University of Sheffield

The Effectiveness of Psychotherapy Delivered in Routine Service Settings

Chris Gaskell

Supervisors:

Dr. Stephen C. Kellett

Dr. Mel Simmonds-Buckley

Dr. Jaime Delgadillo

A report submitted in partial fulfillment of the requirements
for the degree of Doctorate in Clinical Psychology
in the Department of Clinical Psychology

May 25, 2021

This page is intentionally left blank

Preface

Declaration

I declare that this work has not been submitted for any other degree at the University of Sheffield, or any other institution. The work presented is original and all other sources have been referenced accordingly.

This page is intentionally left blank

Word Counts

The word counts for this thesis are as follows:

Literature Review

Excluding tables and references: 7,971

Including tables and references: 12,163

Empirical Project

Excluding tables and references: 7,982

Including tables and references: 11,999

Total

Excluding tables and references: 15,953

Including tables and references: 24,162

This page is intentionally left blank

Lay Summary

Understanding the extent to which psychological therapy is effective, and under which conditions, is important for the provision of mental health treatment. Historically, there has been a large weighting for evidence generated from research methodologies that possess high internal validity (e.g., randomised-controlled trials). Unfortunately, treatments evaluated with these methodologies, usually conducted in University settings, often fail to represent various aspects of routine care settings and treatments which they are purposefully intended for (e.g., lower levels of client heterogeneity and risk, fixed treatment lengths, manualised treatments). Due to this distinction (i.e., efficacy vs. effectiveness), it is subsequently important to establish the extent to which treatments remain effective when delivered within routine care.

Since the turn of the century there has been a marked increase in the amount of evidence from routine care settings regarding the effectiveness of psychological therapy. The first chapter of this thesis was a comprehensive review of this body of literature, focusing on one-to-one, face-to-face psychological treatments within naturalistic settings. 252 studies were identified, comprising a total of 298 different therapy samples. These samples were analysed quantitatively using meta-analysis. The findings demonstrated that psychological therapy produces marked improvements (i.e., large effect-sizes) across each self-reported outcome domain explored (depression, anxiety, general distress). Effect-sizes tended to be larger under certain conditions: (i) studies conducted in the UK/North America, (ii) studies which only include patients who complete treatment, and (iii) studies with low risk of bias. This was the largest review of routine practice studies concerning psychological therapy which has been conducted to date.

The second chapter of this thesis was an evaluation of how patients respond to psychological treatment from a UK tertiary care psychological therapy service. Evidence arising from these settings is highly scarce. Such treatments are intended to be across longer periods and for more distressed/complex presentations than treatments delivered in other care settings (primary and/or secondary care). The current study found that while tertiary care therapy is effective, the amount of change is smaller than available benchmarks. Rates of improvement were suppressed by a sub-set of patients who appeared to not respond to treatment. The three forms of psychological therapy offered were not significantly different in effectiveness. Patient rate of improvement (positive growth) reduced with each session. There was limited evidence to support continuing treatment beyond 100 sessions.

Taken together, the chapters presented provide further evidence that psychological therapy is effective across a wide range of routine care settings. Similar to prior findings, there was limited support for long-form psychological treatments. Support was shown for the potential utility of routine outcome measures as a tool to determine when a patient is no longer benefiting from treatment. The evidence considered in these chapters predominantly concerns self-report outcome measures. Further research is needed to determine consistency of findings through other measures of effectiveness (e.g., clinician-rated change, health-care cost/utilisation, reduction in harm events). Further research is needed which focuses upon how improvements made during therapy are maintained beyond treatment ending.

Acknowledgements

For the project itself, I would like to wholeheartedly thank Dr Steve Kellett, Dr. Mel Simmonds-Buckley, Dr. Jaime Delgadillo for their highly accessible, approachable and responsive support as research supervisors throughout the course of this thesis. In addition, I would like to also thank all collaborators for their valued contributions, in no particular: Gregg, Corrie, Jack, Joe, Suzanne and John.

For sharing and co-surviving what have been the most unusual of three years, I would like to thank each member of my training cohort. The company, learning, and support I have received from this collective has been (for me) the single most important feature of the training process. Particular thanks to Gregg and Niall for regularly reaching out and maintaining connection during an era of marked disconnect. To each of my supervisors during training, I take the utmost pride in drawing on your influences each day: Ben, Dan, Myra, Nigel, Keeley, Mary, Alistair and Kate.

For providing the opportunities that have got me this far: my mum, dad, and two brothers. Regardless of which way I grow, I will endeavor to always stay in touching distance of my roots.

Finally, to Amber, for having always been the person I need; and to Leo, for his unwavering proficiency in being this man's best friend.

This page is intentionally left blank

SYSTEMATIC LITERATURE REVIEW

The Effectiveness of Psychotherapy Delivered in Routine Care Settings: A Systematic Review and Meta-Analysis

Chris Gaskell

Supervisors:

Dr. Stephen C. Kellett

Dr. Mel Simmonds-Buckley

Contents

List of Tables

List of Figures

This page is intentionally left blank

Abstract

Objectives: There has been a substantial increase in the amount of evidence arising from routine (i.e. naturalistic) care settings in the field of psychotherapy in recent decades. This review sought to examine the effectiveness of routinely delivered psychological therapies. **Design:** A pre-registered systematic-review and meta-analysis (CRD42020175235). **Methods:** Random-effects meta-analyses were conducted on studies meeting pre-specified inclusion criteria. Moderator analyses examining methodological, treatment-level and sample-level variables explored between-study heterogeneity. **Results:** The systematic search identified 252 studies (k = 298 samples) for the quantitative synthesis. Of these, 223 studies (k = 263 samples) were eligible for inclusion in the meta-analysis. Results showed large effects for the depression (d = 0.98, [CI 0.9-1.06], p = < 0.001, k = 140), anxiety (d = 0.83 [CI 0.73-0.92], p = < 0.001, k = 84), and global outcome domains (d = 1.01 [CI 0.93-1.08], p = < 0.001, k = 184). Sample completion (completers vs. intention-to-treat), geographical area (continent) and risk of bias were significant moderators of treatment effects. **Conclusions:** This review provides support for the effectiveness of routinely delvered psychological therapy. Findings should be interpreted with caution due to the observational nature of effectiveness studies and also the marked heterogeneity shown across study designs and characteristics.

Keywords:

‘Psychotherapy,’ ‘Effectiveness,’ ‘Naturalistic,’ ‘Routine Outcomes,’ ‘Meta-analysis.’

Practitioner Points:

- Greater depth and consistency of reporting detail is required in routine outcome studies.
- Completer analyses may artificially inflate effect-sizes for outcomes in routine practice.
- There was encouraging evidence that age and ethnicity does not hinder treatment effectiveness; equitable opportunity to access treatment should therefore be provided across these dimensions.

The Effectiveness of Psychotherapy Delivered in Routine Care Settings: A Systematic Review and Meta-Analysis

There is widespread consensus that psychological therapy is an efficacious treatment for a variety of mental health disorders (Lambert, 2013). A substantial proportion of the evidence for these claims originates in reviews of randomised controlled trials (RCTs, e.g., Smith & Glass, 1977). The primary critique of these RCTs is that methods used to enhance experimental control (e.g. homogeneous client groups, random assignment, control groups) mean that the results may not necessarily generalize to routine services that typically treat a heterogeneous patient population (Barkham, Stiles, et al., 2010). Until recently there has been an over-reliance on this form of evidence (i.e. efficacy evidence), exemplified through a seminal review of studies (Roth & Fonagy, 1996) being criticized as being almost exclusively made up of RCTs (Margison et al., 2000). The extent to which efficacious treatments hold up in routine settings (i.e. transportability) remains a contentious debate in psychology (Hunsley & Lee, 2007; Jacobson & Christensen, 1996; Smith & Glass, 1977).

Routine service settings differ substantially to the conditions typically provided for efficacy research (Barkham, Stiles, et al., 2010). Routine services traditionally have higher patient to clinician ratios, higher patient heterogeneity and manage greater levels of clinical risk. Interventions provided within these settings are less standardised, with less frequent use of protocols/manuals and scarce use of integrity/fidelity checks.

The nature of interventions, how they are delivered, and also how they are evaluated within routine settings has changed over time. Evidence of psychological therapy outcome, arising from routine service is ‘effectiveness’ research, also known as ‘practice-based evidence’ (PBE: Barkham, Hardy, et al., 2010). A common critique of PBE is the uncontrolled presence of potentially confounding variables, which may compromise internal validity (Barkham, Stiles, et al., 2010). Since the initial binary distinction of efficacy vs. effectiveness, it has become increasingly recognised that the two approaches can overlap (Stewart & Chambless, 2009), forming an efficacy-effectiveness continuum (Hunsley & Lee, 2007).

Models have been provided to consider how effectiveness and efficacy research may complement each other on this continuum. One example of this is the three-stage 'hour-glass' model of psychological therapy outcomes research (Salkovskis, 1995). Stage one consists of emerging intervention evidence conducted on small numbers of patients in routine practice, often using uncontrolled research designs (e.g. pilot and case studies). Stage two elaborates on promising stage one evidence by investigating the intervention under more tightly controlled efficacy conditions (ideally RCT). Finally stage three involves the transporting of interventions, empirically supported at stage two, to larger practice-based (naturalistic) settings in order to confirm/refute clinical utility. This is the evidence-based practice phase of the hourglass.

Given that the majority of therapy is delivered within routine practice settings it is subsequently necessary to conduct regular reviews of the evidence base generated within these settings (i.e. evidence from stages 1 and 3 of the hour-glass model). Prior reviews have employed meta-analytic approaches in order to aggregate effect-sizes across studies (see Lambert, 2013 for a review). Reviews of effectiveness research generally employ one of two approaches. The representativeness approach (e.g. Shadish et al., 2000; Smith & Glass, 1977; Stewart & Chambless, 2009) uses broad inclusion criteria – including efficacy studies – before rating each study on how much the conditions reported resemble routine services. Representativeness is then assessed for the degree to which it is associated with outcome. An alternative strategy is to restrict inclusion to studies which are highly representative of routine conditions (e.g. Cahill et al., 2010; Wakefield et al., 2021). In the 11 years since the last broad review of therapy effectiveness (Cahill et al., 2010) there has been a considerable increase in the volume of practice-based evidence, thus justifying the need for an updated review. Furthermore, although earlier reviews have quantitatively examined the effectiveness of routinely-delivered therapy, there is scarce evidence on sources of heterogeneity of treatment effects.

Heterogeneity refers to the amount of variability inherent in the aggregated treatment effect-size; and subsequently influences the degree to which findings can be confidently generalized (Kraemer et al., 2006). Patient heterogeneity is generally higher in practice-based studies (compared with RCTs) because of the less frequent use of exclusion criteria. Heterogeneity within meta-analyses of effectiveness research is typically high (e.g. Wakefield et al., 2021). A treatment effect with high heterogeneity may fail to explain potential underlying differences in how different people respond to treatment.

A common approach is to try to reduce the unexplained heterogeneity by measuring moderator variables that may account for a proportion of the heterogeneity. A moderator is a pre-treatment variable that can be used to define subgroups of patients within a larger sample (Kraemer et al., 2006). A moderator of treatment effect is when differential rates of effectiveness between individuals is demonstrated based on prior distinction. Use of moderator variables has been somewhat limited in prior reviews of PBE. For example, Wakefield et al. (2021) reviewed studies conducted in the UK *increasing access to psychological therapy* (IAPT) programme and found that study methodology (intention-to-treat vs. completers analysis) was a significant and consistent moderator of treatment outcomes. The current review sought to measure the influence of a range pre-identified moderator variables. A hypothesis-led approach, based on the extant literature of psychological therapy outcomes was used to select moderator variables.

Moderators can be conceptualised as being at the levels of (i) patient, (ii) treatment, (iii) service or (iv) study methodology. Patient level moderators may include demographics (age, gender, ethnicity etc.) or presenting problem/diagnosis (e.g. Roth & Fonagy, 1996). Treatment level moderators may include therapeutic model (e.g. Roth & Fonagy, 1996), treatment dosage (e.g. Flückiger et al., 2020) or interventionist experience (e.g. Buckley et al., 2006). Service level moderators may include type of service, setting (inpatient vs. outpatient), geographical region, sector (e.g. Wakefield et al., 2021), or service funding structure. Finally, methodological variables may include year of publication, sample analysed (e.g. intention-to-treat, Wakefield et al., 2021), stage of the hour-glass or study methodological quality (e.g. Wakefield et al., 2021). These listed moderators were considered for inclusion in the current review.

When exploring multiple moderator variables it is important to consider the potential interactive relationships that they have. This is possible through a process called multi-variate moderator analysis. Considering only each moderator in isolation will not show how each moderator can become amplified or attenuated when considered with another (Li et al., 2020). The primary barrier to multi-variate moderator analysis is that these methods require a high ratio of studies to co-variates (Borenstein et al., 2009) which are often not available to researchers. A broad review of effectiveness studies is now required which can allow for multi-variate analysis of moderators of treatment effect-size.

Aims

The main objective of the present study was to qualitatively and quantitatively synthesize the evidence on the effectiveness of individual psychological therapy for adults accessing services in routine care. The primary aim was to quantify the effectiveness of psychological treatment delivered in routine services. In doing so, the present study used a liberal conceptualisation of what constitutes as a ‘routine service’ which matches the reality of the inherent heterogeneity shown across routine care settings. This included a systematic search and meta-analysis of studies published prior to the systematic search date. The secondary aim of this review was to explore how a range of moderator variables influence treatment effects. The final aim was to assess the quality of each meta-analytic comparison using the Grading of Recommendations, Assessment, Development and Evaluations (GRADE, Guyatt et al., 2008) process.

Method

Search Strategy and Eligibility

A systematic review and meta-analysis was conducted using the Preferred Reporting Items for Systematic Review and Meta-Analysis guidelines (PRISMA, Moher et al., 2009) following pre-registration on PROSPERO (CRD42020175235).

Inclusion and exclusion criteria are summarized in Table 1 using the PICOS framework (population, intervention, comparator, outcome, setting). Three electronic databases (MEDLINE, CINAHL and PsycInfo) were searched for studies using a pre-developed list of key terms. Terms were selected based on their use in prior reviews of psychotherapy effectiveness (Cahill et al., 2010; Stewart & Chambless, 2009, Appendix A). For inclusion in the current review studies were required to have a methodologically *and* psychologically relevant term in the title or abstract. Psychological relevance was set using ‘psycho-’ (for MEDLINE and CINAHL) or ‘psycho-’/‘therap-’ (PsycInfo). Use of the ‘therap-’ in MEDLINE and CINAHL produced an unmanageable number of irrelevant hits and was subsequently removed. Limiters included ‘adult population’ and ‘English language’ for all available studies (- April 2020). No exclusions were made based on the type of publication.

Studies were required to have included psychological therapy as conducted in a routine/naturalistic setting (i.e. locations where patients are ordinarily seen for therapy, typical referral procedures). Studies were anticipated to be of an observational nature (i.e., open/pilot trials, case series, audit/service evaluation, benchmarking).

Participants

Samples were required to be exclusively adult (Aged 16≥). If the age range for the study sample fell below 16 then the sample was excluded. Treatments could be for psychological disorders or physical health conditions which are associated with psychological distress. No exclusions were imposed regarding diagnosis/presenting problem.

Table 1

Inclusion and exclusion criteria used in the current review, shown using the PICOS framework (population, intervention, comparator, outcome, setting).

Criteria	Inclusion	Exclusion
Population	Sample exclusively aged 16 and above (lower end of sample age range is at least 16).	Adolescent/child samples with a lower age limit below 16.
Intervention	Psychological intervention which includes individual face-to-face psycholgoical therapy (i.e. at least one session).	Samples which indicate that any proportion of patients did not recieve at least one session of individual psychological therapy.
Comparator	Studies with pre and post intervention time points. Post intervention defined here as up to six months following treatment.	(i) Studies which do not report both pre and post intervention time points. (ii) Studies for which the post intervention time point is beyond six months following treatment termination. (iii) Treatment randomisation proceadures.
Outcome	Psychological treatment effectiveness using a validated self-report measurement tool.	Service/settings which do not use a self-report measure of psychological effectiveness. Clinician reported measures were not included in this review.
Setting	Services for which a patient could expect to access psychological therapy (i.e. routine services).	Service/settings that strongly do not appear naturalistic or reflect routine practice.
	(i) Pre-post treatment designs. (ii) Studies which do not use a control condition.	(i) Studies which include a control group. (iii) Studies with N = <6. (iv) Results not available/published in English.

Interventions

Psychological interventions were required to have included a component of individual (i.e. one-to-one) psychological therapy. Study samples which included any proportion of patients who had not received individual psychological therapy (e.g. only group treatment, family therapy) were excluded. Multi-modal treatments which included a component of individual psychological therapy were included.

Comparisons and Multiple Samples

The main outcome of interest was pre-post change for the acute-phase of treatment (outcome measured at treatment termination). Studies which included both pre and post intervention measurement points were included. Post-intervention is defined here as the last session of treatment. For studies which only recorded the post-treatment score at a latter time point (i.e. follow-up), this was coded as the post-intervention score if it was within the first 6-months following treatment ending. Studies with post-treatment measurement beyond the 6-month post-treatment time point were excluded, as this constituted longer-term effects rather than acute-phase effects.

Practice-based studies which employed randomisation/control groups, although offering greater internal validity, are not typical of routine services, and when used in combination with observational practice-based studies pose various methodological and ethical dilemmas (Nordmo et al., 2020). Because of this, various patient sub-groups (e.g. highly distressed patients) are likely to be under represented in studies which use a control group (Philips & Falkenström, in press). For this reason studies which used random allocation or active control groups were excluded.

As a number of included studies reported multiple samples, a standard procedure was developed to support sample extraction. If a pooled study sample was reported then this sample alone was extracted. If only study sub-samples were reported (e.g. CBT vs. behavioural therapy) then each sample was extracted separately if and only if they were independent from each other (i.e., the same patients did not appear in both samples). When both completer and intention-to-treat (ITT) samples were reported the ITT sample only was extracted. When multiple studies used the same or overlapping data-sets then only one was study was included, decided on a case-by-case basis by the current first author.

Outcomes

The outcomes of interest were patient-reported pre-post treatment measures. Outcome studies which employed clinician-rated measures were excluded to reduce heterogeneity. Only validated outcome measures that assessed psychotherapy effectiveness (i.e. not process, predictors, well-being or satisfaction) were included. Three broad outcome domains were employed, for which each study sample could contribute up to one measure. These domains included depression measures (e.g. Beck's Depression Inventory [BDI], Beck et al., 1996), anxiety measures (e.g. Generalised Anxiety Disorder Scale [GAD-7], Spitzer et al., 2006) and general measures of psychological distress and functioning that did not specifically measure depression or anxiety (this broad category could include measures of other forms of distress, such as symptoms of obsessive-compulsive, post-traumatic stress disorder, etc.). For samples which reported multiple measures appropriate for a single outcome domain then a preference system was followed (Appendix B). This system gave priority to global functioning measures (e.g. Outcome Questionnaire-45 [OQ-45], Lambert et al., 2004) and measures which are more frequently used across routine services.

Study Selection

Search results were exported from electronic databases (.ris files) and imported to reference management software (Mendeley, Zotero) for removal of duplicates. Unique results were imported to a web-based program for title/abstract screening using a data-mining approach ('Rayyan,' Ouzzani et al., 2016). All search results were individually screened by the first author using a pre-developed and piloted screening tool (Appendix C). This was performed for title/abstracts, and then full-texts. A sub-sample of articles were screened by a second rater at each screening stage. This included 20% of titles/abstracts (by a trainee clinical psychologist) and 10% of full-texts (by a qualified clinical psychologist). Agreement and inter-rater reliability statistics (Kappa [κ], Cohen, 1960) were used to quantify screening precision. Descriptive classifiers available for interpreting κ were employed (Landis & Koch, 1977), consisting of 'slight' (0-0.2), 'fair' (0.2-0.4), 'moderate' (0.4-0.6), 'substantial' (0.6-0.8), and 'almost perfect' (0.8-1.0). There was substantial reliability ($\kappa = 0.78$) shown at the abstract/title screening stage (1713/1740, 98.45%) and strong reliability ($\kappa = 0.65$) at the full-text screening stage (24/30, 80%).

Additional Papers

Full-text manuscripts from the electronic database search which progressed to data extraction received two additional checks. First, reference lists were scanned for relevant article titles (i.e. backwards reference searching). Second, studies which cited the included articles were scanned (using GoogleScholar) for relevant titles (i.e. forward citation searching). For grey literature, a pragmatic search was conducted using GoogleScholar (terms = "psychotherapy," AND "routine practice" AND "effectiveness") and reviewing the first 50 pages of results.

For the vast majority of studies included in the narrative review (212/252, 84.13%), corresponding authors were contacted via e-mail for additional effect-size information (see 'Effect-size calculation') with a two-week response time. Within the same e-mail, authors were invited to provide/recommend additional papers which they perceived as relevant to the review. This invitation was only performed for studies identified through the electronic database search. Of the information requests made 177 were for authors to provide correlations while 35 were for additional data (e.g. M, SD etc.). E-mail responses were received for 76 authors (35.85%). Data was provided from authors for 41 samples (19.34%).

Extraction and Coding

Data extraction was performed in two phases. First, studies from the systematic database search, and second for all additional studies. To support the data extraction process a standardised extraction sheet was developed using Microsoft Excel. This spreadsheet was tested with a sample of studies (k = 10), and peer-reviewed by the research team. A coding sheet, summarised in Table 2, was developed to provide a uniform system for defining the levels of each moderator variable. Data from a sub-sample of manuscripts (n = 29) were extracted by a second extractor which demonstrated almost perfect reliability (κ = 0.97, agreement = 97.56%). Further detail regarding the data extraction process (i.e., variables and levels) are provided in Appendix D.

Effect-Size Calculation. All analyses were conducted using the R statistical analysis environment (R Core Team, 2020, v 4.0.2). Effect-size calculation and meta-analyses

Table 2

Sumary coding sheet for extracting study information and categorising by level of moderator sub-group. These moderators form the categorical, sub-group variables for the current study.

Moderator	Level	Description
Setting	Outpatient	Sample of patients treated at an out-patient settings.
	Inpatient	Sample of patients treated at either an (i) inpatient; (ii) day hospital; (iii) residential; or (iv) partial hospital setting.
Completion	Completer	Sample of patients who all completed treatment.
	ITT	Sample of patients who used intention-to-treat principles. This is either (i) true ITT; or (ii) modified ITT (i.e. a minimum number of attended sessions).
Sector	University Clinics	Sample of patients seen at (i) University training clinics; or (ii) University based out-patient clinics.
	Primary	Sample of patients seen at a: (i) primary care; (ii) health; (iii) counselling/University counselling; (iv) voluntary ; (v) private [independent or group]; or (vi) employee assistential/occupational health service.
	Secondary	Sample of patients seen at a: (i) secondary care; (ii) CMHTs /CMHC; (iii) tertiary/specialised psychotherapy; (iv) behavioural health/managed care ; or (v) Intensive out-patient setting.
	Inpatient	Sample of patients treated at either an (i) inpatient; (ii) day hospital; (iii) residential; or (iv) partial hospital setting.
Continent	Continents	Continent of study setting, consisting of either: (i) UK; (ii) mainland Europe; (iii) North America; (iv) Asia; (v) Australasia.
Therapy	Dynamic	Therapy or counselling which follows a psychodynamic orientation.
	CBT	Therapy or counselling which follows a cognitive and/or behavioural orientation.
	Counselling	Counselling which is either (i) person-centered; or (ii) orientation not specified.
	Other	Therapy or counselling which (i) has not been mentioned above, or (ii) is not specified/reported in the study manuscript.
Trainee	Unqualified	Interventions exclusively made up of psychology trainees.
	Other	All other samples/studies.
Hour-glass	Stage-1	Methodologies including: (i) pilot or (ii) preliminary effectiveness studies.
	Stage-3	Methodologies including: (i) service evaluatons; (ii) benchmarking; (iii) routine outcome reporting; (iv) predictors of outcome/drop-out.

were performed using the *metafor* (Viechtbauer, 2020), *dmetar* (Harrer et al., 2019a), and *meta* (Schwarzer, 2020) packages.

Paired-samples Cohen's *d* (Standard mean change, Cohen, 1988) was computed for each study sample by dividing the pre-post mean change by the pre-treatment standard deviation (see Figure 1). Sample variance was adjusted using Pearson's *r* in order to account for the inherent violation of independence (i.e. regression to the mean) in pre-post comparisons (Cuijpers et al., 2017). This approach has been advocated for benchmarking of pre-post outcomes studies (Minami et al., 2008).

$$d = \frac{Mean^2 - Mean^1}{SD^1}$$

Figure 1

Formula for standardised mean change. Where 1 is the sample pre-intervention and 2 is the sample following intervention.

Due to the fact that the majority of manuscripts did not report all of the required information for calculating this variant of Cohen's *d*, a hierarchical stepped approach was used to handle the missing information (see Table 3). For studies which reported all of the required information (N, M^1, M^2, SD^1, r) then *d* was calculated without additional consideration. For studies which (commonly) did not report *r* then we e-mailed corresponding authors (two-week response time) to request this information. When unsuccessful *r* was imputed using an empirically supported estimate ($r = .60$, Balk et al., 2012). For studies which did not provide the more fundamental figures (M^1, M^2 or SD^1) but reported a paired samples Cohen's *d* (any variant) then this effect-size was extracted. For studies which did not report fundamental figures *and* did not report a Cohen's *d*, then e-mail requests were sent to corresponding authors. If this was unsuccessful then we applied conversion formulas in situations when studies reported alternative quantitative metrics (e.g. median, range, standard error, ANOVA, regression) to generate means and standard deviations. In situations when all of these steps

Table 3

Hierarchical proceadure for effect-size calculation.

Steps	Scenario	Response
Step 1	Manuscript reports all required information (N, M1, M2, SD1) for preferred d.	Calculate preferred d.
Step 2	Manuscript reports all information apart from Pearson's r.	E-mail corresponding authors to request r.
Step 3	Manuscript does not report the mean or standard deviation but reports paired samples d.	Use the reported d within the manuscript.
Step 4	Manuscript does not report mean, standard deviation, or paired samples d however reports alternative metrics (e.g. median, range, standard error, ANOVA, regression	Estimate the meand and standard deviation by converting available metrics.
Step 5	All above steps attempted without success	Study is not included in meta-analysis but is retained for narrative synthesis.

were unsuccessful/not applicable then the studies in question were removed from the meta-analyses (included in the narrative synthesis only).

