterogeneity is too high, it may be better not to pool the effects sizes of individual studies because they are too different from each other. We calculate the level of heterogeneity with I^2 and its 95% confidence interval (CI), which indicates the level of heterogeneity in percentages (Higgins & Green, 2011). Heterogeneity of 25% is considered to be low, 50% is considered moderate and 75% is considered to be high. The higher the heterogeneity is, the more difficult it is to interpret the pooled effect size.

Type of control group

In all analyses, we differentiated the effects of the different types of control groups (waiting list, care-as-usual, other inactive control group). We used subgroup analyses to examine differences across types of control groups.

Risk of bias

Assessment of risk of bias refers to the 'garbage in, garbage out' principle. Risk of bias indicates possible systematic errors in a study or deviations from the true or actual outcomes. In the current study, we use four items of the Cochrane risk of bias assessment tool

(Higgins *et al.* 2011): adequate generation of allocation sequence, the concealment of allocation to conditions, the prevention of knowledge of the allocated intervention (masking of assessors) and dealing with incomplete outcome data (whether or not intention-to-treat analyses were conducted). Two independent researchers assessed the validity of the included studies and disagreements were solved through discussion. We rated each item as positive (no risk of bias) or negative (there was a risk of bias or risk of bias was unclear).

Masking of patients and therapists was not possible and was therefore not rated in the current study. We also did not rate 'selective outcome reporting' (publication of positive outcomes on one instrument, while the results of another instrument with no or less positive outcomes are not published). This requires that study protocols were published, which is typically not the case in psychotherapy trials. This would have resulted in very few trials in psychotherapy with low risk of bias.

Publication bias

Publication bias can be examined indirectly, based on the assumption that large studies (with many participants) can give more precise estimates of the effect size. The effect sizes in smaller studies are less precise and can divert more from the pooled effect size (Cuijpers, 2016*a*). However, the smaller studies divert from the overall effect size by chance, and therefore they should divert from the mean in both directions, positive and negative. If a meta-analysis finds that in small studies more studies point in the positive direction than in the negative direction, then the small studies with no effects or with negative effects are not published. In this study, we will use a method called 'Duval and Tweedie's trim and fill procedure' to estimate how many studies are missing, and what the effect size is when these missing studies are imputed (Duval & Tweedie, 2000). This can be seen as an indication of the 'file drawer' problem.

For the final effect sizes we found, we also calculated Orwin's fail safe *N*, which indicates the number of studies that have to be found in order to reduce the effect size below the threshold for clinical relevance. As a threshold for clinical relevance, we used the effect size of $g = 0.24$, based on estimates of the 'minimal important difference' (Cuijpers *et al.* 2014).

Results

The unadjusted effects of psychotherapies for adult depression

The flowchart of the inclusion of trials and reasons for excluding studies is given in Appendix B. Selected characteristics of the included studies are given in Appendix C and the references are given in Appendix D. An overview of the characteristics of the studies is given in Appendix E.

The overall pooled effect size for all psychotherapies compared with any control group ($k = 369$) was $g = 0.70$ (95% CI 0.64–0.75), which corresponds with an NNT of 4.18 (Table 1).

In Table 1, we have also given the effect sizes for each of the seven major types of psychotherapy for adult depression (for definitions, see Cuijpers *et al.* 2008*b*). There was no significant difference between the effects of different types of psychotherapy, which is in line with earlier meta-analyses of trials directly comparing different types of therapy (Cuijpers *et al.* 2008*b*; Barth *et al.* 2013).

We also conducted several sensitivity analyses to examine whether the effect size was affected by the design of the studies and the outcome measures. In one analysis, we excluded extreme outliers (effect size $>g = 2.0$). In two other analyses, we included only one effect size per study (because some studies compared more than one psychotherapy to a control group), one in which we only included the highest effect size and one in which we only included the lowest effect size. We also conducted separate meta-analyses that were limited to specific depression measures (that were used in ≥50 comparisons: the HAMD, BDI, BDI-II). As can be seen in Table 1, the effect sizes ranged from $g = 0.61$ to 0.87, which correspond with NNTs from 3 to 5.

Heterogeneity was high in all analyses, and we explored possible sources of heterogeneity. To examine whether characteristics of the patients, the interventions or studies were associated with the effect size and were possible sources of heterogeneity, we first conducted a series of subgroup analyses (Table 1 and Appendix F). The full and the parsimonious models of the multivariate meta-regression analyses can be found in Table 2.

