The effectiveness of psychological interventions delivered in routine practice: Systematic review and meta-analysis

Chris Gaskell*, Melanie Simmonds-Buckley, Stephen Kellett, Corrie Stockton,

Erin Somerville, Emily Rogerson, & Jaime Delgadillo

Clinical and Applied Psychology Unit, University of Sheffield, United Kingdom

Acknowledgments: We would like to thank Dr. Gregg Rawlings for his work in second rating at the screening stage for titles and abstracts.

Competing Interests: The authors have no competing interests to declare that are relevant to the content of this article.

** Corresponding author: Chris Gaskell, University of Sheffield, UK, chris-gaskell@hotmail.co.uk*

Note: supplemental materials can be provided by written request to the corresponding author.

Abstract

Purpose: This review presents a comprehensive evaluation of the effectiveness of routinely delivered psychological therapies across inpatient, outpatient and University-based clinics. **Methods:** This was a pre-registered systematic-review of studies meeting pre-specified inclusion criteria (CRD42020175235). Eligible studies were searched in three databases: MEDLINE, CINAHL and PsycInfo. Pre-post treatment (uncontrolled) effect sizes were calculated and pooled using random effects meta-analysis to generate effectiveness benchmarks. Moderator analyses were used to examine sources of heterogeneity in effect sizes. **Results:** Overall, 252 studies (*k* = 298 samples) were identified, of which 223 (*k* = 263 samples) provided sufficient data for inclusion in meta-analysis. Results showed large pre-post treatment effects for depression (*d* = 0.96, [CI 0.88-1.04], *p* = < 0.001, *k* = 122), anxiety (*d* = 0.8 [CI 0.71-0.9], *p* = < 0.001, *k* = 69), and other outcomes (*d* = 1.01 [CI 0.93-1.09], *p* = < 0.001, *k* = 158). **Conclusions:** This review provides support for the effectiveness of routinely delivered psychological therapy. Effectiveness benchmarks are supplied to support service evaluations across multiple settings.

Keywords: 'Psychotherapy,' 'Effectiveness,' 'Naturalistic,' 'Routine Outcomes,' 'Meta-analysis.'

Introduction

Meta-analyses of clinical trials support the efficacy of psychological interventions for various mental health problems such as depression (Cuijpers et al., 2008), anxiety disorders (e.g., Cuijpers, Sijbrandij, et al., 2014; Mayo-Wilson et al., 2014; Olatunji et al., 2014; Sánchez-Meca et al., 2010; Wolitzky-Taylor et al., 2008), post-traumatic stress disorder (Lewis et al., 2020), obsessive-compulsive disorder (Rosa-Alcázar et al., 2008), eating disorders (Linardon et al., 2017) and other conditions. Grounded in this evidence, clinical guidelines support the use of psychological interventions in routine clinical care (e.g., Chambless & Hollon, 1998; Chambless & Ollendick, 2001; 2011; National institute for Health and Care Excellence, 2011). These guidelines commonly advocate the implementation of empirically supported treatments, closely following the procedures implemented in clinical trials and specified in associated treatment manuals. To this end, competency frameworks have been developed to support the dissemination of empirically supported treatments in routine care and clinical training programmes (e.g., Lemma et al., 2008; Roth et al., 2009; Roth & Pilling, 2008).

Some studies have found similar treatment outcomes when comparing data from efficacy trials and routine practice (e.g., Lutz et al., 2016; Persons et al., 1999). However, there are some reasons to assume that the effects of psychotherapy delivered in routine care settings may differ from those observed in clinical trials. Recent evidence indicates that psychological treatment outcomes are associated with treatment *integrity*, which refers to the competent (skilled) delivery of protocol-driven treatment procedures (Power et al., 2020). However, surveys of clinicians working in routine settings often reveal negative attitudes towards protocol-driven treatment and a lack of adherence to treatment manuals (e.g., Addis & Krasnow, 2000). Hence, the integrity of routinely delivered psychological treatments is unclear, and it probably varies across services (Freedland et al., 2011). Furthermore, the strict selection criteria applied in clinical trials may result in unusually homogeneous samples that

do not reflect the diverse clinical populations typical of routine care settings (e.g., Lambert, 2013; Zimmerman et al., 2002). Previous studies have found systematic differences in the clinical profiles of patients included and excluded from psychotherapy trials (e.g., van der Lem et al., 2012). For these reasons, it is plausible to assume that the effects of routinely delivered therapy may vary across settings and clinical populations, and may not necessarily conform to benchmarks from efficacy trials.

A tradition of practice-based evidence (PBE, Margison et al., 2000) has emerged in recent decades, with numerous studies examining the effects of routinely delivered psychological interventions in various settings. Narrative reviews of PBE generally confirm that moderate-to-large uncontrolled (pre-to-post treatment) effect sizes are observed in routine care settings, supporting the effectiveness of psychotherapy but also demonstrating considerable variability across patient samples, therapists and clinics (e.g., see Barkham et al., 2010; Castonguay et al., 2013, 2021). An inherent limitation of such narrative reviews is that they perform a selective rather than systematic synthesis of available data. Benchmarking studies can be useful to provide general indices of treatment effectiveness, enabling services to evaluate their outcomes relative to efficacy trials (e.g., Minami et al., 2008, McAleavey et al., 2019) or aggregated effect size data from similar clinical services (e.g., Delgadillo et al., 2014). Psychotherapy benchmarking studies tend to report favorable pooled effects sizes, but also show variability in effects across clinics (e.g., Barkham et al., 2001; Connell et al., 2007; Delgadillo et al., 2014; Gyani et al., 2013). Although benchmarking studies help to quantify the expected magnitude of treatment effects observed in routine clinical settings, most are nevertheless circumscribed to small sets of clinics or geographical areas, offering limited insights into possible sources of heterogeneity in treatment outcomes. Systematic reviews and meta-analyses may therefore offer a more comprehensive examination of the effectiveness of routinely delivered treatments.

Some meta-analytic investigations have reported that outcomes from routine practice-based treatments are not as favorable as those delivered in research settings (Weisz et al., 1995). Other meta-analyses suggest that there are no differences in treatment effects when comparing PBE and efficacy studies after controlling for case-mix differences (e.g., Shadish et al., 1997, 2000). However, many of the PBE studies in these meta-analyses applied stringent controls on the treatment procedures (e.g., adherence and competence assessments) – making them more akin to efficacy trials. Hunsley & Lee (2007) reviewed 35 studies and concluded that the completion and improvement rates observed in PBE studies were comparable to efficacy trials. Cahill et al. (2010) reviewed 31 studies, concluding that psychotherapy was most effective for the treatment of common mental disorders, with a pooled uncontrolled effect size of $d = 1.29$. More recently, Wakefield et al. (2021) reviewed 60 studies, of which 47 were eligible for meta-analysis. They reported large uncontrolled effect sizes for depression ($d = 0.87$) and anxiety ($d = 0.88$), and a moderate effect on functional impairment ($d = 0.55$). These meta-analyses show wide variability in treatment effects (i.e., heterogeneity) across studies/samples.

PBE meta-analyses provide some insights into plausible sources of heterogeneity, including methodological (e.g., completers analyses vs. inclusion of patients lost to follow-up) and clinical features (e.g., larger effects for common mental disorders, lower effects for patients with comorbidities and socioeconomic disadvantages, larger effects for lengthier interventions). Nevertheless, these meta-analyses are over a decade old (Cahill et al., 2010; Hunsley & Lee, 2007) or limited to a specific treatment setting (e.g., primary care outpatient services; Wakefield et al., 2021). Further research into the methodological and clinical sources of treatment heterogeneity is needed to better understand why treatment effects vary across samples, and to determine whether or not these effects vary across different treatment settings (e.g., outpatient, inpatient, university-based treatment).

The considerable growth of the PBE literature in the last decade and Implementation of empirically supported treatments across many settings warrants a comprehensive review of treatment outcomes data. The aim of the present study was to systematically review available PBE studies. The objectives of the study were to [1] provide benchmarks of treatment effectiveness using meta-analysis and [2] to examine sources of effect size heterogeneity using pre-specified moderator analyses informed by earlier studies.