Risk of Bias and Methodological Quality Assessment

The Joanna Briggs Institute Quality Appraisal Tool for Case Series (Appendix E, Munn et al., 2020) was used to assess risk of bias for all reviewed studies. The items within this tool were judged by the review team to be of relevance to studies in naturalistic settings. Two items concerning outcome measurement were removed to make an adapted 8-item tool. This is because the review inclusion criteria (validated nomothetic measure of effectiveness) would implicitly mean that every study would meet these criteria. Included study samples were rated as having either met or not met each criteria (yes/no/not sure). The authors of the tool do not provide a scoring classification system or cut-off (Munn et al., 2020). They advise that such decisions should be made by the reviewers who employ them. For this review each 'yes' was given a score of one. Each study subsequently received a cumulative bias score (range = 0-8) with higher scores indicating less risk of bias. All studies were rated by the first author while a sub-sample (23.8%) was second rated by a pair of MSc psychological research methods students (11.9% each). Inter-rater reliability at this stage was substantial ($\kappa = 0.67$, agreement = 86.25%)

The methodological quality of the evidence within each meta-analytic comparison was assessed by three reviewers using guidelines for the *Grading of Recommendations, Assessment, Development and Evaluations* (GRADE, Guyatt et al., 2008). This framework rates evidence quality for each meta-analytic outcome based on included study designs. Individual ratings are initially provided (high, moderate, low or very low) and are then down-graded (or upgraded) through evaluation of five separate criteria: (i) risk of bias within included studies, (ii) inconsistencies in aggregated treatment effect, (iii) indirectness of evidence, (iv) imprecision, and (v) publication bias.

Data Synthesis

Random-effects meta-analyses were used to estimate pooled effect sizes. Pooled and weighted effect-sizes with 95% confidence intervals were calculated for all included study samples. Due to the anticipated high number of studies, forest plots without study details were employed to illustrate the pattern of study effects and the overall pooled estimate. This decision was made to aid visual interpretation as the large volume of included studies meant identifying individual study effect sizes from the plot was challenging[1]. The number of patients needed to treat (i.e. number needed to treat, NNT) in order for one patient to receive a positive outcome was calculated using the method proposed by Kraemer & Kupfer (2006). The extent of between-study heterogeneity was assessed using I^2 (Higgins & Thompson, 2002) and the Q statistic (Cochran, 1954). I^2 was interpreted as low (25-50%), moderate (50-75%) or high (75-100%, Higgins et al., 2003). The impact of publication bias on treatment estimates was visualised 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 fail-safe N (Rosenthal method, Rosenthal, 1979).

Moderator Analyses

Pre-defined moderator analyses were conducted for sub-groups, distinguished by categorical variables while meta-regression was used for continuous variables. There were

[1] Effect-size details for all studies included in the current study are shown in the supplementary material, url link provided in Appendix F.

seven sub-group moderator variables: (i) setting, (ii) type of completion sample, (iii) sector, (iv) region, (v) therapy modality, (vi) experience, and (vii) stage of the treatment evaluation (i.e. hour-glass model). There was eight continuous moderator variables. These included: (i) year of study publication, (ii) average age of sample, (iii) treatment dosage (i.e. number of out-patient sessions), (iv) rate of sample from a minority ethnic background, (v) rate of sample married, (vi) rate of sample in full-time employment, (vii) rate of sample who were female, and (viii) study risk of bias scores (arising from risk of bias assessment). Bonferroni adjustments were applied to each group of moderators, resulting in p-values of .00714 for categorical variables (.05/7) and .00625 for continuous variables (.05/8).

For moderators that produced significant meta-analytic models then between sub-group pairwise comparisons were made. This consisted of inspecting whether sub-group effect-size classifications differed and also whether confidence intervals showed overlap.

The approach to multivariable moderator analysis followed available guidance (Harrer et al., 2019b). Variables were selected based on suspected interactions by the research team. This included completion methodology and mean number of sessions (in influencing effect-size was assessed). Completion analysis, coded as a dummy variable (ITT vs. completion) was first entered into an initial multi-regressive model. A second 'full' model was then built using both completion analysis and mean number of sessions. Differences in these two models were assessed using a log-likelihood ratio rest, with a p-value of p = <.05 required to indicate significant model differences. If models were significantly different then comparisons were made between log-likelihood scores and information criteria statistics. A lower log-likelihood score, and smaller Akaike's-information criteria (AIC) score would indicate improved model fit. Following this, completion methodology and treatment dosage were modeled as interaction terms in the final model. This stage was conducted regardless of whether predictors were significant in isolation within previous models.

Results

Search Results

The systematic search of electronic databases produced 10,503 results. After removal of duplicates, this was reduced to 8,709. Following title/abstract screening there were 325 potentially eligible records remaining. Of these, 30 manuscripts were not available through the first author's institution. E-mails were sent to corresponding authors to request access. This led to the retrieval of a further 8 manuscripts. Articles which remained without full-text manuscripts (22, 6.77%) were excluded from the review on the basis that the full-text was not available for eligibility screening. On completion of the full-text screening process there were 130 studies remaining[2]. All of these articles were included in the narrative (qualitative) synthesis.

Through forward citation searching and backwards reference searching a further 197 articles were identified. Finally, 97 articles were found through grey literature/provided by authors. A PRISMA flow diagram (Schulz et al., 2010) is shown in Figure 2. A break-down of screening decision results are provided in Table 4 with a full-list in Appendix G.

Table 4

Full-text screening decision results for all studies

Decision	Phase 1	Phase 2	Phase 3	Total
Exclude	174	111	54	339
Include	130	79	43	252
No Access	21	7	0	28
Total	325	197	97	619

Note. Difinitions for different phases are as follows:
Phase 1 = studies identified through the electonic database search.
Phase 2 = studies identified through phase 1 reference lists and citation searching.
Phase 3 = grey literature and studies provided by authors contacted during the study.

[2] Due to the high number of studies included in the narrative synthesis and meta-analysis, only in-text citations are reported in the attached bibliography. A complete bibliography is instead available in the supplementary material, url link provided in Appendix F.

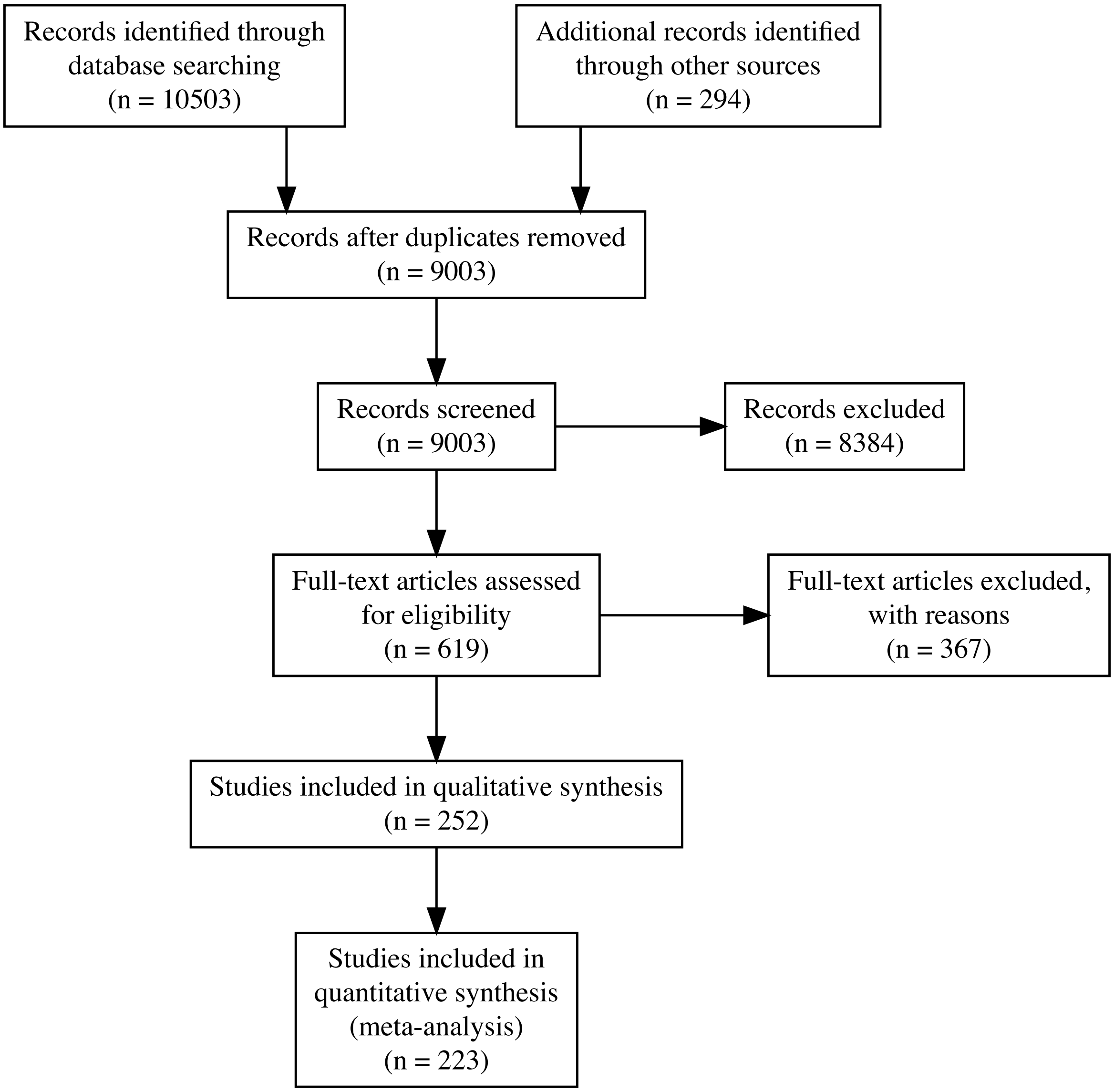

Records identified through database searching (n = 10503)
Additional records identified through other sources (n = 294)
Records after duplicates removed (n = 9003)
Records screened (n = 9003)
Records excluded (n = 8384)
Full-text articles assessed for eligibility (n = 619)
Full-text articles excluded, with reasons (n = 367)
Studies included in qualitative synthesis (n = 252)
Studies included in quantitative synthesis (meta-analysis) (n = 223)

Figure 2

Prisma flow diagram of studies throughout the review.

Narrative Synthesis

All 252 studies included in the data extraction phase were included in the narrative synthesis. Of these studies, 223 were eligible for inclusion in the meta-analysis. As a number of studies provided multiple samples, the total number of samples exceeded the number of studies. There was subsequently 298 samples in the narrative synthesis, while there was 263 samples included in the meta-analysis[3]. Summary statistics for included studies are provided in Table 5.

Methodological Information

Of the samples included in the narrative synthesis, the year of publication ranged from 1984 to 2020. The median publication date was 2013. There were 294 samples reported since 2000, 213 since 2010 and 126 since 2015.

The number of ITT samples, including those that used modified ITT (i.e. studies specifying a minimum number of attended sessions for inclusion) or when ITT could only be assumed was 169. The number of studies that used our more rigorous definition of ITT was 64, while the number of studies that used a completer sample was 118. When distinguishing studies based on the stage of the hour-glass model there was 34 (11.41%) samples at stage-1 and 264 (88.59%) at stage three.

Sample Characteristics

Demographic information was reported for 291 of the 298 samples included in the narrative synthesis. The demographic sample size for samples included in the narrative synthesis ranged from 4 (Sauer-Zavala et al., 2019) to 33,243 (Pybis et al., 2017, CBT sample). The pooled demographic sample size for the narrative synthesis was 233,140. Self-reported gender information was available for the majority of samples (k = 279). Of these samples 144,273 (61.88%) patients were females. When averaging across available percentages, there was a pooled average of 66.00% females. There were 13 exclusively female samples and 2 exclusively male samples. Within studies that reported a mean average age the pooled average

[3] Full details of the study characteristics for all included studies are reported in the supplementary material, url link provided in Appendix F.

age was 35.33 years (range = 19.00 - 60.50).

There were 127 samples that reported the number or percentage of patients from minority ethnic backgrounds. The mean percentage of patients from minority ethnic backgrounds was 23.00%. For marital status, 106 samples reported relevant data with a mean average (patients who were married) of 23.00%. There were 96 samples that reported employment status. The mean percentage of patients in employment across samples was 56.00%.

Table 5

Summary statistics across the pooled sample and also by sector for varying variables.

	Level	Uni Clin	Primary	Secondary	Inpatient	Other	Total
N	Female	5350	95373	14952	5797	22801	144273
	Total	9195	158150	22586	9515	33694	233140
Age	samples	65	77	82	29	7	260
	mean	33	36	35	34	36	176
	min	20	19	21	24	24	109
	max	52	60	52	47	46	258
Sessions	samples	54	64	54	4	6	182
	mean	21	11	14	13	8	69
	min	2	4	1	9	8	24
	max	85	64	64	24	9	247
Setting	Mixed	0	0	0	0	5	5
	Outpatient	68	96	91	0	4	259
	Inpatient	0	0	1	33	0	34
Completion	ITT	48	48	53	16	4	169
	Check	1	2	4	1	1	9
	Completers	19	45	35	16	3	118
Therapy	CBT	43	41	49	14	5	152
	Counselling	0	22	3	0	0	25
	Dynamic	12	9	16	13	0	50
	Other	13	24	24	6	4	71
Hour Glass	Stage-1	4	6	16	7	1	34
	Stage-3	64	90	76	26	8	264
Continent	Asia	4	1	0	0	1	6
	Australasia	5	0	5	0	0	10
	Europe	20	13	14	15	1	63
	N.America	38	32	39	10	4	123
	UK	1	50	34	8	3	96
Continent	**Total**	**68**	**96**	**92**	**33**	**9**	**298**

Service Information

The country that contributed most samples was the USA (*k* = 113), followed by England (*k* = 78), Germany (*k* = 24), Sweden (*k* = 12), and Canada (*k* = 10). These five most well-represented countries accounted for the majority of the included samples (*k* = 237). For continent, when differentiating the UK from mainland Europe, the order of continental representation was North America (*k* = 123), the UK (*k* = 96), mainland Europe (*k* = 63), Australasia (*k* = 10), and Asia (*k* = 6).

For treatment setting, there were 96 (32.21%) samples in the primary care category, 92 (30.87%) in the secondary care category, 33 (30.87%) in the inpatient care category, and 68 (22.82%) from University clinics. There was 9 (3.02%) samples from a combination of sectors (i.e. other)

Treatment Information

In terms of treatment modality, 152 treatments were classified as cognitive and/or behavioural therapies, 50 were dynamic/interpersonal therapy, 25 were classified as non-specific or person-centered counseling, and 71 were classified as other. Treatment duration metrics were reported for the majority of study samples (*k* = 256). The most common treatment duration metric was sessions (or hours, *k* = 225), followed by months (*k* = 12), and then days (*k* = 8). There was no treatment duration metric available for 42 samples. The pooled mean across those studies which reported the mean number of sessions, was 16.30 sessions (range = 1.00-139.30).

There were 62 samples reported as exclusively consisting of unqualified (i.e., trainee) clinicians; while 100 samples reported having at least one unqualified clinician.

Risk of Bias and Methodological Quality

In terms of study risk of bias assessment, mean average bias score across samples included in the narrative review was 5.53 (SD = 1.47, range = 1-8)[4]. The most frequently met criteria was for demographic reporting detail (264/298), followed by service reporting

[4] Total bias score for each individual study is reported in the supplementary material, url link provided in Appendix F.

detail (260/298). The least frequently met criteria was for complete inclusion (i.e. consecutive recruitment and intention-to-treat analysis, 41/298) followed by consecutive inclusion (93/298).

Meta-Analyses

Each outcome domain had a primary meta-analysis. A summary of the primary meta-analyses is shown in Table 6. There was a wide variety of specific measures employed within each meta-analysis (as shown in Appendix B). During the GRADE methodological appraisal process, each of the meta-analysis were initially rated as 'low,' based on the predominant type of study design within the available evidence. Following review of the five GRADE areas these overall ratings were reduced in level to 'very low' based on study limitations and also inconsistency within the available evidence.

Table 6

Findings from the primary meta-analyses.

Variable	k	ES	Lower	Upper	p	I2	Q
Depression	140	0.98	0.90	1.06	< 0.001	98.40	3037.46
Anxiety	84	0.83	0.73	0.92	< 0.001	97.52	1488.88
General	184	1.01	0.93	1.08	< 0.001	98.92	15685.18

Depression

For depression outcomes, (*k* = 140 samples), 10 different outcome measures were used. The most frequently used depression measure was the Beck Depression Inventory (BDI I or II, *k* = 78), followed by the Patient Health Questionnaire (PHQ-9, *k* = 30) and then the Brief Symptom Inventory ([BSI] Depression Index, *k* = 8). The depression meta-analysis had a combined N of 68,077. Individual study effect-sizes are illustrated in the depression forest plot in Figure 3. The pooled effect-size was significant, indicative of a large (*d* = 0.98, [CI 0.9-1.06], *p* = < 0.001, GRADE = very low) reduction in depression symptoms. The number of patients needed-to-treat in order to provide one patient with a positive outcome was 1.95. There was evidence of significant study heterogeneity (I^2 = 98.4%, Q[df = 139] = 3,037.46, *p* = < 0.001). The funnel plot in Figure 4 shows limited visual evidence of asymmetry. The funnel rank correlation test was not significant ($\tau = 0.028$, $p = 0.629$). In contrast, the funnel regression test was significant (Z = 2.665, $p = 0.0077$). The fail-safe N indicating the number of studies reporting no intervention effect that would be required to make the aggregated effect not significant was N=736,945.

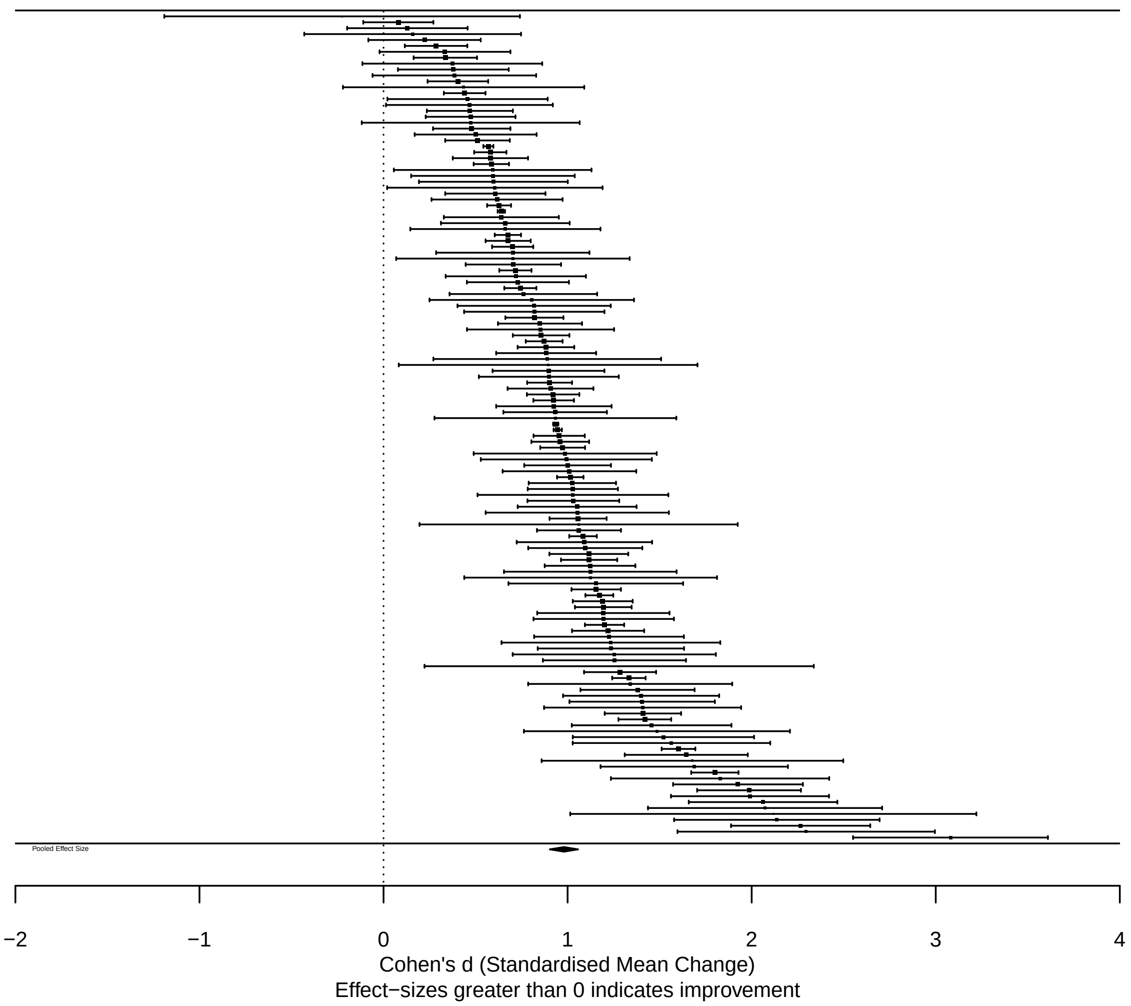


Figure 3

Forest plot of pre-post psychological therapy effect sizes for depression outcomes. Square boxes depict individual study Cohen's d effect sizes, error bars display 95 percent confidence intervals and the diamond represents the pooled estimate effect.

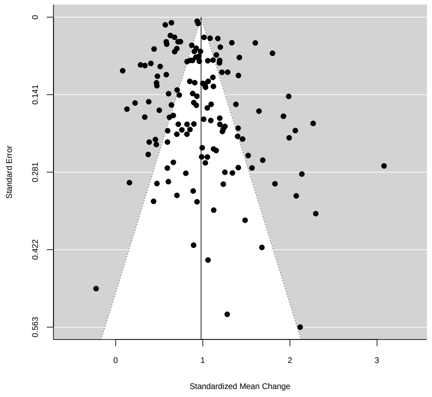


Figure 4

Funnel plot of the distribution of studies reporting pre-post depression outcomes.

Anxiety

For anxiety outcomes, (k = 84 samples), 20 different outcome measures were used. The most frequently used measure was the the Beck Anxiety Inventory (BAI, k = 19), followed by the Generalised Anxiety Disorder (GAD-7, k = 19), and then the Brief Symptom Inventory ([BSI] Anxiety Index, k = 8). The anxiety meta-analysis had a combined N of 26,689. Individual study effect-sizes are illustrated in the anxiety forest plot in Figure 5. The pooled effect-size was significant, indicative of a large (d = 0.83, [CI 0.73-0.92], p = < 0.001, GRADE = very low) reduction in anxiety symptoms. The number of patients needed-to-treat in order to provide one patient with a positive outcome was 2.26. There was evidence of significant study heterogeneity (I^2 = 97.52%, Q[df = 83] = 1,488.88, p = < 0.001). The funnel plot in Figure 6 shows limited evidence of asymmetry. The funnel rank correlation test was not significant (τ = 0.061, p = 0.416). In contrast, the funnel regression test was significant (Z = 3.186, p = 0.0014). The fail-safe N was 155,478

General

For general outcomes, (k = 184 samples), 40 different measures were used. The most frequently used measure was the CORE[5], (k = 40), followed by the Brief Symptom Inventory ([BSI] Global Severity Index, k = 26), the Symptom Checklist 90 ([SCL] Global Severity Index, k = 21) and then the Outcome Questionnaire[6] (k = 14). The meta-analysis for general outcomes had a combined N of 126,734. Individual study effect-sizes are illustrated in the general forest plot in Figure 7. The pooled effect-size was significant, indicative of a large (d = 1.01, [CI 0.93-1.08], p = < 0.001, GRADE = very low) reduction in general symptoms. The number of patients needed-to-treat in order to provide one patient with a positive outcome was 1.91. There was evidence of significant study heterogeneity across the included studies (I^2 = 98.92%, Q[df = 183] = 15,685.18, p = < 0.001). The funnel plot (see Figure 8) shows a degree of asymmetry with clustering to the right of the mid-line. The funnel rank correlation test was significant (τ = 0.228, p = <0.001). In contrast, the funnel regression test was not

[5] This included instances of either CORE-10 and CORE-OM.

[6] This included instances of either OQ-30 or OQ-45.

significant (Z = -0.733, *p* = 0.46). The fail-safe N was 2,018,805.

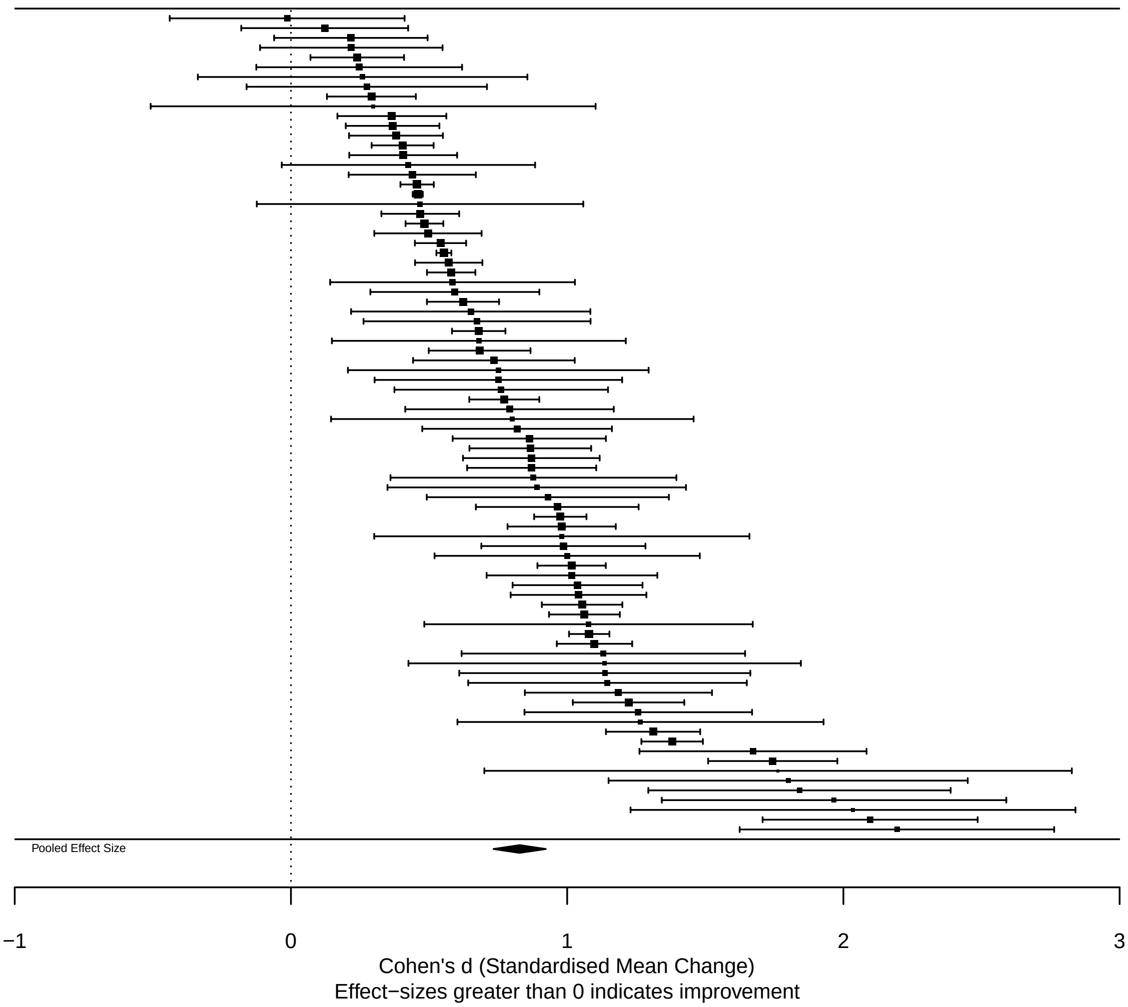


Figure 5

Forest plot of pre-post psychological therapy effect sizes for anxiety outcomes.