There were three characteristics of the trials that were significantly associated with the effect size, in the subgroup analyses, the full meta-regression model and the parsimonious model. The first was the type of control group, with waiting list control groups having significantly larger effect sizes than care-as-usual and other control groups. The second was a risk of bias (entered in the meta-regression model as a continuous variable). For illustrative purposes, we have reported the effect size for each of the scores on the risk of bias tool in Table 1. As can be seen, the effect sizes range from $g = 1.11$ in the studies with the highest risk of bias to $g = 0.46$ in those with the lowest risk. The third variable that was associated with the effect size in all analyses was whether or not

Table 1. *Effects of psychotherapies compared with control groups (k = 369): Hedges' g*[a]

		k	*g*	95% CI	I^2	95% CI	*p*[b]	NNT
All studies		369	0.70	0.64–0.75	76	74–79		4.18
Extreme outliers excluded (*g* > 2.0)		352	0.61	0.57–0.66	65	61–69		4.89
One effect size per study (only highest)		289	0.70	0.65–0.76	79	76–81		4.18
One effect size per study (only lowest)		289	0.64	0.58–0.69	76	73–79		4.63
Only HAMD		103	0.86	0.75–0.97	74	69–78		3.30
Only BDI		128	0.87	0.77–0.98	72	66–76		3.26
Only BDI-II		80	0.68	0.57–0.80	74	68–79		4.32
		Subgroup analyses						
Type of therapy	CBT	192	0.70	0.63–0.77	75	72–78	0.07	4.18
	Behavioural activation	20	0.94	0.66–1.22	75	59–83		2.99
	Interpersonal psychotherapy	25	0.60	0.40–0.80	74	60–82		4.99
	Problem-solving therapy	27	0.77	0.54–1.01	86	80–89		3.74
	Third wave therapies	18	0.77	0.58–0.97	69	46–80		3.74
	Supportive counselling	19	0.58	0.42–0.75	45	0–67		5.18
	Psychodynamic therapy	11	0.40	0.17–0.63	69	31–82		7.91
	Other	57	0.68	0.54–0.82	77	71–82		4.32
Control group	Waiting list	159	0.89	0.80–0.98	73	68–76	≤0.001	3.18
	Care-as-usual	144	0.61	0.53–0.68	77	74–80		4.89
	Other	66	0.51	0.40–0.62	76	70–81		6.01
Country	Western	325	0.63	0.58–0.68	69	65–72	≤0.001	4.71
	Non-Western	44	1.13	0.94–1.33	90	88–92		2.44
Risk of bias score	0 (high risk)	14	1.11	0.87–1.36	23	0–59	≤0.001	2.49
	1	122	0.92	0.79–1.05	76	71–79		3.06
	2	63	0.73	0.60–0.86	72	64–78		3.98
	3	62	0.71	0.58–0.85	85	81–87		4.11
	4 (low risk)	108	0.46	0.41–0.52	58	47–66		6.76

BDI, Beck Depression Inventory; CBT, cognitive behavioural therapy; CI, confidence interval; HAMD, Hamilton Rating Scale for Depression; N_{comp}, number of comparisons; NNT, numbers-needed-to-treat.

[a]According to the random-effects model.

[b]The *p*-values in this column indicate whether the difference between the effect sizes in the subgroups is significant.

the trial was conducted in a Western country, with studies conducted in Western countries reporting significantly lower effects compared with studies in other countries.

The adjusted effects of psychotherapies for adult depression

In Table 3, we have presented the effect sizes for each of the major types of psychotherapy after excluding waiting list control groups, limited to studies with low risk of bias, and after adjustment for publication bias. We excluded studies from non-Western countries, because they were found to be one of the major, independent predictors of the effect sizes (after adjustment for all other predictors). The results for the most examined types of psychotherapy are graphically represented in Fig. 1.

When all types of psychotherapy are taken together, the mean effect size drops from *g* = 0.63 in all studies, to *g* = 0.31 after these adjustments. Only 71 (22%) comparisons were rated as low risk of bias. Comparable results were found for CBT, the best studied type of psychotherapy, with effect sizes dropping from *g* = 0.62 to 0.29, and only 23% with low risk of bias.

