British Journal of Clinical Psychology (2010), 49, 421–453

The British Psychological Society

www.bpsjournals.co.uk

Systematic review of practice-based research on psychological therapies in routine clinic settings

Jane Cahill[1]*, Michael Barkham[2] and William B. Stiles[3]
[1]University of Leeds, UK
[2]University of Sheffield, UK
[3]Miami University, Oxford, Ohio, USA

Purpose. To review the published material on practice-based research and to compare results with benchmarks derived from efficacy studies.

Methods. Electronic and manual searches were carried out up to and including 2008. Studies were screened for content relevance and selected according to specified inclusion criteria. Data were extracted from all studies that met criteria and were quality assessed using an adapted version of a checklist designed for the appraisal of both randomized and non-randomized studies of health care interventions. Studies were synthesized according to (1) the type of problem being treated and (2) study design using descriptive and meta-analytic methods where appropriate.

Results. Psychological treatment conducted in routine clinic settings is effective for a range of client problems, particularly common mental health problems (uncontrolled effect size = 1.29; 95% CI = 1.26–1.33, $N = 10,842$). When benchmarked against data from efficacy studies, practice-based studies yielded effect sizes that fell short of the selected benchmark. In contrast, the practice-based studies achieved the benchmark for percentage of clients meeting a stringent criterion for recovery.

Conclusions. Clients receiving treatment as normally delivered within routine practice report significant relief of symptoms. However, the result of comparisons with efficacy benchmarks is dependent on the outcome index used. Notwithstanding this, substantive factors are also likely to contribute. Therefore, in addition to attending to methodological issues, further work is required to understand the relative contribution of these factors.

A strategic review of psychotherapy services in the UK (Department of Health, 1996) as well as critical scholarly reviews (e.g., Roth & Parry, 1997; Shadish *et al.*, 1997; Shadish, Navarro, Matt, & Phillips, 2000) have emphasized the need for evidence

This paper, which is jointly authored by an editor of the journal, was entirely dealt with by two other members of the Editorial Board.
** Correspondence should be addressed to Jane Cahill, School of Healthcare, Baines Wing, University of Leeds, Leeds LS2 9JT, UK (e-mail: j.l.cahill@leeds.ac.uk).*

DOI:10.1348/014466509X470789

from routine clinical settings to complement the *efficacy* evidence obtained from formal randomized trials. Large practice-based studies (e.g., Stiles, Barkham, Mellor-Clark, & Connell, 2008a) have generated lively debate (e.g., Clark, Fairburn, & Wessly, 2008; Stiles, Barkham, Mellor-Clark, & Connell, 2008b), and the UK's Improving Access to Psychological Therapies programme (Layard, 2006) is making routine practice-based data sets within the National Health Service (NHS) increasingly available. Mindful of the growing interest and the developing evidence-base, we reviewed recent practice-based evidence for psychological therapies.

Shadish and colleagues asked whether psychotherapy outcome studies of differing degrees of clinical representativeness yield differential effectiveness. In a review of 90 psychotherapy outcome studies aggregated using random effect regression analyses, they found that effect sizes (ESs) of both more and less clinically representative studies were essentially comparable, even when potential methodological confounds were corrected (Shadish *et al.*, 1997, 2000). That is, differences in research design quality may not distort estimates of the overall effectiveness of psychological therapies.

From within the spectrum of studies that Shadish and colleagues delineated, we focused on the evidence from routine practice settings and further distinguished *practice-based evidence*, which measures what normally happens in such settings, from *effectiveness evidence*, which can impose research constraints, such as randomization, treatment protocols, manuals, and so forth. While both these approaches are distinct from efficacy research, the research design features are also sufficiently different from each other to warrant the use of separate categories. In comparison to effectiveness evidence, practice-based evidence addresses the next logical step in research on the transportability of efficacious treatments to routine clinical settings.

Hunsley and Lee (2007) reviewed the treatment effectiveness literature relating to the psychological therapies comprising 21 studies of adult disorders and concluded that improvement rates were comparable with those from efficacy benchmarks. To complement this work, we reviewed practice-based evidence reported in the period of 1990–2008. We evaluated results of the practice-based studies against selected benchmarks from efficacy trials (cf. Barkham *et al.*, 2008; Eisen & Dickey, 1996; McEvoy & Nathan, 2007).