Square boxes depict individual study Cohen’s d effect sizes, error bars display 95 percent confidence intervals and the diamond represents the pooled estimate effect.

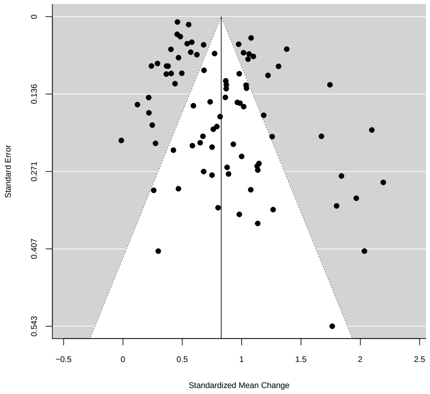


Figure 6

Funnel plot of the distribution of studies reporting pre-post anxiety outcomes.

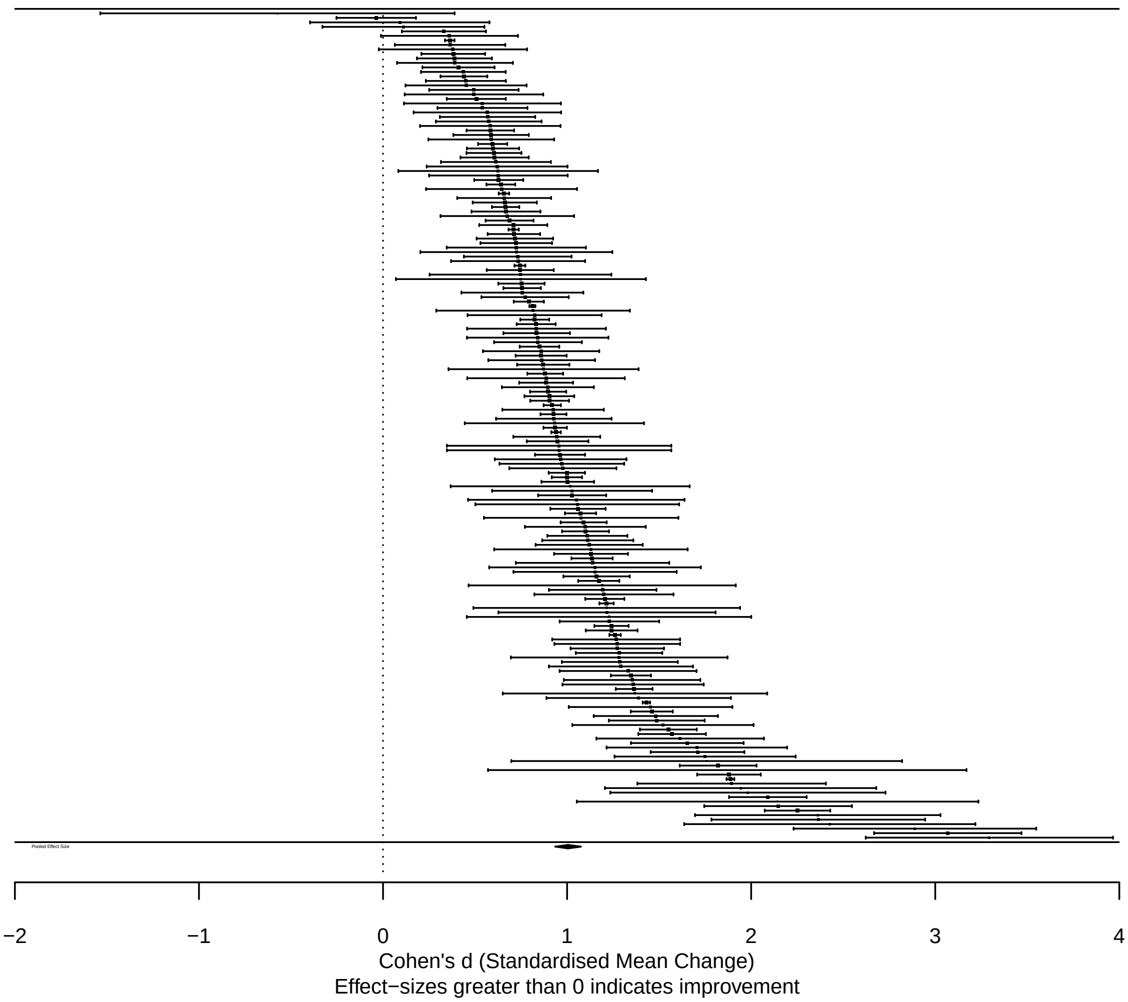


Figure 7

Forest plot of pre-post psychological therapy effect sizes for general outcomes.

Square boxes depict individual study Cohen's d effect sizes, error bars display 95 percent confidence intervals and the diamond represents the pooled estimate effect.

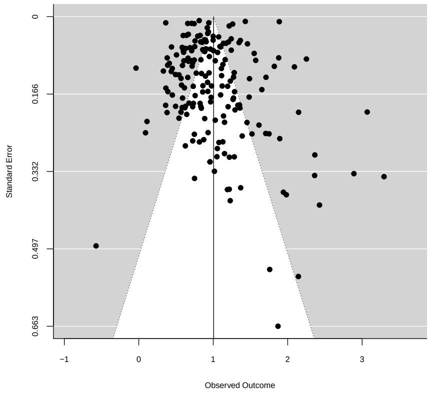


Figure 8

Funnel plot of the distribution of studies reporting pre-post general outcomes.

Moderator Analyses

Univariate Moderators. Categorical moderator analyses (i.e. sub-groups) are reported in Tables 7-9. Of the eight moderators, there were two variables that were significant for all three outcome domains (completion sample and continent). Completer analyses consistently had larger effect sizes compared to ITT analyses across all samples, with no overlap between confidence intervals (CI). For continent, UK and North American studies had larger effect sizes than other continents (mainland Europe, Australasia and Asia), with varying levels of overlap in CI among the other subgroups. To a lesser extent Europe also had larger effect sizes than Australasia and Asia. UK and North American pooled effect-sizes were comparable to each other across domains.

The remaining six moderators were not consistent across outcome domains. Study setting was significant for the anxiety and general domains, although both analyses showed CI overlap. Outpatient samples out-performed inpatient care for anxiety, while the reverse was shown for the general domain. Sector was significant for the anxiety and general domains. Anxiety samples showed greater average outcome within primary and University clinic sectors. For general outcomes, secondary services and University clinics had lower effect sizes than other sectors. Type of therapy was significant for the anxiety and general domains. Anxiety samples which accessed dynamic based therapies had larger effect sizes compared to other interventions (with overlapping CI). For general samples, CBT-based interventions had higher effect sizes compared to the other therapy meta-categories with no CI overlap. Stage of the hour-glass model was significant only for the general domain. Stage-one samples (pilot studies, preliminary evaluations of treatments) had larger effect sizes than stage-three samples, with large overlap in CI. Finally, for experience (i.e. exclusively unqualified samples vs. not), significant results were found for the anxiety and general domains. Anxiety samples exclusively consisting of unqualified clinicians had higher effect sizes than other samples (no CI overlap). For the general domain the reverse was shown, unqualified clinician samples had lower effect sizes than other samples (no CI overlap).

Between-study heterogeneity was also explored using eight continuous variables (see table 10). Neither mean age, proportion of ethnic minority patients, or proportion of married patients were significant for any of the outcome domains. Risk of bias score was significant for all three domains, with higher quality scores linked to larger effects. Year of publication was significant only for anxiety, suggesting that more recent studies produce greater effect-sizes for anxiety. Mean number of sessions was significant only for depression, suggesting that treatment effectiveness increases in line with a greater number of sessions received. Employment was significant only for anxiety, suggesting that studies with greater employment rates show larger effect-sizes. Proportion of female patients was significant for the anxiety domain. Greater female representation was linked with lower anxiety effect-sizes.

Multivariable Moderators. Multivariable moderator analysis was conducted for completion methodology and mean number of sessions. For depression, the full model found completion methodology, but not mean number of sessions to be a significant individual predictor. The overall test of moderators was significant. There was no significant difference in model fit, based on the log-likelihood ratio test. For the interaction model, the overall test of moderators was significant, however neither of the predictor variables or the interaction term were significant in isolation.

For anxiety, the full model found neither completion methodology or mean number of sessions to be a significant individual predictors. The overall test of moderators was not significant. There was no significant difference in model fit, based on the log-likelihood ratio test. For the interaction model, the overall test of moderators was not significant. The individual predictor variables and also the interaction term were not significant in isolation.

For general outcomes, the full model found neither completion methodology or mean number of sessions to be a significant individual predictors. The overall test of moderators was not significant. There was no significant difference in model fit, based on the log-likelihood ratio test. For the interaction model, the overall test of moderators was not significant. The individual predictor variables and also the interaction term were not significant in isolation.

Table 7

Sub-group (categorical) moderator analyses for depression outcomes.

Moderator	Level	k	Effect Size	Confidence Intervals	Q	I2
Random effects model for sector (Q = 5.99, p = 0.2)						
Sector	Primary	31	1.06	0.99 - 1.13	9547771.37	1.00
	Uni. Clinics	29	0.96	0.86 - 1.07	43195.81	1.00
	Secondary	55	0.97	0.91 - 1.04	80785.29	1.00
	Inpatient	15	0.91	0.7 - 1.12	237284.01	1.00
Random effects model for ITT (Q = 13.88, p = <0.001**)						
	ITT	76	0.92	0.88 - 0.97	9728418.18	1.00
Completion	Completers	58	1.09	1.01 - 1.17	220978.90	1.00
Random effects model for setting (Q = 3.82, p = 0.148)						
	Outpatient	115	0.99	0.95 - 1.02	9657600.36	1.00
Setting	Inpatient	16	0.92	0.74 - 1.1	240002.33	1.00
Random effects model for continent (Q = 27.23, p = < 0.001**)						
Continent	N.America	56	0.99	0.9 - 1.08	640246.22	1.00
	UK	43	1.09	1.05 - 1.14	3564834.35	1.00
	Europe	26	0.94	0.82 - 1.05	59071.64	1.00
	Australasia	4	0.67	0.33 - 1	7087.67	1.00
	Asia	5	0.59	0.35 - 0.83	91.78	0.96
Random effects model for therapy modality (Q = 1.5, p = 0.682)						
Therapy	Dynamic	22	1.01	0.82 - 1.19	41858.50	1.00
	Counselling	6	0.89	0.72 - 1.07	3471906.01	1.00
	CBT	88	1.00	0.96 - 1.05	393105.73	1.00
	Other	18	0.98	0.77 - 1.19	307046.63	1.00
Random effect model for training samples (Q = 1.36, p = 0.244)						
	No/NA	116	1.01	0.97 - 1.04	9876336.39	1.00
Trainees	Yes	18	0.89	0.7 - 1.08	76274.10	1.00
Random effects model for hour-glass stage (Q = 1.37, p = 0.242)						
	Stage-3	113	1.00	0.96 - 1.03	9952776.03	1.00
HourGlass	Stage-1	21	0.93	0.82 - 1.04	464.36	0.96

Note. Model Outputs in Bold are significant at either * p = <.05. ** Bonferroni adjustment, p = <.007

Table 8

Sub-group (categorical) moderator analyses for anxiety outcomes.

Moderator	Level	k	Effect Size	Confidence Intervals	Q	I2
Random effects model for sector (Q = 128.47, p = < 0.001**)						
Sector	Primary	21	0.99	0.96 - 1.03	329379.72	1.00
	Secondary	24	0.62	0.55 - 0.69	30702.22	1.00
	Inpatient	8	0.59	0.31 - 0.88	108223.63	1.00
	Uni. Clinics	29	1.00	0.89 - 1.11	32067.93	1.00
Random effects model for ITT (Q = 7.55, p = 0.006*)						
	ITT	57	0.77	0.74 - 0.79	512107.17	1.00
Completion	Completers	26	0.96	0.82 - 1.09	92492.19	1.00
Random effects model for setting (Q = 5.75, p = 0.016*)						
	Outpatient	74	0.84	0.82 - 0.87	435141.60	1.00
Setting	Inpatient	9	0.58	0.37 - 0.79	157933.03	1.00
Random effects model for continent (Q = 26.72, p = < 0.001**)						
Continent	N.America	32	0.90	0.81 - 0.98	230641.72	1.00
	UK	25	0.89	0.8 - 0.98	115765.33	1.00
	Europe	19	0.79	0.67 - 0.91	27933.90	1.00
	Australasia	4	0.61	0.28 - 0.94	3781.78	1.00
	Asia	3	0.59	0.49 - 0.69	3.33	0.40
Random effects model for therapy modality (Q = 105.34, p = < 0.001**)						
Therapy	Dynamic	12	0.92	0.67 - 1.17	11879.61	1.00
	Counselling	2	0.43	0.38 - 0.49	28.37	0.96
	CBT	62	0.86	0.79 - 0.93	174811.70	1.00
	Other	7	0.74	0.47 - 1.02	153478.52	1.00
Random effects model for hour-glass stage (Q = 0.31, p = 0.579)						
	Stage-3	73	0.82	0.79 - 0.84	630326.22	1.00
HourGlass	Stage-1	10	0.87	0.69 - 1.05	639.27	0.99
Random effect model for training samples (Q = 13.8, p = < 0.001**)						
	No/NA	65	0.75	0.72 - 0.78	497245.30	1.00
Trainees	Yes	18	1.12	0.93 - 1.32	59913.29	1.00

Note. Model Outputs in Bold are significant at either * p = <.05. ** Bonferroni adjustment, p = <.007

Table 9

Sub-group (categorical) moderator analyses for general outcomes.

Moderator	Level	k	Effect Size	Confidence Intervals	Q	I2
Random effects model for sector (Q = 45.74, p = < 0.001**)						
Sector	Primary	54	1.10	0.99 - 1.2	39094444.62	1.00
	Secondary	58	0.88	0.86 - 0.9	101238.33	1.00
	Inpatient	24	1.07	0.98 - 1.17	65648.02	1.00
	Uni. Clinics	27	0.81	0.73 - 0.89	25376.01	1.00
Random effects model for ITT (Q = 9.79, p = 0.002**)						
	ITT	89	0.95	0.9 - 1	12412018.93	1.00
Completion	Completers	80	1.09	1.02 - 1.17	10416544.97	1.00
Random effects model for setting (Q = 6.18, p = 0.045*)						
	Outpatient	141	1.00	0.92 - 1.08	104239386.44	1.00
Setting	Inpatient	25	1.06	0.98 - 1.14	67134.33	1.00
Random effects model for continent (Q = 16.45, p = 0.002**)						
Continent	UK	60	1.03	0.89 - 1.18	92811055.56	1.00
	N.America	59	1.03	0.97 - 1.09	4599532.56	1.00
	Europe	41	0.98	0.9 - 1.06	143451.90	1.00
	Australasia	4	0.81	0.72 - 0.9	330.52	0.99
	Asia	5	0.91	0.58 - 1.23	575.15	0.99
Random effects model for therapy modality (Q = 46.9, p = < 0.001**)						
Therapy	CBT	77	1.18	1.12 - 1.23	171878.31	1.00
	Dynamic	34	0.88	0.8 - 0.96	48684.40	1.00
	Counselling	19	0.90	0.8 - 1.01	318722.49	1.00
	Other	39	0.87	0.71 - 1.02	103693471.80	1.00
Random effects model for hour-glass stage (Q = 0.15, p = 0.703)						
	Stage-1	23	1.06	0.86 - 1.26	4506.87	1.00
HourGlass	Stage-3	146	1.02	0.94 - 1.1	104370017.66	1.00
Random effect model for training samples (Q = 20.21, p = < 0.001**)						
	No/NA	144	1.07	0.99 - 1.15	104271844.65	1.00
Trainees	Yes	25	0.76	0.65 - 0.87	58036.50	1.00

Note. Model Outputs in Bold are significant at either * p = <.05. ** Bonferroni adjustment, p = <.007

Table 10

Meta-regression moderator variables (continuous) for depresison, anxiety and general outcome domains.

Domain	Moderator	Mean (range)	k	B	CI	SE	p	Q	R2
Depression	Year (of publication)	(1988 - 2020)	134	0.00	-0.01 - 0	0.00	0.585	0.30	9.14
	Mean age	(19-60 years; M = 36)	122	0.00	0 - 0	0.00	0.751	0.10	23.68
	Sessions (mean)	(1-46 sessions; M = 15)	83	0.01	0 - 0.01	0.00	0.008 *	7.10	18.27
	Ethnicity (% minority)	(0-66%; M = 23%)	61	-0.10	-0.38 - 0.18	0.14	0.482	0.49	0.00
	Martial status (% Married)	(0-73%; M = 35%)	53	-0.10	-0.71 - 0.5	0.31	0.736	0.11	2.36
	Employment (% full-time)	(5-100%; M = 52%)	44	0.37	-0.06 - 0.81	0.22	0.090	2.87	39.11
	Gender (% female)	(0-100%; M = 67%)	127	-0.06	-0.21 - 0.1	0.08	0.476	0.51	7.36
	Risk of Bias (1-10)	(1-8; M = 5.69)	134	0.02	0.01 - 0.04	0.01	<0.001**	13.31	70.63
Anxiety	Year (of publication)	(1999 - 2020)	83	0.02	0.01 - 0.02	0.00	<0.00 **	52.51	0.00
	Mean age	(19-60 years; M = 35)	78	0.00	-0.01 - 0.01	0.00	0.664	0.19	0.00
	Sessions (mean)	(1-46 sessions; M = 16)	52	0.00	0 - 0	0.00	0.997	0.00	0.00
	Ethnicity (% minority)	(0-59%; M = 19%)	40	0.33	-0.19 - 0.85	0.26	0.210	1.57	0.00
	Martial status (% Married)	(3-81%; M = 35%)	35	-0.10	-0.57 - 0.37	0.24	0.678	0.17	0.00
	Employment (% full-time)	(5-100%; M = 60%)	28	1.00	0.59 - 1.42	0.21	<0.001**	22.54	23.16
	Gender (% female)	(0-100%; M = 66%)	78	-0.34	-0.47 - -0.21	0.07	<0.00 **	27.03	18.27
	Risk of Bias (1-10)	(1-8; M = 5.84)	83	0.05	0.02 - 0.09	0.02	0.004 *	8.32	0.00
General	Year (of publication)	(2000 - 2020)	169	0.00	-0.01 - 0.02	0.01	0.870	0.03	0.00
	Mean age	(22-52 years; M = 35)	147	0.00	-0.01 - 0.01	0.01	0.808	0.06	0.00
	Sessions (mean)	(1-65 sessions; M = 15)	102	-0.01	-0.01 - 0	0.00	0.257	1.29	0.00
	Ethnicity (% minority)	(0-70%; M = 25%)	67	-0.47	-1.12 - 0.18	0.33	0.159	1.99	0.00
	Martial status (% Married)	(3-81%; M = 41%)	54	-0.15	-0.48 - 0.17	0.17	0.351	0.87	0.00
	Employment (% full-time)	(0-100%; M = 53%)	59	-0.10	-0.54 - 0.34	0.22	0.651	0.20	6.69
	Gender (% female)	(0-100%; M = 67%)	162	-0.31	-0.74 - 0.12	0.22	0.154	2.04	0.00
	Risk of Bias (1-10)	(1-8; M = 5.69)	169	0.05	0.01 - 0.09	0.02	0.010 *	6.60	39.01

Note. Model Outputs in Bold are significant at either * p = <.05. ** Bonferroni adjustment, p = <.00625

Discussion

The aim of this review was to provide a rigorous and comprehensive evaluation of the effectiveness of psychological therapies delivered in routine practice, and also to explore a range of potential moderators of treatment effectiveness. A broad and inclusive approach was taken, resulting in a large number of eligible studies (*k* = 252) and samples (*k* = 298) for a narrative synthesis. Of these, a large number of studies were also eligible for the meta-analysis (*k* = 223 [88.5%], samples = 263). This review is the largest synthesis of effectiveness studies concerning adult one-to-one psychological therapy conducted to date, expanding on prior reviews in breadth (number of studies, settings, treatment modalities) and depth (meta-analyses, risk of bias and quality appraisals, moderator analyses).

Summary of Findings

The large number of studies included in this review reflects the increase in publication of practice-based evidence; that is, the majority of studies (71.48%) were published since 2010. Consistent with prior reviews of effectiveness, we found large pre-post treatment effects for psychological therapy in the treatment of depression, anxiety and global outcomes (i.e., psychological distress, symptoms and functioning). Method of analysis (ITT vs. completers), study continent, and methodological quality rating were significant moderators across all three treatment domains. A number of additional moderator variables were significant, but not for all domains. There was no evidence of a significant interaction between mean number of sessions and completion methodology. The finding that a large amount of PBE was conducted at stage three of the hour-glass model (add citation again for clarity) is an indication that contemporary services are largely implementing evidence-based practice.

Contribution to the Evidence Base

This review builds on prior reviews of psychological therapy effectiveness in routine care. Consistent with prior reviews, there was strong evidence that psychological therapy leads to clinical improvements across a range of outcomes. The observed large pre-post treatment effect-size for depression outcomes (*d* = 0.98) was consistent with prior effectiveness reviews of depression outcomes reported by Wakefield et al. (2021) (*d* = 0.87) and Hans & Hiller (2013) (*d* = 1.13 [completers]). The large pre-post effect-size for anxiety outcomes (*d* = 0.98) was consistent with that reported by Wakefield et al. (2021) (*d* = 0.88, CI = 0.79-0.97) and the array of large effects-sizes for specific anxiety disorders reported by Stewart & Chambless (2009). Finally, the pre-post treatment effect-size for global outcomes (*d* = 1.01), although somewhat lower than Cahill et al. (2010) (*d* = 1.29) remained within the 'large' effect-size classification. This review expands on prior reviews through utilising a much larger sample of, and more diverse array of, routine services. This was also (to our knowledge) the first effectiveness review to focus on individual (i.e. one-to-one) psychological therapy.

The review found that the majority of individuals accessing psychological therapy across these studies were female. This rate of female over-representation is consistent with findings from other reviews of therapy effectiveness (e.g. 60.2% Wakefield et al., 2021) and global epidemiological studies of mental health prevalence (Seedat et al., 2009).

This review identified three moderator variables as significant across outcome domains. The finding that completer samples had consistently larger effect sizes relative to ITT samples was consistent with prior effectiveness reviews (Hans & Hiller, 2013; Wakefield et al., 2021). The consistency of this finding across a large sample of studies supports prior claims that completer samples may run the risk of providing over-inflated effect-sizes (i.e. type I error, Fergusson et al., 2002).

The finding that continent of study was a significant moderator across domains was a novel finding. Differences in therapy outcomes between continents has, to our knowledge, not been explored in prior outcome studies or meta-analyses. This review found larger effect-sizes for UK and North American studies. Caution in interpretation is required as there was high overlap in CIs. In explaining this finding, 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)[7].

[7] APA: American Psychological Association, NICE: National Institude of Clinical Excellence

The third significant moderator across domains was the continuous variable of risk of bias. Higher effect-sizes were associated with higher quality rating scores. This finding is in contrast to prior evidence which has demonstrated that greater methodological quality is associated with smaller effect sizes (e.g. Wakefield et al., 2021). The discrepancy between this finding and the extant literature is hard to explain, but may be due to differences in appraisal tools/approach to analysis. The appraisal tool in this study was designed for use with case-series designs, with a number of points based on reporting quality as opposed to methodological bias. It is possible that a tool which places greater emphasis upon other areas of bias, such as aspects of internal validity, may have provided a different pattern of results. It is also possible that treating quality rating as a sub-group moderator (i.e. as done by Wakefield et al., 2021) and not a meta-regression variable may have accounted for some of the differences in results.

There was a range of other significant moderator variables that were not significant across all three domains. Stage of the hour-glass model has not previously been explored for its potential influence on effect-size. This variable was significant, but only for the general domain. Higher effect-sizes were demonstrated for preliminary/pilot studies. It is possible that interventions within this earlier stage of development are provided with relatively more resource and impose more internal controls than established, routinely-delivered interventions within benchmarking/evaluation studies.

Therapy modality was a significant moderator for the anxiety and general domains. This finding goes against the well-established *equivalence paradox* in psychotherapy literature; that is, no significant difference in effectiveness between therapeutic models has consistently been shown (Wampold et al., 1997). This review found that, for the general outcome domain, CBT produced higher effect-sizes, with no overlap in CI. Therapy modality was also significant for the anxiety domain however the overlap in CI reduces confidence in identifying a superior treatment. In explaining the superiority of CBT for global outcomes, it is possible that this was due to the inclusion of specific conditions (e.g. PTSD, OCD) for which CBT has a stronger evidence base.

An additional noteworthy finding is the differences shown in outcomes between qualified and unqualified clinicians. This was somewhat surprising as prior outcome studies have consistently found that unqualified clinicians do not produce significantly different effect-sizes to qualified clinicians (e.g. Buckley et al., 2006). Qualified therapists produced significantly higher effect-sizes for the general domain and smaller effect-sizes for the anxiety domain. A potential explanation for this is that unqualified staff are highly supervised and may therefore may be less likely to 'therapeutically drift' (Waller & Turner, 2016) from the identified therapeutic model than qualified clinicians. Training clinicians are also likely to have received more up-to-date training on evidence-based approaches and perhaps are more routinely required to engage in deliberate skills practice. It is not clear however why this set of differences would only apply to anxiety conditions, and why the reverse was shown for general outcomes.

Limitations

There are seven main limitations. First, all studies included in this review were observational by definition and design (i.e. no control group or randomisation). The absence of comparison conditions means that we are unable to rule out alternative explanations for observed effect-sizes such as regression to the mean.

Second, therapies were simply grouped by meta-therapy category. No fidelity/adherence checks were made. This means that we are unable to say with any confidence how much the interventions actually represent intended treatments.

A third limitation is the exclusive focus on self-reported outcome measures. This review of effectiveness therefore is defined by the patient only and does not necessarily extend to effectiveness of routine treatment as defined by the researcher/clinician. Self-report measures are naturally prone to self-perception bias.