For other types of therapy, less evidence was available. The effect size for supportive counselling dropped from *g* = 0.58 to 0.35 after the adjustments, with only seven comparisons with low risk of bias and no waiting list. For problem-solving therapy and psychodynamic therapy, only a handful of studies were available after removal. The effect size for the problem-solving therapy dropped almost 75%, from *g* = 0.77 for all studies, to *g* = 0.20. The effect size for psychodynamic therapy did not significantly differ from zero anymore. For behavioural activation, third wave therapies and interpersonal psychotherapy, only two trials were available after adjustment, which was insufficient for assessing the adjusted effect size.

For the two types of therapy with a sufficient number of studies and an adjusted effect size above the

Table 2. *Standardised regression coefficients of characteristics of studies on psychological treatment of depression: multivariate meta-regression analyses (k = 369)*

		Full model			*Parsimonious model*		
		Coeff	*SE*	*p*	Coeff	*SE*	*p*
Recruitment	Community	Ref.					
	Clinical	−0.15	0.09	0.11			
	Other	0.04	0.09	0.65			
Target group	Adults	Ref.					
	Older adults	0.04	0.11	0.73			
	Student population	0.32	0.16	0.04			
	Women with PPD	−0.14	0.12	0.27			
	General medical	−0.04	0.11	0.73			
	Other	0.05	0.11	0.65			
Diagnosis *v.* cut-off		−0.14	0.07	0.04			
Type	CBT	Ref.					
	Third wave	0.15	0.15	0.32			
	BAT	0.09	0.15	0.55			
	Psychodynamic	−0.14	0.18	0.42			
	IPT	−0.10	0.13	0.42			
	PST	0.09	0.12	0.45			
	Supportive	0.00	0.15	0.99			
	Other	0.01	0.09	0.91			
Format	Individual	Ref.					
	Group	−0.03	0.08	0.70			
	Guided self-help	0.02	0.11	0.87			
	Other/mixed	−0.19	0.14	0.19			
Number of sessions (continuous)		0.00	0.01	0.94			
Control group	Waiting list	Ref.					
	Care-as-usual	−0.12	0.09	0.17	−0.22	0.07	0.002
	Other	−0.27	0.09	0.003	−0.31	0.08	<0.001
Year (continuous)		−0.01	0.00	0.16			
Western *v.* non-Western		0.46	0.11	<0.001	0.42	0.09	<0.001
Risk of bias (continuous)		−0.09	0.03	0.001	−0.11	0.02	<0.001
Intercept		12.76	8.17	0.12	1.08	0.07	<0.001
R^2 analogue		0.24			0.26		

BAT, behavioural activation therapy; CBT, cognitive behavioural therapy; Coeff, regression coefficient; HAM-D, Hamilton Rating Scale for Depression; IPT, interpersonal psychotherapy; *p*, this *p*-value indicates whether the regression coefficient of the sub-groups differ significantly from the reference group; PPD, postpartum depression; PST, problem-solving therapy; Ref, reference group; SE, standard error.

threshold for clinical relevance of $g = 0.24$ (CBT and supportive therapy), we also calculated Orwin's fail safe *N*. This indicates the number of studies needed to reduce the effect size below the threshold of clinical relevance (Table 3). For CBT, 14 studies have to be found (we included 38 studies), and for supportive therapy, five studies have to be found (we included seven).

Heterogeneity was high in most analyses with all trials included, but after exclusion of the trials with waiting lists and risk of bias, heterogeneity dropped to moderate-to-low levels (I^2<50%).

Discussion

We wanted to examine whether psychotherapies for depression were significant after taking problematic features of meta-analyses into account. We found that the effect sizes based on the published literature were large, but very heterogeneous. After exclusion of trials with waiting list control groups and at least some risk of bias, and after adjustment for publication bias, the effect sizes dropped considerably. This was true for all trials together and also for CBT, the best-examined type of therapy. Supportive counselling

Table 3. *Effects of psychotherapies compared with control groups (k = 369), after excluding waiting list control groups, studies with risk of bias and adjustment for publication bias: Hedges' g*[a,b].