Method

We examined research on psychological therapy for adults conducted in routine service settings and published between January 1990 and December 2008.

Search strategy

Our initial strategy involved the search of four major databases: PsycInfo, Cinahl, Medline, and Embase. Search terms are listed in Appendix A. However, a preliminary search of Medline and Embase indicated that these two medical databases, rooted in a paradigm of evidence-based practice, were returning hits that were not relevant to the scope of the present study, and we decided further searching of these databases would not be profitable. We therefore focused on PsycInfo and Cinahl. Within these two databases, three key journals (the journals that returned the most hits relevant to the subject area) were hand-searched: *Journal of Consulting and Clinical Psychology*, *Clinical Psychology: Science and Practice*, and *Journal of Mental Health*. Finally, the reference lists of the identified articles were scanned to pickup any further studies missed by the electronic search. This search strategy returned 12,304 references from the combined electronic databases.

Screening

Studies were first screened for content relevance by title and were included if they referred to any psychological therapy and/or issues around practice-based evidence and effectiveness research. Studies were excluded if they pertained exclusively to medical interventions, non-human data, or efficacy research. This screening yielded 546 references that dealt with issues of effectiveness research and practice-based evidence in routine settings. We next screened these 546 references for content relevance by consulting the abstracts. This left 283 references to which we applied specific inclusion and exclusion criteria listed below. A CONSORT diagram is presented in Figure 1.

Selection

The first two authors independently reviewed the 283 abstracts using the criteria described below. Full text copies of the articles were consulted as needed. Disagreements were resolved by discussion. This procedure identified 31 eligible studies to be included in the quality review.

Inclusion criteria

We selected all studies published in English that examined or reported the effect of a therapy for adult clients conducted by qualified staff in routine clinic settings and assessed using outcome measures relevant to the problem(s) being treated.

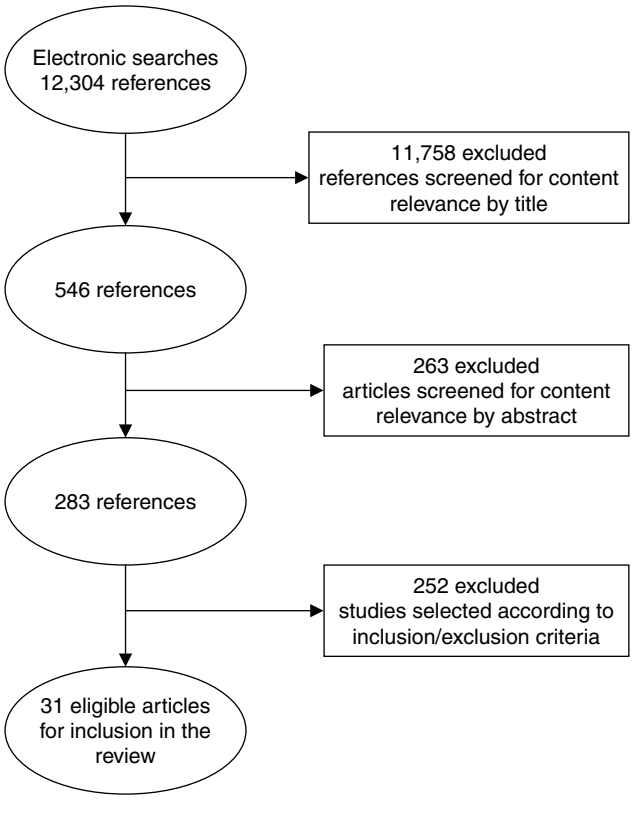

Figure 1. CONSORT diagram of electronic search strategy.

These included pre–post-follow-up designs, aggregated data from multiple service settings, and measure development studies that reported outcome data.

Exclusion criteria

Studies were excluded on the following grounds: (1) child/adolescent populations; (2) studies in which psychological outcomes were not measured or studies using a non-psychological intervention; (3) randomized controlled trials and controlled clinical trials conducted in routine clinic settings; (4) studies using manualized or protocol interventions/treatments; (5) studies relating to self-help/computer treatments: these models constitute a qualitatively different therapeutic model that is not being evaluated in this study; (6) studies dealing with carers of people with mental illness; and (7) reviews/meta-analyses or other non primary research.