A fourth limitation concerns aspects of methodological precision. Because this review was conducted as part of a doctoral thesis the resources required to double rate all aspects of the project were not available. A substantial proportion of this large review was therefore done by a single clinician. In an effort to overcome this limitation integrity checks were conducted at several stages, with each showing substantial to near perfect reliability. A related strength is that the current review attempted to contact a large number of authors to request additional unreported information (Pearson's *r*) to improve precision of sample variance. It is possible however that because a large proportion of data remained unavailable, and was subsequently imputed, that this may have introduced bias. The risk of this happening was high as the response rates from authors was generally low. This was likely to have been influenced by the pragmatic decision to use a single e-mail template, with minimal tailoring. A further point on precision is that statistical interpretation of subgroup differences using confidence intervals is a somewhat conservative method; statistical differences may therefore not represent clinically meaningful differences

A fifth limitation concerns the risk of bias tool employed in this study. The tool employed was a brief measure with many of the items based on study reporting detail. This tool was selected for its perceived relevance to uncontrolled treatment studies. We would offer caution around any interpretations of risk of bias as there are many aspects of bias that this tool did not measure (e.g. fidelity, outcomes assessors).

A sixth limitation concerns the search strategy. It is highly unlikely that the search strategy used captured every available study. A 'complete' review of effectiveness research is not likely to be feasible, however we feel that the current reviews gives an adequate range and depth of effectiveness research with which to make tentative interpretations regarding the field of effectiveness research.

A final limitation is that the current review used broad outcome domains. This did not account for whether the outcome measured was of primary interest for change. This is difficult to achieve as many studies report multiple measures, and without a specified primary measure.

Implications for Research, Policy & Practice

To provide further understanding around the effectiveness of routinely delivered therapy future research should: (a) include fidelity and competency measures to confirm whether treatments delivered resembled treatment intended; (b) routinely assess outcomes at follow-up to establish maintenance of gains; (c) provide greater representation of therapy outcomes from non-western countries/services; and (d) explore variability in outcome among different clinicians.

In terms of policy and practice, the following implications are considered. First, the need for development of reporting standards for practice-based evidence. The marked variation in how studies report details around the sample and intervention make comparisons and replication difficult. For example ethnicity rates were only reported for 127 samples (42.62%). This prevents accurate calculation of ethnicity rates across services/studies. Simply calculating the average rate of representation across those studies which do report statistics is not a valid approach as it is does not account for why studies omit ethnicity rates. Potential reasons include clinician/researcher oversight in reporting, or alternatively a marked lack of ethnic representation/access in these services/studies. There was also a lack of endeavor from studies to contextualize demographic utilization rates in terms of how representative they are of the populations/communities that they are intended to serve. Future practice-based studies of therapy effectiveness should routinely report all relevant rates of patient demographics and also quantify how proportionate they are of communities served.

Second, this study found no evidence of differential outcome based on ethnicity, age, or marital status through meta-regression. This provides further support for the need to provide fair and equitable access of psychological therapy across the dimensions of age, ethnicity and marital status as there is no evidence that they impede effectiveness.

Third, routine recording of outcomes maintained at follow-up points should be enabled through necessary service commissioning of follow-up reviews/assessments. The body of evidence presented here concerns improvements made at the end of treatment. While follow-up was not included in this review, it was frequently apparent to reviewers that follow-up was rarely reported within studies. This information is necessary to determine the durability of improvements made during treatment.

Fourth, in light of differential outcomes demonstrated between qualified and unqualified clinicians (e.g. unqualified producing greater outcomes for anxiety) a review of training needs may be required for clinicians at different levels of experience.

Conclusion

This review provides substantial support for the effectiveness of psychological therapy as delivered in routine settings across a range of outcomes. Continent, method of analysis, and risk of bias score were significant moderators across all outcome domains. A key limitation of this review, and potentially the wider literature is the highly western-centric representation and reliance upon observational pre-post study designs. Nevertheless, for patients seeking help for psychological distress in routine services, there is growing evidence that interventions provided are clinically effective. The challenge for routine service delivery and associated effectiveness research is now to demonstrate the durability of this acute phase effect.

Appendix

Appendix A. Systematic search terms

Table 11

List of search terms and limiters for systematic database search

Effectiveness Study Term	Psychological Relevence Term	Limiters
’Practice based evidence’	Psycho* OR Therap [PsycInfo]	English Language
’Routine practice’	Psycho* [CINAHL and MEDLINE]	Adult Sample
Benchmarking		
Transportability		
Transferability		
Clinical* representat		
’External valid* N0 findings		
Applicab* N0 findings		
Applicab* N0 intervention*		
’Empiric* support*’ N0 treatment*		
’Empiric* support*’ N0 intervention*		
’Clinical* Effective*’		
Dissem* N0 treatment*		
Dissem* N0 intervention*		
’Clinical Practice’ N0 intervention*		
’Clinical Practice’ N0 treatment*		
’Service deliv*’ N0 intervention*		
’Service deliv*’ N0 treatment*		
’Clinical* effective*’ N2 evaluat*		
’Service deliv*’ N0 evaluat*		
Transporting		
’Managed care setting’		
Uncontrolled		
’Community clinic’		
’Community mental health centre’		
’Clinic setting’		
’Service setting’		

Appendix B. Preference system for outcome measures

Preference system for outcome measures

Because of the heterogeneity of outcome measures which could fit within the 'general' category the following hierarchy was used: (1) global measures of psychological distress (e.g. CORE-OM, SCL-90); (2) mono-symptomatic measures (e.g. Y-BOCS, EDE-Q). If a study used more than one measure at the same point in the hierarchy then we used the measure that had been most frequently employed in studies reviewed prior. Below is the final table of outcome measures used in the general category.

Table 12

Frequency of outcome measures with at least three occurances

Measure	n	Domain
BDI-II	44	**Depression**
BDI	34	
PHQ-9	30	
BSI (Depression)	8	
SCL (Depression)	8	
CESD-10	5	
HADS (Depression)	4	
DASS (Depression)	3	
BAI	19	**Anxiety**
GAD-7	19	
BSI (Anxiety)	8	
HADS (Anxiety)	7	
SCL (Anxiety)	6	
PSWQ	5	
DASS (Anxiety)	3	
CORE-OM	35	**General**
BSI-GSI	26	
SCL (Global)	22	
OQ-45	13	
PCL	12	
WSAS	7	
Y-BOCS	7	
EDEQ	6	
BHM	5	
GHQ	4	
CORE-10	3	
SF-36	3	

Note.

Abbreviations: Beck's Depression Inventory (BDI); Patient Health Questionnaire-9 (PHQ-9); Brief Symptom Invetory (BSI); Symptom Checklist 90 Revised (SCL90R); Centre for Epidemiological Studies Depression Scale (CESD10); Depression Anxiety and Stress Scale (DASS); Hospial Anxiety & Depression Scale (HADS) Short Form-36 (SF36); Beck's Anxiety Inventory BAI); Generalised Anxeity Disorder-7 (GAD7); Penn-State Worry Questionnaire (PSWQ); CORE Outcome Measurement (CORE-OM); Outcome Questionnaire-45 (OQ45); PTSD Checklist (PCL); Work and Social Adjustment Scale (WSAS); Yale-Brown Obsessive Compulsive Scale (Y-BOCS); Eating Disorder Examination Questionnaire (EDEQ); Behavioural Health Measure (BHM); General Health Questionnaire (GHQ); Short Form-36 (SF36).

Appendix C. Systematic review screening tool

Table 13

List of search terms and limiters for systematic database search

Criteria	Theme	Notes
Is there a psychological intervention	Psychological interventionists	No exclusions should be made based on the interventionist.
Is there a psychological intervention	Multi-component or multi-disciplinary	Multi-component or multi-disciplinary interventions which include individual psychology components should be included.
Is there a psychological intervention	Non-psychological interventions	Interventions which use only medical, alternative (e.g. yoga, acupuncture), physical (exercise, physio) or occupational therapy should be excluded.
Primary piece of quantitative research	Review papers	Meta-analyses or systematic reviews should be excluded.
Primary piece of quantitative research	Secondary Analysis	Secondary analysis papers will be removed in favor of the initial/primary publication.
Primary piece of quantitative research	Overlapping Study Samples	If studies have overlapping participants (but otherwise eligible) then the study with the larger sample size will be preferred unless there is a strong reason otherwise
Is the sample exclusively adults	Age	If evidence of participants (any proportion) under the age of 16 then the study/sample should be excluded.
Is the sample exclusively adults	Adults	Mention of "child" or "adolescent" participants is grounds for sample exclusion.
Individual psychotherapy	Mode of delivery	If any proportion of the sample have only received group intervention (and therefore no individual psychotherapy) then exclude the study.
Individual psychotherapy	Couples and family therapy	If any proportion of the sample have only received couples/family therapy then the sample should be excluded
Individual psychotherapy	Conjoint therapy	If study/sample involves participants receiving multiple treatments then this is acceptable as long as one of the interventions is individual psychotherapy.
Individual psychotherapy	By proxy (carer/family)	If any proportion of the sample have only received intervention by proxy (i.e. through family or carers only) then the study sample should be excluded.
Effectiveness outcomes	Effectiveness	A measurement of treatment effectiveness if required. Studies which only use non-effectiveness based measures (e.g. process, satisfaction, cost-effectiveness, satisfaction, QoL) should be excluded.
Effectiveness outcomes	Pre-Post	Pre-post comparisons are required. If samples only report the results of interventions that are not finished (e.g. 5 sessions of treatment) then this is acceptable.
Effectiveness outcomes	Follow-up	If the post treatment observation is indicated by follow-up then this should not be more than 6 months
Effectiveness outcomes	Validated measures	Study effectiveness measures should use a psychometrically validated questionnaire (or adapted from)
Effectiveness outcomes	Questionnaire studies	Studies which seek to validate instruments in routine services are acceptable as long as pre-post data is available
Effectiveness outcomes	Self-report	Only self-report measures are accepted. If any proportion of the sample only receive a clinician rates measure then the sample should be excluded.
Face-to-face	Format	If any proportion of the sample have only received telephone or internet based treatment then the sample should be excluded
Sample Size	Sample Size	Study samples should be at least 6 however analysed samples may be less than this.

Not naturalistic	Control group	If the study uses any form of control group (e.g. TAU, WLC) then the study should be excluded. The only exception is if the control group is a historic naturalistic dataset.
Not naturalistic	Randomisation	If any use of random assignment is used then the study should be excluded.
Not naturalistic	Recruited from routine settings	If a study "recruits from" a routine setting, but the intervention is reported to take place in a non-routine treatment setting then the study should be excluded.
Not naturalistic	Use of internal controls	No exclusion should be made based on financial incentive for participation; research/assessor involvement; competency/adherance measures; protocol based treatments; or qualification of interventionist
Design	Design	No exclusions based on design other than that which related to prior criteria (i.e. exclusion of control groups or studies without pre-post data).

Appendix D. Further information for extraction and coding process

Sample Characteristics

There was high variability of demographic reporting for each study (e.g. gender, age, ethnicity etc.). For demographic information, the (i) mean age of each sample was extracted, and then (when reported) the number and percentage of: (ii) female, (iii), minority ethnic group, (iv) full-time employed, (v) and married patients. Each of these variables were summarised by averaging across mean averages for studies which reported this information.

Methodological Information

For methodological information the type of completion analysis used was extracted. Samples were coded as either true ITT (everyone had an equal chance of inclusion), modified ITT, or completers. The stage of the hour-glass model was also recorded for each effectiveness study. Samples were rated as either stage-1 (pilot and preliminary effectiveness studies) or stage-3 (evaluation/benchmarking studies studies). The region (country and continent) was recorded; studies from the UK were separated from mainland Europe, due to the high volume of effectiveness research originating in the UK.

Service Information

The type of service and associated sector were extracted for each study. As there were a large number of different sectors represented, a grouping system clustered similar sectors together. Services from primary care, health settings, counseling, and voluntary services were collated into a ‘primary’ sector category. Services delivering interventions for more specialist, complex or enduring presentations were grouped into a ‘secondary’ category. This included specialist/tertiary therapy services/clinics, community mental health teams/centers, and intensive out-patient services. University based services (either training clinics or counseling centers) were assigned to a ‘University clinics’ category. Finally, inpatient, day hospital and partial hospital services were grouped into a ‘inpatient’ category. Whether or not study interventionists consisted of clinicians in training was also recorded as a separate variable. We defined clinicians in training as staff training towards a professional psychology training course (i.e. clinical psychology interns/students, training psychiatrists or assistant psychologists). Staff who were not psychologists or qualified therapists, but who had a core profession (e.g. nurses, social workers) were not recognised as unqualified interventionists.

Treatment Information

The treatment delivered was recorded for each study. Treatments were then assigned to a broad meta-therapy category, including: (i) cognitive and/or behavioural, (ii) dynamic/interpersonal, (iii) person-centered counseling (or counseling without a specified orientation), or (iv) other/non-specified. The average number of sessions was also recorded. For studies that reported the mean number of sessions then this was the metric extracted. For studies that alternatively used a time metric (days/weeks/months/years) then a uniform metric was applied (i.e. conversion to days). There was subsequently two possible dosage metrics, sessions of treatment and treatment days. If studies reported sample dosage, but with an alternative measure of central tendency (i.e. median) then this was converted to mean average.

Appendix E. Quality appraisal tool

Table 14

Adaprted version of ’The Joanna Briggs Institute Critical Appraisal tools: Checklist for Case Series’.

Criteria	Definition
1	Clear criteria for inclusion (if any form of inclusion criteria is provided and clearly stated).
2	Consecutive inclusion (need to be an explicit statement of being ’consecutive’ or ’all patients between the dates of’.
3	Complete inclusion (fulfill previous criteria + true intention-to-treat analysis performed).
4	Clear reporting of demographics (2 out of the following: gender, age, ethnic, marital status, employment).
5	Clear reporting of treatment information (2 out of the following: details regarding the interventionist, type of treatment, number of sessions).
6	Post-outcome clearly reported (means and standard deviations are available).
7	Site/Clinic reporting details (i.e. brief statement about the nature of the host service/s).
8	Appropriate statistical analyses (measure of effect-size reported [e.g. cohen’s d, hedge’s g, reliable change])).

Appendix F. Supplementary material link

The supplementary material for the current review is available at: `https://osf.io/p9sx5/?view_only=b906293276f54850824a1bcb86d47440`

Included in the supplementary material is a complete bibliography, including all studies which featured in the qualitative synthesis and meta-analysis.

Appendix G. Systematic search exclusion reasons

Table 15

Frequency of exclusion reasons from the systematic search.

ExclusionReason	SecondaryReason	n
No effectiveness data	Aggregated outcomes measure areas	1
	Clinician rated measured	1
	No effectiveness measure	21
	No pre-post	22
	No self-report	20
	No validated measure	3
No individual psychotherapy	By proxy	2
	Family/couples	17
	Group therapy	67
No primary data	Book chapter	1
	No psychology intervention	1
	Overlap with other study	19
	Review/meta	8
	Secondary analysis	12
No psychotherapy	No psychology intervention	11
Not adult population	Not adult population	1
	Not adult Population	4
Not face-to-face	Format	2
Not naturalistic	Control group	21
	Randomisation	33
	Recruited from routine settings	2
	Stage II	4
Sample size	Sample size	5
No English full-text	No English full-text	1
No Psychotherapy	No psychology intervention	1
Original Study Used	Overlap with other study	5

References

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

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., Hardy, G. E., & Mellor-Clark, J. (2010). *Developing and delivering practice-based evidence: A guide for the psychological therapies*. Wiley-Blackwell.

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). Wiley-Blackwell. https://doi.org/10.1002/9780470687994.ch2

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck Depression InventoryII - PsycNET. *APA PsycTests*. https://doi.org/10.1037/t00742-000

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 (Eds.). (2009).

Introduction to meta-analysis. John Wiley & Sons.

Buckley, J. V., Newman, D. W., Kellett, S., & Beail, N. (2006). A naturalistic comparison of the effectiveness of trainee and qualified clinical psychologists. *Psychology and Psychotherapy: Theory, Research and Practice*, *79*(1), 137–144. https://doi.org/10.1348/147608305X52595

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

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*(4), 364–368. https://doi.org/10.1017/S2045796016000809

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

Fergusson, D., Aaron, S. D., Guyatt, G., & Hébert, P. (2002). Post-randomisation exclusions: The intention to treat principle and excluding patients from analysis. *BMJ*, *325*(7365), 652–654. https://doi.org/10.1136/bmj.325.7365.652

Flückiger, C., Wampold, B. E., Delgadillo, J., Rubel, J., Vîslă, A., & Lutz, W. (2020). Is there an evidence-based number of sessions in outpatient psychotherapy? A comparison of naturalistic conditions across countries. *Psychotherapy and Psychosomatics*, *89*(5), 333–335. https://doi.org/10.1159/000507793

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

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].

Harrer, M., Cuijpers, P., Furukawa, Toshi. A., & Ebert, David. D. (2019b). Multiple Meta-Regression. In *Doing meta-analysis in R: A hands-on guide*.

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

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

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

Jacobson, N. S., & Christensen, A. (1996). Studying the effectiveness of psychotherapy. How well can clinical trials do the job? *The American Psychologist*, *51*(10), 1031–1039. https://doi.org/10.1037/0003-066X.51.10.1031

Kraemer, H. C., Frank, E., & Kupfer, D. J. (2006). Moderators of treatment outcomes: Clinical, research, and policy importance. *JAMA*, *296*(10), 1286–1289. https://doi.org/10.1001/jama.296.10.1286

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, *59*(11), 990–996. https://doi.org/10.1016/j.biopsych.2005.09.014

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 behavior change* (6th ed., pp. 169–218). John Wiley & Sons, Incorporated.

Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The Outcome Questionnaire-45. In M. E. Maurish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment: Instruments for adults* (Vol. 3, pp. 191–234). Lawrence Erlbaum Associates Publishers.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Li, X., Dusseldorp, E., Su, X., & Meulman, J. J. (2020). Multiple moderator meta-analysis using the R-package Meta-CART. *Behavior Research Methods*, *52*(6), 2657–2673. https://doi.org/10.3758/s13428-020-01360-0

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

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S. J., & Kircher, J. C. (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. *Journal of Consulting and Clinical Psychology*, *76*(1), 116–124. https://doi.org/10.1037/0022-006X.76.1.116

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*, b2535. https://doi.org/10.1136/bmj.b2535

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

NICE. (2011). *Common mental health problems: Identification and pathways to care*. National Institute for Clinical Excellence (NICE).

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

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan: A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Philips, B., & Falkenström, F. (in press). What research evidence Is valid for psychotherapy research? *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.625380

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), Article 215. 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/

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., & Fonagy, P. (1996). *What works for whom?: A critical review of psychotherapy research*. Guilford.

Salkovskis, P. M. (1995). Demonstrating specific effects in cognitive and vehavioural therapy. In *Research foundations for psychotherapy practice* (pp. 191–228). Wiley.

Sauer-Zavala, S., Ametaj, A. A., Wilner, J. G., Bentley, K. H., Marquez, S., Patrick, K. A., Starks, B., Shtasel, D., & Marques, L. (2019). Evaluating transdiagnostic, evidence-based mental health care in a safety-net setting serving homeless individuals. *Psychotherapy*, *56*(1), 100–114. https://doi.org/10.1037/pst0000187

Schulz, K. F., Altman, D. G., Moher, D., & the CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*(1), 18. https://doi.org/10.1186/1741-7015-8-18

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

Seedat, S., Scott, K. M., Angermeyer, M. C., Berglund, P., Bromet, E. J., Brugha, T. S., Demyttenaere, K., de Girolamo, G., Haro, J. M., Jin, R., Karam, E. G., Kovess-Masfety, V., Levinson, D., Medina Mora, M. E., Ono, Y., Ormel, J., Pennell, B.-E., Posada-Villa, J., Sampson, N. A., … Kessler, R. C. (2009). Cross-national associations between gender and mental disorders in the world health organization world mental health surveys. *Archives of General Psychiatry*, *66*(7), 785–795. https://doi.org/10.1001/archgenpsychiatry.2009.36

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

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. https://doi.org/10.1037/0003-066X.32.9.752

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

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

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

Waller, G., & Turner, H. (2016). Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy*, *77*, 129–137. https://doi.org/10.1016/j.brat.2015.12.005

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially, "all must have prizes". *Psychological Bulletin*, *122*(3), 203–215. https://doi.org/10.1037/0033-2909.122.3.203

EMPIRICAL PROJECT

Effectiveness of Tertiary Care Outpatient Psychological Interventions; A Benchmarking Study

Chris Gaskell

Supervisors:

Dr. Stephen C. Kellett

Dr. Mel Simmonds-Buckley

Dr. Jaime Delgadillo

Contents

Appendix **98**

References **123**

List of Tables

List of Figures

This page is intentionally left blank

Abstract

Background: When patients are not responsive to primary care interventions then they can be referred to further tiers of the stepped-care system (i.e. to secondary/tertiary care). However, evidence regarding the effectiveness of tertiary care psychological therapy is very scarce. **Objectives:** To explore the effectiveness of psychological interventions delivered in a tertiary care psychotherapy service using equivalent service benchmarks. **Methods:** A retrospective analysis of psychotherapy outcomes on the Outcome Questionnaire-45 (OQ-45) over a 10 year period (2011-2021) in a tertiary care psychotherapy service based in the United Kingdom. The service delivered three interventions; cognitive behavioural, cognitive analytic and psychoanalytic. Rates of effectiveness were calculated at the service level and also for different treatment modalities using pre-post treatment effect sizes (Cohen's d) and clinical recovery indices. Trajectories of change were examined using growth curve modeling. **Results:** Baseline distress on the OQ-45 was higher than comparative norms (M = 102.57, SD = 22.79, N = 364). The average number of sessions was 48.68 (SD = 42.14, range = 5-335). There was a small pre-post effect small effect small effect (*d* = 0.46, 95% CI = 0.37-0.55) that was lower than available OQ-45 and tertiary benchmarks. Mean change between different treatments was comparable (overlap between confidence intervals). The recovery rate was 10.16% and 29.95% made a reliable improvement. Change in OQ-45 score over time was best explained using a nonlinear (cubic) time trend and mean change between the three treatments was comparable. **Conclusions:** Patients receiving tertiary care psychotherapy in the present sample had higher baselines distress and outcomes were suppressed accordingly. Suggestions are made regarding the role of tertiary care psychotherapy in mental health services.

Keywords: ‘Psychotherapy,’ ‘effectiveness,’ ‘tertiary-care,’ ‘growth curves.’

Practitioner Points:

- There are a sub-sample of tertiary care patients who do not respond to treatment. For these patients there is limited evidence for extending therapy beyond 100 sessions.

- Outcomes monitoring within supervision may provide a suitable means of determining when treatment should end.
- There was limited evidence to favor one psychological modality over another; however evidence was promising for CAT.

Effectiveness of Tertiary Care Outpatient Psychological Interventions; A Benchmarking Study

Psychological therapies are an essential part of public healthcare in the UK, available through the National Health Service (NHS, Department of Health, 2004). There is considerable support for the efficacy of psychological therapy from controlled trials (Roth & Fonagy, 1996). The effectiveness of such therapies in routine or naturalistic settings has also been repeatedly supported via meta-analyses (e.g. Cahill et al., 2010; Stewart & Chambless, 2009; Wakefield et al., 2021). While there is now consensus that psychological therapy is an effective treatment for a range of disorders, questions remain regarding which factors and conditions serve to enhance or hinder effectiveness. One factor considered to strongly influence therapy outcome is the context (or sector) in which treatment is delivered and the complexity of the presenting problems (Firth et al., 2020; Lambert et al., 2001; Paley et al., 2008; Smith & Glass, 1977).

Sectors

Within the NHS, health-care delivery is organised through a range of tiers, otherwise known as sectors. Primary care, which accounts for the largest proportion of UK therapy delivery, offers interventions which are brief or ‘time-limited’ for individuals who have mild-to-moderate (low intensity psychological intervention indicated) and moderate-to-severe (high intensity psychological intervention indicated) levels of common mental health difficulties. Patients with greater complexity or who have not responded to treatment in primary care are referred to secondary or tertiary care services. These populations with chronic common mental problems have been described as being ‘treatment resistant’ and requiring specialist psychological interventions (Taylor et al., 2012). Despite the differences between care sectors regarding therapy delivery, the body of supporting evidence for NHS psychological therapy is overwhelmingly based upon primary and secondary care services. There is a relative lack of evidence for the effectiveness of tertiary care services and how they compare to other sectors. Tertiary care therapy services have been characterized as being in high demand, but lacking both resources and supporting practice-based evidence (Warden et al., 2008). The three primary reasons for lack of evidence are (a) services are few in number, (b) outcome studies failing to label when therapies are delivered in these services and (c) when tertiary care outcomes are reported in such studies there is often high levels of missing data (up to 95%, Firth et al., 2020). The final reason also motivates researchers to omit tertiary care samples from multi-sector analyses due to concerns around selection bias (see Stiles et al., 2006 for an example).

Tertiary Care Services

Tertiary care therapy services are highly underrepresented in the UK, are few in number, cover wide geographical regions and offer more resource intensive interventions than those provided within primary care. For example, cognitive-behavioural psychotherapies, although provided within primary care, are often available for longer durations within tertiary care services. Examples of other treatments provided in tertiary care services include: dynamic interpersonal therapy (DIT, Douglas et al., 2016), intensive short-term dynamic interpersonal therapy (ISTDP, Johansson et al., 2014), cognitive analytic therapy (CAT, Ryle & Kerr, 2020), psychodynamic-interpersonal therapy (PIT, Paley et al., 2008) and psychoanalytic psychotherapy (Warden et al., 2008).

Those few studies which have provided evidence for tertiary outcomes have tended to analyse outcomes for a small number of patients. Paley et al. (2008) explored the effectiveness of psychodynamic interpersonal therapy with 47 tertiary care patients and Douglas et al. (2016) explored the effectiveness of dynamic interpersonal therapy with 28 patients. There are no other published instances of practice-based evidence for UK tertiary care therapy and this stands in comparison to UK primary care therapy datasets samples of greater than 100,000 patients (e.g., Delgadillo et al., 2016, N = 110,415).

Outside of the UK, there has been greater representation of tertiary care practice-based evidence in larger samples. This originates in Canadian outcome studies exploring the effectiveness of intensive short-term dynamic psychotherapy (ISTDP) in a single tertiary care psychotherapy service (Abbass et al., 2008; Johansson et al., 2014; Lilliengren et al., 2020; Nowoweiski et al., 2020). Each of these studies has provided support for the effectiveness of psychological therapy delivered within tertiary care services. For example, Johansson et al. (2014) observed a large reduction in general psychological distress (*d* = 0.87, N = 412). Caution is required in generalising these findings to UK tertiary care services as there are various differences between UK and Canadian services (e.g. funding structure, clinician training) and further potential differences between patient groups (e.g. presenting problem, treatment received). While there are also a small number of tertiary care therapy samples from other countries, they have generally focused on specialist populations, such as chronic fatigue (Heins et al., 2011; Worm-Smeitink et al., 2016) and autism (Blainey et al., 2017).

