Cognitive behavior therapy for adult depressive disorders in routine clinical care: A systematic review and meta-analysis

1. Introduction

Depressive disorders are very common in the general population, second only to anxiety disorders (Kessler et al., 2005). A recent large (N = 36,309) epidemiological study (Hasin et al., 2018) from the USA, using DSM-5 (APA, 2013) criteria, found a lifetime prevalence of 20.6%; significantly more common in females (26.1%) than in males (14.7%). The most recent (2019) version of the WHO Global Burden of Disease study (GBD Diseases and Injuries Collaborators, 2020) showed that depressive disorders are the highest of the mental disorders, ranked 6th overall (measured as disability-adjusted life-years) and had increased by 61% since 1990.

Using DSM-IV diagnostic criteria (APA; 1994) the National Comorbidity Survey-Replication in the US (Kessler et al., 2003) found that 72.1% of participants with major depressive disorder had at least one comorbid disorder. The most common was any anxiety disorders (59.2%) followed by impulse control disorders (30%), and substance use disorders (24%). A recent study from the Netherlands (Kan et al., 2021) found very similar results with an overall comorbidity rate of 71.7%. The study by Hasin et al. (2018) using DSM-5 criteria found a comorbidity rate of 57.9% for substance use disorder, 37.3% for anxiety disorders, and 31.9% for personality disorders.

The type of psychological treatment with the strongest research support is cognitive behavior therapy (CBT), which often is used as an umbrella concept for various therapies with similar theoretical basis and clinical procedures. In the present meta-analysis, we will focus on the four variants of CBT receiving a strong recommendation by the following three organizations: American Psychological Association (2019), Australian Psychological Society (2018), and National Institute for Health and Care Excellence (2022) in the UK. These treatments are CT/CBT (Beck et al., 1979), Guided self-help CBT (GSH; Andersson et al., 2005), Behavior Activation (BA; Martell et al. 2001), and Problem-solving therapy (PST; D'Zurilla, 1986). In a recent comprehensive meta-analysis of 309 RCTs, covering 15 different forms of psychotherapy for depression, Cuijpers et al. (2020) reported the following effect sizes (*g*) against various control conditions: CT 0.95, GSH 0.97, BA 1.05, and PST 0.90. Thus, all four treatments yielded large controlled effect sizes.

Randomized controlled trials (RCTs) are the gold standard for evaluating the efficacy of interventions. In RCTs, scientists employ stringent inclusion and exclusion criteria, and they deliver the treatment according to specific treatment manuals by highly trained therapists to maximize internal validity. In addition, the fidelity of treatment is checked, and assessments are done by independent assessors. Studies with such characteristics are often more feasible to conduct in academic contexts (e.g., university research clinics). It has been argued that the context characteristics of academic studies (e.g., patients, therapists, skills level, independent assessments) may not be representative of routine clinical care (e.g., Kazdin, 2008; Tolin et al., 2015; Westen & Morrison, 2001). It is also usually assumed that patients in the real-world clinical settings are more difficult to treat, compared to those in research clinics, as the former have more severe psychopathology, and higher psychiatric and somatic comorbidity (Stewart & Chambless, 2009). Consequently, the field has called for investigation of the outcome of psychological treatment conducted within routine clinical care (Chambless & Hollon, 1998; Chambless & Ollendick, 2001). Such studies are generally known as effectiveness studies although a clear consensus on their definition and characteristics is not currently available (Hans & Hiller, 2013a). The outcome of a study is probably more representative of the real-world setting if it is conducted within routine clinical care (e.g., mental health centers, outpatient clinics) where the treatment is provided by practicing clinicians employed at the clinic to referred patients (Hunsley & Lee 2007; Stewart & Chambless, 2009). Furthermore, effectiveness studies may use various designs such as experimental (RCT), quasi-experimental, or pre-post (Stewart & Chambless, 2009).

Previous meta-analyses of effectiveness studies have mainly focused on anxiety disorders (e.g., Hans & Hiller, 2013a; Stewart & Chambless, 2009) or a mix of common mental disorders (e.g., Gaskell et al., 2022; Wakefield et al., 2021). The meta-analysis coming closest to a focus on depressive disorders is that of Hans and Hiller (2013b), which included studies of samples having both depressive and anxiety disorders, as long as the majority of patients suffered from depression. Other inclusion criteria were non-randomized studies of face-to-face CBT for at least 6 sessions, and participant age 16-65 years. The meta-analysis included 34 studies and found a pre-post ES (*d*) for depressive severity of 1.13 for completers and 1.06 for intention-to-treat (ITT) analyses, and 0.67 for general anxiety measures. There was no significant difference in outcome between individual and group treatment, but individual treatment (M 42.0%) had significantly higher attrition than group treatment (M 16.7%). In a benchmark analysis against five efficacy RCTs the effectiveness studies yielded inferior effects; however, statistical tests were not carried out for this analysis.

The present meta-analysis differs from that of Hans and Hiller (2013b) in the following important ways: we included (a) only studies in which all participants had been diagnosed with a depressive disorder, (b) both face-to-face and guided self-help CBT since research (Carlbring et al., 2018) has shown that guided self-help yields effects equivalent to face-to-face therapy, (c) both randomized and non-randomized (pre-post) studies since effectiveness studies can be RCTs, (d) studies of participants over 65 years of age since CBT can be used irrespective of the participants' age, and (e) we did not apply a minimum number of therapy sessions.

The present meta-analysis has three aims: First, to examine the effectiveness of CBT for depressive disorders in routine clinical care regarding the primary depression measure as well as secondary measures of general anxiety and quality of life. Second, to evaluate methodological stringency and risk of bias in the effectiveness studies and investigate potential moderators of treatment outcome. Third, to examine how CBT delivered in routine clinical care does in comparison with efficacy studies for depression. We predicted that there would be no significant differences in ES between effectiveness and efficacy studies for any of the outcomes, based on the results of previous meta-analyses in adults (Hans & Hiller, 2013a; 2013b; Öst et al., 2022; Stewart & Chambless, 2009) and youth (Wergeland et al., 2021).

2. Method

The protocol for this meta-analysis was pre-registered at PROSPERO with ID MASKED. The meta-analysis was conducted according to the PRISMA guidelines (Liberati et al., 2009), and reported according to AMSTAR 2 (Shea et al., 2017), see online Supplement 1 and 2. The meta-analysis was designed according to the PICOS acronym in the following way:

• *Population:* adults ≥ 18 years of age with a diagnosis of Major Depressive Disorder or Dysthymia.

• *Intervention:* CT/CBT, Guided self-help CBT, BA, PST, delivered in routine clinical care.

• *Comparison:* within-group change, i.e., pre vs. post-data and pre vs. follow-up data.

• *Outcome:* primary (depressive symptoms) and secondary (general anxiety and quality of life) measures.

• *Study design:* RCTs and pre-post/Non-randomized studies of intervention (NRSI).