	All studies				No waiting list				Low risk of bias				Adjusted for publication bias			
	k	*g*	95% CI	I^2	95% CI	*k*	*g*	95% CI	I^2	95% CI	*k*	*g*	95% CI	*k*	*g*	95% CI	k^{oc}
All studies	325	0.63	0.58–0.68	69	65–72	179	0.51	0.45–0.58	71	67–75	71	0.38	0.32–0.44	46	26–59	84	0.31	0.24–0.38	35
CBT	165	0.62	0.55–0.69	64	57–69	84	0.47	0.39–0.54	59	46–67	38	0.36	0.27–0.45	45	13–62	44	0.29	0.18–0.39	14
Supportive	18	0.58	0.41–0.76	48	0–69	14	0.48	0.34–0.62	22	0–58	7	0.41	0.27–0.55	0	0–58	10	0.35	0.21–0.49	5
PST	27	0.77	0.54–1.01	86	80–89	13	0.58	0.25–0.92	90	85–93	6	0.24	0.08–0.40	35	0–73	7	0.20	0.03–0.37	[d]
Dynamic	10	0.37	0.13–0.61	68	25–82	8	0.30	0.06–0.54	69	17–84	5	0.43	0.10–0.77	74	3–88	7	0.22	−0.13–0.58	[d]
BAT	17	0.81	0.57–1.04	55	9–73	8	0.74	0.47–1.01	35	0–70	2	[e]					[e]		
Third wave	16	0.67	0.60–0.78	0	0–45	5	0.71	0.48–0.95	0	0–64	2	[e]					[e]		
IPT	23	0.54	0.36–0.71	61	32–74	20	0.42	0.28–0.56	30	0–58	2	[e]					[e]		

BAT, behavioural activation therapy; CBT, cognitive–behavioural therapy; CI, confidence interval; IPT, interpersonal psychotherapy; PST, problem-solving therapy.
[a]According to the random-effects model.
[b]Studies outside North America, Europe and Australia were excluded from these analyses.
[c]Orwin's fail safe *N*: the number of studies that have to be identified to reduce the pooled effect size below the threshold for clinical relevance of $g = 0.24$ (Cuijpers *et al.* 2014).
[d]These effect sizes are already below the threshold of $g = 0.24$ for clinical relevance and therefore Orwin's fail safe *N* was not calculated.
[e]Too few studies were available to calculate effect sizes.

and problem-solving therapy had small effect sizes, based on a small number of trials. Psychodynamic therapy was no longer significant. For other types of therapy including problem-solving therapy, interpersonal psychotherapy and behavioural activation, insufficient evidence was available to analyse whether they have significant effects.

It is clear that meta-analyses that include all published trials, without taking into account the problems of meta-analyses, heavily overestimate the effects of psychotherapies. This makes it highly probable that Smith & Glass (1977) in their early meta-analysis considerably overestimated the effects of psychotherapy. However, we found the overall effect of psychotherapy was small, but significant. Can this be considered as evidence that Eysenck was wrong and that psychotherapy is effective with effects that go beyond spontaneous recovery? Not completely. We still have sources of possible bias that have not been taken into account in these analyses. Selective outcome reporting is one of these. There is also the problem that patients and therapists cannot be blinded for the condition they have been assigned to, while it is known from medication trials that blinding has an effect on patients (Cuijpers *et al.* 2015). There may also be other sources of bias that have not yet been described and examined, but may affect the effect sizes. Because the resulting effect sizes are relatively small, only a small change in the overall estimate, caused by an unknown source of bias, can make this non-significant.

We previously made a rough estimate of the threshold for clinical relevance in treatments of depression and found that an effect size of $g = 0.24$ may be the threshold for clinical relevance (Cuijpers *et al.* 2014). The effect sizes we found for the four therapies with sufficient studies after adjustment ranged from $g = 0.20$ to $g = 0.35$. Problem-solving therapy and psychodynamic therapy did not cross the threshold for clinical relevance, and the effect sizes for CBT and supportive therapy were close to this threshold. Orwin's fail safe *N* indicated relatively low numbers of studies that have to be identified in order to reduce the effect sizes of CBT and supportive therapy below this threshold.

So, the best evidence we currently have does suggest that psychotherapies for depression work, and overall the effects can be considered clinically relevant, especially for CBT and non-directive counselling. However, at the same time, the possibility that psychotherapies do not have effects that are larger than spontaneous recovery cannot be excluded.