The Type and Duration of Interventions

One of the main differences between NHS care sectors is the treatment provided. There has been an overwhelming amount of attention within the psychotherapy literature exploring differential rates of effectiveness between therapy modalities. Reviews of efficacy trials have consistently shown that bona fide psychotherapeutic interventions tend to have highly similar treatment outcomes (Wampold et al., 1997); this common finding has since been referred to as the equivalence paradox (Rosenzweig, 1936). It has subsequently been argued that excessive attention to comparing different 'brands of therapy' (Johns et al., 2019) has scarcely moved the area forward. While this finding has shown to be robust when averaging across multiple populations/studies there is evidence that in specific circumstances certain treatments perform better than others (e.g., Delgadillo & Gonzalez Salas Duhne, 2020). As tertiary care services in the UK are considered to represent a distinct group of patients with more severe, enduring disorders, it could be that these more complex patients have significant differential responses to particular treatments. Knowledge of differential response to treatment could inform treatment selection decisions; this is particularly relevant for patients accessing tertiary care treatments, as they have typically accessed and not responded to previous psychological treatments (Taylor et al., 2012). No study has yet explored potential differences between bona fide psychotherapeutic interventions within a UK tertiary care service.

Studying how treatment duration influences effectiveness can provide valuable information regarding ‘dose-response’ relationships (Howard et al., 1986). Such information is very pertinent to tertiary care provision due to the lengthier treatments offered. A review of the dose-response literature for psychological therapy in routine settings found strong support for a curvi-linear relationship between sessions and effectiveness (Robinson et al., 2020). In other words, effectiveness shows a negatively accelerating rate of change. The review estimated that an optimal dosage for routine settings is 4-26 sessions, but that this will vary based on setting, population, and outcome measure. The review also found there to be scarce, inconclusive evidence for chronic/severe patient samples and little evidence to support long-term therapy beyond 30 sessions. This is pertinent to tertiary care services as longer treatment durations are provided on the notion that more complex presentations require longer treatment durations. Some support for this claim has been provided through early studies demonstrating that patients with chronic and characterological symptoms may require longer treatment durations in order to reach a comparable response rate (Howard et al., 1986) and that interpersonal problem resolution may lag behind symptom improvement (Kopta et al., 1994). There is no empirical evidence available for optimal treatment dosage within UK tertiary services, and only limited evidence for services which provide long form treatments (Robinson et al., 2020).

To summarise, there is a distinct lack of evidence for the effectiveness of UK tertiary care psychotherapy services using a validated outcome measures and with an adequately sized sample of tertiary care patients. This study aimed to primarily provide a quantifiable benchmark of tertiary care effectiveness and establish rates of recovery and deterioration. The secondary aim was to explore how change occurs over time (i.e. growth trajectories) and compare outcomes between three routinely delivered therapies. It was hypothesized that (1) service effectiveness outcomes would be comparable to benchmarks, (2) there would be no significant difference in pre-post outcomes between treatments and (3) outcome trajectories would be comparable across the treatments.

Method

Design

A retrospective analysis of naturalistic therapy outcomes data collected from a tertiary care/specialist psychotherapy service.

Service

Historical data from a tertiary level Specialist Psychotherapy Service (SPS) in a UK city in the North of England spanning a 10-year period (2011-2021) was accessed. Patients are referred to SPS from a variety of referrers, including primary care therapy services, secondary care therapy services/community mental health teams, or general practitioners. Patients referred have enduring and complex mental health difficulties and have often not responded to previous psychological treatments within primary and/or secondary care therapy services. The staff team at SPS is multi-disciplinary, with a range of core professions (clinical psychologists, psychotherapists and psychiatrists) and training professionals represented. Many of the treating clinicians have received further training and accreditation in specific models of psychotherapy. Supervision is provided internally and externally, with each clinician receiving clinical supervision matched on therapeutic modality. The treatments offered include: cognitive-behaviour therapy (CBT), cognitive-analytic therapy (CAT), psycho-analytic therapy (PAT) and eye movement desensitization and reprocessing (EMDR). Clinicians offer a single modality. Patients are allocated to treatments based on team/clinician judgment of model suitability. The length of treatment varies between therapeutic modalities, with fixed contract lengths agreed at the start of treatment. Generally, CBT and EMDR offer treatment for six-months or more, CAT offers either 8, 16 or 24 sessions and PAT offers contacts of 1-2 years. PAT treatments are offered as either individual or group-based. For some patients, multiple treatments are provided within a single care episode, and within a specific modality. For example, patients may receive consecutive treatments of CBT, or alternatively patients who receive one-to-one PAT may progress to group treatment. Patients are offered follow-up appointments to review progress. The uptake of follow-up appointments is optional and determined collaboratively by clinician and patient. Follow-up analyses were not explored within the current study. The self-report outcome measure used in this study was the Outcome Questionnaire-45 (OQ-45, Appendix A)

Patients

From the outset of the evaluation (2011) four teams within SPS enrolled in the outcomes evaluation and have subsequently contributed to the current study. These teams included (i) the Personality Disorder Team, (ii) the Anxiety and/or and Post-Traumatic Stress Disorder (PTSD) Team, (iii) the Focused Depression Team, and finally (iv) the Obsessive-Compulsive Disorder (OCD) and/or Body Dysmorphic-Disorder (BDD) Team. Since implementation, the latter three teams have been merged into a single team, known as the Mood, Anxiety and Post-Traumatic Stress Related Disorders Team (MAPPS). Regardless of service accessed, patients are required to be: (i) aged 16 years or older, (ii) not currently experiencing a mental health crisis, (iii) and have previously accessed psychological intervention from primary/secondary care services.

Sample Size

Data collection spanned 10 years (February 2011 - January 2021). There were 4,203 OQ-45 administrations collected from 1027 patients, treated by 53 therapists. When limiting data to the first care episode (referral to discharge) there were 3,198 OQ-45s remaining. There were 2639 OQ-45s across 364 patients with at least two OQ-45s (i.e. for recovery and effectiveness analyses) and at least one treatment session. There were 298 patients with at least three OQ-45 measures and which could be used in the growth-curve analysis. Differences between treatment groups were explored for a range of demographic variables.

Procedure

Outcome Monitoring and Data Extraction

Patients complete an outcome measure at various stages of therapy. This includes: (i) assessment; (ii) therapy sessions; (iii) final therapy appointment; and (iv) follow-up appointments. In order for the author to analyse the SPS outcomes data an application was made to an NHS research ethics committee. As the outcomes database originally contained very limited information regarding co-variates (i.e. no demographic, treatment, or service usage information) a request (section 251) was made to access additional information held within patient electronic health records. This application was made to the NHS Confidentiality Advisory Group (CAG) through the Health Research Authority (HRA). More specifically, permission to access, view and analyse information regarding demographic (gender, age, employment, ethnicity), treatment (dosage, treatment received, clinician, completion status), and service usage (others services sought prior and since) variables. Following a full set of approvals (Appendix B) the author accessed relevant IT systems training. Extraction utilised an automated extraction process, supported by a data analyst and clinical governance officer within SPS. Instances when information could not be extracted using this approach were rare. When this did happen, manual data extraction was conducted by the author. In line with guidance set out by the Confidentiality Advisory Group (CAG) an opt-out poster was placed in the SPS waiting room and the SPS website during April 2021 for four weeks (Appendix C). At the point of extraction there were no patient opt-out requests. All extracted information was saved to a carbon copy of the outcomes database and stored in line with data storage requirements using a secure VPN, provided by the host health-care trust (Sheffield Health & Social Care NHS Foundation Trust). Once extraction was complete it was stripped of unique person identifiers. The spreadsheet was then anonymity checked by an SPS clinician (SK) before being transferred from NHS secure servers to the designated data-handler (M-SB) within the University of Sheffield (sent via password encrypted e-mail). The database remained password encrypted throughout with access only available to the author and the data-handler.

Outcome Measurement

The OQ-45 is a 45 item self-report measure of global psychological distress (Lambert et al., 2004). Each item provides a 5-point Likert scale ranging from 0 (never) to 4 (almost always) with the cumulative score across items (maximum = 180) providing a global distress score. US normative data for community and clinical populations (Lambert et al., 1996) has provided a clinical cut-off point of 64. Patients who score above 63 are considered to be within the clinical range, while scores above 85 are considered to be very high (Lambert, 2004). The reliable change score is 14 (Lambert, 2004). The psychometric properties of the OQ-45 are well documented (Lambert et al., 1996). For convergent validity, the measure has shown to have moderate to high correlations with a range of widely used measures of therapy effectiveness (Lambert et al., 1996). The OQ-45 has good reliability/internal consistency (r = .93, Lambert et al., 1996) and is sensitive to change (Lambert et al., 1996; Vermeersch et al., 2000). The OQ-45 has been validated in numerous countries however there is, to the author's knowledge, no clinical or community normative data for the UK. Embedded within the OQ-45 are three additional sub-scales. *Symptom distress* (range = 100, cut-off = 36, reliable change = 10) is designed to map onto symptoms of common mental health disorders (i.e., anxiety and affective disorder symptoms, Lambert, 2004). *Interpersonal relationships* (range = 1-36, cut-off = 15, reliable change = 8) explores complaints of conflict, loneliness and family difficulty. Finally, *Social role* (range = 44, cut-off = 12, reliable change = 7) is the extent to which the individual patient experiences difficulties relating to occupational and functional independence.

Analysis

For patients who had multiple recorded care episodes, we used the first care episode recorded within the OQ-45 database. Care episode is defined as the time between the points of referral and discharge. For some patients, a single care episode included multiple treatments within a single modality. As concurrent treatments (based on the information available) could not be accurately disentangled, we analysed change across care episodes. Change was considered by comparing the first and final OQ-45 treatment session (exclusive of follow-up) within a single care episode. Within this approach the first recorded OQ-45 within the care episode was used as the baseline score while the final recorded score was used as the end score. This represented a last observation carried forward approach so that patients who dropped out prematurely, or who did not complete an OQ-45 in the final appointment were not excluded from the analysis. This provided a more conservative estimate of effect size. As patients complete the OQ-45 prior to treatment sessions, we only measured change in patients who had received at least two treatment sessions, to ensure that at least some treatment had been delivered. This sample was used for analysis of effectiveness and recovery rates. Longitudinal multi-level (linear mixed) modeling was employed to examine trajectories of change. Patients with a minimum of three OQ-45 assessments, and of which at least one represented a treatment appointment were included. This was to allow for the possibility of higher-order polynomial change trajectories (e.g. quadratic, cubic).

Effectiveness

Pooled full and sub-scale effect-sizes were calculated using Cohen's *d* (standardised mean change, Cohen, 1988) using the formula advocated in benchmarking studies (Minami et al., 2008). This approach subtracted the pooled end-of-therapy score from the pooled pre-treatment score, before dividing it by the pre-treatment standard deviation. Regression to the mean was accounted for by adjusting confidence intervals by the pre-post correlation (Pearson's *r*) between pre and post-treatment measures. Effect-sizes were interpreted as 'small' (0.2–0.5), 'moderate' (0.5–0.8) or 'large' (> .08). Effect-sizes were calculated for each treatment modality (cognitive-analytic, cognitive-behavioral [including EMDR] and psycho analytic).

Benchmarking

A benchmarking approach was used to compare the effectiveness of tertiary care psychotherapy to other services and sectors (Minami et al., 2008). Benchmarking compares clinical outcome data to established reference points from efficacy trials or practice-based outcome studies (Delgadillo et al., 2014; Minami et al., 2008). This allows services to compare their performance against services which are similar in design, or against aggregated study benchmarks (Department of Health, 2004).

To benchmark service intake scores, a range of US OQ-45 intake severity comparators were used to create a rounded comparison. These included: (i) an employee assistance program (Lambert et al., 1996), (ii) a University out-patient clinic (Lambert et al., 1996), (iii) a community health centre (Lambert et al., 1996), and (iv) acute short-stay inpatients (Doerfler et al., 2002).

For pre-post change, a variety of effectiveness benchmarks were used. Findings from a large Canadian tertiary care outpatient outcome study (Johansson et al., 2014) were the tertiary-care benchmark. In order to better contextualize outcomes, an additional OQ-45 benchmark was developed that included N=13 studies that had employed the OQ-45. Cohen's *d* effect-sizes were then entered into a random effects meta-analysis to provide an aggregated OQ-45 effectiveness benchmark. Further details regarding the meta-analytic benchmark is provided in 'additional analysis detail' (Appendix D).

Recovery

This study calculated categorical rates of improvement and deterioration using reliable and clinically significant change indices (Jacobson & Truax, 1992) for the whole service and the three treatments. A *reliable change* (improvement or deterioration) was indicated when a patient met a minimum pre-determined change score based on a magnitude of change that exceeded measurement error. A *clinical change* change was indicated when a patient moved from above the threshold for clinical distress to below. Patients who meet criteria for both reliable improvement and clinical change were labeled as 'recovered.' Post-treatment status was classified as either: (i) recovery, (ii) reliable improvement, (iii) no reliable change, or (v) reliable deterioration. Patients who scored below the clinical threshold at baseline assessment (i.e., no initial clinical status) were unable to achieve full recovery. Statistical equations are provided in Equations 1 and 2. Recovery rates for each of the three treatments were then compared to a set of established OQ-45 recovery benchmarks (Hansen & Lambert, 2002). These benchmarks represent recovery rates for pre-post psychological therapy for a range of different US care sectors (employee assistential program, community mental health centre, health maintenance organisation). We selected the pooled service benchmark and the community mental health centre benchmark for perceived relevance to the current study.

Dose Response

In order to determine how number of sessions impacts upon statistical and clinical change, dose-response comparisons were made. Unfortunately, as OQ-45s were not always administered at the start of input, this means that the total number of sessions often differed from the number of sessions measured. For dose-response calculations, number of sessions measured (i.e. last OQ-45 session number - first OQ-45 session number) was used. Patients were ordered by number of sessions measured and then split into 10 groups (i.e. deciles). Dosage groups were then compared on the basis of statistical change (Cohen's *d*) and clinical change (recovery rates). Patients who do not respond to treatment (i.e. no change *or* deterioration) were plotted (bar plot) to visualise differences between groups.

Growth Curve Modelling

Individual growth trajectories (also termed growth curves) were developed for the OQ-45 total score and each of the sub-scales. Growth curve modeling allows for exploration of change trajectories, variance in change, and the factors that influence change. Each of these can be considered at a within-person and/or between person level. The approach is robust to missing and unbalanced data inherent in practice-based datasets and which was particularly the case for the current study. Available guidance for conducting growth curve modeling was followed throughout (Singer & Willett, 2003). The hierarchical data structure for the current study was repeated face-to-face sessions (level-1) nested within individual patients (level-2). Further details regarding maximum likelihood and power calculation are provided in Appendix D. The statistical equations for calculating individual growth curves are shown in Equation 3.

Time. As the OQ-45 administration number did not necessarily correspond to session number, this was not a viable temporal predictor. There were two alternative temporal variants which were considered for the current study. The number of *days* since the first recorded OQ-45 and the number of face-to-face *sessions* (i.e., contacts) since the first recorded session (in that care episode). Sessions was preferred for the reason that this could more precisely place the first recorded OQ-45 within the context of service input, particularly in situations when the first recorded OQ-45 did not represent the first treatment session. Sessions is also in fitting with the wider literature employing sessions as a measure of treatment dosage. Session number was centered on the first session (i.e. first session = 0). Additional sessions variables were created to assess polynomial and log-linear trends. This included sessions^2 (quadratic), sessions^3 (cubic), and the natural logarithm of sessions (log-linear). A log(sessions+1) transformation was used to adjust for time scores of 0.

Model Building

There were four stages of model building: i) co-variance structure selection; ii) time-trend selection, iii) unconditional models, and (iv) conditional models. Unconditional models are a set of models which do not include predictor variables (other than sessions). These include means models (i.e. intercept only), random intercepts models, and random slopes models.

Covariance structure. In order to select a best-fitting covariance structure, random slopes unconditional models were estimated fitting a series of alternative covariance structures (standard, unstructured, compound symmetry, auto-regressive1, topeplitz). The unstructured covariance structure failed to converge and was subsequently discarded. The remaining structures were nested in order of complexity by descending the degrees of freedom for each model (standard, topeplitz, compound symmetry, auto-regressive1). Each model was compared to the previous to determine if there was an improved goodness-of-fit. Goodness-of-fit statistics included the Akaike's information criterion (AIC), bayesian information criterion (BIC) and -2 log-likelihood statistic, with lower scores indicating improved model fit. log-likelihood ratio tests (i.e. Chi-squared test) compare adjacent pairwise models, however these were not applicable for model combinations that possess identical degrees of freedom (i.e. standard vs toeplitz or compound symmetry vs auto-regression 1). For this reason, goodness-of-fit statistics alone were used to select covariance structure. Topeplitz provided the best fit for total OQ-45 score, symptom distress and interpersonal relationships (table shown in Appendix E). A standard covariance structure provided the best fit for social role.

Time form. Random slopes models were developed using a series of linear, log-linear and higher order polynomial models (quadratic and cubic), while retaining optimal covariance structures. As linear and log-linear models are not nested, they were compared using goodness-of-fit statistics only. The best-fitting model was then compared against quadratic and cubic trends using likelihood ratio tests. A cubic form best fit OQ-45 total score, inter-personal and symptom distress while a log-linear form best fit social role. Time form fixed effects and goodness-of-fit statistics are shown in Appendix F.

Unconditional models. For unconditional models, the fixed effects of intercept (initial status) and slope (sessions) were assessed for significance in a series of models (i.e., means model, random intercepts, random slopes). Allowing intercepts and slopes to vary provided significant improved fit for each outcome variable (as shown in Appendix G). Final unconditional and conditional models subsequently utilised random intercepts and slopes, optimal covariance structures and optimal time trends. As topeplitz covariance structures are fit using generalised least squares there were no random effects to report. Growth curves for unconditional models were visualised using scatter plots.

Conditional models. As each of the four dependent variables found a significant main effect for time (i.e. sessions), predictor variables were considered for inclusion in model iterations. Therapy modality was the only predictor variable included in conditional models. As the CAT treatment group was much smaller than the CBT and PAT groups it was decided to merge CAT and PAT to form a single analytic treatment group. The associated hypothesis for the current study was that there would be no significant effect for therapy modality and no significant time:modality interaction. There were no significant differences between the cognitive and analytic groups for average baseline distress (p = 0.177) or number of sessions (p = 0.115); therefore no adjustments were necessary.

Statistical Software

All analyses were performed using R (R Core Team, 2020, v 4.0.2). Multi-level modeling was conducted using the nlme (Pinheiro et al., 2020) package while growth-curve plots were developed using ggplot2 (Wickham, 2016). Effect-sizes (Cohen's *d*) and random-effects meta-analysis were computed using the metafor package (Viechtbauer, 2010), while forest plots were made using (Gordon & Lumley, 2020).

Equations

$$
\begin{aligned}
RCI &= \frac{(pre) - (posttreatment)}{S_{diff}} \\
S_{diff} &= \sqrt{2S2/E} \\
S_E &= SD\sqrt{1 - r_{XX}}
\end{aligned}
\tag{1}
$$

Equation 1 is the formula for calculation of reliable change.

$$
CSC = \frac{(SD_1)(M_2) + (SD_2)(M_1)}{SD_1 + SD_2} \tag{2}
$$

Equation 2 is the formula for calculation of the clinical cut-off.

$$
\begin{aligned}
&\text{Level 1} \\
&\quad Yij = \beta_{0j} + \beta_{1j} Observation_{ij} + R_{ij} \\
&\text{Level 2} \\
&\quad \pi_{0i} = \gamma_{00} + \gamma_{01} Therapy_j + \zeta_{0i} \\
&\quad \pi_{1i} = \gamma_{10} + \gamma_{11} Therapy_j + \zeta_{1i} \\
&\text{ with} \\
&\quad \left(\frac{U_{oj}}{U_{1j}}\right) \sim \mathcal{N}\left(\frac{0, \tau_{00}^2 \tau_{01}}{0, \tau_{01}, \tau_{10}^2}\right) \\
&\text{ and} \\
&\quad U_{oj} \sim \mathcal{N}\left(0, \sigma^2\right)
\end{aligned}
\tag{3}
$$

Equation 3 is the formula for growth curves.

Results

Sample

For the 364 patients included in the effectiveness and recovery analysis patients had an average of 5.64 OQ-45 administrations (SD = 4.41, range = 2-29) over a care episode duration of 138.7 weeks on average (SD = 64.69, range = 15.57- 424.86). This was likely to have been influenced by periods of time on treatment waiting lists, however this could not be accounted for as such information was not available for all patients. Of the included patients there were only 14 (3.85%) with care periods shorter than 12 months.

Mean number of sessions within the first recorded care episode was 48.68 (SD = 42.14, range = 5-335) and across all SPS care episodes a totaled mean of 62.47 sessions (SD = 61.54, range = 5-503). There was 354 (97.25%) patients that received at least 10 sessions, 308 (84.62%) received at least 20 sessions, 175 (48.08%) received at least 40 sessions, 22 (6.04%) received at least 100 sessions, 10 (2.75%) received at least 150 sessions, and 8 (2.2%) attended for over 200 sessions.

PAT treatments were the lengthiest (64.56, SD = 62.38, range = 5-335), followed by CBT (45.58, SD = 33.33, range = 5-292), and then CAT (28.83, SD = 10.87, range = 9-63). There were significant differences between modalities regarding duration of treatment (F[2, 361], = 10.64, p = <0.001). More specifically, PAT delivered significantly more sessions than CBT (p = 0.001) or CAT (p = <0.001). There was no significant difference in number of sessions between CBT and CAT (p = 0.089).

The mean average baseline score across patients included in the effectiveness/recovery analyses was 102.57 (SD = 22.79). There was no significant difference in baseline distress between treatments (see Table 1). CBT (M = 104.04, SD = 22.86) was marginally higher than CAT (M = 102.27, SD = 19.32) and PAT (M = 98.44, SD = 23.42). CBT was the most frequently delivered treatment (n = 248), followed by PAT (n = 86), and then CAT (n = 30). Completion status (i.e., completer or premature drop-out) was not routinely recorded and therefore could not be used in the current study. In terms of clinicians, there was 47 unique clinicians recorded.

Sample Characteristics

Sample characteristics are displayed in Table 1. Almost a third of referrals came from general practitioners (31.32%). The average age across participants was 42.16 (SD = 11.78) and there was a greater representation of female patients (60.71%). The vast majority of patients identified as white British (82.69%). A substantial proportion of patients did not have a recorded employment status (36.26%). There was more people in paid/unpaid employment (26.65%) than not employed (18.68%). In terms of marital status, 39.84% were married or in a settled relationship and 38.74% were single. Significant differences were shown between differing demographic groups for: (i) baseline symptom distress, (ii) number of treatment sessions, (iii) care episode duration, and (iv) employment status.

Table 1

Sample charachteristics of patients included in the current study broken down by treatment modality.

		CBT	CAT	PAT	Total	p-value
Patients	N	248	30	86	364	
Age	Mean	42.61	39.67	41.74	42.16	0.404
	SD	11.63	11.24	12.37	11.78	
	Range	18-74	21-61	17-73	17-74	
Baseline OQ-45 Severity	Total Mean	104.04	102.27	98.44	102.57	0.145
	Total SD	22.86	19.32	23.42	22.79	
	SD Mean	65.78	63.53	60.17	64.27	<0.05
	SR Mean	16.12	17.2	16.72	16.35	0.494
	IR Mean	22.58	22.73	21.97	22.45	0.757
Sessions in Care Period	Mean	45.58	28.83	64.56	48.68	<0.05
	SD	33.33	10.87	62.38	42.14	
Weeks in Care Period	Weeks	143.82	93.17	141.64	138.7	<0.05
	Range	16-382	41-162	28-425	16-425	
Gender	Female	145 (58.47%)	19 (63.33%)	57 (66.28%)	221 (60.71%)	0.422
	Male	103 (41.53%)	11 (36.67%)	29 (33.72%)	143 (39.29%)	

Table 1

Sample charachteristics of patients included in the current study broken down by treatment modality. (continued)

		CBT	CAT	PAT	Total	p-value
Ethnicity	White British	207 (83.47%)	26 (86.67%)	68 (79.07%)	301 (82.69%)	0.087
	Any other	11 (4.44%)	1 (3.33%)	3 (3.49%)	15 (4.12%)	
	Not Stated	12 (4.84%)	0 (0.00%)	3 (3.49%)	15 (4.12%)	
	Black	9 (3.63%)	2 (6.67%)	3 (3.49%)	14 (3.85%)	
	Asian	8 (3.23%)	0 (0.00%)	3 (3.49%)	11 (3.02%)	
	White Other	1 (0.40%)	1 (3.33%)	6 (6.98%)	8 (2.20%)	
Employment	Not Known/Other	92 (37.10%)	8 (26.67%)	32 (37.21%)	132 (36.26%)	<0.05
	Employed	55 (22.18%)	9 (30.00%)	33 (38.37%)	97 (26.65%)	
	Unemployed	53 (21.37%)	8 (26.67%)	7 (8.14%)	68 (18.68%)	
	Sick/Disabled	31 (12.50%)	3 (10.00%)	3 (3.49%)	37 (10.16%)	
	Student	14 (5.65%)	2 (6.67%)	7 (8.14%)	23 (6.32%)	
	Retired	3 (1.21%)	0 (0.00%)	4 (4.65%)	7 (1.92%)	
Marital Status	Married or Settled	99 (39.92%)	14 (46.67%)	32 (37.21%)	145 (39.84%)	0.661
	Single	96 (38.71%)	12 (40.00%)	33 (38.37%)	141 (38.74%)	
	Other	37 (14.92%)	1 (3.33%)	15 (17.44%)	53 (14.56%)	
	Divorsed/Seperated	16 (6.45%)	3 (10.00%)	6 (6.98%)	25 (6.87%)	
Referrer	Other	120 (48.39%)	13 (43.33%)	40 (46.51%)	173 (47.53%)	0.819
	GP	72 (29.03%)	12 (40.00%)	30 (34.88%)	114 (31.32%)	
	Psychiatry	32 (12.90%)	3 (10.00%)	11 (12.79%)	46 (12.64%)	
	IAPT	17 (6.85%)	2 (6.67%)	3 (3.49%)	22 (6.04%)	
	External Hospital	3 (1.21%)	0 (0.00%)	2 (2.33%)	5 (1.37%)	
	Dental	4 (1.61%)	0 (0.00%)	0 (0.00%)	4 (1.10%)	

Note. SD = Symptom Distress, SR = Social Role, IR = Interpersonal

Benchmarking; intake and outcomes

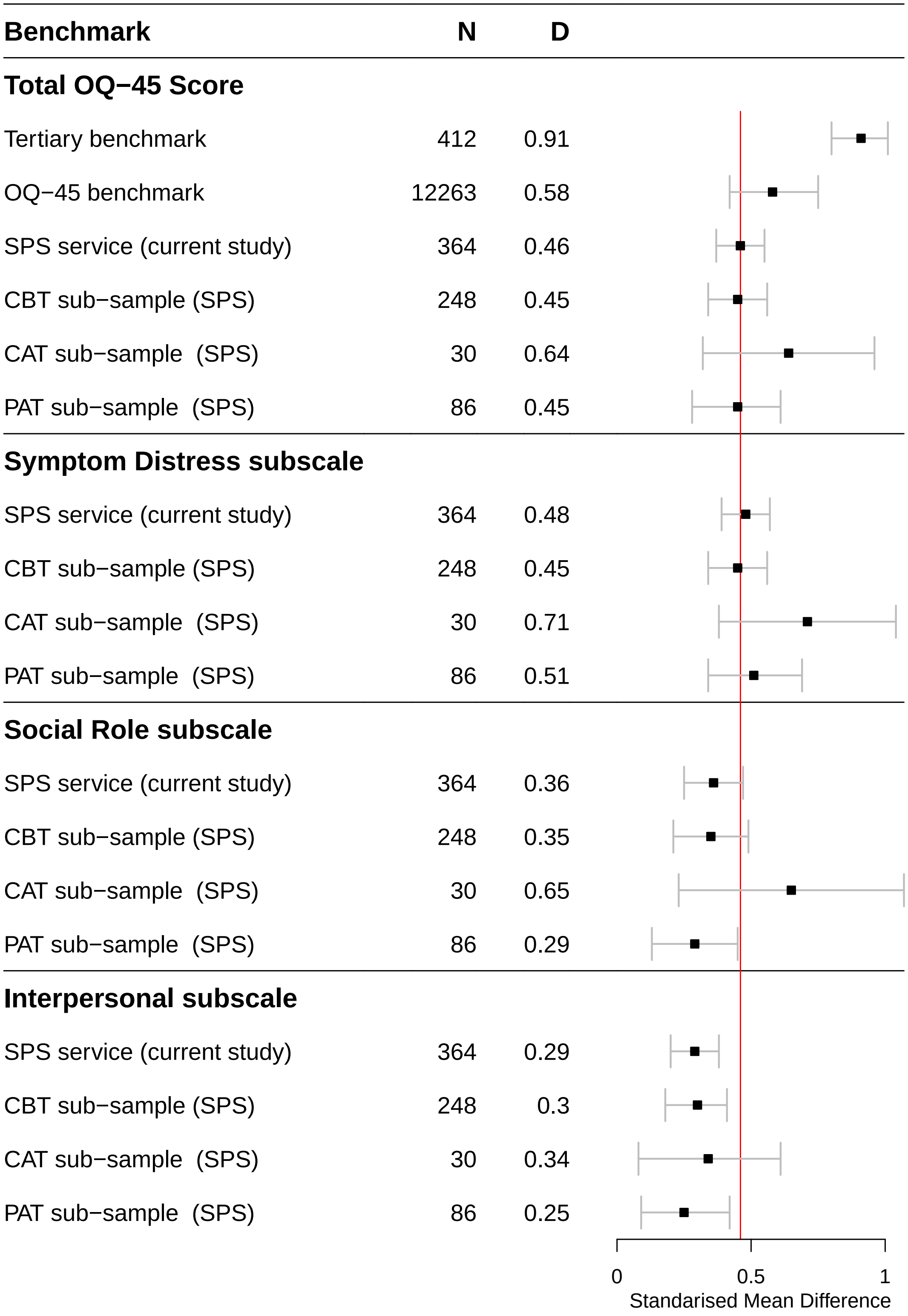
The SPS intake score (mean = 102.57, SD = 22.79) was markedly higher than any of the OQ-45 baseline distress benchmarks (73.02-89.17). The overall SPS effect-size was small (d = 0.46, 95% CI = 0.37- 0.55). For specific treatments, CAT produced the largest effect size (medium, d = 0.64, 95% CI = 0.32- 0.96) and although this was greater than both CBT (d = 0.45, 95% CI = 0.34- 0.56) and PAT (d = 0.45, 95% CI = 0.28- 0.61), the CAT sample was much smaller.