Treatment of duplicate data

In instances where the same dataset had been used more than once, duplicate studies were removed and the study selected for inclusion was the one that represented the most recent use of the data. Where the same dataset had been used for different purposes, the article selected for inclusion was the one that most closely matched practice-based research on psychological therapies (as specified in the inclusion criteria).

Data extraction

Data were extracted by the first author from full text copies of the 31 studies that met the above criteria and organized by the problem being treated (see Table 1). ESs were reported, where possible, and were either extracted from the reports or computed from study data.

Quality assessment of studies

Methodological quality was assessed by an adapted version of a checklist designed for the appraisal of both randomized and non-randomized studies of health care interventions (Downs & Black, 1998). The checklist was adapted from the original to make it more applicable to the tenets of practice-based research and is presented in Appendix B. The checklist provided an overall score for quality as well as subscale scores. There were four dimensions: (1) *reporting* (11 items) assessed to what extent the information provided in the article was sufficient to allow the reader to make an unbiased assessment of the study findings; (2) *external validity* (11 items) assessed whether the findings from the study could be generalized to the wider population; (3) *internal reliability* (5 items) examined biases in the measurement of the treatment and outcome; and (4) *internal reliability confounding (selection) bias* (5 items) addressed issues relating to confounding factors and selection bias. The first author rated the full text copy of each article and an independent rater who was a doctorate-level practitioner rated a random subsample of six studies. The overall pairwise agreement level for these ratings was $\kappa = .59$. Landis and Koch (1977) proposed a classification of the agreement levels according to the value of Kappa whereby values above .81 are considered excellent, .61–.80 good, .41–.60 moderate, .21–.40 fair, .0–.20 poor, and below 0.0 very poor.

Table 1. Practice-based studies on psychological therapies in routine clinic settings

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Psychosis					
Farhall and Cotton (2002)	Single group Pre-post	*Clients*: 33 referrals with a diagnosis of psychosis and the presence of persisting hallucinations and/or delusions. *Therapists*: 11 psychologists employed in the area mental health services. *Setting*: Australia	Global Assessment of Functioning (GAF) Referral information and baseline observations record End of therapy record	Cognitive-behavioural therapy intervention	13
Common mental health problems					
Baker *et al.* (2002)	Non equivalent groups design Pre-post; 6 month, 1 year, and 2 year follow-up	*Clients*: Patients within the Dorset primary care counselling service. Counselled group: $N = 1,724$; mean age = 38.7 ($SD = 13.7$); 26% males. Waiting list group: $N = 367$; mean age = 38.0; $SD = 14.2$; 25% males. Patients did not have a long history of emotional problems or any formal psychopathology. Patients' distress was required to be acute and in response to some clearly definable situation. *Therapists*: Minimum of two years training and some practical experience required. No information on Ns. *Setting*: UK	Symptomatology: DSSI; Self esteem: RSE; Quality of life: QOL	Generic counselling in general practice. Counsellors usually offered eight session contracts with patients	26
Barkham *et al.* (2001)	Aggregated group Pre-post outcome data	*Clients*: 224 completer patients from 6 secondary care sites. Mean age 40.9 years ($SD = 15.2$). 61.6% female. *Therapists*: Practitioners from six secondary care sites. Mean number of patients per site was 37.3 ($SD = 47.9$; range = 11–34). *Setting*: UK	CORE-OM including subjective well-being Symptoms; functioning, risk	Clinical psychology, counselling, psychotherapy, psychiatry	24

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Booth, Cushway, and Newnes (1997)	Single group Pre-post	*Clients*: 51 clients receiving counselling at one of 15 general practices. 80% were female. *Therapists*: Six female counsellors who saw from 3 to 18 clients each. Counsellors had between 2 and 19 years of counselling experience and all had/were in the process of gaining accredited qualitificaitons in counselling. Five described approach as humanistic/eclectic and one as psychodynamic. *Setting*: UK	Quality of life scale (QOL)	Between 2 and 18 counselling sessions with mean of 7 sessions	17
Borkovec, Echemendia, Ragusea, and Ruiz (2001)	Aggregated group Pre-post	*Clients*: 220 clients. Two-thirds were women. Average age 39 years. Most were white/Caucasian. The mode common disorder was depression/anxiety, followed by mood disorders and anxiety disorders. *Therapists*: 77 therapists. Average age was 48 years, slight majority was male. Most were white/Caucasian. Therapists saw an average of 23 clients per week. Cognitive therapy made the largest contribution to therapist practice (33.9%), followed by psychodynamic (19.6%), behavioural (18.6%) family systems (10.5%), experiential (9.2%). Nearly three quarters had a PhD degree and 15% had master's degrees. *Setting*: USA	Compass measures: (a) Current life functioning assessing difficulties in six areas of self-management, work/school, relationships, family relationships, social relations, and health/ grooming (b) current symptoms assessing anxiety, mania/hypomania, depression, OCD, phobia and (c) ratings of therapist.	Psychotherapy delivered across a range of naturalistic settings	19