Methods

Search strategy and study selection

The present study followed good practice guidelines for systematic reviews (PRISMA, Page et al., 2021) and meta-analyses of psychotherapy studies (MAP-24, Flückiger et al., 2018). A review protocol was pre-registered in the PROSPERO database (CRD42020175235)[1]. Literature searches were carried out without any restrictions on date of publication up to the search date (April 2020). Inclusion criteria were: (a) studies reporting outcomes for routinely delivered treatments (i.e., not as part of efficacy trials); (b) all adult sample (no patients under 16); (c) employed a *psychological treatment* (i.e., driven by psychological theory and intended to be therapeutic [Spielmans & Flückiger, 2018], as inferred or described by study manuscripts); and (d) conducted face-to-face. Studies were excluded if they: usedI) family/group treatments, (f) were not available in English; (g) did not employ a self-report measure of treatment effectiveness[2]; (g) did not provide sufficient data to calculate pre-post

[1] Protocol available at: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=175235

[2] The authors recognise that use of the term *effectiveness* may be somewhat misleading. The pre-post (uncontrolled) methodology which forms the body of evidence in this review is unable to disentangle treatments effects from other potential causes of change (e.g., regression to the mean, placebo). Observed change in symptoms may therefore not exclusively represent treatment effectiveness. We have opted to retain use of this term

treatment effect sizes; or (f) emplo7isualize7ed7tion procedures or control groups. A more detailed table of inclusion/exclusion criteria is available in supplementary Table 1.

The search strategy had three phases. Phase one was a systematic search of three electronic literature databases (MEDLINE, CINAHL and PsycInfo) via EBSCO using a pre-registered list of key terms. Methodological terms included: *practice-based evidence*, *routine practice*, *benchmarking*, *transportability*, *transferability*, *clinically representative*, *managed care setting*, *uncontrolled*, *external validity*, *applicable findings*, *empirically supported*, *dissemination,* and *clinical effectiveness evaluation*. These terms were informed by prior reviews of psychotherapy effectiveness (Cahill et al., 2010; Stewart & Chambless, 2009). *Effectiveness* and *evaluation* were not used as single word terms due to producing unmanageable numbers of irrelevant records. For the psychologically relevant term: *psycho** OR *therap** was used for PsycInfo while *psycho** alone was used for MEDLINE and CINAHL (*therap** was removed from MEDLINE/CINAHL due to producing an unmanageable number of irrelevant records). Limiters included *adult population* and *English language*. No exclusions were made based on the type of publication. Key term combinations and Boolean operators are reported in supplementary Table 2. Phase two included a manual search of reference lists, and forward citation searching (using Google Scholar) for studies identified in phase one. Titles relevant to the current review were identified by the first author. Finally, phase three was a grey literature search using the terms *psychotherapy* AND *routine-practice* AND *effectiveness* in Google Scholar.

After removal of duplicates, titles and abstracts of potentially eligible studies were screened by the first author using a pre-developed and piloted screening tool. Sub-samples

---

within the current review because it has consistently and frequently been used as such in the extant literature (e.g., Lambert, 2013; Nordmo et al., 2020).

were screened by a second coder at each stage (20% at the stage of title screening; 10% at the stage of full-text screening). Percentage agreement and inter-rater reliability statistics (Kappa [κ], Cohen, 1960) indicated good reliability (κ = 0.78, 1713/1740, 98.45%) in the first stage and adequate reliability (κ = 0.65, 24/30, 80%) in the second stage. After the selection process was completed, corresponding authors for eligible studies were contacted via email to request additional recommendations for potentially eligible studies, and to request additional statistical information to calculate effect sizes. E-mail responses were received from 76 authors and additional data was provided for 41 samples.

Data extraction

There were three separate outcome domains (and subsequently three meta-analyses) for '*depression'*, '*anxiety'* and '*other'* outcomes. The latter category consisted of general psychological distress scales, measures of functioning/quality of life, or diagnosis-specific outcome scales (e.g., obsessive-compulsive disorder, etc.)8isualize8edised extraction sheet was developed and pilot-tested with a sample of studies (*k* = 10). When multiple samples were reported in the same study, effect-sizes across these samples were aggregated to reduce bias of statistical dependency (Gleser & Olkin, 2009; Hoyt & Del Re, 2018). To avoid loss of information (e.g., aggregating sub-samples that are distinct based on levels of a moderator), study samples were disaggregated for moderator analyses (Cooper, 1998). Studies with overlapping datasets (e.g., reanalysis of the same sample) were only included once in the meta-analysis. Samples which performed an intention-to-treat (ITT) analysis were preferred to completer samples due to being less prone to attrition bias (Jüni et al., 2001); so the ITT data was extracted for studies that reported both ITT and completer analyses. As extraction of multiple study effect-sizes within a single domain (e.g., depression) threatens statistical dependency (Borenstein et al., 2021) we selected a single effect-size per domain (Card, 2015; Cuijpers, 2016), using a preference system (defined a priori, supplementary material).

Reliability of coding for effect-size data was computed using a second coder for a sub-sample of manuscripts (n = 29) demonstrating almost perfect reliability across all values (κ = 0.97, agreement = 97.56%) and perfect reliability for effect-size values (κ = 1.00). Key categorical and numerical variables extracted from manuscripts for moderator analyses are reported in Table 1. For sample severity, the decision was made to cluster university counselling centers in the 'mild' severity category due to prior research finding normative data of UK University students comparable to primary care samples (Connell et al., 2007).

Risk of bias and quality assessment

The Joanna Briggs Institute Quality Appraisal Tool for Case Series (Munn et al., 2020) was used to assess risk of bias. Eight criteria primarily focusing upon manuscript reporting detail were used. Criteria included manuscript reporting of: (i) patient inclusion criteria, (ii) service description, (iii) treatment description, (iv) sample characteristics, (v) outcome data, (vi) effect-size calculation, (vii) consecutive patient recruitment, and (viii) inclusion of patients lost to follow-up in statistical analysis. Each item was coded as either met or not met (including not clear) by the first author for each sample. A sub-sample (23.8%) was rated independently by two other reviewers (11.9% each). The pooled agreement was 84.17% (κ = 0.62).

Statistical analysis

All analyses were conducted using the R statistical analysis environment (R Core Team, 2020, v 4.0.2). We calculated 9isualize9edised mean change (SMC: Becker, 1988) for included studies using the *metafor* package. This approach divides the pre-post mean change score by the pretreatment standard deviation with a sampling variance adjustment using the correlation between the pre-treatment and post-treatment measures (Morris, 2008). When unavailable, Pearson's *r* was imputed using an empirically derived estimate (*r* = .60, Balk et

al., 2012). Aggregation of samples/sampling errors was conducted using the *aggregate* function of *metafor* using standard inverse-variance weighting.

Random effects meta-analyses were performed using the *metafor* (Viechtbauer, 2020), *dmetar* (Harrer et al., 2019a), and *meta* (Schwarzer, 2020) packages. Forest plots were used10isualizelise pre-post treatment effects sizes across samples. Effect size heterogeneity was assessed using I^2 (Higgins & Thompson, 2002) and the Q statistic (Cochran, 1954). Publication bias was examined using funnel plots and assessed statistically using rank correlation tests (Begg & Mazumdar, 1994), Egger's regression test for funnel plot asymmetry (Egger et al., 1997), and the fail-safe N (Rosenthal method, Rosenthal, 1979).

Moderator analyses were based on a set of moderator variables selected a priori, following evidence from prior reviews. Subgroup variables included: (i) *analysis* (inclusion of patients lost to follow-up), (ii) *geographical region*, (iii) *severity* (mild, moderate, severe, university[3]) (iv) *treatment modality*, (v) *experience* (unqualified [i.e., trainees] vs. qualified therapists) (vi) *stage of treatment development* (preliminary study vs. routine evaluations), and (vii) *sample size* (small, medium, large). Continuous meta-regression variables included (i) *publication year*, (ii) average *age* of sample, and (iii) percentage of samples who identified as *female*. All moderators were included in meta-regression which was based on a mixed effects (i.e., multilevel) model (Borenstein et al., 2021) with weighted estimation (inverse-variance weights).

Finally, we developed effect size benchmarks to support the evaluation of effectiveness across four broad settings: outpatient services, inpatient services and university

[3] University clinics refers to university managed clinics treating communities beyond the student population. University counselling centres that are more specifically targeted at the student population are included within the mild category.

counselling services [i.e., student population] and university psychotherapy clinics [non-student population]). Informed by previous benchmarking studies (Delgadillo et al., 2014), pooled effect sizes (using random effects meta-analyses) were stratified into quartiles to differentiate between low effectiveness (bottom 25%), average effectiveness (middle 50%) and high effectiveness benchmarks (top 25%).

Results

Search results

The PRISMA diagram in Figure 1 presents a summary of the study selection process. Overall, 10,503 records were identified, of which 252 manuscripts were eligible for inclusion and 223 (samples $k = 263$) had sufficient information to be included in the meta-analysis. Summary statistics are provided in Table 2.