The majority of the studies representing the meta-analytic benchmark (total sample = 12,263) came from the USA (n = 8), with the remaining studies coming from Switzerland (n = 2), Norway (n = 1) and Israel (n = 1). Only two studies were larger in sample size than the SPS effectiveness/recovery sample (Baldwin et al., 2009; Goldberg et al., 2016). Details of outcomes studies included in the OQ-45 benchmark are reported in Appendix H. The aggregated OQ-45 pre-post therapy effect-size was medium (*d* = 0.58, *k* = 13, CI = 0.42-0.75 *p* = < .001). Two studies were statistical outliers (Goldberg et al., 2016; Lunnen et al., 2008). These studies were kept in the OQ-45 effectiveness benchmarks as preliminary sensitivity removal did not substantially alter the effect-size (*d* = 0.60, *k* = 11, 95% CI = 0.48-0.73, *p* = < .001).

In comparing the SPS effect-size and CI to the meta-analytic benchmark and separate tertiary care benchmark, it is shown that both benchmark effect-sizes (OQ-studies and tertiary care) exceeded the SPS CI region, indicating inferior effectiveness for the current study outcomes. CIs for each specific SPS treatment modality overlapped with the SPS pooled effect-size indicating that no specific treatment was superior. SPS effect-sizes (total and subscales) and the selected tertiary and OQ-45 benchmarks are shown in Figure 1.

Figure 1

Forest plot of pre-post therapy effect sizes for the current study sample and selected effectiveness benchmarks. The square boxes depict individual Cohen's d effect sizes and error bars display 95 percent confidence intervals. The red horizontal line represents the Cohen's d effect-size for the pooled SPS sample.



Benchmark	N	D
Total OQ−45 Score		
Tertiary benchmark	412	0.91
OQ−45 benchmark	12263	0.58
SPS service (current study)	364	0.46
CBT sub−sample (SPS)	248	0.45
CAT sub−sample (SPS)	30	0.64
PAT sub−sample (SPS)	86	0.45
Symptom Distress subscale		
SPS service (current study)	364	0.48
CBT sub−sample (SPS)	248	0.45
CAT sub−sample (SPS)	30	0.71
PAT sub−sample (SPS)	86	0.51
Social Role subscale		
SPS service (current study)	364	0.36
CBT sub−sample (SPS)	248	0.35
CAT sub−sample (SPS)	30	0.65
PAT sub−sample (SPS)	86	0.29
Interpersonal subscale		
SPS service (current study)	364	0.29
CBT sub−sample (SPS)	248	0.3
CAT sub−sample (SPS)	30	0.34
PAT sub−sample (SPS)	86	0.25

Recovery

Recovery rates for SPS overall and each treatment are shown in Table 2 and are visualised in a Jacobson plot (Figure 2). The plot shows a dense congestion of people within the ‘no change’ and ‘improved’ regions. There were 18 (4.95%) patients who fell within the non-clinical range at baseline and so these could not reach all criteria required for clinical recovery. In terms of post treatment status, 37 (10.16%) patients recovered, 109 (29.95%) patients made a reliable improvement and 29 (7.97%) patients made a reliable deterioration. There were only small discrepancies shown in recovery rates between modalities. In comparison to the pooled benchmark, less patients recovered and more had no reliable change. Rates of recovery and no-change were very similar between SPS and the community mental health centre benchmark. By comparison SPS showed less deterioration and greater rates of reliable improvement. There was a significant difference in response rates between SPS and the overall recovery benchmarks (p = .001), but no difference to the more closely related community mental health centre benchmarks (p = 0.098).

Table 2

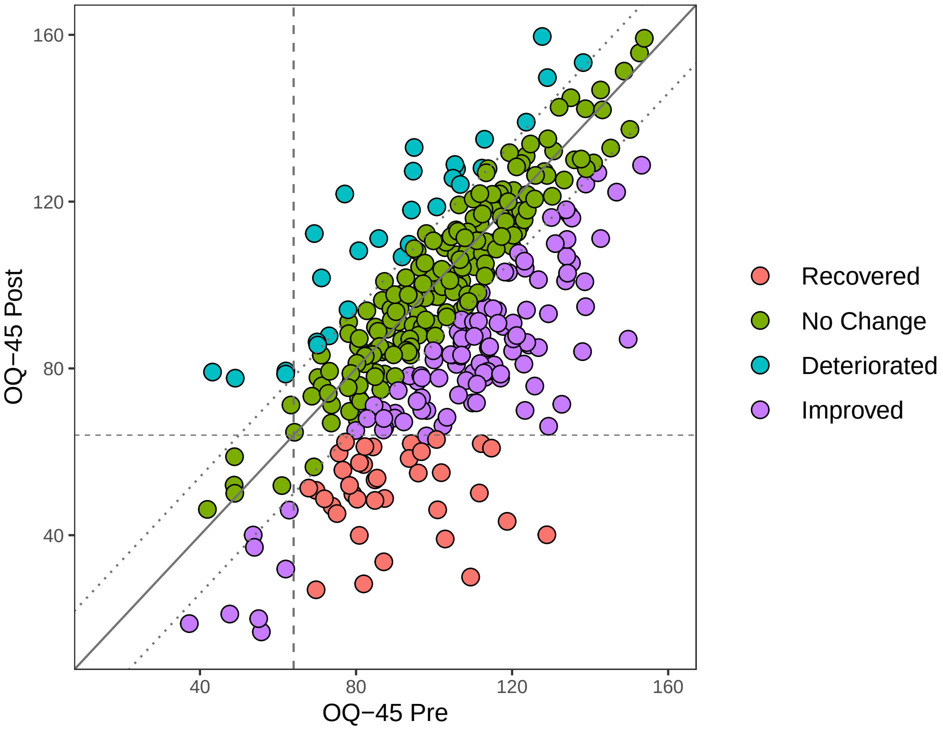
Rates of reliable change, recovery and deterioration for the current study sample and for selected bencmarks (Hansen and Lambert 2002).

Study		Recovered	No Change	Deterioration	Improved	Total
Hansen (2002)	Total	681 (14.3%)	2709 (56.9%)	377 (7.9%)	994 (20.9%)	4761
	CMHC	31 (8.6%)	219 (60.6%)	37 (10.2%)	74 (20.5%)	361
SPS	Total	37 (10.16%)	189 (51.92%)	29 (7.97%)	109 (29.95%)	364
	CBT	22 (8.87%)	128 (51.61%)	23 (9.27%)	75 (30.24%)	248
	PDT	11 (12.79%)	46 (53.49%)	4 (4.65%)	25 (29.07%)	86
	CAT	4 (13.33%)	15 (50.00%)	2 (6.67%)	9 (30.00%)	30

Note. 18 patients fell within in the non-clinical range at baseline. No change = no reliable changes Improved = Reliable Improvement Deterioration = reliable Deterioration

Figure 2

Jacobson plot to show the rates of patient response. Points to the right of the vertical dashed line represent patients who started treatment as clinically distressed. Points beneath the horizontal dashed line represent patients who finished treatment in the non-clinically distressed range.



Dose Response

Figure 3 shows a bar plot for treatment response broken down by number of sessions. Patients with very short treatment durations (i.e. less than 10) were highly unlikely to have responded to treatment. Response generally increases in-line with sessions until approximately 40 sessions. It is important to note that this graph does not determine at which point within treatment patients responded. A similar trend was shown for statistical change (Table 3), Cohen's *d* pre-post effect-sizes were particularly small for patients receiving less than 10 sessions, increases with number of sessions, until the point of approximately 40 sessions.

Figure 3

Rates of response to tertiary care therapy, based on number of sessions received. Response here is the sum of patients who showed reliable improvement. Bars show rate of improvement by number of sessions received. The red line/text denotes the cumulative number of patients who had improved by the number of sessions received.

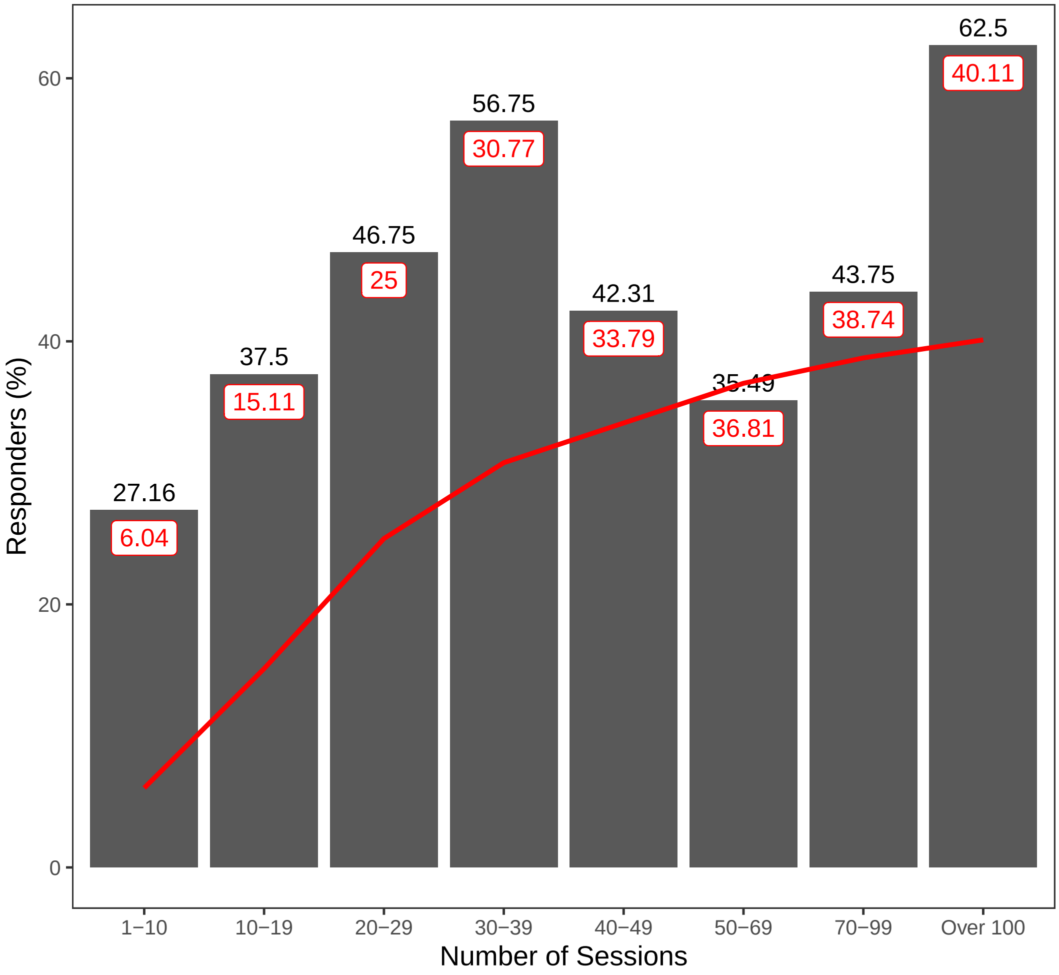


Table 3

Non-cumulative, differential rates of statistical and clinical change based on different dosage groups.

Sessions	n	d	ci	Recovered	No Change	Deteriorated	Improved
1-10	81	0.15	-0.04-0.35	4 (4.94%)	50 (61.73%)	9 (11.11%)	18 (22.22%)
10-19	88	0.51	0.31-0.71	8 (9.09%)	48 (54.55%)	7 (7.95%)	25 (28.41%)
20-29	77	0.63	0.41-0.85	14 (18.18%)	32 (41.56%)	9 (11.69%)	22 (28.57%)
30-39	37	0.94	0.58-1.3	5 (13.51%)	16 (43.24%)	0 (0.00%)	16 (43.24%)
40-49	26	0.54	0.17-0.91	0 (0.00%)	14 (53.85%)	1 (3.85%)	11 (42.31%)
50-69	31	0.25	-0.07-0.57	3 (9.68%)	18 (58.06%)	2 (6.45%)	8 (25.81%)
70-99	16	0.67	0.17-1.16	1 (6.25%)	9 (56.25%)	0 (0.00%)	6 (37.50%)
Over 100	8	0.46	-0.2-1.12	2 (25.00%)	2 (25.00%)	1 (12.50%)	3 (37.50%)

Growth Curves

There were 298 patients with at least three OQ-45 measures and which could be used in the growth-curve analysis. Fixed effects and goodness-of-fit statistics for unconditional and conditional models are shown in Table 4. Growth curves for unconditional models are shown in Figure 4 while conditional model growth curves are in Appendix I.

For the OQ-45 total score final unconditional model there was a significant main effect for intercept (initial score, γ = 104.18, F = 2100, p = <0.001). For sessions, the significant time trends included linear (γ = -0.5, F = 38.04, p = <0.001), quadratic (γ = 0.005, F = 25.15, p = <0.001), and cubic (γ = -0.00001, F = 21.39, p = <0.001), In other words, OQ-45 total score dropped by 0.5 per session, however this was gradually reversed by the curvi-linear terms. Examination of the growth curve for total score (Figure 4) reveals that improvements begin to dissipate at approximately the 150th session. In the conditional model (including treatment modality), there was no significant main effect for treatment modality (γ = -3.1, F = 6.42, p = 0.149), or treatment:sessions interaction (γ = -0.07, F = 1.88, p = 0.171) however the conditional model did provide a significantly improved model fit (χ^2 = 8.24, p = 0.016).

As shown in Table 4, the fixed effect for therapy modality (represented by ‘Analytic’) and the modality:session interaction (not displayed) was not significant for any of the OQ-45 sub-scales. Conditional models for symptom distress and social role demonstrated significant improved model fit when compared to unconditional models however this was not the case for inter-personal. Log-linear (social role) and cubic (symptom distress, interpersonal) growth curves are illustrated in growth curve plots in Figure 4.

Table 4

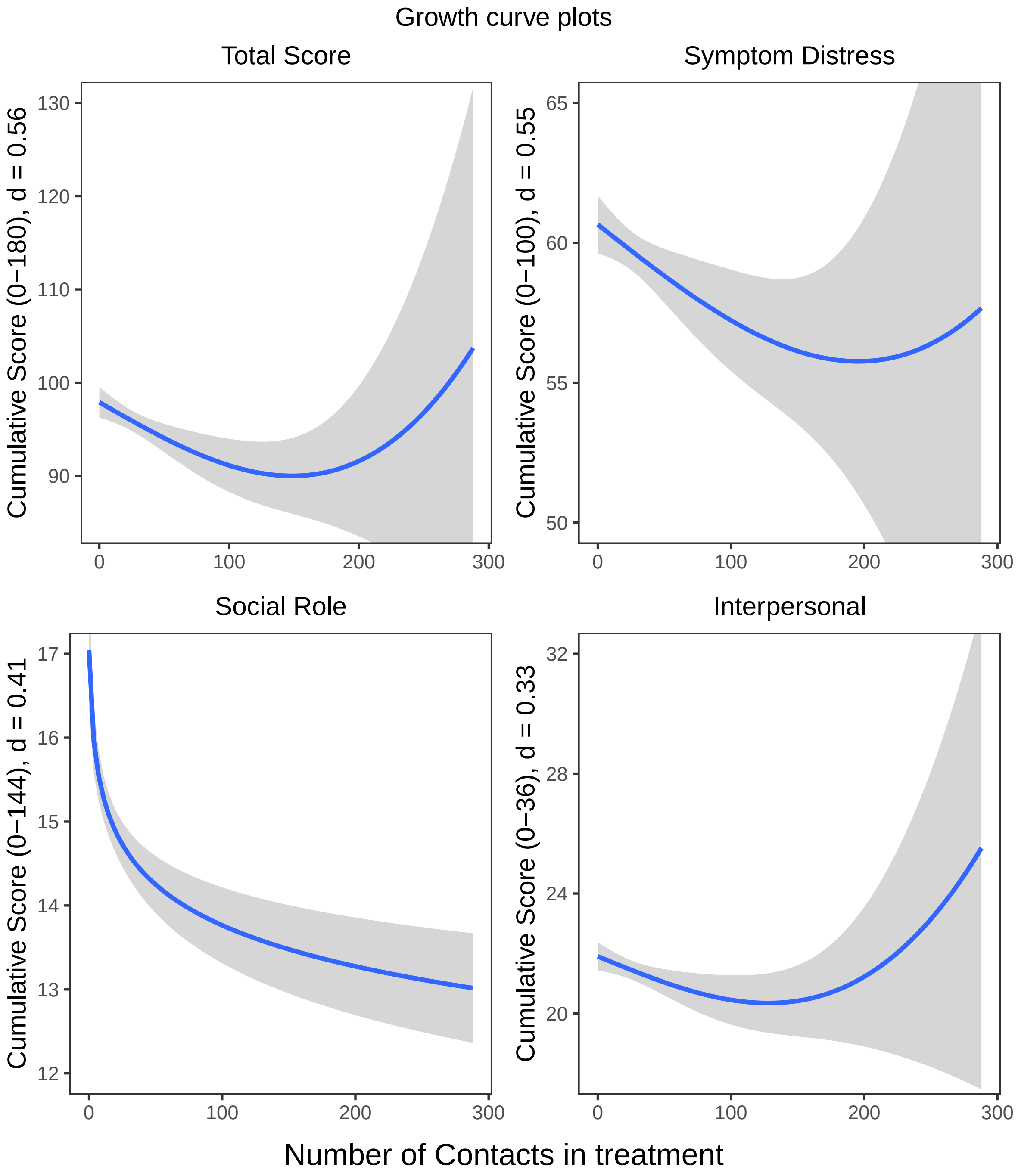
Fixed effects and goodness-of-fit statistics for optimal unconditional and final conditional models for the OQ-45 and each of the three sub-scales

	Total Score		Symptom Distress		Social Role		Inter-Personal	
	OQ Cubic	OQ Cond.	SD Log	SD Cond.	SR Log	SR Cond.	IR Quad	IR Cond.
Fixed Effects								
Intercept	104.180*	105.453*	64.725*	65.403*	17.668*	17.847*	22.973*	23.151*
	(1.497)	(1.693)	(0.962)	(1.246)	(0.407)	(0.515)	(0.434)	(0.496)
Linear/Log	-0.496*	-0.494*	-0.293*	-0.275*	-0.949*	-1.021*	-0.103*	-0.102*
	(0.057)	(0.058)	(0.036)	(0.070)	(0.129)	(0.165)	(0.017)	(0.017)
Quadratic	0.005*	0.005*	0.003*	0.003*			0.001*	0.001*
	(0.001)	(0.001)	(0.000)	(0.001)			(0.000)	(0.000)
Cubic	-0.000*	-0.000*	-0.000*	-0.000*			-0.000*	-0.000*
	(0.000)	(0.000)	(0.000)	(0.000)			(0.000)	(0.000)
Analytic		-3.102		-1.107		-0.425		-0.424
		(2.134)		(1.641)		(0.768)		(0.634)
Goodness of fit								
AIC	17216.0	17211.7	15199.3	15193.4	12110.7	12114.2	12166.7	12168.6
BIC	17261.2	17268.2	15244.5	15255.6	12144.6	12159.4	12211.9	12225.1
Log-Likelihood	-8599.99	-8595.87	-7591.66	-7585.72	-6049.34	-6049.08	-6075.34	-6074.32
p value		0.016		0.008		0.774		0.362

* $p < 0.05$

Figure 4

Growth curves for optimal unconditional models. Blue lines represent growth curve trajectories. Grey shaded regions represent 95% confidence interval regions. Trends line types represent cubic for OQ-45 total score, interpersonal and symptom distrss. Social role is represented by a log-linear trend.



Discussion

The aim of this study was to explore the effectiveness of psychological therapy delivered within a tertiary care therapy service using a benchmarking approach to contextualise intake distress and outcomes produced. The baseline severity of initial distress was higher in the current study than any of the available OQ-45 benchmarks. The pooled pre-post therapy effect-size for the current study was small and also smaller than the two employed benchmarks from tertiary care services and OQ-45 outcome studies. The three treatment modalities showed equivalent rates of effectiveness evidenced through CI overlap with SPS pooled effect-size. SPS individual patient response rates for reliable improvement and recovery were comparable to a US OQ-45 community mental health center benchmark (Hansen & Lambert, 2002). Growth curve trajectories demonstrated that OQ-45 scores reduced by 0.5 points per session, but that this rate declined in line with significant curvi-linear trends (quadratic and cubic). There was no significant difference in change over time between treatment modalities.

Contribution to the Evidence Base

Because tertiary care services differ markedly from primary/secondary care services (i.e. longer treatment, specialist interventions, marked distress and delivering therapy to those that have not responded to prior treatments), it is necessary that the effectiveness of such services is assessed. Prior UK tertiary care outcome studies have been particularly small (n = <50, Douglas et al., 2016; Paley et al., 2008) making it hard to generalize results to larger populations, comment on service design or make practice recommendations.

The current study found greater initial distress status than a range of benchmarks from other service sectors which have employed the OQ-45. This finding provides support to the notion that tertiary care patients tend to present with greater levels of psychological distress and are therefore possibly a different clinical population. This was to the extent that patients in the current study showed, on average, higher levels of distress than US acute/short stay hospital inpatients (Doerfler et al., 2002). As the current study was the first (to our knowledge) UK study to employ the OQ-45, it is not clear as to whether the elevated scores in the current study are solely due to differences in service/sector. It is possible that differences are in part due to a broader difference between UK and US samples and societies. For example is has been demonstrated that out-patient and community norms are higher in the UK than the US for a number of measures of psychological distress (Francis et al., 1990; Ryan, 2007). Further use of the OQ-45 in other UK services/sectors is necessary in order to provide a more valid comparison of baseline distress between UK service sector patients. As the available US initial distress benchmarks were more than 15 years old, it is also possible that these may not reflect contemporary levels of initial distress.

This study demonstrated the effectiveness of tertiary care therapy using a much greater sample size than has previously been reported in the UK. Effectiveness within this study was based on the first care period for each patient. For an unknown proportion of patients this will have included treatments delivered consecutively, and with periods of waiting between treatments. Effectiveness should therefore be considered to represent tertiary care episodes, and not necessarily single treatments.

The finding that non-linear trends best fit each of the outcome variables was consistent with the dose-response evidence base (Bone et al., 2021; Howard et al., 1986; Robinson et al., 2020). The time trends of best fit for the OQ-45 total, symptom distress and interpersonal relationships (all cubic) suggest that patients who remain in therapy for very long periods are at risk of starting to deteriorate. More specifically, patients with very long treatments (i.e. over 100) had a smaller pooled effect-size, but without raised rates of deterioration. This suggests that patients engaged in long treatments tend to either (i) make smaller over-all improvements, and with that improvements occurring in earlier stages of treatment; or, (ii) can be classed as 'non-responders.' For social role (log-linear), improvements were rapid in the early stage of treatment, while further improvements were made with a negatively decelerating rate. These findings are consistent with the body of literature demonstrating that the majority of improvements made within therapy are achieved within the early stages (Bone et al., 2021; Robinson et al., 2020). Practically, this means that patients may show limited added benefit from long forms of psychotherapy.

The relatively high rate of patients who (based on the OQ-45) did not respond to treatment (deterioration and no change combined = 59.89%) may be due to a range of factors which are considered to be more prevalent within complex client groups (e.g. compulsion to repeat, van der Kolk, 1989). These factors may have been also due to reasons beyond the therapy room that were not measured in the current study such as the influence of social disadvantage, life circumstances or chronicity of prior traumas (Finegan et al., 2018). These potential reasons for non-response should be empirically explored within future tertiary care therapy studies exploring predictors of treatment response/non-response.