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Brown and Jones (2005)	Aggregated group Repeated measurement	*Clients*: 9,608 adults receiving psychotherapy in a manged care environment over a time period of approximately 4 years. *Therapists*: 7,000 clinicians *Setting*: USA	Depression symptoms OQ-30	Psychotherapy	15
Conway, Audin, Barkham, Mellor-Clark, and Russell (2003)	Single group Repeated assessment	*Clients*: Five consecutive group cohorts representing 30 patients with pre- and post-therapy data. 20 patients were female. Mean age 35 (SD – 12) *Therapists*: *Setting*: UK	Interpersonal Problems: IIP-32; Psychiatric Symptomatology: BSI; General Health Status Measure: SF-36	Brief time-intensive multi-modal group therapy. 12 weeks.	25
Evans, Connell, Barkham, Marshall, and Mellor-Clark (2003)	Aggregated group Pre-post	*Clients*: $N = 6,610$ clients from 33 NHS primary care services *Therapists* *Setting*: UK	CORE-OM including subjective well-being; symptoms; functioning, risk.	Primary care counselling services	23
Gibbard and Hanley (2008)	Single group Pre-post	*Clients*: 1,098 clients from a primary care counselling service in Lancashire. 72% were female and 95% were white. Mean age 41. Sample consisted of all those accepted into therapy over a 5-year period.	CORE-OM including subjective well-being; symptoms; functioning, risk.	Primary care counselling service offering person centered counselling.	22

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
		Therapists: 12 counsellors and 17 students whose therapeutic approach was person-centered counselling. Experience ranged from newly qualified to BACP accredited. *Setting*: UK		Initial contract of 6 sessions that could be extended to 12.	
Gilbert, Barkhamm, Richards, and Cameron (2005)	Aggregated group Pre-post	*Clients*: 2,205 patients for whom demographic data was available. 69% were women and 86.3% were from white European ethnic backgrounds. Mean age 36.2 years ($SD = 12.2$). *Therapists*: Primary care mental health practitioners *Setting*: UK	CORE system CORE-OM	Primary care mental health service: delivered in 54 GP practices and provides mental health assessment, brief psychological interventions (up to six sessions) and onward referral to specialist services as required.	25
Gordon and Graham (1996)	Single group Pre-post; 4 month follow-up	*Clients*: 95 clients receiving treatment from a counselling service operating over three separate general practices in one urban locality (Eastleigh, Hampshire) *Therapists*: Three counsellors employed within the psychology services of a health care trust. Each had qualifications and experience approaching B.A.C. accreditation levels. *Setting*: UK	Psychiatric symptoms: SCL-90-R; Anxiety and Depression: HADS; Problems: Effect on Life Scale (EOL); Problem-type: Problem Rating Scales; Satisfaction: questionnaires designed for GP and clients	Brief 6 session counselling.	21
Hansen and Lambert (2003);	Aggregated group Pre-post	*Clients*: 4,761 patients from standard treatment settings within the USA *Therapists*: No information *Setting*: USA	Psychological symptoms: OQ-45	Psychotherapy across a variety of settings.	15

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Hirsch, Jolley, and Williams (2000)	Single group Pre-post	*Clients*: All those who attended for at least two sessions at a direct referral clinical psychology service in London ($N = 98$). Anxiety and or depression was identified by the therapist as the primary problem. 67% female, mean age 67 years. *Therapists*: No information *Setting*: UK	Depression symptoms: BDI; Anxiety symptoms: BAI; Global Assessment Of Functioning: GAF	Cognitive-behavioural therapy. Number of sessions flexible but typically between 4–18 sessions over 2–12 months.	21
Kates, Crustolo, Farrar, and Nikolaou (2002)	Single group Pre-post	*Clients*: 3,550 referrals to a Canadian primary care counselling programme. 13% under the age of 18 years and 8% over the age of 65 years. Major problems were anxiety, depression, and family problems. *Therapists*: 41 counsellors occupying 23 full time positions. Each counsellor saw an average of 161 cases yearly. Counsellors included registered nurses (25%), social workers with masters degrees (50%) or bachelors degrees (15%), PhDs psychologists (2%), or other degrees (8%). *Setting*: Canada	GHQ Centre for Epidemiological Studies Depression Rating Scale		18