Study characteristics

Eligible studies were published between 1984 and 2020 (median = 2013, $k = 294$ published ≥ 2000). Of these, 169 samples included patients lost to follow-up ($k = 118$, 56.72% completers). Most studies were from the USA ($k = 113$, 37.92%), England ($k = 78$, 26.17%), Germany ($k = 24$, 8.05%), Sweden ($k = 12$, 4.02%) and Canada ($k = 10$, 3.36%). These five most represented countries accounted for most of the included samples ($k = 237$, 79.53%).

Sample characteristics

Sample characteristics were reported for 291 samples, with a cumulative N of 233,140 patients (mean = 838.63, median = 81.5, range – 4 - 33,243, IQR = 224.5). The prevalence of female participants was 61.88% (N = 144,273, $k = 279$) with 13 all-female samples and 2 all-male samples. The mean average sample age was 35.33 years (range = 19–00 - 60.50).

Across studies which provided information, 23.00% of patients were from ethnic minorities (k = 127), 37.00% were married (k = 106), and 23.00% were in employment (k = 96).

Treatment characteristics

Most samples evaluated cognitive-behavioral interventions (k = 152, 51.01%) while 50 samples evaluated psychodynamic (16.78%), and 25 samples evaluated counselling (8.29%; other = 71, 23.82%). Counselling interventions were interventions described simply as 'counselling' by study authors (with no further treatment information) or 'person-centered counselling' interventions. Interventions termed 'counselling' but described in a way that fit closely with of the other treatment modalities (e.g., cognitive-behavioral counselling) was assigned to the more specific treatment modality group. For symptom severity, 96 (32.21%) samples came from services treating mild conditions, 92 (30.87%) from services treating moderate conditions, 33 (11.07%) from services treating severe conditions, and 68 (22.82%) from university psychotherapy clinics (not counselling centers) that treated a wide spectrum of conditions from mild-to-severe (other, k = 9, 3.02%). Treatment dosage, when reported (k = 256) was in hours/sessions (k = 225), months (k = 12) or days (k = 8). The pooled (non-weighted) average dose (hours) was 16.30 sessions (median = 13.00, range = 1.00-139.30, IQR = 11.00). A total of 62 (20.81%) samples reported that treatment was delivered exclusively by trainees, while 100 (35.58%) samples reported having at least one trainee.

Risk of bias

In order of satisfactory criteria (e.g., the criterion under evaluation was met), the following risk of bias domains were assessed: demographic reporting detail (264/298, agreement = 98.33%, κ = 0.88), service reporting detail (260/298, agreement = 85%, κ = 0.31), study outcome reporting details (240/298, agreement = 83.33%, κ = -0.03), intervention reporting detail (234/298, agreement = 85%, κ = 0.32), service inclusion criteria (214/298, agreement =

90%, κ = 0.64), appropriate use of analysis (214/298, agreement = 70%, κ = 0.26), complete inclusion (i.e. consecutive recruitment and inclusion of those lost to follow-up, 41/298, agreement = 85%, κ = 0.45), and consecutive inclusion (93/298, agreement = 76.67%, κ = 0.51).

Meta-analyses

The random-effects meta-analysis for depression outcomes (k = 140, N = 68,077), across 10 unique measurement tools was statistically significant (p = < 0.001), indicative of a large pre-post treatment (d = 0.96,==CI = 0.88-1.04) reduction in depression severity. There was a large magnitude of statistically significant heterogeneity (I^2 = 97.94%, Q[df = 121] = 2,677.37, p = < 0.001). The funnel plot (Figure 2) shows limited visual evidence of asymmetry. The funnel rank correlation test was not statistically significant (τ = 0.061, p = 0.46) however the funnel regression test was statistically significant (Z = 2.13, p = 0.033). The fail-safe N was 515,853.

The random-effects meta-analysis for anxiety outcomes (k = 84, N = 26,689, measurement tools = 20) was statistically significant (p = < 0.001), indicative of a large (d = 0.80, CI = 0.71-0.90) reduction in symptom severity. Heterogeneity was large and statistically significant (I^2 = 97.51%, Q[df = 68] = 1,328.96, p = < 0.001). The funnel plot shows limited evidence of asymmetry. The funnel rank correlation test was not significant (τ = 0.009, p = 0.888). In contrast, the funnel regression test was statistically significant (Z = 2.533, p = 0.011). The fail-safe N was 121,899.

The random-effects meta-analysis for other outcomes (k = 184, N = 126,734, measurement tools = 40) was statistically significant (p = < 0.001), indicative of a large (d = 1.01, CI = 0.93-1.09) reduction in severity of indices of distress. Heterogeneity was large and statistically significant (I^2 = 99.06%, Q[df = 157] = 15,330.32, p = < 0.001). The funnel plot

shows a degree of asymmetry with clustering to the right of the mid-line. The funnel rank correlation test was statistically significant (τ = 0.208, p = <0.001). In contrast, the funnel regression test was not significant (Z = 3.697, p = <0.001). The fail-safe N was 1,695,607.

Moderator analyses

Multivariable meta-regressions were conducted for each of the three outcome domains (Tables 3-5). After controlling for other moderators, the depression meta-regression found a significant effect for geographical region, therapist experience and type of analysis. UK samples had larger effect sizes compared to samples from Asia; effects sizes in samples treated by qualified staff members were larger than those observed in samples exclusively consisting of trainees; and samples excluding patients lost to follow-up (i.e., completer analyses) had larger effect sizes compared to intention-to-treat analyses. For anxiety outcomes, UK studies had larger effect sizes than studies from mainland Europe; mild severity samples had larger effect sizes than samples of patients with moderate or severe symptoms; and cognitive-behavioural interventions had larger effect sizes than counselling interventions. Finally, for other outcomes, the only significant moderator indicated that cognitive-behavioural interventions had larger effect sizes than psychodynamic interventions and unspecified (i.e., other) interventions.

Benchmarking data

Pooled effect-sizes for low, average and high performing services are shown in Table 6, organized according to setting (outpatient services, inpatient services, university counselling services [i.e., student population] and university psychotherapy clinics [non-student population]). Although the effect size estimates for each benchmark vary across settings, confidence intervals consistently overlapped, indicating similar levels of symptom-changes across the performance strata (low, average, high). The exception to this is the low

performance benchmark for anxiety measures which were significantly larger in university psychotherapy clinics (d = 0.51) and significantly smaller in inpatient services (d = 0.13) by comparison to outpatient services (d = 0.37).

Discussion

This review provides a comprehensive quantitative review of the effectiveness of psychological treatments delivered in routine care settings. Overall, 252 studies (samples k = 298) were identified, of which 223 (88.5%, k = 263) were included in the meta-analysis. Consistent with prior psychotherapy effectiveness reviews, we found large uncontrolled (pre-post treatment) effect sizes (d = 0–80 - 1.01) across multiple outcome domains (depression, anxiety, and general psychological distress).

Consistent with previous meta-analyses of PBE (e.g., Cahill et al., 2010; Hunsley & Lee, 2007; Wakefield et al., 2021), we observed wide variability in effect sizes across studies and large (>90%) indices of heterogeneity across outcome domains. The large number of samples included in this review enabled us to carry out adequately-powered moderator analyses to better understand potential sources of heterogeneity. For depression outcomes, smaller effect sizes were found for samples in Asia (compared to the UK), and in treatments delivered by trainees (i.e., compared to qualified professionals). For anxiety outcomes, smaller effect sizes were found for treatments delivered in mainland Europe (compared to the UK), services treating patients with moderate or high levels of severity (compared to mild severity), and counselling interventions (compared to cognitive-behavioural interventions). For other outcomes, only therapy modality was significant. Psychodynamic and unspecified interventions produced smaller effect-sizes (compared to cognitive-behavioural interventions). To some extent, these results are consistent with and support clinical guidelines that recommend cognitive-behavioural therapy as a first-line intervention, prior to

considering other treatment modalities (National Institute for Health and care Excellence, 2011). However, caution is advised when interpreting these between-therapy comparisons using uncontrolled data from observational studies, as they could be explained by other unmeasured factors such as relevant case-mix differences between patients (e.g., socioeconomic status, personality, comorbid physical illnesses, etc.). Studies that control for case-mix variables using individual patient data find that there are no significant differences in treatment effects when comparing different treatment modalities (e.g., Pybis et al., 2017). Furthermore, as found in a previous meta-analysis (Wakefield et al., 2021), completers analyses tended to produce inflated (biased) effect sizes by comparison to intention-to-treat (more conservative and stringent) analyses.