It is not clear as to why social role was the only sub-scale that continued to make improvements (albeit at a reducing rate), or why interpersonal and symptom distress did not. This pattern is different to patterns identified in prior empirical studies (Schilling et al., 2020; White et al., 2015); these have included (i) alliance and motivational difficulties; (ii) social support and life event difficulties; and (iii) indistinguishable patterns. This pattern may be more characteristic of patients who access UK tertiary care therapy. The finding that interpersonal problems was a predictor of not responding to treatment was consistent with prior evidence (Probst et al., 2020). Two of the therapies delivered (i.e. PAT and CAT) focus on interpersonal issues, but this does not seem to have made a great difference.

The overall effect-size observed in the current study was smaller than that reported by Paley et al. (2008). The overall effect-size was also less than that shown in the Johansson et al. (2014) Canadian tertiary care service. As these studies used different outcome measures, it was not possible to say whether the two studies were comparable on baseline distress. These measures may also differ in sensitivity to change. There is a need for more tertiary care outcome studies which are comparable in size and design to the current study in order to provide a more valid set of benchmarks. The overall effect-size for this study was likely to have been suppressed by patients with long treatments who do not respond (hence the cubic finding).

Based on the growth-curve plots, there was evidence that growth became negative from the point of approximately 150 sessions for OQ-45 total score. This is consistent with the literature which shows a lack of added value for very long treatment durations (Bone et al., 2021; Robinson et al., 2020). Negative growth was influenced by two sub-scales with cubic trends (symptom distress and interpersonal). In other words, patients in the latter stages of long treatments show particularly elevated rates of interpersonal distress and symptom distress. While tertiary care psychological therapy will, in many cases, intend to work on inter-personal difficulties and symptoms distress, those who stay in treatment for very long periods (over 100 sessions) make limited additional improvements, despite the theoretical grounding of the models often being used.

There was a significant difference between SPS recovery rates and the OQ-45 recovery benchmarks. More specifically, less patients in the current sample made full recovery but more made a reliable improvement. It is a likely that this difference is a reflection of the higher initial baseline distress in the SPS sample (i.e., a greater amount of change is required to fall under the clinical threshold to meet full recovery). When compared to the community mental health center sub-group of the OQ-45 recovery benchmarks, which was suspected to be most comparable to SPS, there was no significant difference in recovery rates. Taken together with the above effect-sizes (Cohen's *d*), the current study found that tertiary care patients showed smaller average (within-group) change, but comparable rates of individual change to other service sectors. These comparisons do not consider an array of potential differences in treatments (e.g., number of sessions, clinician experience, heterogeneity in the clinical presentations and complexity of the patient samples).

The current study found limited evidence for significant differential rates of effectiveness between different psychological modalities. CAT showed the highest indices of average (within-group) change (Cohen's *d*), however the accompanying confidence intervals overlapped with the effect-sizes of both of the other treatment modalities. The potential of CAT in tertiary care may be the relatively brief treatment contract and the relational focus. A recent meta-analysis of CAT (Hallam et al., 2021) showed its differential efficacy in typically complex client groups. In this study, both CAT and CBT showed superior levels of 'service efficiency' based on providing comparable clinical effectiveness, but within the boundaries of significantly shorter treatment contacts. It is possible that PAT treatments were skewed by a small number of patients with very long care periods. The comparable rates of effectiveness (Cohen's *d* and response rates) between the three modalities provides further support to the equivalence paradox; that is, bone fide psychotherapies delivered in routine care tend not to significantly differ in terms of effectiveness.

Limitations

There were nine additional main study limitations: (i) by focusing on pre-post change it is not possible to say to what extent the observed results are due to the allocated treatment or due to other potential confounds (e.g. regression to the mean, spontaneous remission, co-occurring pharmacotherapy); (ii) as outcome measurement was not conducted at every session, the level of detail regarding change over time is limited in granularity; (iii) the pre-post effect-size calculation did not include patients who were accepted for treatment but who did not attend the first appointment (i.e. treatment rejectors), this study may therefore have over-inflated effect-sizes; (iv) concurrent treatments (e.g. pharmacological treatment, other involved services) were not known, and which may account for a range of within-group differences in effectiveness; (v) the absence of treatment fidelity measures means that the study is unable to assess the degree to which treatments were delivered as intended; (vi) frequency of clinical supervision was not reported; (vii) the lack of UK normative data for the OQ-45 means that recovery rates may be inaccurate, (viii) findings were exclusively based upon a single self-report measurement tool, it was therefore not possible to comment upon effectiveness that may be derived through other methods (e.g. clinician rated, change in health-care utilisation, change in harm events). While there are analysis methods available to calculate change indices and cut-off scores using only the current available sample (e.g. Morley & Dowzer, 2014) these methods are less precise. Finally, the current study was purely quantitative and therefore little is known about the patient experience of the changes recorded.

Implications: Research, Policy & Practice

The finding that initial distress was higher than benchmarks for other sectors fits with the intention that tertiary care services cater for patients with particularly high levels of distress. The current study found that tertiary care therapy services produce significant change over time with small rates of therapeutic effectiveness. Effect-sizes are likely to be suppressed by a sub-group of patients who appear to not respond to psychological treatment, despite considerable numbers of sessions. On average, growth appears to become negative from approximately 150 sessions. The identification of a group of patients that are ‘treatment non-responders’ in a larger sample of patients referred to a tertiary service for people that have been previously ‘treatment resistant,’ means that tertiary services need to identify ‘non-responders’ at the earliest possibility. Whilst services want to help (or possibly ‘rescue’), there may be some patients for whom psychological pain is enduring and help is ineffective and this means alternative management strategies need to be developed and tested for this population. Despite this, it should be noted that other indices of effectiveness (e.g. rates of harm, post-treatment rates of health service utilisation) were not considered and which may provide alternative justifications for treatment continuation. Utilising available information regarding not-on-track status during clinical supervision and patient feedback may allow for consideration of earlier, necessary treatment termination, and which may subsequently improve the efficiency of treatments (de Jong et al., 2021).

Future Research

Further practice-based evidence is clearly required in the area of tertiary care psychotherapy. Such studies are particularly needed to provide more representative UK benchmarking comparisons. Studies should strive to employ a session-by-session outcome measurement approach that can reduce data missingness and clinician bias. Future studies comparing differential rates of modality effectiveness should aim to provide more adequate sample numbers across treatment comparison groups. Mixed methods would also be useful. Future studies should also employ fidelity measures to ensure treatments compared are treatments intended. Future studies should pay attention to other indices of treatment success (e.g. health care utilisation, harm events). Attention should be paid to differential rates of effectiveness between therapists (i.e. the role of ‘therapist effects’) however this would need a much larger sample. Finally, psychometric validation studies for the OQ-45 have taken place in various countries outside the US (e.g., Jong et al., 2007; Wennberg et al., 2010), but not yet in the UK; this is an important next step.

Conclusions

The current study explored the effectiveness of psychological therapy within a tertiary care therapy service. Patient initial distress was higher than a range of service benchmarks, indexing a highly distressed clinical population. Despite delivery of relatively lengthy interventions, the size of pre-post effectiveness outcome was small. The study effect-size was lower than a meta-analytic benchmark of studies employing the OQ-45 and a separate benchmark for a Canadian tertiary care service. Effectiveness was seemingly reduced by a number of patients receiving very long treatments and who do not improve during these interventions. Cohen's *d* and response rates were generally equivalent across different treatment modalities. Change in tertiary patients occurred over time and with a cubic form. Further evaluations of effectiveness of tertiary care psychological therapy services is required and particularly the development of service delivery models that can be responsive to lack of change in patients whilst retaining fidelity to the theory and methods of the interventions delivered.

Appendix

Appendix A. The Outcome-Questionnaire 45 (Lambert et al., 2004).

Appendix B. Documents of approval for the current project. Included here are key pages indicating approval.

1. Preliminary research ethics committee approval (REC) and subsequent full REC (two pages).
2. Health Research Authority (HRA) approval (two pages).
3. Approval for section 5 of the health service (Control of Patient Information) Regulations 2002 ('section 251 support') to process confidential patient information without patient consent (four pages).
4. University scientific approval (one page).
5. Research sponsor letter from University (two pages).
6. Local research approval from the host NHS trust (one page).

NHS
Health Research
Authority

North West - Greater Manchester West Research Ethics Committee
Barlow House
3rd Floor
4 Minshull Street
Manchester
M1 3DZ

Please note: This is an acknowledgement letter from the REC only and does not allow you to start your study at NHS sites in England until you receive HRA Approval

17 January 2020

Mr Chris Gaskell
Clinical Psychology Unit
Cathedral Court
1 Vicar Lane
S1 2LT

Dear Mr Gaskell

Study title:	**Effectiveness of Differing Psychotherapies Offered in a Specialist Psychotherapy Service – a benchmarking study.**
REC reference:	**19/NW/0753**
IRAS project ID:	**273884**

Thank you for your letter of 09 January 2020. I can confirm the REC has received the documents listed below and that these comply with the approval conditions detailed in our letter dated 17 December 2019

Documents received

The documents received were as follows:

Document	*Version*	*Date*

Covering letter on headed paper		09 January 2020
IRAS Application Form [IRAS_Form_09012020]		09 January 2020

Approved documents

The final list of approved documentation for the study is therefore as follows:

Document	*Version*	*Date*
Covering letter on headed paper		09 January 2020
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [University Insurance]		31 October 2019
IRAS Application Form [IRAS_Form_14112019]		14 November 2019
IRAS Application Form [IRAS_Form_09012020]		09 January 2020
IRAS Application Form XML file [IRAS_Form_14112019]		14 November 2019
IRAS Checklist XML [Checklist_14112019]		14 November 2019
Other [Second Supervisor CV]		31 October 2019
Research protocol or project proposal [Research Protocol]		31 October 2019
Summary CV for Chief Investigator (CI) [Chief Investigator CV]		31 October 2019
Summary CV for supervisor (student research) [Lead Supervisor CV]		31 October 2019
Validated questionnaire [Outcome Questionnaire - 45]		

You should ensure that the sponsor has a copy of the final documentation for the study. It is the sponsor's responsibility to ensure that the documentation is made available to R&D offices at all participating sites.

19/NW/0753 **Please quote this number on all correspondence**

Yours sincerely

Gemma Warren

E-mail: nrescommittee.northwest-gmwest@nhs.net

Copy to: *Mr Chris Gaskell*
Ms Yiwei Harland, Sheffield Health & Social Care NHS Foundation Trust

Lead Nation England: HRA.Approval@nhs.net

Ymchwil Iechyd
a Gofal Cymru
Health and Care
Research Wales

NHS
Health Research
Authority

Mr Chris Gaskell
Trainee Clinical Psycholoigst
Sheffield Health & Social Care NHS Foundation Trust
Clinical Psychology Department
Cathedral Court
Vicar Lane
S1 2LT

Email: approvals@hra.nhs.uk

15 March 2021

Dear Mr Gaskell

HRA and Health and Care Research Wales (HCRW) Approval Letter

Study title:	**Effectiveness of Differing Psychotherapies Offered in a Specialist Psychotherapy Service – a benchmarking study.**
IRAS project ID:	**273884**
REC reference:	**19/NW/0753**
Sponsor	**The University of Sheffield**

I am pleased to confirm that **HRA and Health and Care Research Wales (HCRW) Approval** has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications received. You should not expect to receive anything further relating to this application.

Please now work with participating NHS organisations to confirm capacity and capability, in line with the instructions provided in the "Information to support study set up" section towards the end of this letter.

How should I work with participating NHS/HSC organisations in Northern Ireland and Scotland?

HRA and HCRW Approval does not apply to NHS/HSC organisations within Northern Ireland and Scotland.

If you indicated in your IRAS form that you do have participating organisations in either of these devolved administrations, the final document set and the study wide governance report (including this letter) have been sent to the coordinating centre of each participating nation. The relevant national coordinating function/s will contact you as appropriate.

Please see IRAS Help for information on working with NHS/HSC organisations in Northern Ireland and Scotland.

How should I work with participating non-NHS organisations?

HRA and HCRW Approval does not apply to non-NHS organisations. You should work with your non-NHS organisations to obtain local agreement in accordance with their procedures.

What are my notification responsibilities during the study?

The standard conditions document "*After Ethical Review – guidance for sponsors and investigators*", issued with your REC favourable opinion, gives detailed guidance on reporting expectations for studies, including:

- Registration of research
- Notifying amendments
- Notifying the end of the study

The HRA website also provides guidance on these topics, and is updated in the light of changes in reporting expectations or procedures.

Who should I contact for further information?

Please do not hesitate to contact me for assistance with this application. My contact details are below.

Your IRAS project ID is **273884**. Please quote this on all correspondence.

Yours sincerely,

Mark Sidaway
Approvals Specialist

Email: approvals@hra.nhs.uk

Copy to: *Dr Andrew Thompson*

NHS
Health Research Authority

Skipton House
80 London Road
London
SE1 6LH

Tel: 020 797 22557
Email: cag@hra.nhs.uk

10 March 2021

Dr Stephen Kellett
Clinical Psychology Unit
Cathedral Court
1 Vicar Lane
S1 2LT

Dear Dr Kellett

Application title:	**Effectiveness of Differing Psychotherapies Offered in a Specialist Psychotherapy Service – a benchmarking study.**
CAG reference:	**20/CAG/0017**
IRAS project ID:	**273884**
REC reference:	**19/NW/0753**

Thank you for submitting a **research** application under Regulation 5 of the Health Service (Control of Patient Information) Regulations 2002 (’section 251 support’) to process confidential patient information without consent.

Supported applications allow the controller(s) of the relevant data sources, if they wish, to provide specified information to the applicant for the purposes of the relevant activity without being in breach of the common law duty of confidence. Support provides a lawful basis to allow the information to be processed by the relevant parties for the specified purposes without incurring a breach of the common law duty of confidence only. Applicants must ensure the activity remains fully compliant with all other relevant legislation.

The role of the Confidentiality Advisory Group (CAG) is to review applications submitted under these Regulations and to provide advice to the Health Research Authority on whether application activity should be supported, and if so, any relevant conditions. This application was considered at the precedent set CAG meeting held on 14 February 2020. The application was considered via the Precedent Set process under criteria 4. Time limited access to undertake record linkage/validation and to anonymise the data

This outcome should be read in conjunction with the provisional support letter dated 02 March 2020.

Health Research Authority decision

The Health Research Authority, having considered the advice from the Confidentiality Advisory Group as set out below, has determined the following:

1. The application, to allow the applicant, who is not a member of the direct care team, to access confidential patient information from the Specialist Psychotherapy within Sheffield Health and Social Care NHS Foundation Trust in order to extract an anonymised dataset for analysis at the University of Sheffield, is conditionally supported, subject to compliance with the standard and specific conditions of support.

Please note that the legal basis to allow access to the specified confidential patient information without consent is now in effect.

Context

Purpose of application

This application from the University of Sheffield set out the purpose of medical research that seeks to determine how effective and durable psychotherapy is for patients presenting for specialist tertiary care psychotherapy.

Tertiary level psychotherapy services, often regional centres, cater for particularly complex and enduring presentations. This is because patients using these services have been non-responsive to interventions offered in Primary and Secondary care. Little research evidence is available regarding the effectiveness of tertiary care psychotherapy services and is therefore under-represented with regards to evidence compared to primary and secondary services. The applicants will carry out an in-depth statistical analysis of a pre-existing routine outcome data-set collected by a Specialist Psychotherapy Service (SPS) at Sheffield Health and Social Care NHS Foundation Trust, to collect evidence for local tertiary care services to establish how tertiary care services compare to existing benchmarks set by randomised controlled trials.

Patients accessing the SPS are invited to complete an outcome measure, the OQ-45, at the start of treatment, monthly, at the end of treatment and as follow-up from psychotherapy at the SPS. The changes in the questionnaire score will be used as an indicator of the effectiveness of an intervention. Patient outcomes data, along with other relevant demographic information, age, gender, employment status, is routinely recorded within an electronic database. Patients who indicate they do not wish to share their clinical records for health purposes, or decline questionnaire participation, are omitted from this process. The applicants will use this dataset for analysis. They will also request individual electronic patient records in order to extract additional information on medication use, previous psychological therapy input, diagnosis, carecluster. The dataset will be imported into the service outcomes database. Support is sought as the student investigator, who is not part of the direct care team, will carry out the anonymisation of the dataset, before it is exported to the University of Sheffield through the data-gatekeeper identified for this study. Following verification that data is fully anonymised, data will be forwarded to the research team for data analysis.

A recommendation for class 1,4 5 and 6 support was requested to cover access to the relevant unconsented activities as described in the application.

Confidential patient information requested

The following sets out a summary of the specified cohort, listed data sources and key identifiers. Where applicable, full datasets and data flows are provided in the application form and relevant supporting documentation as this letter represents only a summary of the full detail.

Cohort	Patients aged between 18 and 80, who have attended a specialist psychotherapy service. The applicants anticipate that 1000 patients will be included.
Data sources	1. Electronic patient records at Sheffield Health and Social Care NHS Foundation Trust 2. The outcomes evaluation database at Sheffield Health and Social Care NHS Foundation Trust
Identifiers required for linkage purposes	1. Name 2. NHS Number 3. Date of birth
Identifiers required for analysis purposes	No identifiers will be retained for analysis purposes

Confidentiality Advisory Group advice

This letter summarises the outstanding elements set out in the provisional support letter, and the applicant response. The applicant response was considered by a sub-committee of the CAG.

1. **Members asked that a data flow diagram was provided, clarifying the "NHS Electronic records' which will be accessed. Members also queried whether primary care data would be included and whether the data source used was complete enough to provide the required information.**

The applicant provided a data flow diagram.

The applicant clarified that the NHS electronic records required will be those held by Sheffield Health & Social Care NHS Foundation Trust. Primary care data will not be accessed, and the applicants anticipated that the required information will be reported within Sheffield Health & Social Care NHS Foundation Trust secondary care clinical notes. The applicants confirmed that they expected that this data would be sufficient enough to run the proposed regression analyses.

The CAG noted a concern with step 2 in the data flow diagram. This states that 'This will include notes from the Specialist Psychotherapy Service specifically, and also notes from any other SHSC NHS Trust service that patients have accessed. This information will then be incorporated into the 'extrac_V1' spread sheet.' It was not clear whether a truly anonymised dataset would be created, which was safe to share freely with others. If there was a possibility that the dataset remained disclosive, then the dataset would need to be described as 'de-identified' throughout the dataflow, as a signal that the output continues to require protection.

2. **Patient and public involvement must be undertaken to explore the views of patients around the use of their confidential patient information. The Group suggested contacting a mental health charity, such as MIND, to facilitate this. Feedback from the patient and public involvement is to be provided to the CAG for review. Details should be provided around the format of the activity, the demographics of those involved and the information which was provided together with an overview of the feedback which was provided.**

The applicants explained that they had approached MIND, who declined to be involved as they determined that the Sheffield branch did not provide a suitable platform for conducting patient and public involvement. The applicants discussed the study with two expert-by-experience consultants, who worked for Sheffield Health & Social Care NHS Trust. Feedback from this discussion was provided.

The applicants also planned to conduct on-going patient and public involvement during the course of the study, including presenting the study at a patient involvement group, SHSC Sun:Rise. The applicants noted that this meeting may be delayed due to the Covid-19 pandemic.

3. **A poster offering an opt out is to be displayed in the Sheffield service for a period of 4-6 weeks before data extraction takes place. The poster needs to contain information on how patients can register their dissent. Notices should also be placed on appropriate websites.**

The applicants provided a poster, to be displayed in the host service for 4-6 weeks prior to the planned extraction. This poster was refined using feedback from patient and public involvement will input from the Communications Department at (Sheffield Health & Social Care NHS Trust. The applicants also planned to use the poster content in a patient notice to be placed on the SHSC Specialist Psychotherapy Service website.

Confidentiality Advisory Group advice conclusion

The CAG agreed that the minimum criteria under the Regulations appeared to have been met, and therefore advised recommending support to the Health Research Authority, subject to compliance with the specific and standard conditions of support as set out below.

Specific conditions of support

1. The dataflow diagram needs to be revised to describe the dataset as ‘de-identified.’

2. Favourable opinion from a Research Ethics Committee. **Confirmed 17 January 2020.**

3. Confirmation provided from the IG Delivery Team at NHS Digital to the CAG that the relevant Data Security and Protection Toolkit (DSPT) submission(s) has achieved the ‘Standards Met’ threshold. See section below titled ‘security assurance requirements’ for further information. Confirmed: **2019/20** DSPT **for Sheffield Health and Social Care NHS Foundation Trust has been reviewed as ‘Standards Met’** (confirmed by email to CAG inbox on 8 March 2021).

As the above conditions have been accepted and met, this letter provides confirmation of final support. I will arrange for the register of approved applications on the HRA website to be updated with this information.

The University Of Sheffield.

Department Of Psychology.
Clinical Psychology Unit.

Doctor of Clinical Psychology (DClin Psy) Programme
Clinical supervision training and NHS research training & consultancy.

Clinical Psychology Unit
Department of Psychology
University of Sheffield
Floor F, Cathedral Court
1 Vicar Lane
Sheffield
S1 2LT

Dr A R Thompson, Clinical Training Research Director
Please address any correspondence to Amrit Sinha
Research Support Officer
Telephone: 0114 2226650
Email: a.sinha@sheffield.ac.uk

6th January 2020

To: Research Governance Office

Dear Sir/Madam,

RE: Confirmation of Scientific Approval and indemnity of enclosed Research Project

Project title: **Effectiveness of Differing Psychotherapies Offered in a Specialist Psychotherapy Service – a benchmarking study.**

Investigators: Chris Gaskell (DClin Psy Trainee, University of Sheffield); Dr Jaime Delgadillo & Dr Steve Kellett (Academic Supervisors University of Sheffield)

I write to confirm that the enclosed proposal forms part of the educational requirements for the Doctoral Clinical Psychology Qualification (DClin Psy) run by the Clinical Psychology Unit, University of Sheffield.

Three independent scientific reviewers usually drawn from academic staff within the Psychology Department have reviewed the proposal. Review includes appraisal of the proposed statistical analysis conducted by a statistical expert based in the School of Health and Related Research (ScHARR). Where appropriate an expert in qualitative methods is also appointed to review proposals.

I can confirm that approval of a proposal is dependent upon all necessary amendments having been made to the satisfaction of the reviewers and I can confirm that in this case the reviewers are content that the above study is of sound scientific quality. Consequently, the University will if necessary indemnify the study and act as sponsor.

Given the above, I would remind you that the Department already has an agreement with your office to exempt this proposal from further scientific review. However, if you require any further information, please do not hesitate to contact me.

Yours sincerely

Jaime Delgadillo

The
University
Of
Sheffield.

Department Of Psychology.
Clinical Psychology Unit.

Doctor of Clinical Psychology (DClin Psy) Programme
Clinical supervision training and NHS research
:training & consultancy.

Clinical Psychology Unit
Department of Psychology
University of Sheffield
Cathedral Court
Sheffield
S1 2LT
UK

Address: Chris Gaskell
Clinical Psychologist
Department of Psychology
Cathedral Court

Clinical Psychology Unit
Department of Psychology
Cathedral Court
Sheffield

Date: 27.11.2020

Telephone: 0114 22 26650
Email: a.sinha@sheffield.ac.uk

Project title: **Effectiveness of Differing Psychotherapies Offered in a Specialist Psychotherapy Service – a benchmarking study.**

URMS number: ***164633***

Dear Chris,

The University has reviewed the following documents:

1. A University approved URMS costing record;
2. Confirmation of independent scientific approval;
3. Confirmation of independent ethics approval.

All the above documents are in place. Therefore, the University now **confirms** that it is the project's research governance sponsor and, as research governance sponsor, **authorises** the project to commence any non-NHS research activities. Please note that NHS R&D/HRA approval will be required before the commencement of any activities which do involve the NHS.

You are expected to deliver the research project in accordance with the University's policies and procedures, which includes the University's Good Research & Innovation Practices Policy: www.shef.ac.uk/ris/other/gov-ethics/grippolicy, Ethics Policy: www.sheffield.ac.uk/ris/other/gov-ethics/ethicspolicy and Data Protection Policies: www.shef.ac.uk/cics/records

Your Supervisor, with your support and input, is responsible for providing up-to-date study documentation to all relevant sites, and for monitoring the project on an ongoing basis. Your Head of Department is responsible for independently monitoring the project as appropriate. The project may be audited during or after its lifetime by the University. The monitoring responsibilities are listed in **Annex 1**.

Yours sincerely

Jaime Delgadillo

Dr Jaime Delgadillo

Director of Research Training, Clinical Psychology Unit

cc. Academic Supervisor/s Jaime Delgadillo & Steve Kellett
Head of Department/School: Elizabeth Milne

Authorisation When Using This Organisation Information Document as An Agreement

(when used as an Agreement, the Participating NHS Organisation is a "Party" to the Agreement and the Sponsor is a "Party" to the Agreement – collectively the "Parties").

Authorisation on behalf of Participating NHS / HSC Organisation
It is not intended that this confirmation requires wet-ink signatures, or a passing of hard copies between the Sponsor and participating NHS / HSC organisation. Instead, Sponsors are expected to accept confirmation by email from an individual empowered by the Participating NHS / HSC Organisation to agree to the commencement of research (including any budgetary responsibility, where the study involves the transfer of funds).

^ Authorised on behalf of Participating NHS / HSC Organisation by:

Name	Yiwei Harland
Job Title	Research Governance Officer
Organisation Name	Sheffield Health & Social Care NHS FT
Date	19 March 2021

Appendix C. Opt-out poster for patients attending the specialist psychotherapy service. Poster advised and approved by CAG.

NHS
Sheffield Health
and Social Care
NHS Foundation Trust

Your Questionnaires (OQ-45)

Improving Psychological Care

The questionnaires you complete helps the service to see how effective psychological therapy is. We are planning to put this information (for all our patients) together so that we can look for patterns. We hope that this will help us to work out:

- Which therapy is best for which problems?
- How many appointments are needed for therapy to work?
- When therapy works, how long do the effects last for?

Knowing the answers to these questions would help the service to plan therapy with future people accessing therapy.

We now have enough information to begin looking at patterns. We have teamed up with researchers from the University of Sheffield who are specially trained to analyse such information.

One researcher will be required to access mental health notes for everyone who has accessed therapy here over the past 10 years.
No information that can identify you (i.e., names, NHS number etc.) will be taken from the service. We therefore can ensure that your anonymity will be maintained.

If you would prefer to opt-out of this evaluation programme, then please e-mail Chris.Gaskell@shsc.nhs.uk.

Proud to care in Sheffield

Appendix D. Further details regarding study analyses.

Meta-analytic benchmark

Benchmarks were calculated for the total OQ-45 score, as well as each of the subscales. Effect-sizes (including 95% confidence intervals [CI]) for SPS (total sample and each therapy modality separately) and the described benchmarks were plotted to a forest-plot for visual comparison. Inspection of confidence interval overlap was performed to enable pairwise comparisons of effectiveness equivalence between SPS and available benchmarks. Three equivalence categories were employed: (i) *equivalent effectiveness* was when the benchmark effect-size was contained by SPS CIs; (ii) *superior effectiveness* was when the SPS CI region exceeded the SPS effect-size; finally, (iii) *inferior effectiveness* was when the SPS CI region fell beneath the SPS effect-size.