Is this problem limited to psychotherapies for depression? It is well known that the problems of waiting lists, risk of bias and publication bias also exist in

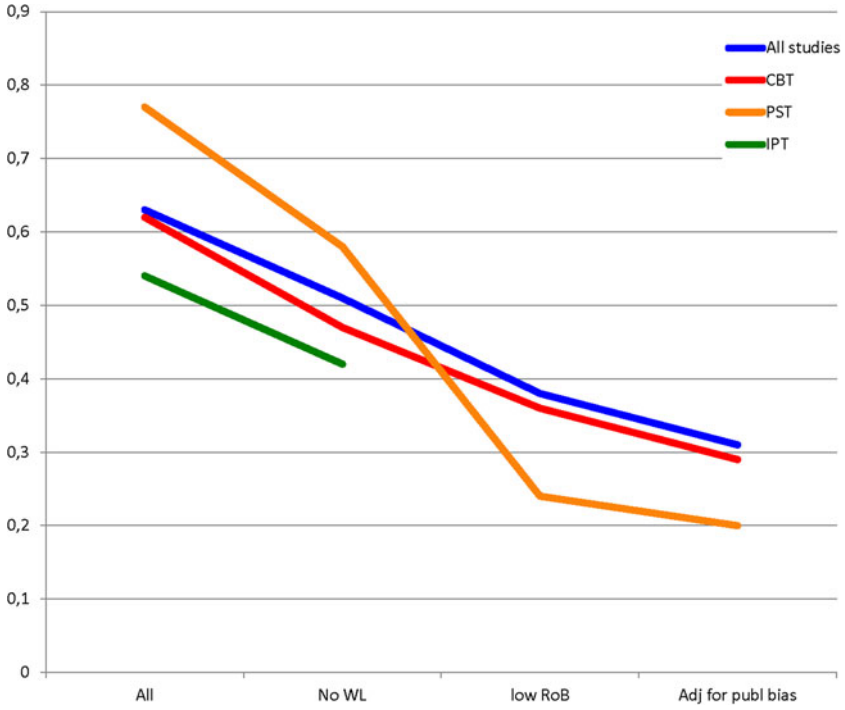

Fig. 1. Effect sizes of the three best examined types of psychotherapy for adult depression, after removal of waiting list controlled studies, after removal of studies with at least some risk of bias and after adjustment for publication bias[a]. ([a]For IPT insufficient studies were available to calculate effect sizes after removal of studies using waiting list control groups.)

meta-analyses of psychotherapies for other mental disorders. For example, in a recent meta-analysis, we found that by far the majority of trials on CBT for social anxiety disorder, generalised anxiety disorder and panic disorder used a waiting list control group (Cuijpers *et al.* 2016*a*). In this study, we also found that the effect sizes after removing these trials dropped considerably, just as in the current meta-analysis. However, whether our results are also true for other mental disorders is an empirical question and should be examined.

It should be noted that the discussion whether treatments for depression are effective is not limited to psychotherapies. Comparable methodological problems have been found for antidepressant medication. For example, one meta-analysis found that across medication trials, 52% of patients respond to antidepressants, while 38% respond to placebo (Levkovitz *et al.* 2011). This difference of 14% between antidepressant and placebo is not adjusted for methodological problems. For example, publication bias has been found to reduce the effects of meta-analyses with about 25% (Turner *et al.* 2008). The majority of trials also do not report the randomisation methods, just as in psychotherapy trials (Jakobsen *et al.* 2017); and there are also indications that patients know whether they are in the medication or placebo condition because of experienced or lack of side effects (Moncrieff *et al.* 1998). This breaks the blind and may affect the effect size. Because of these methodological issues, some authors suggest that antidepressants are not effective at all or only at levels that are not clinically relevant (Moncrieff & Kirsch, 2015).

We found that only 20% of the trials that were conducted in Western countries had low risk of bias and did not use a waiting list control group. This means that 80% of the trials are at risk for not resulting in reliable estimates of the effect size of psychotherapies. These studies can heavily overestimate the true effect, give a good estimate of the true effect, or in some cases may underestimate the effect size. But it is clear that the largest part of the literature on the effects of psychotherapy does not reflect the true effects of psychotherapies for depression.