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Lucock *et al.* (2003)	Single group pre, post, and 6 month follow-up	*Clients*: 2,885 clients referred to a multiprofessional adult psychological therapies service. Service covers the communities of Wakefield and the Pontefract area made up of ex-mining communities. Relatively high unemployment and relatively low ethnic minority population. *Therapists*: Service comprises clinical psychologists, specialist psychotherapy team, counsellors, nurse therapists, cognitive behaviour therapists and an art therapist. *Setting*: UK	Depressive, anxiety, life/social functioning symptoms: CORE-OM; Depression symptoms: BDI; Interpersonal Problems: IIP-32	Range of therapies including cognitive-behavioural therapy, psychodynamic and psychoanalytic therapies, person centered approaches, and integrative therapies such as cognitive analytical therapy.	16
Mellor-Clark, Connell, Barkham, and Cummins (2001)	Aggregated group Pre-post	*Clients*: 2,042 clients from over 200 general practice settings offering counselling Mean age 38 years ($SD = 13.3$). 71% female *Therapists*: Approximately 150 counsellors providing counselling to clients in over 200 practice settings. *Setting*: UK	CORE-OM	Primary care counselling. Average 6 sessions	24
Minami *et al.* (2008)	Aggregated group	*Clients*: 12,743 patients receiving psychotherapy treatment for adult clinical depression in a managed care environment. 70% female. Mean age 40 years ($SD = 11.20$).	Depression symptoms OQ-30	Psychotherapy treatment: range of approaches. Treatment approach was not mandated or monitored.	21

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
		Therapists: 7,539 treatment providers providing treatment over a period of approximately 5 years. Providers were licensed in their jurisdictions and held master's degree or higher in: counselling/clinical psychology; marriage and family therapy; clinical social work; psychiatry; nursing. *Setting*: USA			
Nettleton *et al.* (2000)	Single group Pre-post	*Clients*: 131 patients in three GP practices in a rural setting. 72% were female. *Therapists*: One counsellor trained in the Gestalt approach and accredited by confederation of Scottish Counseling Agencies. *Setting*: UK	Adapted general well-being index	Counselling. Patients were offered 6 sessions with a further 6 if necessary. Later in the study number of sessions left to discretion of counsellor	19
Shepherd *et al.* (2005)	Aggregated group Pre-post	*Clients*: 3,687 adults with a wide range of problems from a culturally diverse inner London borough referred to primary care psychologists and counsellors. *Therapists*: Clinical and counselling psychologists, and part time counsellors in approximately 20 GP practices in culturally diverse inner London borough. Level of experience varies from newly qualified to highly experienced. *Setting*: UK	Global distress: CORE-OM	Number of different approaches including cognitive behavioural, psychodynamic, person-centered, systemic, integrative, and others. 6–8 sessions offered	23

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Smith, Sexton, and Bradley (2005)	Aggregated group Pre-post	*Clients*: 143 clients receiving individual counselling in either agency/organizational settings or private practice. 60.8% were female. Mean age 31.5 years. Primary presenting problems were couple conflict (25.2%), mood disorders (22.2%), situational based (16%), family conflict (13%), and anxiety (12%). Clients had significant symptom distress problems – comparable to that of distressed clinical populations. *Therapists*: 26 counsellors. 84% were female with an average of 13.7 years counselling experience. 56.7% were Caucasian. *Setting*: USA	OQ-45	Individual psychotherapy/counselling	20
Snell, Mallinc krodt, Hill, and Lambert (2001).	Single group Pre – 1 year follow-up	*Clients*: 199 clients who received counselling at a US university counselling centre. 63% were female. Mean age 26 years. 58% were single, 26% married; 7% lived with partner; 9% separated/divorced. *Therapists*: Counsellors including 11 PhD psychologists, 3 *post doc* fellows; 4 interns, 11 students, 7 non-psychologist staff members. No therapist saw more than nine clients. *Setting*: USA	Psychological symptoms OQ-45 CASPER: Computerized intake assessment of target complaints	Counselling offered at university counselling centre	24