The finding of large clinical improvements during psychotherapy and across outcomes was consistent with prior meta-analyses of psychotherapy effectiveness for depression outcomes (Hans & Hiller, 2013; Wakefield et al., 2021), anxiety outcomes (Stewart & Chambless, 2009; Wakefield et al., 2021), and other indices of psychological distress and functioning (Cahill et al., 2010). Pooled uncontrolled effect-sizes were smaller than that reported by Cahill et al. (2010) (d = 1.29), although this may reflect differences in the focus of the reviews (e.g., Cahill et al., 2010 included group treatments) or the changing distribution of geographical representation (i.e., more studies from non-UK/North American countries). Large clinical improvements are also consistent with many meta-analyses of psychotherapy controlled trials (e.g., Cuijpers, Sijbrandij, et al., 2014; Cuijpers et al., 2008; Mayo-Wilson et al., 2014; Olatunji et al., 2014).

It is possible that there are continental differences in models of training, service structures, therapy provision and emphasis on evidence-based practice which underlie the observed differences in pooled effect-sizes between continents. This is consistent with UK and US clinical guidance recommending delivery of empirically supported treatments (APA,

2006; NICE, 2011). It is possible that the service policy context in the UK places greater emphasis on the delivery of treatment with high fidelity to empirically supported treatment protocols, and this may explain the relatively larger effect sizes in this geographical location, since high integrity is associated with better treatment outcomes and especially for anxiety treatment outcomes (Power et al., 2022). Despite these differences, all continents demonstrated positive change for all outcomes (d = 0–59 - 1.10) supporting the *universality hypothesis* (i.e., that psychotherapy is assumed to work across cultures; Flückiger et al., 2018).

Consistent with several prior meta-analytic reviews (e.g., Cuijpers, Turner, et al., 2014; Driessen et al., 2010; Furukawa et al., 2017), symptom severity did not predict effectiveness of treatment for depression. For anxiety outcomes, services categorized as treating mild conditions consistently had larger effect sizes. It is possible that classifying by type of service provided an imprecise proxy for sample severity and therefore future research should explore severity as a continuous variable in routine settings.

Limitations

The most notable critique of this review is that it is based exclusively on evidence from observational studies. We are unable to rule out alternative explanations for observed effect sizes (placebo effects, spontaneous remission [Posternak & Miller, 2001; Whiteford et al., 2012]) and subsequently the observed effect sizes in this review cannot be directly compared to efficacy trials. Nevertheless, pooled effect sizes from observational studies serve as a valuable data source for benchmarking of routine care and quality improvement initiatives (e.g., Clark et al., 2018; Delgadillo et al., 2014; Gyani et al., 2013).

A key design limitation concerns statistical dependency. Efforts to avoid statistical dependency included: (i) taking one sample measure per domain, (ii) aggregating multiple

unique study samples within a single domain, and (iii) extracting one measurement tool per study, per construct (i.e., preference system). These approaches have well-documented limitations (Borenstein et al., 2021; Hoyt & Del Re, 2018; Van den Noortgate et al., 2013). A preferable approach would have been to model dependency using a multi-level analysis (Van den Noortgate et al., 2013, 2015) or through robust variance estimation and should be considered for future replications. Use of robust-variance estimation would avoid the need to assign outcomes to a restrictive number of outcome domains. This would also circumvent the need to adopt a highly heterogeneous "other" outcome domain, which for the current review included both diagnosis specific and global distress-based measures.

An additional limitation concerns the inherent limitations of the risk-of-bias assessment tool which was selected for this study a priori. It could be argued that this tool primarily indexes manuscript reporting detail and not necessarily risk of bias. Future reviews of effectiveness could consider assessing methodological rigour using other available rating tools (e.g., see Munder & Barth, 2018).

Due to resource constraints and the large number of included studies, the systematic search, data extraction and risk-of-bias ratings were not performed completely in duplicate. For the subsample of full texts screened by two coders there was a strong, but imperfect, agreement/reliability (80%, $\kappa = 0.65$). Similarly, not extracting data or assessing RoB in duplicate is problematic due to risk of imprecise estimates of treatment effect and RoB (Armijo-Olivo et al., 2014). An additional limitation surrounds coding decisions for moderator variables. Therapy modality was coded from manuscript self-definition. The degree to which treatments truly resembled treatment code (or treatment intended) is not clear. It was also apparent during extraction that very few practice-based studies report fidelity/adherence checks. As this becomes more routinely reported opportunities for modelling differences based on adherence/competence/integrity will become available. The

use of categorical moderator levels to differentiate samples at the study level may also have provided imprecise proxies for moderator levels. For example, patient severity would preferably be modelled through meta-regression at the patient level to account for the heterogeneity within samples as it has been shown that university counselling center samples have numerous highly distressed individuals (Xiao et al., 2017). Future studies investigating these moderator variables at the patient level (e.g., through individual participant data meta-analysis) would help to shed light on this.

The search strategy is unlikely to have identified every available study. Search terms were based on prior reviews and omitted several terms that were found to produce an unmanageable number of records (e.g., "effectiveness", "evaluation"). Despite this, the current reviews gives an adequate range and depth of effectiveness research with which to make tentative interpretations regarding the field of psychotherapy effectiveness research. A final caveat is the decision to focus exclusively on self-report measures of effectiveness. Meta-analytic evidence has demonstrated significant differences between self-report and clinician rated measures of clinical improvement (Cuijpers et al., 2010). Future research is therefore needed to see if the pooled effect-sizes from this study are consistent with clinician-rated measures of effectiveness in routine settings.

Conclusions

This review provides support for the effectiveness of psychological therapy as delivered in routine settings across a range of outcomes. Overall, the effects of psychotherapy appear to generalize well to diverse clinical settings, contexts, and populations. Nevertheless, it is evident that treatment effects vary considerably across services, and this review provides performance benchmarks to support routine service evaluation and practice development initiatives.

Table 1:

Summary coding sheet for extracting study information. These moderators form the subgroup and continuous variables moderator variables for the current study.

Categorical variables

- **Setting:** the study was (i) *out-patient*, (ii) *inpatient* or (iii) *mixed.*
- **Analysis:** samples (i) *included* or (ii) *excluded* (completers) patients lost to follow up.
- **Severity:** was determined through a stratification of studies based on characteristics of the service (similar to the approach used by de Jong et al., 2021). (i) *Mild services* included primary care, physical health, university counselling, voluntary, private (independent or group) and employee assistance programmes; (ii) *Moderate services* included secondary care, community mental health centers, specialist psychotherapy centers, managed care settings, or intensive outpatient programmes; (iii) *severe services* represented inpatient samples; and (iv) *university* included university outpatient and training clinics (which are known to vary in the severity of sample).
- **Treatment modality:** Treatments were coded as (i) *cognitive-behavioral* or (ii) *psychodynamic* based on manuscript self-designation (i.e., if the manuscript described treatment as CBT, then that was coded). In the absence of these terms, modality of best-fit was decided using treatment descriptions. Treatments that could not be confidently allocated to these groups were coded as (iii) *counselling* (e.g., person-centred, undefined) or (iv) *other*. Treatments that did not describe treatment modality were rated as other.
- **Continent:** Studies were coded as North America, United Kingdom (UK), mainland Europe, Australasia, or Asia. The UK was separated from Europe because of the high representation of outcomes research coming from the UK.
- **Intervention development stage:** Studies were coded as (i) *preliminary studies* (i.e., testing novel treatments or treatment iterations) or (ii) *routine evaluations*.
- **Experience:** Samples for which treatment delivery was exclusively by (i) *trainees*, or (ii) *qualified professionals*
- **Measurement tool:** Measures that were represented at least ten times in the meta-analysis were entered as subgroups
- **Sample Size:** Following the approach of Barth et al. (2013), studies were coded as small (N=<25), medium (N=25-50), or large (N=50+).

Continuous variables

- **Age:** Sample mean average age.
- **Year:** of publication.
- **Female participants**: Sample rate (%).