2.1. Literature search

We found the studies through a systematic and comprehensive literature search of electronic databases and scanning of the included articles' reference lists. The search was applied to Ovid MEDLINE, Embase OVID, and PsycINFO from the start of the databases to October 6, 2021. An updated search was done on September 21, 2022. The list of search terms utilized to identify potential studies was generated by all authors in collaboration with a university librarian, who conducted the database searches. We used the following search terms to search the databases: (Cognitive therapy; behav* therapy; cognitive behav* therapy; cognitive behav* treatment; CBT; behav* activation; Problem solving therap*) AND (Major depressive disorder*; MDD ; depression; depressive disorder*; dysthymic disorder*; affective disorder*; mood disorder) AND (open study; clinical study; community trial; intervention study; Pre post study; randomized controlled trial) AND (outpatient clinics; community mental health services; effectiveness; routine care; regular care, community clinic*) AND adults. The full search strategy for the databases are shown in the online Supplement 3.

Pairs of authors read the abstracts of all the papers from this initial search to decide whether a study warranted a more detailed reading. The full-text articles were retrieved if there was any indication in the text that the target group of patients had received a specified evidence-based cognitive-behavioral treatment in a routine clinical care setting. This meant that we were over-inclusive in the first stage of the screening process. We also checked the reference list in the retrieved articles and in meta-analyses, to see whether any other articles that might fulfill the inclusion criteria were identified. In total, 227 full-text articles were considered for inclusion. We used the inclusion and exclusion criteria detailed below to reach a final decision of the inclusion of an article. The full-text articles were read by pairs of authors, and any disagreements were resolved by consensus discussion among the authors and/or consultation with the first author. Twenty-eight studies met the inclusion criteria and were included in the present meta-analysis. These studies comprised a total of 35 treatment conditions since some studies compared two CBT conditions.

2.1.1. Inclusion criteria

To be included in the review and meta-analysis, a study had to:

1. Be published, or in press, in an English language journal.

2. Have participants diagnosed with Major Depressive Disorder or Dysthymia according to DSM (III and later) or ICD (10 or 11).

3. Be testing a form of CBT (face-to-face or guided self-help), BT, BA, or PST.

4. Have participants referred for treatment through usual clinical routes, e.g., a health professional, instead of being recruited via media advertisements.

5. Be an effectiveness study, i.e., carried out in a routine care setting, such as community mental health centers.

6. Have therapists who are practicing mental health clinicians for whom provision of service is a substantial part of a job, e.g., clinical psychologists or social workers (Shadish et al., 2000).

7. Have a treated sample consisting of at least 10 participants.

8. Have a minimum participant age of 18.

9. Provide data for a standardized and validated continuous measure of depression.

2.1.2. Exclusion criteria

1. The study is a secondary analysis of a previously published study. However, separate follow-up studies to the basic study are included to provide follow-up data.

2. The study evaluates a service where the results for individual disorders *cannot* be extracted.

3. The study tests a combination of CBT and pharmacological treatment, and all participants in that condition receive both treatments.

2.2. Potential categorical moderators

For a potential categorical moderator to be included in the analysis, at least 70% of the studies needed to provide information on the moderator. This proportion was decided on to ensure that the information extracted was representative of the entire body of studies.

Type of study was either an RCT (when a CBT condition was compared with some control/comparison condition), a pre-post trial (when only a CBT-condition was used in the study), or a NRSI (when there was a comparison between two conditions to which participants had not been randomized). Statistical analysis was categorized as intent-to-treat (ITT; if all randomized or starting participants were included in the statistical analysis) or as completers (if dropouts were deleted). Risk of bias (RoB) was based on a summary evaluation of the domains rated for the different designs (see below) and the studies were categorized as low, moderate, or high RoB. The treatment format could either be individual or group.

2.3. Potential continuous moderators

We used the following continuous measures as potential moderators when at least 70% of the studies provided information of the variable: year of publication, number of therapists in the study, mean age, pre-treatment severity (calculated as a percentage by dividing the sample mean with the maximum score possible of the rating scale applied, e.g., if the sample mean on the Beck Depression Inventory was 32.3 this was divided by 63 = 51.3%), methodology score (see below), hours of treatment, and percent attrition in the treatment condition. To be counted as a dropout, the patient had to fulfill the inclusion criteria, be offered the treatment, have accepted it, and participated in at least the first session but less than the number of sessions defined as completion of treatment. A coding scheme and a scoring manual including the variables of interest were developed. The data extraction and categorizations were done independently by pairs of authors and any disagreements were solved after a consensus discussion.

2.4. Methodological quality

2.4.1. The Psychotherapy Outcome Study Methodology Rating Scale (POMRS)

The scale was developed by Öst (2008) and consists of 22 items rated on a 0-2 scale. It has the following items: 1. Clarity of sample description, 2. Severity/chronicity of the disorder, 3. Representativeness of the sample, 4. Reliability of the diagnosis in question, 5. Specificity of outcome measures, 6. Reliability and validity of outcome measures, 7. Use of blind evaluators, 8. Assessor training, 9. Assignment to treatment, 10. Design, 11. Power analysis, 12. Assessment points, 13. Manualized, replicable, specific treatment programs, 14. Number of therapists, 15. Therapist training/experience, 16. Checks for treatment adherence, 17. Checks for therapist competence, 18. Control of concomitant treatments, 19. Handling of attrition, 20. Statistical analyses and presentation of results, 21. Clinical significance, and 22. Equality of therapy hours (for non-WLC designs only). The total score has a maximum of 44 points, but since all items did not apply to all studies, the total score was recalculated as a percentage of the maximum score possible for the individual study. The internal consistency of the scale was good with a McDonald’s ω of 0.80. The inter-rater reliability of the scale (between the first and the last author), based on 20% randomly selected and blindly rated studies, was ICC = 0.93 (95% CI 0.47-0.99, $p = 0.008$), which according to Cicchetti (1994) is excellent.

2.4.2. Risk of bias

When assessing RoB, we used the Cochrane Collaboration tool for RCTs (Sterne et al., 2019). The domains rated were: the randomization process, missing outcome data, measurement of the outcome, and selection of the reported result. The domain deviations from intended interventions was not rated since therapists and patients in psychotherapy studies cannot be blind regarding the treatment applied. The RCT-studies were classified into the categories high, some concern, or low risk of bias. For NRSI and pre-post studies the Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I; Sterne et al., 2016) was applied. The following domains were judged: confounding, selection of participants, classification of interventions, missing data, measurement of outcome, and selection of the reported result. The NRSI and pre-post studies were classified into the categories low, moderate, serious or critical RoB. When the results across these different study designs were judged to be at similar risk of bias, these classifications were combined into one: low, moderate (some concerns) and high (serious) RoB. The inter-rater reliability of the overall RoB-ratings (between the first and the last author), based on 20% randomly selected and blindly rated studies, was Cohen's kappa = 1.0, *p* = 0.008, which according to Cicchetti (1994) is excellent.