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Stiles *et al.* (2003)	Single group Repeated assessment	*Clients*: 135 clients from 3 main bases of a large UK NHS Trust across the Wakefield Metropolitan District. Mean age 37.1 years ($SD = 11.0$). 70% were women. *Therapists*: 33 therapists including 11 clinical psychologists, 4 consultant clinical psychologists, 1 consultant psychotherapist, 11 counsellors, 2 counselling psychologists, and 4 nurse therapists. All had training in psychological therapy and at least 1 year post qualification experience. 23 (70%) were women. *Setting*: UK	Depression and Anxiety symptoms: CORE measures; BDI BAI Sudden gains: (SGs); Interpersonal Problems: IIP-32	A variety of treatment approaches including cognitive therapy, psychodynamic therapy, gestalt therapy, cognitive analytic therapy, transactional analysis, and other integrative therapies. Treatment duration was variable	23
Stiles, Barkham, Mellor-Clark, and Connell (2006)	Non equivalent groups Pre-post	*Clients*: 1,309 clients receiving counselling psychotherapy in primary care UK settings during a 3 year period. 70.7% were female. 2.8% were under 20 years, 19.6% aged 20–29, 29.7% aged 30–39, 24.7% aged 40–49, 15.2% aged 50–59, and 8.0% aged over 60. Over half the patients were taking prescribed psychotropic medications at the start of therapy. *Therapists*: 251 therapists each saw 1–29 patients. 15 of these therapists saw 10 or more of the patients. *Setting*: UK	Depression and Anxiety symptoms: CORE measures.	Three targeted approaches; CBT, cognitive behavioural and/or cognitive/ behavioural; PCT, person centered; PDT, psychodynamic and/or psychoanalytic.	24

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Stiles *et al.* (2008)	Non equivalent groups Pre-post	*Clients*: 5,613 clients receiving counselling psychotherapy in primary care UK settings during a 3 year period. 70.7% were female. Mean age = 40.7 (*SD* = 1.27). Over half the patients were taking prescribed psychotropic medications at the start of therapy. *Therapists*: 399 therapists who each saw 3–154 patients. 90 of the therapists saw 20 or more of these patients and 145 therapists saw 3 or fewer of these patients. *Setting*: UK	Depression and anxiety symptoms: CORE measures	Three targeted approaches; CBT, cognitive behavioural and/or cognitive/ behavioural; PCT, person centered; PDT, psychodynamic and/or psychoanalytic.	25
Wampold and Brown (2005)	Aggregated groups Repeated measurement	*Clients*: 6,146 patients receiving psychotherapy in a managed care environment over a time period of approximately 2 years. 72% were women, mean age 39.8 (*SD* = 10.8). Mean number of sessions was 10.63 (*SD* = 8.08). Primary diagnoses were depression, adjustment disorders, and anxiety. *Therapists*: 581 therapists. 72% were women. Mean age 51.5 years (*SD* = 14.87). Mean number of years experience 21.2 (*SD* = 7.65). Mean number of patients seen was 9.68 (*SD* = 5.61). *Setting*: USA	Psychological symptoms/quality of life; OQ-30	Psychotherapy	21
Westbrook and Kirk (2005)	Single group Pre-post	*Clients*: 1,276 patients who completed a course of therapy at an Oxford Adult Mental Health Psychology Department. Mean age 35.1 years (*SD* = 11.8). 68.3% were women. *Therapists*: No information *Setting*: UK	Depression: BDI; Anxiety: BAI; Individual target problems.	Specialized CBT service – on average clients received 13 sessions of treatment.	23