Table 2:

Summary statistics across the pooled sample and by sample severity.

		university	mild	moderate	severe	other	total
N	N	9195	158150	9515	22586	33694	233140
	k	58	88	32	92	8	278
	mean	158.53	1797.16	297.34	245.50	4211.75	838.63
	median	93.50	121.00	61.00	63.00	93.00	81.50
	iqr	162.50	935.00	347.00	107.50	1999.75	224.50
Females	N	5350	95373	5797	14952	22801	144273
Age	*k*	65.000	77.00	29	82	7	260
	mean	33.78	36.53	34.80	35.55	36.24	35.33
	min	20.50	19.00	24.30	21.52	24.52	19.00
	max	52.29	60.50	47.49	52.00	46.10	60.50
Sessions	*k*	54	64	4	54	6	182
	mean	21.00	11.26	13.75	14.67	8.55	15.13
	min	2.15	4.00	9.00	1.00	8.00	1.00
	max	85.33	64.90	24.00	64.00	9.52	85.33
	median	14.77	8.18	11.00	11.15	8.15	13.00
	iqr	13.55	8.60	3.750	10.00	1.20	9.98
Setting	mixed	0	0	0	0	5	5
	outpatient	68	96	0	91	4	259
	inpatient	0	0	33	1	0	34
Continent	Asia	4	1	0	0	1	6
	Australasia	5	0	0	5	0	10
	Europe	20	13	15	14	1	63
	America	38	32	10	39	4	123
	UK	1	50	8	34	3	96
Analysis	inclusion	48	48	16	53	4	169
	completers	19	45	16	35	3	118
therapy modality	cognitive-behavioural	43	41	14	49	5	152
	counselling	0	22	0	3	0	25
	psycho-dynamic	12	9	13	16	0	50
	other	13	24	6	24	4	71
treatment stage	preliminary	4	6	7	16	1	34
	evaluations	64	90	26	76	8	264

Table 3:

Multi-moderator analyses for depression outcomes

k = 124, Tau^2 = 0.17[SE = 0.02], I^2 = 99.99%, R^2 = 19.28%

	Moderator Level	*d*	SE	Z	P	CI
Intercept		1.22	0.17	7.20	<.0001	0.88~1.56
Region	UK	Ref				
	North America	-0.04	0.10	-0.34	0.733	-0.24~0.17
	Mainland Europe	-0.25	0.13	-1.94	0.053	-0.50~0.00
	Asia	-0.62	0.24	-2.62	0.001*	-1.09~-0.16
	Australasia	-0.49	0.26	-1.87	0.062	-1.01~0.02
Severity	Mild	Ref				
	Moderate	-0.13	0.11	-1.17	0.241	-0.34~0.09
	Severe	-0.10	0.15	-0.70	0.482	-0.39~0.18
	university (mild-to-severe)	0.20	0.15	1.34	0.180	-0.01~0.93
Therapy modality	Cognitive-behavioural	Ref				
	Psychodynamic	0.07	0.11	0.62	0.540	-0.15~0.28
	Counselling	-0.27	0.23	-1.19	0.236	-0.71~0.18
	Other	-0.04	0.12	-0.33	0.742	-0.28~0.12
Treatment Stage	preliminary studies	Ref				
	routine evaluations	-0.12	0.13	-0.99	0.324	-0.37~0.12
Analysis	Includes lost to follow up	Ref				
	Completers	0.16	0.13	2.01	0.045*	0.00~0.32
experience	Qualified Staff	Ref				
	Trainees	-0.29	0.14	-2.06	0.039*	-0.57~-0.01
sample size	Large	Ref				
	Medium	-0.15	0.10	-1.44	0.151	-0.35~0.05
	Small	-0.21	0.11	-1.83	0.069	-0.43~0.02
Publication year		-0.001	0.01	-0.19	0.851	-0.01~0.01
Sample age		-0.004	0.01	-0.67	0.503	-0.02~0.01
% Female		0.13	0.23	0.56	0.574	-0.33~0.59

Note. * = < .05

Table 4:

Multi-moderator analyses for anxiety outcomes

k = 78, Tau^2 = 0.13[SE = 0.02], I^2 = 99.95%, R^2 = 40.55%

	Moderator Level	*d*	SE	Z	P	CI
Intercept		1.24	0.22	5.59	<.0001	0.80~1.67
Region	UK	Ref				
	North America	-0.13	-0.13	-0.91	0.363	-0.40~0.14
	Mainland Europe	-0.35	0.15	-2.37	0.018*	-0.63~-0.06
	Asia	-0.55	0.29	-1.87	0.061	-1.13~0.026
	Australasia	-0.32	0.24	-1.30	0.194	-0.79~0.16
Severity	Mild	Ref				
	Moderate	-0.41	0.15	-2.71	0.007*	-0.70~-0.11
	Severe	-0.49	0.19	-2.56	0.011*	-0.86-0.11
	university (mild-to-severe)	0.03	0.17	0.20	0.838	-1.21~0.45
Therapy modality	Cognitive-behavioural	Ref				
	Psychodynamic	0.00	0.14	0.01	0.989	-0.27~0.28
	Counselling	-0.64	0.30	-2.16	0.031*	-1.23~-0.06
	Other	-0.64	0.16	-0.41	0.368	-0.39~0.25
Treatment Stage	Preliminary studies	Ref				
	Routine evaluations	-0.13	0.16	-0.81	0.421	-0.45~0.19
Analysis	Includes lost to follow up	Ref				
	Completers	0.15	0.12	1.28	0.120	-0.08~0.38
experience	Qualified Staff	Ref				
	Trainees	0.08	0.16	0.50	0.614	-0.23~0.39
sample size	Large	Ref				
	Medium	0.15	0.13	1.11	0.267	-0.11~0.40
	Small	-0.01	0.12	-0.07	0.942	-0.23~0.22
Publication year		0.01	0.01	1.71	0.088	-0.00~0.03
Sample age		-0.01	0.01	-1.16	0.248	-0.03~0.01
% Female		-0.22	0.37	-0.59	0.555	-0.94~0.50

Note. * = < .05

Table 5:

Multi-moderator analyses for other outcomes

k = 153, Tau^2 = 0.24[SE = 0.03], I^2 = 100%, R^2 = 21.44%

	Moderator Level	*d*	SE	Z	P	CI
Intercept		1.13	0.17	6.60	<.0001	0.80~1.47
Region	UK	Ref				
	North America	0.17	0.11	1.59	0.111	-0.04~0.39
	Mainland Europe	0.04	0.12	0.32	0.752	-0.20~0.27
	Asia	0.03	0.27	0.10	0.924	-0.50~0.55
	Australasia	-0.16	0.31	-0.49	0.626	-0.76~0.46
Severity	Mild	Ref				
	Moderate	-0.14	0.11	-1.23	0.220	-0.36~0.08
	Severe	-0.21	0.14	-0.12	0.901	-0.30~0.26
	University (mild-to-severe)	0.49	0.17	-1.25	0.210	-0.54~0.12
Therapy modality	Cognitive-behavioural	Ref				
	Psychodynamic	-0.25	0.11	-2.23	0.026*	-0.47~-0.03
	Counselling	-0.16	0.18	-0.86	0.387	-0.51~0.12
	Other	-0.39	0.11	-3.47	0.001*	-0.60~-0.17
Treatment Stage	Preliminary studies	Ref				
	Routine evaluations	0.06	0.13	0.45	0.650	0.20~0.32
Analysis	Includes lost to follow up	Ref				
	Completers	0.14	0.09	1.59	0.111	-0.03~0.31
experience	Qualified Staff	Ref				
	Trainees	-0.30	0.16	-1.90	0.058	-0.61~0.01
sample size	Large	Ref				
	Medium	-0.01	0.12	-0.09	0.925	-0.24~0.22
	Small	-0.06	0.12	-0.49	0.626	-0.30~0.18
Publication year		0.00	0.01	0.46	0.646	-0.01~0.02
Sample age		-0.00	0.01	-0.56	0.576	-0.02~0.01
% Female		-0.14	0.23	-0.62	0.534	-0.59~0.31

Note. * = < .05

Table 6:

Benchmarks for routine services based on individual study sample quartiles.

		Outpatient	Inpatient	UCC	Uni Clinics
Top 25%	Depression	*d* = 1.68 [1.53-1.83]	*d* = 1.34 [1.16-1.52]	*	*d* = 1.77 [1.50-2.03]
	Anxiety	*d* = 1.56 [1.38-1.73]	*d* = 1.07 [1.04-1.09]	*	*d* = 1.80 [1.57-2.02]
	other	*d* = 1.70 [1.54-1.86]	*d* = 1.67 [1.37-1.97]	*d* = 1.47 [1.24-1.69]	*d* = 1.14 [1.10-1.18]
Average	Depression	*d* = 0.94 [0.90-0.97]	*d* = 0.98 [0.81-1.15]	*	*d* = 0.91 [0.87-0.95]
	Anxiety	*d* = 0.84 [0.78-0.89]	*d* =0.67 [0.42-0.92]	*	*d* = 0.95 [0.87-1.02]
	other	*d* = 0.92 [0.89-0.96]	*d* = 1.04 [0.96-1.11]	*d* = 0.94 [0.84-1.03]	*d* = 0.86 [0.77-0.94]
Low 25%	Depression	*d* =0.46 [0.41-0.52]	*d* = 0.38 [0.26-0.5]	*	*d* = 0.40 [0.27-0.54]
	Anxiety	*d* = 0.37 [0.33-0.42]	*d* = 0.13 [0.03-0.29]	*	*d* = 0.51 [0.44-0.57]
	other	*d* = 0.49 [0.43-0.54]	*d* = 0.58 [0.46-0.69]	*d* = 0.64 [0.61-0.67]	*d* = 0.41 [0.23-0.59]

Note. *cannot be computed due to too few samples.