Maximum likelihood estimation

All models developed in this study used maximum likelihood estimation in order to allow for the more conservative assumption that data missingness represents data missing at random (MAR, Mallinckrodt et al., 2001) as opposed to data being missing completely at random (MCAR).

Power calculation

There is no definitive rule for the number of (or ratio of) observations:cases required to fit growth curve models. Sample size ratios which have been suggested (e.g. Raudenbush & Bryk, 2002) tend to represent more traditional hierarchical linear models (i.e. not longitudinal) which generally have fewer level-1 observations per level-2 group (Kwok et al., 2008). Maas & Hox (2005) suggest that 30 level-2 units (cases) are sufficient for reasonably unbiased regression estimates however this does not specify the number of level-1 observations. The only consensus for power calculation in growth curves analysis is that ‘more is better.’ Studies applying growth curve techniques to psychological therapy longitudinal data have used as few as 13 level-2 units (i.e. patients, Kolly et al., 2015). We specified a-priori that 150 cases (i.e. level-2) with at least three observations each would be sufficient for the current study.

Appendix E. Comparison of of different covariance structures.

Table 5

Fixed effects and goodness-of-fit statistics for covariance structures: OQ-45 total score and symptom distress

	Total Score				Symptom Distress			
	1.Stand	1.Toep	1.Comsym	1.Auto1	2.Stand	2.Toep	2.Comsym	2.Auto1
OQ Score (Intercept)	103.960*	100.992*	100.832*	99.975*	65.056*	62.890*	62.886*	62.341*
	(1.337)	(1.435)	(1.335)	(1.385)	(0.874)	(0.920)	(0.863)	(0.890)
Contacts (Slope)	-0.383*	-0.150*	-0.148*	-0.119*	-0.255*	-0.089*	-0.088*	-0.073*
	(0.041)	(0.025)	(0.013)	(0.027)	(0.028)	(0.016)	(0.008)	(0.017)
AIC	17366.8	17257.6	17522.8	17394.5	15327.4	15234.5	15535.2	15345.7
BIC	17400.7	17291.5	17545.4	17417.1	15361.3	15268.4	15557.8	15368.3
LogLiklihood	-8677.411	-8622.795	-8757.386	-8693.255	-7657.714	-7611.228	-7763.600	-7668.829

* p < 0.05

Table 6

Fixed effects and goodness-of-fit statistics for covariance structures: Social role and interpersonal relationships

	Social Role				Inter-Personal			
	3.Stand	3.Toep	3.Comsym	3.Auto1	4.Stand	4.Toep	4.Comsym	4.Auto1
OQ Score (Intercept)	16.347*	15.865*	15.851*	15.602*	22.726*	22.288*	22.342*	22.084*
	(0.309)	(0.320)	(0.293)	(0.295)	(0.395)	(0.413)	(0.387)	(0.396)
Contacts (Slope)	-0.063*	-0.029*	-0.029*	-0.020*	-0.059*	-0.030*	-0.032*	-0.023*
	(0.009)	(0.006)	(0.004)	(0.006)	(0.009)	(0.007)	(0.004)	(0.008)
AIC	12168.7	12176.0	12231.7	12382.4	12308.2	12186.2	12351.3	12334.5
BIC	12202.6	12209.9	12254.3	12405.0	12342.1	12220.1	12373.9	12357.1
LogLiklihood	-6078.362	-6081.990	-6111.844	-6187.195	-6148.124	-6087.124	-6171.635	-6163.244

* p < 0.05

Appendix F. Comparison of dfferent time trends.

Table 7

Fixed effects and goodness of fit statistics to determine most suitable time trend: Total score and symptom distress.

	Total Score				Symptom Distress			
	1.Log	1.Linear	1.Quadratic	1.Cubic	2.Log	2.Linear	2.Quadratic	2.Cubic
Intercept	105.709*	100.992*	102.880*	104.180*	65.563*	62.890*	63.918*	64.725*
	(1.642)	(1.435)	(1.480)	(1.497)	(1.048)	(0.920)	(0.951)	(0.962)
Rate of Change								
Linear/Log	-3.301*	-0.150*	-0.302*	-0.496*	-1.915*	-0.089*	-0.172*	-0.293*
	(0.401)	(0.025)	(0.039)	(0.057)	(0.252)	(0.016)	(0.025)	(0.036)
Quadratic			0.001*	0.005*			0.001*	0.003*
			(0.000)	(0.001)			(0.000)	(0.000)
Cubic				-0.000*				-0.000*
				(0.000)				(0.000)
AIC	17228.0	17257.6	17235.2	17216.0	15209.8	15234.5	15218.3	15199.3
BIC	17261.9	17291.5	17274.8	17261.2	15243.7	15268.4	15257.9	15244.5
LogLiklihood	-8607.98	-8622.79	-8610.62	-8599.99	-7598.92	-7611.23	-7602.16	-7591.66
Ratio Test			0.022	<0.001			0.011	<0.001

* $p < 0.05$

Table 8

Fixed effects and goodness of fit statistics to determine most suitable time trend: Social role and interpersonal.

	Social Role				Inter-Personal			
	3.Log	3.Linear	3.Quadratic	3.Cubic	4.Log	4.Linear	4.Quadratic	4.Cubic
Intercept	17.668*	16.347*	16.519*	16.753*	23.071*	22.288*	22.716*	22.973*
	(0.407)	(0.309)	(0.314)	(0.325)	(0.480)	(0.413)	(0.427)	(0.434)
Rate of Change								
Linear/Log	-0.949*	-0.063*	-0.078*	-0.106*	-0.599*	-0.030*	-0.065*	-0.103*
	(0.129)	(0.009)	(0.010)	(0.014)	(0.120)	(0.007)	(0.012)	(0.017)
Quadratic			0.000*	0.001*			0.000*	0.001*
			(0.000)	(0.000)			(0.000)	(0.000)
Cubic				-0.000*				-0.000*
				(0.000)				(0.000)
AIC	12110.7	12168.7	12159.8	12154.4	12178.4	12186.2	12173.7	12166.7
BIC	12144.6	12202.6	12199.4	12199.6	12212.3	12220.1	12213.2	12211.9
ICC	0.7	0.7	0.7	0.7				
LogLiklihood	-6049.34	-6078.36	-6072.91	-6069.21	-6083.19	-6087.12	-6079.85	-6075.34
Ratio Test			<0.001	<0.001			0.010	<0.001

* $p < 0.05$

Appendix G. Comparison of unconditional models

Table 9

Fixed effects and goodness of fit statistics to determine most unconditional growth model.

	Total Score			Symptom Distress			Social Role			Inter-Personal		
	1.MM	1.RI	1.RS	2.MM	2.RI	2.RS	3.MM	3.RI	3.RS	4.MM	4.RI	4.RS
Fixed Effects												
Intercept	97.252*	100.832*	100.992*	60.745*	62.886*	62.890*	15.142*	15.851*	16.347*	21.569*	22.342*	22.288*
	(1.283)	(1.335)	(1.435)	(0.831)	(0.863)	(0.920)	(0.276)	(0.293)	(0.309)	(0.374)	(0.387)	(0.413)
Rate of Change												
Linear		-0.148*	-0.150*		-0.088*	-0.089*		-0.029*	-0.063*		-0.032*	-0.030*
		(0.013)	(0.025)		(0.008)	(0.016)		(0.004)	(0.009)		(0.004)	(0.007)
Goodness of Fit												
AIC	17645.4	17522.8	17257.6	15648.9	15535.2	15234.5	12287.9	12231.7	12168.7	12418.7	12351.3	12186.2
LogLiklihood	-8819.69	-8757.39	-8622.79	-7821.45	-7763.6	-7611.23	-6140.95	-6111.84	-6078.36	-6206.35	-6171.63	-6087.12
p-value		<0.001	<0.001		<0.001	<0.001		<0.001	<0.001		<0.001	<0.001
BIC	17662.3	17545.4	17291.5	15665.8	15557.8	15268.4	12304.8	12254.3	12202.6	12435.7	12373.9	12220.1

* $p < 0.05$

MM = Means Model, RI = Random Intercepts Model, RS = Random Slopes and Intercepts Model

Appendix H. Study charachteritics for OQ-45 outcomes studies included in the random effects meta-analytic benchmark

Table 10

Study characteristics for OQ-45 outcome studies identified through the systematic search and included in the random-effect meta-analysis.

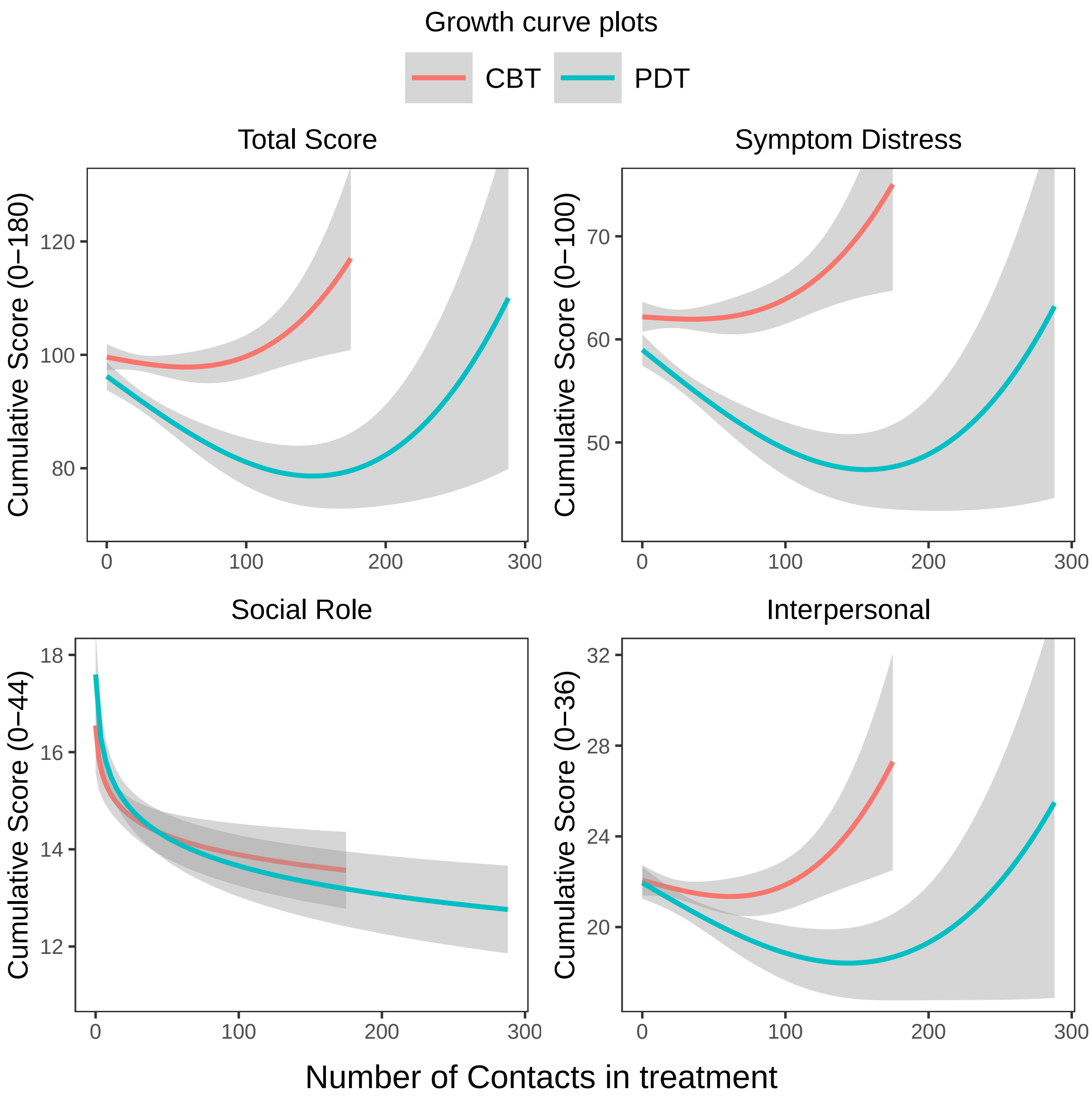
Study	Country	Service	Analysis	Therapy	Sessions	Problem
Baldwin (2009)	USA	University counselling centre	Modified ITT	Not reported	6.46	Various
Bradshaw (2009)	USA	Cmhc	Completers	Psychodynamic	26.00	Moderate-to-severe mental health problems
Galili-Weinstock (2018)	Israel	University outpatient clinic	Completers	Psychodynamic	22.10	Various
Goldberg (2016)	USA	University counselling centre	Modified ITT	Various	8.09	NA
Haugen (2017)	USA	World trade centre responder clinic	ITT	Integrated	26.06	Trauma
Kaplinski (2014)	USA	University training clinic	Modified ITT	Not stated	10.73	Various
Kolly (2015)	Switzerlanc	University hospital	Assume ITT	Cbt or psychodynamic	NA	Personality disorder
Kramer (2013)	Switzerland	Diverse outpatient settings	Completers	Various	7.80	Moderate-to-severe mental health problems

Lunnen (2008)	USA	Cmhcs	Modified ITT	Various	5.75	Moderate-to-severe mental health problems
Prout (2016)	USA	University training clinic	Modified ITT	Various	8.17	Various
Ronnestad (2019)	Norway	Private practice	Completers	Various	64.90	Check
Roseborough (2006)	USA	Outpatient mental health centre	Assume ITT	Pdt	64.00	Various
Wiseman (2014)	Israel	University counselling centre	Completers	Psychodynamic psychotherapy	NA	Various

Appendix I. Growth curve plots for conditional modles (i.e. treatment modality).

Figure 5

Growth curves for conditional models, seperated by treatment modality. Coloured lines represent growth curves. Grey shaded regions represent 95% confidence intervals. Trends line types represent (i) cubic for total score, interpersonal and symptom distress; and (ii) log-linear for social role.



References

Abbass, A. A., Joffres, M. R., & Ogrodniczuk, J. S. (2008). A naturalistic study of intensive short-term dynamic psychotherapy trial therapy. *Brief Treatment and Crisis Intervention*, *8*(2), 164–170. `https://doi.org/10.1093/brief-treatment/mhn001`

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose-effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology*, *77*(2), 203–211. `https://doi.org/10.1037/a0015235`

Blainey, S. H., Rumball, F., Mercer, L., Evans, L. J., & Beck, A. (2017). An evaluation of the effectiveness of psychological therapy in reducing general psychological distress for adults with autism spectrum conditions and comorbid mental health problems. *Clinical Psychology & Psychotherapy*, *24*(6), 474–484. `https://doi.org/10.1002/cpp.2108`

Bone, C., Delgadillo, J., & Barkham, M. (2021). A systematic review and meta-analysis of the good-enough level (GEL) literature. *Journal of Counseling Psychology*, *68*(2), 219–231. `https://doi.org/10.1037/cou0000521`

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`

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

de Jong, K., Conijn, J. M., Gallagher, R. A. V., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, *85*, 102002. https://doi.org/10.1016/j.cpr.2021.102002

Delgadillo, J., Asaria, M., Ali, S., & Gilbody, S. (2016). On poverty, politics and psychology: The socioeconomic gradient of mental healthcare utilisation and outcomes. *The British Journal of Psychiatry*, *209*(5), 429–430. https://doi.org/10.1192/bjp.bp.115.171017

Delgadillo, J., & Gonzalez Salas Duhne, P. (2020). Targeted prescription of cognitivebehavioral therapy versus person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology*, *88*(1), 14–24. https://doi.org/10.1037/ccp0000476

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

Department of Health. (2004). *Organising and delivering psychological therapies*. Department of Health.

Doerfler, L. A., Addis, M. E., & Moran, P. W. (2002). Evaluating mental health outcomes in an inpatient setting: Convergent and divergent validity of the OQ-45 and BASIS-32. *The Journal of Behavioral Health Services & Research*, *29*(4), 10.

Douglas, A., Ablett-Tate, N., & Chadd, N. (2016). Dynamic interpersonal therapy in an NHS tertiary level specialist psychotherapy service. *Psychoanalytic Psychotherapy*, *30*(3), 223–239. https://doi.org/10.1080/02668734.2016.1198415

Finegan, M., Firth, N., Wojnarowski, C., & Delgadillo, J. (2018). Associations between socioeconomic status and psychological therapy outcomes: A systematic review and meta-analysis. *Depression and Anxiety*, *35*(6), 560–573. https://doi.org/10.1002/da.22765

Firth, N., Saxon, D., Stiles, W. B., & Barkham, M. (2020). Therapist effects vary significantly across psychological treatment care sectors. *Clinical Psychology & Psychotherapy*, *27*(5), 770–778. https://doi.org/10.1002/cpp.2461

Francis, V. M., Rajan, P., & Turner, N. (1990). British community norms for the Brief Symptom Inventory. *British Journal of Clinical Psychology*, *29*(1), 115–116. https://doi.org/10.1111/j.2044-8260.1990.tb00857.x

Goldberg, S. B., Miller, S. D., Nielsen, S. L., Rousmaniere, T., Whipple, J., Hoyt, W. T., & Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology*, *63*(1), 1–11. https://doi.org/10.1037/cou0000131

Gordon, M., & Lumley, T. (2020). *Forestplot: Advanced forest plot using 'grid' graphics*. https://CRAN.R-project.org/package=forestplot

Hallam, C., Simmonds-Buckley, M., Kellett, S., Greenhill, B., & Jones, A. (2021). The acceptability, effectiveness, and durability of cognitive analytic therapy: Systematic review and meta-analysis. *Psychology and Psychotherapy*, *94*(1), 8–35. https://doi.org/10.1111/papt.12286

Hansen, N. B., & Lambert, M. J. (2002). An evaluation of the doseresponse relationship in naturalistic treatment settings using survival analysis. *Mental Health Services Research*, *5*(1), 1–12.

Heins, M. J., Knoop, H., Lobbestael, J., & Bleijenberg, G. (2011). Childhood maltreatment and the response to cognitive behavior therapy for chronic fatigue syndrome. *Journal of Psychosomatic Research*, *71*(6), 404–410. https://doi.org/10.1016/j.jpsychores

.2011.05.005

Howard, K., Kopta, S., Krause, M., & Orlinsky, D. (1986). The doseeffect relationship in psychotherapy. *American Psychologist*, *41*(2), 159–164.

Jacobson, N. S., & Truax, P. (1992). Clinical significance : A statistical approach to defining meaningful change in psychotherapy research. In *Methodological issues & strategies in clinical research* (pp. 631–648). American Psychological Association. https://doi.org/10.1037/10109-042

Johansson, R., Town, J. M., & Abbass, A. (2014). Davanloo's intensive short-term dynamic psychotherapy in a tertiary psychotherapy service: Overall effectiveness and association between unlocking the unconscious and outcome. *PeerJ*, *2*, Article e548. https://doi.org/10.7717/peerj.548

Johns, R. G., Barkham, M., Kellett, S., & Saxon, D. (2019). A systematic review of therapist effects: A critical narrative update and refinement to Baldwin and Imel's (2013) review. *Clinical Psychology Review*, *67*, 78–93. https://doi.org/10.1016/j.cpr.2018.08.004

Jong, K. de, Nugter, M. A., Polak, M. G., Wagenborg, J. E. A., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: A cross-cultural validation. *Clinical Psychology & Psychotherapy*, *14*(4), 288–301. https://doi.org/10.1002/cpp.529

Kolly, S., Kramer, U., Maillard, P., Charbon, P., Droz, J., Fresard, E., Berney, S., & Despland, J.-N. (2015). Psychotherapy for personality disorders in a natural setting: A pilot study over two years of treatment. *The Journal of Nervous and Mental Disease*, *203*(9), 735–738. https://doi.org/10.1097/NMD.0000000000000356

Kopta, S. M., Howard, K. I., Lowry, J. L., & Beutler, L. E. (1994). Patterns of symptomatic recovery in psychotherapy. *Journal of Consulting and Clinical Psychology*, *62*(5), 1009–1016. https://doi.org/10.1037/0022-006X.62.5.1009

Kwok, O.-M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures. *Rehabilitation Psychology*, *53*(3), 370–386. https://doi.org/10.1037/a0012765

Lambert, M. J. (2004). *Administration and scoring manual for the OQ-45.2 (Outcome questionnaire)*. OQ Measures, LLC.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the outcome questionnaire. *Clinical Psychology & Psychotherapy*, *3*(4), 249–258. https://doi.org/10.1002/(SICI)1099-0879(199612)3:4%3C249::AID-CPP106%3E3.0.CO;2-S

Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The Outcome Questionnaire-45. In M. E. Maurish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment: Instruments for adults* (Vol. 3, pp. 191–234). Lawrence Erlbaum Associates Publishers.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology*, *69*(2), 159–172. https://doi.org/10.1037/0022-006X.69.2.159

Lilliengren, P., Cooper, A., Town, J. M., Kisely, S., & Abbass, A. (2020). Clinical- and cost-effectiveness of intensive short-term dynamic psychotherapy for chronic pain in a tertiary psychotherapy service. *Australasian Psychiatry*, *28*(4), 414–417. https://doi.org/10.1177/1039856220901478

Lunnen, K. M., Ogles, B. M., & Pappas, L. N. (2008). A multiperspective comparison of satisfaction, symptomatic change, perceived change, and end-point functioning. *Professional Psychology: Research and Practice*, *39*(2), 145–152. https://doi.org/10.1037/0735-7028.39.2.145

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-1881.1.3.86

Mallinckrodt, C. H., Clark, W. S., & David, S. R. (2001). Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics*, *11*(1-2), 9–21. https://doi.org/10.1081/BIP-100104194

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S. J., & Kircher, J. C. (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. *Journal of Consulting and Clinical Psychology*, *76*(1), 116–124. https://doi.org/10.1037/0022-006X.76.1.116

Morley, S., & Dowzer, C. (2014). *Manual for the Leeds Reliable Change Indicator: Simple Excel(tm) applications for the analysis of individual patient and group data.*

Nowoweiski, D., Abbass, A., Town, J., Keshen, A., & Kisely, S. (2020). An observational study of the treatment and cost effectiveness of intensive short-term dynamic psychotherapy on a cohort of eating disorder patients. *Journal of Psychiatry and Behavioral Sciences*, *3*(1), Article 1030.

Paley, G., Cahill, J., Barkham, M., Shapiro, D., Jones, J., Patrick, S., & Reid, E. (2008). The effectiveness of psychodynamic-interpersonal therapy (PIT) in routine clinical practice: A benchmarking comparison. *Psychology and Psychotherapy: Theory, Research and Practice*, *81*(2), 157–175. https://doi.org/10.1348/147608307X270889

Pinheiro, J., Bates, D., & R-core. (2020). *Nlme: Linear and nonlinear mixed effects models*. https://svn.r-project.org/R-packages/trunk/nlme/

Probst, T., Kleinstäuber, M., Lambert, M. J., Tritt, K., Pieh, C., Loew, T. H., Dahlbender, R. W., & Delgadillo, J. (2020). Why are some cases not on track? An item analysis of the Assessment for Signal Cases during inpatient psychotherapy. *Clinical Psychology & Psychotherapy*, *27*(4), 559–566. https://doi.org/10.1002/cpp.2441

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Robinson, L., Delgadillo, J., & Kellett, S. (2020). The dose-response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy Research*, *30*(1), 79–96. https://doi.org/10.1080/10503307.2019.1566676

Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry*, *6*(3), 412–415. https://doi.org/10.1111/j.1939-0025.1936.tb05248.x

Ryan, C. (2007). British outpatient norms for the Brief Symptom Inventory. *Psychology and Psychotherapy: Theory, Research and Practice*, *80*(2), 183–191. https://doi.org/10.1348/147608306X111165

Ryle, A., & Kerr, Ian. B. (2020). The main features of CAT. In *Introducing cognitive analytic therapy* (2nd ed, pp. 9–29). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119375210.ch2

Schilling, V. N. L. S., Zimmermann, D., Rubel, J. A., Boyle, K. S., & Lutz, W. (2020). Why do patients go off track? Examining potential influencing factors for being at risk of psychotherapy treatment failure. *Quality of Life Research*. https://doi.org/10.1007/s11136-020-02664-6

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. https://doi.org/10.1037/0003-066X.32.9.752

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

Stiles, W. B., Barkham, M., Twigg, E., Mellor-Clark, J., & Cooper, M. (2006). Effectiveness of cognitive-behavioural, person-centred and psychodynamic therapies as practised in UK National Health Service settings. *Psychological Medicine*, *36*(4), 555–566. https://doi.org/10.1017/S0033291706007136

Taylor, D., Carlyle, J., McPherson, S., Rost, F., Thomas, R., & Fonagy, P. (2012). Tavistock Adult Depression Study (TADS): A randomised controlled trial of psychoanalytic psychotherapy for treatment-resistant/treatment-refractory forms of depression. *BMC Psychiatry*, *12*(1), 60. https://doi.org/10.1186/1471-244X-12-60

van der Kolk, B. A. (1989). The compulsion to repeat the trauma: Re-enactment, revictimization, and masochism. *Psychiatric Clinics of North America*, *12*(2), 389–411. https://doi.org/10.1016/S0193-953X(18)30439-8

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment*, *74*(2), 242–261. https://doi.org/10.1207/S15327752JPA7402_6

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://www.jstatsoft.org/v36/i03/

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

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially, "all must have prizes". *Psychological Bulletin*, *122*(3), 203–215. https://doi.org/10.1037/0033-2909.122.3.203

Warden, S., Ricketts, T., Saxon, D., Houghton, S., St. Ledger, S., Curran, J., & Fitzgerald, G. (2008). When to offer cognitive behavioural or psychoanalytic psychotherapy in an integrated psychotherapy service: Are everyday allocation decisions theoretically congruent? *Counselling and Psychotherapy Research*, *8*(2), 102–109. https://doi.org/10.1080/14733140801972604

Wennberg, P., Philips, B., & Jong, K. de. (2010). The Swedish version of the Outcome Questionnaire (OQ-45): Reliability and factor structure in a substance abuse sample. *Psychology and Psychotherapy: Theory, Research and Practice*, *83*(3), 325–329. https://doi.org/10.1348/147608309X478715

White, M. M., Lambert, M. J., Ogles, B. M., Mclaughlin, S. B., Bailey, R. J., & Tingey, K. M. (2015). Using the assessment for signal clients as a feedback tool for reducing treatment failure. *Psychotherapy Research*, *25*(6), 724–734. https://doi.org/10

.1080/10503307.2015.1009862

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Worm-Smeitink, M., Nikolaus, S., Goldsmith, K., Wiborg, J., Ali, S., Knoop, H., & Chalder, T. (2016). Cognitive behaviour therapy for chronic fatigue syndrome: Differences in treatment outcome between a tertiary treatment centre in the United Kingdom and the Netherlands. *Journal of Psychosomatic Research*, *87*, 43–49. https://doi.org/10.1016/j.jpsychores.2016.06.006