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Wolgast, Lambert, and Puschner (2003)	Single group Pre-post	*Clients*: 788 clients receiving individual psychotherapy at a University Counselling centre. 67% were female. Mean age 23 (range 18 to 64). Clients presented with wide range of problems, primarily mood or anxiety disorder. *Therapists*: 19 qualified and 24 student psychotherapists. Qualified therapists saw 51% of clients. Qualified and student therapists endorsed a variety of theoretical orientations including psychoanalytic (10%), psychodynamic (5%), integrative (55%), and eclectic (30%). *Setting*: Europe	Depression: CESD; General symptoms: GHQ	Individual counselling; couple and family treatment; group counselling	19
Panic disorder					
Hahlweg *et al.* (2001)	Single group pre-post	*Clients*: 416 PD with agoraphobia patients having pre-post data treated in 3 Clinical Psychology Clinics in Germany. 67% women, mean age 35.6 years ($SD = 8.9$). *Therapists*: 52 therapists of the Christoph-Dornier Foundation of Clinical Psychology. All were diploma psychologists (roughly equivalent to a master's degree. 72% were female with training in behaviour therapy. (therapists were inexperienced, 17 had medium experience and 26 were experienced. Fifteen therapists treated 1 or 2 patients; 16 treated 3 to 7 patients, 14 treated 8 to 14 patients; 8 treated 15 or more patients (range = 15–24). *Setting*: Europe	Anxiety symptoms: BAI; Agoraphobic and anxiety cognitions: ACQ; Bodily symptoms: BSQ; Avoidance behaviour: MI; MIA; MIB; Depression symptoms: BDI; SCL-90-R; Subjective rating of improvement.	High-density cognitive behavioural *in vivo* exposure lasting 4–10 days, during which patients are expected to confront the feared situations for several hours per day.	22

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Houghton and Saxon (2007)	Single group pre-post	*Clients*: 191 patients with an anxiety disorder: PD (with/without agoraphobia); specific phobias; social anxiety; generalized anxiety disorder; PTSD; OCD; hypochondria). 57% women. *Therapists*: two mental health nurses with post-registration training in CBT. Setting: UK	Psychological distress: CORE-OM; Anxiety/phobias: Fear Questionnaire; Satisfaction: Client Satisfaction Questionnaire	Brief psycho-educational course for anxiety disorders delivered in four 90 min classes. 24 patients in each class, classes held weekly, patients could attend classes in any order	16
Lincoln *et al.* (2003)	Single group pre – 6 week follow-up	*Clients*: 217 patients with social phobia undergoing treatment in the Christoph-Dornier-Foundation for Clinical Psychology (CDS). All patients had received a diagnosis of social phobia as primary disorder according to criteria listed in DSM-III. Mean age of sample was 33.7 years ($SD = 10.3$); 57% patients were male. *Therapists*: 57 diploma psychologists with training in behavioural therapy who were doctoral students of the CDS. Inexperienced therapists treated 22% patients, medium experience therapists treated 43% patients, high experience therapists treated 35% patients. *Setting*: Europe	General Impairment: SCL-90-R; questions on life satisfaction: FLZ M; social phobia: subscale interpersonal sensitivity of the SCL-90-R; SPS; SIAS; self rating of impairment due to social phobia; related fears and avoidance: BSQ; ACQ; Anxiety subscale of SCL-90-R; Depression: BDI; Depression subscale of SCL-90-R; Rating of improvement	Patients were treated with high density in vivo exposure supplemented by cognitive interventions.	23

Table 1. (*Continued*)

Author and year	Design	Population, setting, and sample size	Outcomes	Psychological therapy/treatment	Quality rating
Bulimia nervosa					
Thompson-Brenner and Westen (2005)	Practice network approach Random sample of US clinicians provided data on 145 patients	*Clients*: 145 patients all female. 86.2% patients met DSM-IV criteria for BN. Average age was 28.5 ($SD = 10.2$) *Therapists*: 86.7 sample were psychologists; 66.4 were female. 37.3 described their orientation as self reported CBT, 33.8% self-reported psychodynamic, and 28.9% purely eclectic or other. Mean years clinical experience was 16.1 ($SD = 7.9$). Average clinician reported caseload that included 10.4% ($SD = 15.5$) patients with eating disorders. Average hours worked per week was 28.6 ($SD = 12.9$). *Setting*: USA.	Global outcome (5 GAF items) and eating disorder (2 GAF items): Global Assessment of Functioning (GAF; 1994). Improvement on eating symptoms: inferential ratings; Global improvement: inferential ratings; Complete remission of binge eating and purging: objective yes/no rated assessments	Cognitive behavioural treatment: average 69 sessions Psychodynamic treatment: average 117 sessions.	23