UCC: University Counselling Centres; *d*: uncontrolled, pre-to-post treatment effect size [95% confidence intervals]

University clinics refers to university managed clinics treating communities beyond the student population. University counselling centres that are more specifically targeted at the student population are included within the mild category.

Figure 1:

Prisma flow diagram of studies throughout the review.

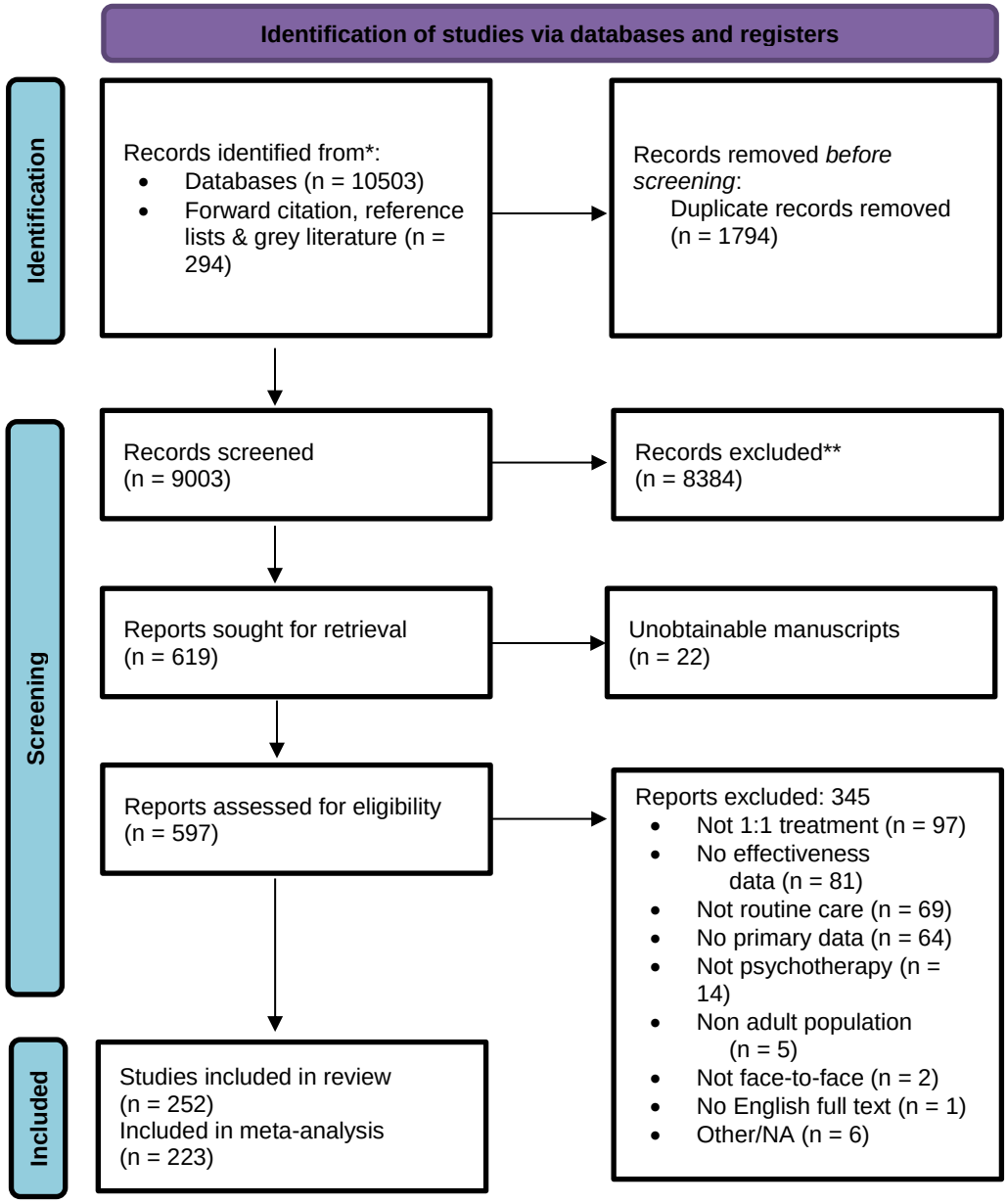
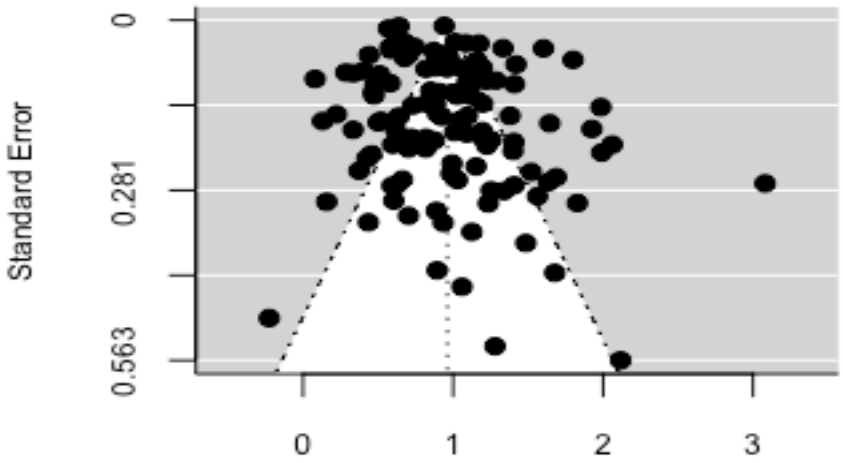


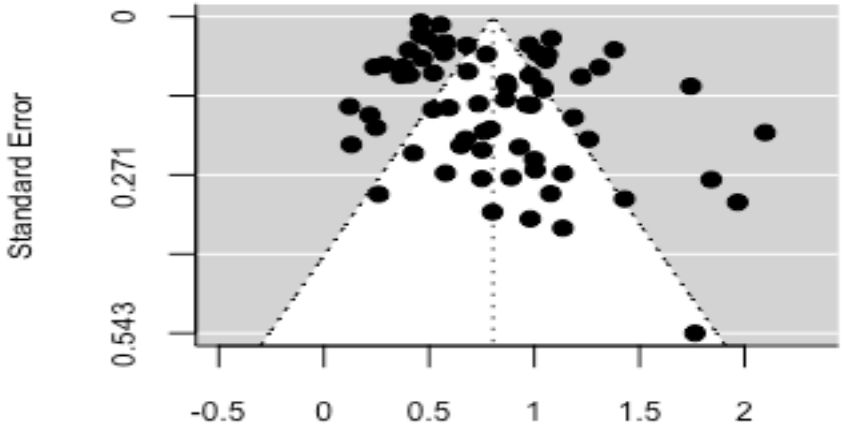
Figure 2:

Funnel plots displaying the distribution of studies reporting pre-post outcomes for (i) depression, (ii) anxiety, and (iii) miscellaneous outcomes.