2.5. Effect size measures

In this meta-analysis we used the pre-post and pre-follow-up effect size as outcome measure. We extracted data on both primary (depression symptoms) and secondary measures (general anxiety and quality of life) from the studies. Psychotherapy researchers (e.g., Kazdin, 2022) have argued for broadening the assessment from the focus of depressive symptoms to also include measures of daily functioning and IsHak et al. (2011) called quality of life the ultimate outcome measure of interventions in depressive disorders.

2.5.1. Primary outcome measures

In a meta-analysis of 70 RCTs on treatments for depression Cuijpers et al. (2010) reported a significantly higher ES for independent assessor (IA) ratings than for patients' self-ratings of depressive symptoms. Thus, we used both types of measures in this meta-analysis to investigate if this difference also pertains to effectiveness studies. All nine studies using IA-ratings applied the Hamilton Depression Rating Scale (HDRS; Hamilton, 1960) in its 17-item version and the assessors in all studies were blind as to the treatment patients had received. Twenty-five studies used self-rating scales; Beck Depression Inventory; BDI (Beck et al., 1961) in nine studies, BDI-II (Beck & Steer, 1993) in nine, the Montgomery-Åsberg Depression Rating Scale (MADRS-S; Svanborg & Åsberg, 1994) in three, the Center for Epidemiological Studies Depression scale (CES-D; Radloff, 1977) in two, and the Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001) in two studies.

2.5.2. Secondary outcome measures

Since anxiety disorders are common comorbidities in depression, we extracted data on anxiety symptoms, which were provided by 8 (29%) of the included studies. Beck Anxiety Inventory (Beck et al., 1988) was used in three studies, Hospital Anxiety and Depression Scale (Zigmond & Smith, 1983) in two, Generalized Anxiety Disorder-7 items (Spitzer et al., 2006) in two, and Penn State Worry Questionnaire (Meyer et al., 1990) in one study.

Quality of life was reported by 9 studies (32%). The Quality of Life Enjoyment and Satisfaction Questionnaire (Endicott et al., 1993) was used in two, the World Health Organization Quality of life scale (WHOQOL Group 1998) in two, the Quality of Life Inventory (Frisch et al., 1992) in one, the EuroQol - 5 Dimension (EQ-5D; the EuroQol Group, 1990) in one, the RAND-36 (van der Zee & Sanderman, 1993) in one, and the Health-related quality of life (Short Form-36 MCS; Ware et al., 1993) in one study.

2.6. Meta-analysis

Within-group ES was calculated as (M_{pre} – M_{post})/SD_{pre} according to the recommendation by Lakens (2013), since there is good reason to assume that the interventions influence not only the means but also the standard deviations. The mean ES was computed by weighting each ES by the inverse of its variance. We used intent-to-treat data when a study provided those, if not completer data had to be used.

Before pooling, the effect sizes were screened for statistical outliers, defined as being outside M ± 2SD. At the post-treatment assessment, one (2.4%) of the ESs was an outlier, and at follow-up assessment, there was also one (4.0%). For these ESs, *winsorizing* (Lipsey & Wilson, 2001) was used by reducing the outliers to the exact value of M+2SD. The *Comprehensive Meta-Analysis v. 4* (CMA; Borenstein et al., 2022) software was used for all analyses and to correct for small sample sizes Hedges' *g* was calculated. A random effects model was used since it cannot be assumed that the ESs come from the same population. Lipsey (1990) described an empirically developed rule-of-thumb for considering an ES as small (<0.32), moderate (0.33-0.55), and large (0.56-1.20). Also, Sawilowsky (2009) denoted effect sizes as very large (1.20-1.99) and huge (≥ 2.00).

Proportion of attrition was calculated in CMA. The values of the individual studies were transformed using logit transformation and the meta-analytics was done on the transformed proportions using the random effects model. Then the pooled proportion and its 95% confidence interval was back-transformed to a proportion (according to recommendations by Barendregt et al., 2013; Barker et al., 2021).

Heterogeneity among ES's was assessed with the *Q*-statistic and the 95% prediction interval, which is the interval which 95% of all comparable studies falls within. The possibility of publication bias was analyzed with funnel plot, Egger's regression intercept (Egger et al., 1997), and the trim-and-fill method of Duval and Tweedie (2000). Moderator analyses of continuous variables were carried out with meta-regression using the random effects model and for categorical variables with subgroup analysis using the mixed effect model.

2.7. Efficacy studies for comparison

One of the most recent comprehensive meta-analyses of evidence-supported psychotherapies for adult depression (Cuijpers et al., 2020) was consulted to obtain the efficacy studies we used to compare the effects of CBT effectiveness studies with. From this meta-analysis, we listed the RCTs of cognitive-behavioral treatments having a strong recommendation by the treatment guidelines APS (2018), APA (2019), and NICE (2022), and where participants were required to fulfill diagnostic criteria for a depressive disorder rather than just a cut-off score on a rating scale.

Since the Cuijpers et al. (2020) meta-analysis included both efficacy and effectiveness studies we first checked for any effectiveness studies missed in our data base search. Then we deleted those RCTs that were included in our body of effectiveness studies. This resulted in 52 efficacy RCTs for our comparison and the references of these are listed in the online Supplement 5. This type of benchmarking, where effectiveness and efficacy studies are directly compared in a meta-analysis software has previously been done in three similar meta-analyses on effectiveness studies in children and adolescents (Riise et al., 2021; Wergeland et al., 2021; Wergeland et al., 2022) and three in adults (Öst et al., 2022; MASKED a, b).

Using the same procedure as for the effectiveness studies, we extracted data of the primary continuous outcome at post- and follow-up assessment for the efficacy studies. In order to compare the two categories of studies on background and treatment variables, we also extracted data on mean age, proportion of women, pre-treatment severity (% of the maximum score on the continuous measure), comorbidity (% of the sample having at least one comorbid disorder), medication (% of the sample that at pre-treatment was prescribed an antidepressant), treatment time (in 60 min units), and attrition rate (% dropping out of those patients who participated in at least one session). Other variables were reported in too few studies to be included as background variables. Since some of the result tables will entail many statistical tests, we used the Holm-Bonferroni correction to control the family-wise error rate (Jaccard & Guilamo-Ramos, 2002).

2.8. Power analysis

In the overall comparison of effectiveness and efficacy studies, we have the following number of studies and treatment conditions, which is the unit of analysis: effectiveness studies 28/35 and efficacy studies 52/61, for a total number of 80 studies and 96 treatment conditions with an average of 66 participants per condition. According to the formulas for power analysis in meta-analyses by Valentine et al. (2010), we would have 100% power to detect an ES of 0.20, when assuming that the heterogeneity of effect sizes will be high.

3. Results

3.1. Description of the studies

Figure 1 shows a flowchart of the inclusion of studies in the present meta-analysis. For references to the included studies, see online Supplement 4.