It should be noted that the studies with risk of bias are not necessarily low-quality trials. The methods for doing trials have changed over time. For example, for a long time, it was not the habit to describe the exact methods of randomisation in papers. This does not imply that these studies did not do this correctly. However, it is unclear whether this happened correctly, and the trials that do report that this randomisation was done correctly can now be considered to give the best evidence.

The current meta-analysis has some important limitations that should be noted. First, we excluded studies with waiting list control groups. But studies with care-as-usual control groups have also problems, mainly because it can differ considerably what care-as-usual means (Mohr *et al.* 2009; Gold *et al.* 2017). This introduces more heterogeneity in the meta-analyses. Another problem is that the methods we used to impute unpublished studies, based on an asymmetrical funnel plot, has been criticised (van Assen *et al.* 2015), and because this is based on a statistical method, it can never be an accurate estimate of the true publication bias. However, in a recent study, we examined how many grants on psychotherapy for depression funded by the National Institutes of Health (NIH) were not published and how much the mean effect size found for these studies was reduced (Driessen *et al*. 2015*b*). We found that the effect size dropped with a comparable rate as was found with indirect methods based on the funnel plot. Finally, we only looked at short-term outcomes of psychotherapies, while there are indications that they do have long-term effects, compared, for example, with antidepressants (Karyotaki *et al.* 2015).

Despite these limitations, it seems safe to conclude that the effects of psychotherapies for depression are considerably overestimated when all studies are included and no further analyses of risk of bias, publication bias and type of control group are done. After adjustment for these methodological issues, the effect sizes found for psychotherapies drop considerably from moderate to small. For several types of psychotherapy for adult depression, insufficient evidence is available to judge whether they result in clinical relevant effects, because too few low-risk studies were available, including problem-solving therapy, interpersonal psychotherapy and behavioural activation. The effect size for psychodynamic treatments did not surpass the threshold for clinical significance and was not significant after adjusting for risk of bias. Whether other sources of bias that were not examined further reduce the effects of psychotherapies cannot be said at this moment. Thus, it remains questionable whether Eysenk was truly right or wrong.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/S2045796018000057.

Acknowledgements

None.

Financial support

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of Interest

None.

References

Barth J, Munder T, Gerger H, Nuesch E, Trelle S, Znoj H, Juni P, Cuijpers P (2013). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLoS Medicine* **10**, e1001454.

Bergin AE, Lambert MJ (1971). The evaluation of therapeutic outcomes. In *Handbook of Psychotherapy and Behavior Change* (ed. SL Garfield and AE Bergin), pp. 139–189. Wiley: New York.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009). *Introduction to Meta-Analysis*. Wiley: Chichester, UK.

Cuijpers P (2016*a*). Are all psychotherapies equally effective in the treatment of adult depression? The lack of statistical power of comparative outcome studies. *Evidence Based Mental Health*, ebmental–2016.

Cuijpers P (2016*b*). *Meta-Analyses in Mental Health Research; A Practical Guide*. Vrije Universiteit: Amsterdam, NL. Available at: https://indd.adobe.com/view/5fc8f9a0-bf1e-49d3-bf5f-a40bfe5409e0

Cuijpers P, van Straten A, Warmerdam L, Andersson G (2008*a*). Psychological treatment of depression: a meta-analytic database of randomized studies. *BMC Psychiatry* **8**, 36.

Cuijpers P, van Straten A, Andersson G, van Oppen P (2008*b*). Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology* **76**, 909–922.

Cuijpers P, Driessen E, Hollon SD, van Oppen P, Barth J, Andersson G (2012). The efficacy of non-directive supportive therapy for adult depression: a meta-analysis. *Clinical Psychology Review* **32**, 280–291.

Cuijpers P, Turner EH, Koole SL, van Dijke A, Smit F (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety* **31**, 374–378.

Cuijpers P, Karyotaki E, Andersson G, Li J, Mergl R, Hegerl U (2015). The effects of blinding on the outcomes of psychotherapy and pharmacotherapy for adult depression: a meta-analysis. *European Psychiatry* **30**, 685–693.

Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJ (2016*a*). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry* **15**, 245–258.

Cuijpers P, Donker T, Weissman MM, Ravitz P, Cristea IA (2016*b*). Interpersonal psychotherapy for mental health

problems: a comprehensive meta-analysis. *American Journal of Psychiatry* **173**, 680–687.