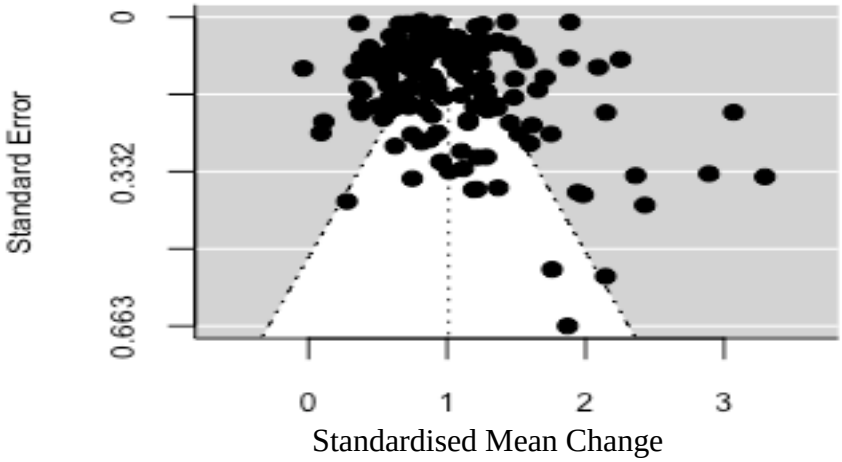
Depression



Anxiety



Miscellaneous



References

Addis, M. E., & Krasnow, A. D. (2000). A national survey of practicing psychologists' attitudes toward psychotherapy treatment manuals. *Journal of Consulting and Clinical Psychology*, *68*(2), 331–339. https://doi.org/10.1037/0022-006X.68.2.331

APA. (2006). Evidence-based practice in psychology. *Presidential Task Force on Evidence-Based Practice. American Psychologist*, *61*, 271–285. https://doi.org/www.apa.org/pubs/journals/features/evidence-based-statement.pdf

Armijo-Olivo, S., Ospina, M., Costa, B. R. da, Egger, M., Saltaji, H., Fuentes, J., Ha, C., & Cummings, G. G. (2014). Poor Reliability between Cochrane Reviewers and Blinded External Reviewers When Applying the Cochrane Risk of Bias Tool in Physical Therapy Trials. *PLOS ONE*, *9*(5), e96920. https://doi.org/10.1371/journal.pone.0096920

Balk, E. M., Earley, A., Patel, K., Trikalinos, T. A., & Dahabreh, I. J. (2012). Empirical assessment of within-arm correlation imputation in trials of continuous outcomes. *Methods Research Reports*, *12*(13).

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K., & McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*(2), 184–196. https://doi.org/10.1037/0022-006X.69.2.184

Barkham, M., Stiles, W. B., Lambert, M. J., & Mellor-Clark, J. (2010). Building a rigorous and relevant knowledge base for the psychological therapies. In M. Barkham, G. E. Hardy, &

J. Mellor-Clark (Eds.), *Developing and Delivering Practice-Based Evidence* (pp. 21–61). John Wiley & Sons Ltd. https://doi.org/10.1002/9780470687994.ch2

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical & Statistical Psychology*, *41*(2), 257–278. https://doi.org/10.1111/j.2044-8317.1988.tb00901

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101. https://doi.org/10.2307/2533446

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to Meta-Analysis*. John Wiley & Sons.

Cahill, J., Barkham, M., & Stiles, W. (2010). Systematic review of practice-based research on psychological therapies in routine clinic settings. *The British Journal of Clinical Psychology*, *49*(4), 421–453. https://doi.org/10.1348/014466509X470789

Card, N. A. (2015). *Applied Meta-Analysis for Social Science Research*. Guilford Publications.

Castonguay, L. G., Barkham, M., Jeong Youn, S., & Page, A. C. (2021). Practice-based evidence findings from routine clinical settings. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (Seventh, pp. 191–222). Wiley.

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavey, A. A. (2013). Practice-orienIrach: Approaches and applications. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (Sixth, pp. 85–133). Wiley.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*(1), 7–18. https://doi.org/10.1037//0022-006x.66.1.7

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Empirically Supported Psychological Interventions: Controversies and Evidence*, *52*(1), 685–716. https://doi.org/10.1146/annurev.psych.52.1.685

Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., & Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): An analysis of public data. *The Lancet*, *391*(10121), 679–686. https://doi.org/10.1016/S0140-6736(17)32133-5

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129. https://doi.org/10.2307/3001666

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Connell, J., Barkham, M., & Mellor-Clark, J. (2007). CORE-OM mental health norms of students attending university counselling services benchmarked against an age-matched primary care sample. *British Journal of Guidance & Counselling*, *35*(1), 41–57. https://doi.org/10.1080/03069880601106781

Cooper, H. M. (1998). *Synthesizing Research: A Guide for Literature Reviews*. SAGE.

Cuijpers, P. (2016). *Meta-analyses in mental health research: A practical guide*. University of Amsterdam.

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*(6), 768–778. https://doi.org/10.1016/j.cpr.2010.06.001

Cuijpers, P., Sijbrandij, M., Koole, S., Huibers, M., Berking, M., & Andersson, G. (2014). Psychological treatment of generalized anxiety disorder: A meta-analysis. *Clinical Psychology Review*, *34*(2), 130–140. https://doi.org/10.1016/j.cpr.2014.01.002

Cuijpers, P., Turner, E. H., Mohr, D. C., Hofmann, S. G., Andersson, G., Berking, M., & Coyne, J. (2014). Comparison of psychotherapies for adult depression to pill placebo control groups: A meta-analysis. *Psychological Medicine*, *44*(4), 685–695. https://doi.org/10.1017/S0033291713000457

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: A meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology*, *76*(6), 909–922. https://doi.org/10.1037/a0013075

Delgadillo, J., McMillan, D., Leach, C., Lucock, M., Gilbody, S., & Wood, N. (2014). Benchmarking routine psychological services: A discussion of challenges and methods. *Behavioural and Cognitive Psychotherapy*, *42*(1), 16–30. https://doi.org/10.1017/S135246581200080X

Driessen, E., Cuijpers, P., Hollon, S. D., & Dekker, J. J. M. (2010). Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-analysis. *Journal of Consulting and Clinical Psychology*, *78*(5), 668–680. https://doi.org/10.1037/a0020570

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Flückiger, C., Del Re, A. C., Barth, J., Hoyt, W. T., Levitt, H., Munder, T., Spielmans, G. I., Swift, J. K., Vîslă, A., & Wampold, B. E. (2018). Considerations of how to conduct meta-analyses in psychological interventions. *Psychotherapy Research, 28*(3), 329–332. https://doi.org/10.1080/10503307.2018.1430390

Freedland, K. E., Mohr, D. C., Davidson, K. W., & Schwartz, J. E. (2011). Usual and unusual care: Existing practice control groups in randomized controlled trials of behavioral interventions. *Psychosomatic Medicine, 73*(4), 323–335. https://doi.org/10.1097/PSY.0b013e318218e1fb

Furukawa, T. A., Weitz, E. S., Tanaka, S., Hollon, S. D., Hofmann, S. G., Andersson, G., Twisk, J., DeRubeis, R. J., Dimidjian, S., Hegerl, U., Mergl, R., Jarrett, R. B., Vittengl, J. R., Watanabe, N., & Cuijpers, P. (2017). Initial severity of depression and efficacy of cognitivebehavioural therapy: Individual-participant data meta-analysis of pill-placebo-controlled trials. *British Journal of Psychiatry, 210*(3), 190–196. https://doi.org/10.1192/bjp.bp.116.187773

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (Second, pp. 357–376). Russell Sage.

Gyani, A., Shafran, R., Layard, R., & Clark, D. M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy*, *51*(9), 597–606. https://doi.org/10.1016/j.brat.2013.06.004

Hans, E., & Hiller, W. (2013). 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*(1), 75–88. https://doi.org/10.1037/a0031080

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, David. D. (2019a). *Dmetar: Companion R package for the guide 'doing meta-analysis in R'.* [R Package].

Higgins, J., & Thompson, S. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hoyt, W. T., & Del Re, A. C. (2018). Effect size calculation in meta-analyses of psychotherapy outcome research. *Psychotherapy Research*, *28*(3), 379–388. https://doi.org/10.1080/10503307.2017.1405171

Hunsley, J., & Lee, C. M. (2007). Research-informed benchmarks for psychological treatments: Efficacy studies, effectiveness studies, and beyond. *Professional Psychology: Research and Practice*, *38*(1), 21–33. https://doi.org/10.1037/0735-7028.38.1.21

Jüni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of controlled clinical trials. *BMJ*, *323*(7303), 42–46.

Lambert, M. J. (2013). The efficacy and effectiveness of psychotherapy. In M. J. Lambert & A. E. Bergin (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavi*^{or} *change* (6th ed., pp. 169–218). John Wiley & Sons, Incorporated.

Lemma, A., Roth, A. D., & Pilling, S. (2008). *The competences required to deliver effective Psychoanalytic/ Psychodynamic Therapy*. Research Department of Clinical, Educational and Health Psychology, University College London.