3.1.1. Background data

Table 1 displays background data for the included studies. Some studies involved more than one CBT intervention, therefore Table 1 contains 35 treatment conditions from 28 studies. The total number of participants receiving CBT in these studies was 3734 (range 12-1203). The majority of the 28 studies was done in Europe (n = 21), followed by North America (n = 5), Asia (n = 1), and Australia (n = 1). Twenty (71.4%) of the studies were randomized controlled trials. The proportion of patients declining the offered treatment was reported in 19 studies (67.8%) and the mean was 4.9% (*SD* 12.9; range 0-18.3%). Only nine studies reported the proportion of participants having had at least one prior episode of depression and the mean was 66.3%. Another three studies reported the mean number of previous depression episodes, which were 4, 5, and 6, respectively. Pre-treatment severity of depression (calculated as the percentage of the maximum score of the rating scale applied) had a mean of 45.3% (*SD* 3.6; range 21.9-65.6%).

In the 35 treatment conditions a majority of participants were women (67.3%, *SD* 12.7), and the mean age was 38.7 (*SD* 2.8) years. Only 13 conditions (37%) reported comorbidity and often in an unsystematic way. Of these, 54.5% of the participants had at least one comorbid disorder, and 46.7% of the participants were taking antidepressants at the time of inclusion, based on reports from 27 of the 35 conditions (77.1%).

3.1.2. Treatment data

Table 2 displays treatment data for the included studies. Individual treatments were delivered in 30 conditions (86%) and group treatments in 5. The number of therapists providing treatment for the CBT conditions varied from 1 to 25, with a mean of 6.4 (*SD* 5.6). The majority of the therapists were clinical psychologists (n = 15; 60%), followed by a mixed team of professions (n = 9), and psychiatric nurses (n = 1). The mean number of therapy sessions was 12.8 (*SD* 4.4). Summarized as hours of treatment, the mean was 12.1 (*SD* 5.6). Attrition was specified for 34 conditions (97%) with a mean of 25.1% (*SD* 13.1%), and a range of 0-62%. Follow-up data were reported in 21 of the conditions (60%) and the mean number of months after post-assessment for these was 7.9 (*SD* 3.0) ranging from 4 to 12 months. Lastly, 80% of the conditions used intent-to-treat analysis and 20% used completer analysis.

3.2. Methodological data

3.2.1. Methodology ratings

The mean of the research methodology (POMRS) score (% of the maximum possible score for the individual study) was 51.5 (*SD* 10.8), equivalent to a raw score of 22.7 points. Limiting the analysis to only RCTs showed a mean of 54.9 (*SD* 9.9).

3.2.2. Risk of bias

The risk of bias classification is presented in Supplement 6. Five studies (25%) of the 20 RCTs had a high RoB and 15 had some concerns. Concerning the 8 NRSI/pre-post studies, three (38%) had a high and five a moderate RoB. Merging the two categories of studies generated eight studies (29%) with high and 20 with moderate RoB.

3.3. Meta-analysis

3.3.1. Attrition

Using treatment condition (k = 34) as the unit of analysis, the attrition rate was 25.1% (95% CI 20.9-29.9, z = 8.93, p < 0.0001), which was significantly heterogeneous. RCTs (k =21, 24.9%) and pre-post/NSRI trials (k = 13, 25.4%) did not differ significantly (Q_{between} = 0.01, p = 0.92) in this respect. Also, there was no significant difference (Q_{between} = 1.34, p = 0.25) between individual (k = 29, 26.4%) and group treatment (k = 5, 16.0%).

3.3.2. Primary measure

Table 3 (upper part) displays the ESs for the primary depression measure at post-treatment and follow-up assessment, which was carried out on average 8 months after post-treatment assessment. Some studies used both independent assessor ratings and self-ratings of depression and for these studies, the mean ES was used in the analysis so that each condition only contributed with one ES to the overall mean. The mean ES at post-treatment, was very large (1.51) and significantly heterogeneous, indicated by the Q-values and the 95% PI. The

ES at follow-up, had increased somewhat (1.71). If we just include studies that have follow-up assessment, the ES was 1.55 at post and 1.71 at follow-up, indicating that the treatment effects were maintained.

Publication bias. Egger's regression intercept showed a significant *t*-value (2.92, p = 0.006) and the Duval and Tweedie trim-and-fill method recommended trimming 12 studies to the left of the mean, which would have reduced the *g*-value to 1.27 (95% CI 1.12-1.41). Hence, publication bias is a problem regarding within-group ES for these effectiveness studies.

3.3.3. Moderator analyses

Since mean ES (Table 3) was significantly heterogeneous, moderator analyses were carried out and Table 4 presents the results for the categorical variables. Holm-Bonferroni correction was applied, since there are four variables being analyzed, and no variable showed a significant moderation effect. Studies using intent-to-treat and completer analysis, yielded similar ESs. Risk of bias was not a significant moderator; high RoB studies had a nominally lower ES than studies with moderate RoB, which from a methodological point of view, is important.

Continuous variables, reported from at least 70% of the studies, were analyzed with the meta-regression module in the CMA program using the random effects analyses. Seven variables were analyzed, and after applying the Holm-Bonferroni correction two generated a significant point estimate. Number of therapists providing treatment in the condition was a significant negative moderator, i.e., a lower number of therapists was associated with a higher ES (k = 24, point estimate = -0.0729, z = -3.46, p = 0.0005). Also, year of publication was a negative moderator, i.e., more recent publications were associated with lower ES (k = 35, point estimate = -0.0021, *z* = -2.73, *p* = 0.0062). The following variables were not significant moderators: mean age of the sample, hours of treatment, methodological quality, pre-treatment depression severity, and percent using antidepressant medication.

3.3.4. Secondary measures

Table 3 (middle part) presents the ESs for general anxiety measures at post-treatment and follow-up. Only 11 (31.4%) of the treatment conditions had data on general anxiety, which means that the results should be interpreted with caution. The post-treatment ES was large, 0.71, significantly different from zero, and with high heterogeneity. The follow-up ES was somewhat lower at 0.67, but also significantly different from zero. If we restrict the analysis to studies having follow-up assessment the post ES was 0.62 and the follow-up 0.67, indicating that the effect on general anxiety measures was maintained on average 8 months after the treatment. Publication bias analysis showed that Egger's regression intercept (*t* = 0.72) was not significant and Duval and Tweedie's trim-and-fill method did not suggest trimming any study. Thus, publication bias is not an issue for anxiety measures in the present meta-analysis.

The lower part of Table 3 contains the results for quality of life measures. Thirteen conditions (37.1%) had data on these measures, which means that caution is advised when interpreting the data. The post-assessment ES was large, 0.78, which was significantly different from zero and with high heterogeneity. At follow-up the mean was somewhat lower, 0.54 and heterogeneity was moderate. Confining the analysis to studies having follow-up assessment, the post ES was 0.64. The publication bias analysis yielded a non-significant *t*-value (0.43) for Egger's regression intercept and no study was suggested to trim.