Driessen E, Hegelmaier LM, Abbass AA, Barber JP, Dekker JJM, Van HL, Jansma EP, Cuijpers P (2015*a*). The efficacy of short-term psychodynamic psychotherapy for depression: a meta-analysis update. *Clinical Psychology Review* **42**, 1–15.

Driessen E, Hollon SD, Bockting CLH, Cuijpers P, Turner EH (2015*b*). Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of us national institutes of health-funded trials. *PLoS ONE* **10**, e0137864.

Duval S, Tweedie R (2000). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56**, 455–463.

Eysenck HJ (1952). The effects of psychotherapy: an evaluation. *Journal of Consulting Psychology* **16**, 319–324.

Furukawa TA (1999). From effect size into number needed to treat. *The Lancet* **353**, 1680.

Furukawa TA, Noma H, Caldwell DM, Honyashiki M, Shinohara K, Imai H, Chen P, Hunot V, Churchill R (2014). Waiting list may be a nocebo condition in psychotherapy trials: a contribution from network meta-analysis. *Acta Psychiatrica Scandinavica* **130**, 181–192.

Glass GV. (1977). 9: Integrating findings: the meta-analysis of research. *Review of Research in Education* **5**, 351–379.

Gold SM, Enck P, Hasselmann H, Friede T, Hegerl U, Mohr DC, Otte C (2017). Control conditions for randomised trials of behavioural interventions in psychiatry: a decision framework. *The Lancet Psychiatry* **4**, 725–732.

Higgins JPT, Green S (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011].* The Cochrane Collaboration.

Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savović J, Schulz KF, Weeks L, Sterne JAC (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* **343**, d5928.

Hunt M. (1997). *How Science Takes Stock: The Story of Meta-Analysis*. Russell Sage Foundation: New York.

Jakobsen JC, Katakam KK, Schou A, Hellmuth SG, Stallknecht SE, Leth-Møller K, Iversen M, Banke MB, Petersen IJ, Klingenberg SL, Krogh J, Ebert SE, Timm A, Lindschou J, Gluud C (2017). Selective serotonin reuptake inhibitors versus placebo in patients with major depressive disorder. A systematic review with meta-analysis and Trial Sequential Analysis. *BMC Psychiatry* **17**, 58.

Karyotaki E, Kleiboer A, Smit F, Turner DT, Pastor AM, Andersson G, Berger T, Botella C, Breton JM, Carlbring P, Christensen H, de Graaf E, Griffiths K, Donker T, Farrer L, Huibers MJH, Lenndin J, Mackinnon A, Meyer B, Moritz S, Riper H, Spek V, Vernmark K, Cuijpers P (2015). Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. *Psychological Medicine* **45**, 2717–2726.

Laupacis A, Sackett DL, Roberts RS (1988). An assessment of clinically useful measures of the consequences of treatment. *The New England Journal of Medicine* **318**, 1728–1733.

Levkovitz Y, Tedeschini E, Papakostas GI (2011). Efficacy of antidepressants for dysthymia: a meta-analysis of placebo-controlled randomized trials. *The Journal of Clinical Psychiatry* **72**, 509–514.

Luborsky L, Singer B, Luborsky L (1975). Comparative studies of psychotherapies: is it true that everyone has won and all must have prizes? *Archives of General Psychiatry* **32**, 995–1008.

Mohr DC, Spring B, Freedland KE, Beckner V, Arean P, Hollon SD, Ockene J, Kaplan R (2009). The selection and design of control conditions for randomized controlled trials of psychological interventions. *Psychotherapy and Psychosomatics* **78**, 275–284.

Moncrieff J, Kirsch I (2015). Empirically derived criteria cast doubt on the clinical significance of antidepressant-placebo differences. *Contemporary Clinical Trials* **43**, 60–62.

Moncrieff J, Wessely S, Hardy R (1998). Meta-analysis of trials comparing antidepressants with active placebos. *The British Journal of Psychiatry* **172**, 227–231.

Smith ML, Glass GV (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist* **32**, 752–760.

Smith ML, Glass GV, Miller TI (1980). *The Benefits of Psychotherapy*. Johns Hopkins University Press: Baltimore, Maryland.

Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *The New England Journal of Medicine* **358**, 252–260.

van Assen MA, van Aert R, Wicherts JM (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods* **20**, 293–309.