Lewis, C., Roberts, N. P., Andrew, M., Starling, E., & Bisson, J. I. (2020). Psychological therapies for post-traumatic stress disorder in adults: Systematic review and meta-analysis. *European Journal of Psychotraumatology*, *11*(1), 1729633. https://doi.org/10.1080/20008198.2020.1729633

Linardon, J., Wade, T. D., de la Piedad Garcia, X., & Brennan, L. (2017). The efficacy of cognitive-behavioral therapy for eating disorders: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *85*(11), 1080–1094. https://doi.org/10.1037/ccp0000245

Lutz, W., Schiefele, A.-K., Wucherpfennig, F., Rubel, J., & Stulz, N. (2016). Clinical effectiveness of cognitive behavioral therapy for depression in routine care: A propensity score based comparison between randomized controlled trials and clinical practice. *Journal of Affective Disorders*, *189*(1), 150–158. https://doi.org/10.1016/j.jad.2015.08.072

Margison, F. R., Barkham, M., Evans, C., McGrath, G., Clark, J. M., Audin, K., & Connell, J. (2000). Measurement and psychotherapy: Evidence-based practice and practice-based

evidence. *British Journal of Psychiatry*, *177*(2), 123–130. https://doi.org/10.1192/bjp.177.2.123

Mayo-Wilson, E., Dias, S., Mavranezouli, I., Kew, K., Clark, D. M., Ades, A. E., & Pilling, S. (2014). Psychological and pharmacological interventions for social anxiety disorder in adults: A systematic review and network meta-analysis. *The Lancet Psychiatry*, *1*(5), 368–376. https://doi.org/10.1016/S2215-0366(14)70329-3

McAleavey, A.A., Youn, S.J., Xiao, H., Castonguay, L.G., Hayes, J.A., & Locke, B.D. (2019). Effectiveness of routine psychotherapy: Methods matters. Psychotherapy Research, 29,139-156

Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. S. (2008). Using clinical trials to benchmark effects produced in clinical practice. *Quality and Quantity, 42*(4), 513-525. https://doi.org/10.1007/s11135-006-9057-z

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Munder, T., & Barth, J. (2018). Cochrane's risk of bias tool in the context of psychotherapy outcome research. *Psychotherapy Research*, *28*(3), 347–355. https://doi.org/10.1080/10503307.2017.1411628

Munn, Z., Barker, T. H., Moola, S., Tufanaru, C., Stern, C., McArthur, A., Stephenson, M., & Aromataris, E. (2020). Methodological quality of case series studies: An introduction to

the JBI critical appraisal tool. *JBI Evidence Synthesis*, *18*(10), 2127–2133. https://doi.org/10.11124/JBISRIR-D-19-00099

National Collaborating Centre for Mental Health. (2011). *Common Mental Health Disorders: Identification and Pathways to Care* (Vol. 123). RCPsych Publications.

National institute for Health and Care Excellence. (2011). *Common mental health problems: identification and pathways to care* (CG90). https://www.nice.org.uk/guidance/cg123/chapter/1-guidance

Nordmo, M., Sønderland, N. M., Havik, O. E., Eilertsen, D.-E., Monsen, J. T., & Solbakken, O. A. (2020). Effectiveness of open-ended psychotherapy under clinically representative conditions. *Frontiers in Psychiatry*, *11*, 384. https://doi.org/10.3389/fpsyt.2020.00384

Olatunji, B. O., Kauffman, B. Y., Meltzer, S., Davis, M. L., Smits, J. A. J., & Powers, M. B. (2014). Cognitive-behavioral therapy for hypochondriasis/health anxiety: A meta-analysis of treatment outcome and moderators. *Behaviour Research and Therapy*, *58*, 65–74. https://doi.org/10.1016/j.brat.2014.05.002

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*, *18*(3), e1003583. https://doi.org/10.1371/journal.pmed.1003583

Persons, J. B., Bostrom, A., & Bertagnolli, A. (1999). Results of randomized controlled trials of cognitive therapy for depression generalize to private practice. *Cognitive Therapy and Research, 23*(5), 535–548. https://doi.org/10.1023/A:1018724505659

Synthesis Methods, 10(3), 330–342. https://doi.org/10.1002/jrsm.1354

Posternak, M. A., & Miller, I. (2001). Untreated short-term course of major depression: A meta-analysis of outcomes from studies using wait-list control groups. *Journal of Affective Disorders, 66*(2-3), 139–146. https://doi.org/10.1016/s0165-0327(00)00304-9

Power, N., Noble, L., Simmonds-Buckley, M., Kellett, S., Stockton, C., Firth, N., & Delgadillo, J. (in press). Associations between treatment adherence-competence-integrity (ACI) and adult psychotherapy outcomes: a systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology, 90*(5), 427–445. https://doi.org/10.1037/ccp0000736

Pybis, J., Saxon, D., Hill, A., & Barkham, M. (2017). The comparative effectiveness and efficiency of cognitive behaviour therapy and generic counselling in the treatment of depression: evidence from the 2nd UK National Audit of psychological therapies. *BMC psychiatry, 17*(1), 1-13. https://doi.org/10.1186/s12888-017-1370-7

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rosa-Alcázar, A. I., Sánchez-Meca, J., Gómez-Conesa, A., & Marín-Martínez, F. (2008). Psychological treatment of obsessivecompulsive disorder: A meta-analysis. *Clinical Psychology Review, 28*(8), 1310–1325. https://doi.org/10.1016/j.cpr.2008.07.001

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Roth, A., Hill, A., & Pilling, S. (2009). *The competences required to deliver effective Humanistic Psychological Therapies*. London: University College London.

Roth, A., & Pilling, S. (2008). Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behavioural and Cognitive Psychotherapy*, *36*(2), 129–147. https://doi.org/10.1017/S1352465808004141

Sánchez-Meca, J., Rosa-Alcázar, A. I., Marín-Martínez, F., & Gómez-Conesa, A. (2010). Psychological treatment of panic disorder with or without agoraphobia: A meta-analysis. *Clinical Psychology Review*, *30*(1), 37–50. https://doi.org/10.1016/j.cpr.2009.08.011

Schwarzer, G. (2020). *Meta: General package for meta-analysis*. https://CRAN.R-project.org/package=meta

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crits-Christoph, P., Hazelrigg, M. D., Jorm, A. F., Lyons, L. C., Nietzel, M. T., Robinson, L., Prout, H. T., Smith, M. L., Svartberg, M., & Weiss, B. (1997). Evidence that therapy works in clincally representative conditions. *Journal of Consulting and Clinical Psychology*, *65*(3), 355–365. https://doi.org/10.1037/0022-006X.65.3.355

Shadish, W. R., Navarro, A. M., Matt, G. E., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*(4), 512–529. https://doi.org/10.1037/0033-2909.126.4.512

Spielmans, G. I., & Flückiger, C. (2018). Moderators in psychotherapy meta-analysis. *Psychotherapy Research*, *28*(3), 333–346. https://doi.org/10.1080/10503307.2017.1422214

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*(4), 595–606. https://doi.org/10.1037/a0016032

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, *47*(4), 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

van der Lem, R., de Wever, W. W., van der Wee, N. J., van Veen, T., Cuijpers, P., & Zitman, F. G. (2012). The generalizability of psychotherapy efficacy trials in major depressive disorder: An analysis of the influence of patient selection in efficacy trials on symptom outcome in daily practice. *BMC Psychiatry*, *12*(1), 192. https://doi.org/10.1186/1471-244X-12-192

Viechtbauer, W. (2020). *Metafor: Meta-analysis package for r*. https://CRAN.R-project.org/package=metafor

Wakefield, S., Kellett, S., Simmonds-Buckley, 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), 1–37. https://doi.org/10.1111/bjc.12259

Weisz, J. R., Weiss, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin*, *117*(3), 450–468. https://doi.org/10.1037/0033-2909.117.3.450

Whiteford, H., Harris, M., Mckeon, G., Baxter, A., Pennell, C., Barendregt, J., & Wang, J. (2012). Estimating remission from untreated major depression: A systematic review and meta-analysis. *Psychological Medicine*, *43*, 1–17. https://doi.org/10.1017/S0033291712001717

Wolitzky-Taylor, K. B., Horowitz, J. D., Powers, M. B., & Telch, M. J. (2008). Psychological approaches in the treatment of specific phobias: A meta-analysis. *Clinical Psychology Review*, *28*(6), 1021–1037. https://doi.org/10.1016/j.cpr.2008.02.007

Xiao, H., Carney, D. M., Youn, S. J., Janis, R. A., Castonguay, L. G., Hayes, J. A., & Locke, B. D. (2017). Are we in crisis? National mental health and treatment trends in college counseling centers. *Psychological Services, 14*(4), 407–415. https://doi.org/10.1037/ser0000130

Zimmerman, M., Mattia, J. I., & Posternak, M. A. (2002). Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *American Journal of Psychiatry*, *159*(3), 469–473. https://doi.org/10.1176/appi.ajp.159.3.469