3.4. Efficacy-effectiveness comparison

3.4.1. Background and treatment variables

Table 5 presents a comparison of effectiveness and efficacy studies on some background and treatment variables. Using the Holm-Bonferroni correction there was no variable on which the two types of studies differed significantly.

3.4.2. Effect size on the primary outcome measure

Table 6 shows the ES for the two types of studies. The upper part includes all studies and has data on the mean ES across measures if both assessor and self-ratings were used in the study. At post-treatment as well as follow-up assessment, the mean ESs were very large for both effectiveness and efficacy studies, with no significant difference between them. Then follows the results for self-ratings, which also show very large ESs for both types of studies, with no significant difference between them. Then the table shows the results for assessor ratings with a significantly higher ES for efficacy compared to effectiveness studies at post-assessment, but the difference is no longer significant at follow-up.

A comparison of the ESs for assessor and self-ratings for the effectiveness studies shows that the assessor ES is somewhat higher than the self-rating ES both at post (1.70 vs. 1.52) and at follow-up assessment (1.90 vs. 1.70). When restricting the comparison to the six studies using both types, we found similar results: 1.94 for assessor and 1.76 for self-ratings ($Q_{between}$ = 0.78, *p* = 0.37) at post-assessment.

3.4.3. Comparison of RCTs only

To evaluate if the results for all studies in Table 6 were unreasonably affected by pre-post/NSRI trials, we repeated the analyses with only RCTs. The results are presented in the lower part of Table 6. Neither at post-assessment, nor at follow-up assessment, were there any significant differences between effectiveness and efficacy studies regarding ESs.

4. Discussion

The first aim of this meta-analysis was to examine the effectiveness of CBT for depressive disorders in routine clinical care regarding the primary depression measure as well as secondary measures of general anxiety and quality of life. Results showed that CBT yielded very large ESs (Hedges' *g*) for reduction in depression both at post-treatment (1.51) and at follow-up (1.71). In comparison, the meta-analysis by Hans and Hiller (2013b) found a post-treatment *g*-value of 1.13 for completer and 1.06 for intent-to-treat, giving a weighted mean of 1.11. The ES of 1.51 in the present meta-analysis is significantly higher ($t(78) = 5.47$, $p<0.0001$) than the one reported by Hans and Hiller (2013b). It is also in line with pre-post ESs for efficacy studies reported by other meta-analyses (Cristea et al., 2017; Johnsen & Thimm, 2018; Rubin & Yu, 2017). This is reassuring to clinicians in terms of choosing CBT for the treatment of depression in real-word clinical settings, despite the fact that more recent publications were associated with lower ES.

Regarding the secondary measure of general anxiety our ES of 0.71 corroborates that of 0.67 found by Hans and Hiller (2013b). Both meta-analyses found lower ES for the general anxiety measure compared to the depression measures. This is probably due to the fact that the pre-treatment severity of a secondary problem is not as high as for the primary disorders, which means that the room for improvement is smaller and, thus, the ES will be lower. Also, the CBT variants tested in the included studies focus on depressive symptoms and not on comorbid disorders like anxiety.

The quality of life measure yielded a pre-post mean ES of 0.78, which corroborates the ES of 0.63 reported in a meta-analysis of CBT for depression (Hofmann et al., 2017). As for general anxiety, the ES for quality of life is about half of what we found for depression (1.51) and the same was the case for Hofmann et al. who reported an ES of 1.30 for depression. The finding of a favorable outcome on general anxiety and quality of life without being targeted in treatment is encouraging, especially as higher quality of life after cognitive therapy predicted less relapse and recurrence of depression during a 2-year follow-up phase (Vittengl et al., 2021). The ES for quality of life was higher and more heterogeneous at post-treatment compared to follow-up. As an increase in quality of life may axiomatically be related to maintenance of outcome in the treatment of depression, further research is warranted to investigate this relationship, and whether different approaches in CBT for depression may lead to higher quality of life.

The second aim was to evaluate methodological stringency and risk of bias in the effectiveness studies and investigate potential moderators of treatment outcome. The mean POMRS score was 51.7%, corresponding to a raw score of 22.7, which is very similar to the 22.9 found in a recent meta-analysis of effectiveness studies in OCD (Öst et al., 2022). Future effectiveness studies could improve their methodological quality in various ways, e.g., by using trained and blinded assessors, having three or more properly trained therapists in the study and analyzing the therapist effect on the outcome, carrying out a power analysis before starting the study, including a follow-up assessment at least one year after post-assessment, and allocate equal number of therapy hours to these in RCTs comparing two active treatments.

The risk of bias ratings showed that 8 (29%) of the studies had an overall rating of high RoB, whereas 20 (71%) were rated as moderate. High risk of bias in such a proportion of the studies presents a limitation, although the subgroup analysis showed that the RoB-categorization was not a significant moderator of the ES. Studies with high RoB actually had a nominally lower ES than those with moderate RoB. However, lack of studies with low risk of bias was a significant limitation, and it stresses the urgency of improved quality of future studies-by using randomized controlled designs, pre-registering the treatment protocol, presenting results for all outcome measures, and applying intent-to-treat analysis.

The overall attrition in the effectiveness studies was 25.1%, which is very similar to the 24.6% reported in Hans and Hiller's (2013b) meta-analysis. The latter found a significant difference between individual (42.0%) and group treatment (16.7%). Our results were in the same direction without being significant; individual (26.7%) and group treatment (16.0%); however, the low number of group treatment conditions (n = 5) means that this result must be interpreted with caution. As the outcome of individual and group CBT for depression are comparable, the low attrition in group formats that also increase access to treatment for a larger number of patients might be an important factor to consider in routine care settings.

The subgroup analysis of categorical variables did not yield any significant moderator when using the Holm-Bonferroni correction. From a methodological point of view, it is reassuring that type of study (RCT or pre-post trial) and type of statistical analysis (ITT or completers) did not significantly moderate the ES. These results corroborate those of three meta-analyses of effectiveness studies in children (Riise et al., 2021; Wergeland et al., 2021; 2022) and three in adults (Öst et al., 2022; MASKED a, MASKED b). Risk of bias (high or moderate) was also not a significant moderator, which confirms the results of the meta-analyses of Öst et al. (2022), MASKED a, and MASKED b. All these reports indicated that the category of studies with high RoB had the lowest ES and in MASKED (a) the difference was significant. Thus, although a substantial proportion of included studies have high RoB, they do not seem to inflate the pooled ES.

Regarding continuous variables the number of therapists carrying out the treatment in the CBT conditions was a significant negative moderator. The finding that lower number of therapists was associated with higher ES could mean that studies with just one or two therapists are more similar to specialist treatment than to routine care. The same result was obtained by Wergeland et al. (2021) in a meta-analysis of effectiveness studies for internalizing disorders in youth, but when an outlier study was removed this moderator was no longer significant. Year of publication was also a significant negative moderator, i.e., later years were associated with lower ES. The same results were obtained by Johnsen and Friborg (2015) who found that the effects of CBT for depression declined significantly over time in a meta-analysis covering 70 RCTs from the 1970's to the 2010's. However, this meta-analysis has been criticized on methodological grounds (Cristea et al., 2017; Ljótsson et al, 2017). Also, the same research group did not find year of publication to be a significant moderator in a meta-analysis of mindfulness-based cognitive therapy (Thimm & Johnsen, 2020), and even that the effect size increased significantly over time in a meta-analysis of group CBT for depression (Johnsen & Thimm, 2018). Thus, it is possible that our result for this moderator is a random finding and should be interpreted with caution.

Although not designated as a moderator, we analyzed assessor ratings and self-ratings separately (Table 6) and found a *g*-difference of 0.18 in favor of assessor ratings. Previous meta-analyses of efficacy studies in depression have reported similar results for within-group

ES; Johnsen & Friborg (2015) 0.16, Johnsen & Thimm (2018) 0.23, and Rubin & Yu (2017) 0.33, and Cuijpers et al. (2010) found a similar between-group difference of 0.21. One explanation for this effect is that the pre-treatment SD relative to the M for assessor ratings is somewhat smaller compared to that ratio for self-ratings, and since the SD_{pre} is used as the denominator in the formula to calculate within-group ES it will lead to a higher ES. Thus, using self-report measures in routine care yield more conservative effect sizes.

The third aim was to examine how CBT delivered in routine clinical care does in comparison with efficacy studies for depression. First, effectiveness and efficacy studies were compared on seven background and treatment variables and there were no significant differences when applying the Holm-Bonferroni correction. However, with a *p*-value of 0.01 samples in effectiveness studies had a higher mean proportion of participants having prescribed antidepressants (ADM) at the inclusion of the respective studies. The difference in medication rate is probably due to the fact that only one (3.6%) of the effectiveness studies compared CBT with ADM, and such comparison require that CBT-patients do not have ADM concurrently, whereas eight (15.4%) of the 52 efficacy studies did so, a significant difference ($p = 0.046$, Fisher's exact probability test, one-tailed).

The effectiveness and efficacy studies had very large and similar ESs at post- (1.51 vs. 1.71) and follow-up assessment (1.71 vs 1.85), representing a non-significant difference. The same result was found for self-report measures, but for assessor ratings there was a significant difference in favor of efficacy studies (2.44 vs. 1.70) at post-treatment, but no longer at follow-up (2.31 vs. 1.90). One possible explanation for this difference is that 100% of the effectiveness studies using assessors reported that they were blind to the patients' treatment condition, whereas the proportion was 72% in the efficacy studies. When we confined the analyses to RCTs only we found the same non-significant differences, both at post-treatment and at follow-up. These results differ from Hans and Hiller (2013b) who found that effectiveness studies yielded lower ESs than efficacy studies. However, in their benchmarking they only used five selected efficacy studies and did not do statistical tests of possible differences. We used 52 studies from the Cuijpers et al. (2020) meta-analysis, which had all patients diagnosed with depressive disorders and treated with one of the four CBTs having strong recommendations in current clinical guidelines. Therefore, we believe our comparison is more stringent and valid and, thus, disapproves the notion of Ivory tower of research and non-generalization of outcome of efficacy studies to the real-world settings. The equivalent effects for effectiveness and efficacy studies are corroborated by a number of meta-analyses in anxiety disorders (Hans & Hiller, 2013a; Öst et al., 2022; Stewart & Chambless, 2009; Wergeland et al., 2021).

The present meta-analysis has some limitations that should be considered. Included studies are limited to those published, or in-press reports, in English-language peer-reviewed journals. Studies published in other languages could have provided additional information about the effectiveness of CBT for depressive disorders in adults. Unpublished studies could have introduced bias as it could have been easier to identify unpublished studies from more recent compared to earlier years. This is not a limitation *per se*, but our selected effectiveness studies span 32 years (1990-2022) and the efficacy studies 33 years (1984-2017). We had planned to include analysis of remission rates, but as the studies used different definitions and criteria for remission it was not possible to include this as an outcome variable. Also, the use of pre-post ES in meta-analyses can contribute to biased outcomes; still, for evaluation of improvement found in routine clinical care versus efficacy studies, these analyses are considered informative (Cuijpers et al., 2017). Finally, lack of studies with low risk of bias introduces a significant limitation, and we cannot say anything about the ES for studies with low RoB. Future studies should consider this challenge and employ the strategies available to overcome this limitation. The present meta-analysis also has some methodological strengths. Researchers worked in pairs and independently screened abstracts, read full-text articles, extracted information from the included studies, and any disagreements were solved in consensus discussions. The large number of studies included in the meta-analysis resulted in a high power to detect a small effect size. Finally, the ratings of methodological quality and risk of bias were done with excellent inter-rater reliability.

Clinical implications from the present meta-analysis are that when treating depressive disorders in routine clinical care therapists should use one of the cognitive behavioral treatments (CT, GSH-CBT, BA, PST) that have received strong recommendations in current clinical guidelines, since the effects were equivalent to those in efficacy studies. Furthermore, therapists should have at least a basic training in the therapy method, should use it during enough sessions as described by the developers of the treatment in question, and should evaluate the therapy effect by applying validated self-report measures of depression, general anxiety, and quality of life as used by the studies in this meta-analysis.

Future research regarding the effectiveness of CBT in routine clinical care should apply this type of meta-analysis comparing effectiveness and efficacy studies for other mental disorders, e.g., eating disorders, bipolar disorder, schizophrenia, alcohol use disorders, sleep disorders, and personality disorders. Also, follow-up assessment should be an integrated part of effectiveness studies to evaluate long-term maintenance of the outcomes. The first study of this kind (von Brachel et al., 2019) presented a follow-up assessment 5-20 years (M = 8) after the treatment of 263 patients with anxiety and depression disorders who had received CBT in routine care. Not only had the treatment effect been maintained at this long-term follow-up but there was a further improvement which is encouraging, but more studies on long-term follow-up are needed.

References

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders (DSM-IV)* (4th ed.). Washington, DC: American Psychiatric Press.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5)* (5th ed.). Washington, DC: American Psychiatric Publishing.

American Psychological Association. (2019). *Clinical practice guideline for the treatment of depression across three age cohorts*. Retrieved from https://www.apa.org/depression-guideline.

Andersson, G., Bergström, J., Holländare, F., Carlbring, P., Kaldo, V., & Ekselius, L., (2005). Internet-based self-help for depression: randomised controlled trial. *British Journal of Psychiatry, 187*, 456–461. https://doi.org/10.1192/bjp.187.5.456.

Australian Psychological Society (APS). (2018). *Evidence-based psychological interventions in the treatment of mental disorders: A literature review* (4th ed.). Retrieved from https://www.psychology.org.au.

Barendregt, J. J., Doi, S. A., Lee, Y. Y., Norman, R. E., & Vos, T. (2013). Meta-analysis of prevalence. *Journal of Epidemiology and Community Health, 67,* 974–978. doi:10.1136/jech-2013-203104

Barker, T. H., Migliavaca, C. B., Stein, C., Colpani. V., Falavigna, M., Aromataris, E., & Munn, Z. (2021). Conducting proportional meta-analysis in different types of systematic reviews: A guide for synthesisers of evidence. *BMC Medical Research Methodology, 21,* e189. doi: 10.1186/s12874-021-01381-z

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology, 56,* 893–897. doi: 10.1037//0022-006x.56.6.893

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy for depression.* New York: Guilford Press.

Beck, A. T., & Steer, R. A. (1993). *Beck Depression Inventory - Manual*. San Antonio: The Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. E., & Erbaugh, J. K. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571. doi: 10.1001/archpsyc.1961.01710120031004

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2022). *Comprehensive meta-analysis (version 4).* Englewood, NJ: Biostat.

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders:

an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy, 47*, 1-18. doi: 10.1080/16506073.2017.1401115

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18. doi:10.1037/0022-006X.66.1.7

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52,* 685–716. doi:10.1146/annurev.psych.52.1.685

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290. doi.org/10.1037/1040-3590.6.4.284.

Cristea, J. A., Stefan, S., Karyotaki, E., David, D., Hollon, S. D., & Cuijpers, P. (2017). The effects of cognitive behavioral therapy are not systematically falling: a revision of Johnsen and Friborg (2015). *Psychological Bulletin, 143*, 326–340. http://dx.doi.org/10.1037/bul0000062

Cuijpers, P., Karyotaki, E., de Wit, L., & Ebert, D.D. (2020). The effects of fifteen evidence-supported therapies for adult depression: A meta-analytic review. *Psychotherapy Research, 30*, 279–293. https://doi.org/10.1080/10503307.2019.1649732

Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review, 30,* 768–778. doi:10.1016/j.cpr.2010.06.001.

Cuijpers, P., Weitz, E., Cristea, I. A., & Twisk, J. (2017). Pre-post effect sizes should be avoided in meta-analyses. *Epidemiology and Psychiatric Sciences, 26,* 364–368. doi.org/10.1017/S2045796016000809

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. doi: 10.1111/j.0006-341X.2000.00455.x

D'Zurilla, T. J. (1986). *Problem-solving therapy: A social competence approach to clinical intervention.* New York: Springer.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634. doi.org/10.1136/bmj.316.7129.469.

Endicott, J., Nee, J., Harrison, W., & Blumenthal, R. (1993). Quality of Life Enjoyment and Satisfaction Questionnaire: a new measure. *Psychopharmacology Bulletin, 29,* 321–326. doi: 10.1111/j.1365-2850.2011.01735.x

EuroQol Group. (1990). EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy, 16,* 199–208. doi: 10.1016/0168-8510(90)90421-9.

Frisch, M. B., Cornell, J., Villanueva, M,, & Retzlaff, P. J. (1992). Clinical validation of the Quality of Life Inventory: a measure of life satisfaction for use in treatment planning and outcome assessment. *Psychological Assessment, 4,* 92-101. https://doi.org/10.1037/1040-3590.4.1.92

Gaskel, C., Simmonds-Buckley, M., Kellett, S., Stockton, C., Somerville, E., Rogerson, E., &·Delgadillo, J. (2022). The effectiveness of psychological interventions delivered in routine practice: Systematic review and meta-analysis. *Administration and Policy in Mental Health and Mental Health Services Research*, https://doi.org/10.1007/s10488-022-01225-y

GBD 2019 Diseases and Injuries Collaborators (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet, 396*, 1204–1222. doi: 10.1016/S0140-6736(20)30925-9.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23*, 50–55. doi: 10.1136/jnnp.23.1.56.

Hans, E., & Hiller, W. (2013a). A meta-analysis of nonrandomized effectiveness studies on outpatient cognitive behavioral therapy for adult anxiety disorders. *Clinical Psychology Review, 33*, 954–964. doi.org/10.1016/j.cpr.2013.07.003.

Hans, E., & Hiller, W. (2013b). Effectiveness of and dropout from outpatient cognitive behavioral therapy for adult unipolar depression: A meta-analysis of nonrandomized effectiveness studies. *Journal of Consulting and Clinical Psychology, 81*, 75–88. doi: 10.1037/a0031080.

Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, J., Stohl, M., & Grant, B. F. (2018). Epidemiology of adult *DSM-5* major depressive disorder and its specifiers in the United States. *JAMA Psychiatry*, 75, 336-346. doi:10.1001/jamapsychiatry.2017.4602

Hofmann, S. G., Curtiss, J., Carpenter, J. K., & Kind, S. (2017). Effect of treatments for depression on quality of life: a meta-analysis, *Cognitive Behaviour Therapy, 46*, 265-286. doi: 10.1080/16506073.2017.1304445

Hunsley, J., & Lee, C. M. (2007). Research-informed benchmarks for psychological treatments, efficacy studies, effectiveness studies, and beyond. *Professional Psychology: Research and Practice, 38*, 21–33. doi: 10.1037/0735-7028.38.1.21

IsHak, W. W., Greenberg, J. M., Balayan, K., Kapitanski, N., Jeffrey, J., Fathy, H., Fakhry, H., Rapaport, M., & Hyman, M. (2011). Quality of life: The ultimate outcome measure of interventions in major depressive disorder. *Harvard Review of Psychiatry, 19*, 229-239. doi: 10.3109/10673229.2011.614099

Jaccard, J., & Guilamo-Ramos, V. (2002). Analysis of variance frameworks in clinical child and adolescent psychology: issues and recommendations. *Journal of Clinical Child & Adolescent Psychology, 31*, 130-146. doi:10.1207/S15374424JCCP310115

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19. doi: 10.1037/0022-006X.59.1.12

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: a meta-analysis. *Psychological Bulletin, 141*, 747–768. http://dx.doi.org/10.1037/bul0000015

Johnsen, T. J. & Thimm, J. C. (2018). A meta-analysis of group cognitive–behavioral therapy as an antidepressive treatment: Are we getting better? *Canadian Psychology, 59,* 15–30. https://doi.org/10.1037/cap0000132

Kan, K., Lokkerbol, J., Jörg, F.,·Visser, E., Schoevers. R. A., & Feenstra, T. L. (2021). Real-world treatment costs and care utilization in patients with major depressive disorder with and without psychiatric comorbidities in specialist mental healthcare. *PharmacoEconomics, 39*, 721–730. https://doi.org/10.1007/s40273-021-01012-x

Kazdin, A. E. (2008). Evidence-based treatment and practice: new opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63*, 146–159. doi: 10.1037/0003-066x.63.3.146

Kazdin, A. E. (2022). Expanding the scope, reach, and impact of evidence-based psychological treatments. *Journal of Behavior Therapy and Experimental Psychiatry, 76,* e101744. https://doi.org/10.1016/j.jbtep.2022.101744

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A. J., Walters, E. E,, & Wang, P.S. (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R)**.** *JAMA, 289*, 3095-3105. doi: 10.1001/jama.289.23.3095.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*, 593-602. doi:10.1001/archpsyc.62.6.593

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine, 16,* 606–613. doi: 10.1016/j.genhosppsych.2015.11.005

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas. *Frontiers in Psychology, 4,* e863. doi. org/10.3389/fpsyg.2013.00863.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P., ... Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *British Medical Journal, 339,* b2700. doi.org/10.1136/bmj.b2700.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA, US: Sage Publications, Inc.

Ljótsson, B., Hedman, E., Mattsson, S., & Andersson, E. (2017). The effects of cognitive–behavioral therapy for depression are not falling: a re-analysis of Johnsen and Friborg (2015). *Psychological Bulletin, 143,* 321–325. http://dx.doi.org/10.1037/bul0000055

Martell, C. R., Addis, M. E., & Jacobson, N. S. (2001). *Depression in context: Strategies for guided action*. New York: Norton.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy, 28,* 487–495. doi: 10.1016/0005-7967(90)90135-6.

National Institute for Health and Care Excellence (2022). *Depression in adults: treatment and management.* NICE guideline. Retrieved from www.nice.org.uk/guidance/ng222

Öst, L-G. (2008). Efficacy of the third wave of behavioral therapies: A systematic review and meta-analysis. *Behaviour Research and Therapy, 46*, 296–321. doi: 10.1016/j.brat.2007.12.005

Öst. L-G., Enebrink, P., Finnes, A, Ghaderi, A., Havnen, A., Kvale, G., Salomonsson, S. & Wergeland, G.J. (2022). Cognitive behavior therapy for obsessive-compulsive disorder in routine clinical care: A systematic review and meta-analysis. *Behaviour Research and Therapy, 159,* e104170. doi: 10.1016/j.brat.2022.104170

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401. https:// doi. org/ 10. 1177/01466 21677 00100 306

Riise, E.N., Wergeland, G. J. H., Njardvík, U., & Öst, L-G. (2021). Cognitive behavior therapy for externalizing disorders in children and adolescents in routine clinical care: A systematic review and meta-analysis. *Clinical Psychology Review*, *83,* e101954. doi: 10.1016/j.cpr.2020.101954.

Rubin, A., & Yu, M. (2017). Within-group effect size benchmarks for cognitive–behavioral therapy in the treatment of adult depression. *Social Work Research, 41,* 136-144. https://doi.org/10.1093/swr/svx011

Sawilowsky, S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8, e26. doi: 10.22237/jmasm/1257035100

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: a meta-analysis. *Psychological Bulletin, 126*, 512-529. doi: 10.1037/0033-2909.126.4.512

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., .. Henry, D. A. (2017). Amstar 2: A critical appraisal tool for systematic reviews that include randomized or non-randomised studies of healthcare interventions, or both. *British Medical Journal, 358*, j4008. doi.org/10.1136/bmj.j4008.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Lowe, B., (2006). A brief measure for assessing generalized anxiety disorder - the GAD-7. *Archives of Internal Medicine, 166,* 1092–1097. https://doi.org/10.1001/archinte.166.10.1092.

Sterne, J. A. C. et al. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised atudies of intervenyion. *British Medical Journal, 355*, i4919. doi: 10.1136/bmj.i4919

Sterne, J. A. C., Savovic, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., ….Higgins, J. P. T. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *British Medical Journal, 366*, l4898. doi:10.1136/bmj.l4898

Stewart, R. E., & Chambless, D. L. (2009). Cognitive-behavioral therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies. *Journal of Consulting and Clinical Psychology, 77*, 595–606. doi.org/10.1037/a0016032.

Svanborg, P., & Åsberg, M. (1994). A new self-rating scale for depression and anxiety states based on the comprehensive psychopathological rating scale. *Acta Psychiatrica Scandinavica, 89*, 21–28. doi: 10.1111/j.1600-0447.1994.tb01480.x.

Thimm, J. C., & Johnsen, T. J. (2020). Time trends in the effects of mindfulness-based cognitive therapy for depression: A meta-analysis. *Scandinavian Journal of Psychology, 61*, 582–591. doi: 10.1111/sjop.12642

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: recommendations for a new model. *Clinical Psychology: Science and Practice, 22,* 317–338. doi: 10.1111/cpsp.12122

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35,* 215-247. doi:10.3102/1076998609346961

van der Zee, K. I., & Sanderman, R. (1993). *Assessing the general health condition using the RAND-36: a manual*. Northern Center for health questions: Groningen.

von Brachel, R., Hirschfeld, G., Berner, A., Willutzki, U. Teismann, T., Cwik, J-C., Velten, J. Schulte, D., & Margraf, J. (2019). Long-term effectiveness of cognitive behavioral therapy in routine outpatient care: A 5- to 20-year follow-up study. *Psychotheraåy & Psychosomatics, 88*, 225–235. doi: 10.1159/000500188

Vittengl, J. R., Jha, M. K., Minhajuddin, A., Thase, M. E., & Jarrett, R. B., (2021). Quality of life after response to acute-phase cognitive therapy for recurrent depression. *Journal of Affective Disorders, 278*, 218-225. https://doi.org/10.1016/j.jad.2020.09.059

Wakefield, S., Kellett, S., Simmonds-Buckle, M., Stockton, D., Bradbury, A., & Delgadillo, J. (2021). Improving Access to Psychological Therapies (IAPT) in the United Kingdom: A systematic review and meta-analysis of 10-years of practice-based evidence. *British Journal of Clinical Psychology, 60,* 1–37. doi:10.1111/bjc.12259

Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 health survey: manual and interpretation guide*. Boston: The Health Institute, New England Medical Center.

Wergeland, G. J. H., Riise, E. N., & Öst, L-G. (2021). Cognitive behavior therapy for internalizing disorders in children and adolescents in routine clinical care: A systematic review and meta-analysis. *Clinical Psychology Review, 83,* e101918. https://doi.org/10.1016/j.cpr.2020.101918

Wergeland, G. J. H., Posserud, M.-B., Fjermestad, K., Njardvik, U., & Öst, L.-G. (2022). Early behavioral interventions for children with autism spectrum disorder in routine clinical care: A systematic review and meta-analysis. *Clinical Psychology: Science and Practice*, http://dx.doi.org/10.1037/cps0000106

Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69,* 875–899. doi: 10.1037//0022-006X.69.6.875

WHOQOL Group (1998). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychological Medicine, 28*, 551–558. doi: 10.1017/s0033291798006667

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica, 67*, 361–370. https://doi.org/10.1111/j.1600-0447.1983. tb09716.x.