Note. ACQ, Agoraphobic Cognition Questionnaire; BAI, Beck Anxiety Inventory; BDI, Beck Depression Inventory; BSI–GSI, Brief Symptom Inventory – General severity Index; BSI–PSI: Brief Symptom Inventory – Positive Symptom Total Index; BSQ, Body Sensations Questionnaire; CESD, Centre for Epidemiological Studies Depression Rating Scale; COMPASS, Core Battery Outcome Measures; CORE-OM, Clinical Outcomes in Routine Evaluation – Outcome Measure; DSSI, Bedford and Foulds Delusion-Symptoms – States Inventory; FACE, Functional Analysis of CORE Assessment; FQ, Fear Questionnaire; GAF, Global Assessment of Functioning; GHQ, General Health Questionnaire; HAD, Hospital Anxiety and Depression Scale; HRSD, Hamilton Rating Scale for Depression; IIP-32, Inventory of Interpersonal Problems; MI, Mobility Inventory; MIA, Mobility Inventory when alone; MIB, Mobility Inventory when with other person; OQ-30, Outcome questionnaire – 30.1; OQ-45: Outcome Questionnaire 45; PANAS, Positive and Negative Affect; RSE, Rosenberg Self Esteem Scale; SCL-90-R, Symptom Checklist-90-Revised; SF-36, Short-Form 36; SPS, Social Phobia Scale; SIAS, Social Interaction Anxiety Scale.

Data analysis

Studies were classified according to (1) the type of problem being treated and (2) study design. Due to the heterogeneity of methods and outcome measures represented in the sample, we used descriptive methods and/or meta-analytic methods as appropriate.

ESs were produced using Stata software (StataCorp, 2001) which was also used to produce selected graphs of the ESs. Where study data permitted, all continuous outcomes were translated to ESs by dividing the difference in mean values pre- and post-therapy by the pre-treatment standard deviation. Only matched pre- and post-treatment data were used to compute ESs. All ESs were taken directly from data and figures provided in the published papers of the study. ES data were available for 14 of the 31 studies in the quality review. Where studies used multiple outcomes, the analysis used the primary outcome measure if identified in the study or the authors selected the most widely used and validated measure. Stata was used to calculate an overall estimate of treatment effect with 95% confidence intervals (positive estimates representing results favouring post-treatment assessments). As for ESs, reliable and clinically significant improvement (RCSI) rates were taken from data and figures provided in the published papers of the study.

Meta-analyses

Two analyses were conducted in the current study: (1) meta-analysis of ESs and (2) summary of rates of RCSI. Of the 31 studies included in the quality review, only a portion supplied the necessary data to be included in the meta-analyses: 14 were able to be included in the ES meta-analysis and 14 in the RCSI meta-analysis. It should be noted that the 14 studies supplying ES and RCSI data were not the same 14 studies although a number supplied the data for calculation of both ES and RCSI estimates ($N = 10$).

In selecting the studies to be statistically combined in the ES meta-analysis, due consideration was given to factors such as differences in study characteristics that would be likely to substantially affect the outcome such as population factors and treatment delivery context. All ES meta-analyses used a fixed effects model. This model was chosen because it makes the assumption that there is one single average effect (the effectiveness of psychological treatment) and that the studies we combined came from a population measuring this effect – the fixed effect. The current predominant use of fixed effects in studies led us to adopt this approach for the purposes of making direct comparisons with the efficacy benchmarks. However, we also present parallel analyses using a random effects model where the minimum criterion regarding the number of available studies (k), set as five, is met (Hedges & Vevea, 1998). The primary argument for using a random effects model is that the results better enable generalizability of findings to the relevant population. The ES meta-analyses report the I2 measure of heterogeneity.

Statistical significance gives limited indication of a therapy's effect in that a significant result could be achieved by a small change in an outcome measure providing that change was experienced by most clients (Mullin, Barkham, Mothersole, Bewick, & Kinder, 2006). To safeguard against this effect, Jacobson and colleagues (Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991) espoused reporting rates of reliable and clinically significant change which yields four categories: (1) RCSI which requires that a client change by an amount greater than might expected by chance or measurement error and that the client begin treatment in the clinical population and end treatment in the non-clinical population; (2) reliable improvement only in which

a client's score changes by an amount greater than might expected by chance or measurement error; (3) no reliable change in which the change is within the bounds of measurement error; and (4) reliable deterioration in which the change score worsens by an amount greater than might expected by chance or measurement error. Rates were drawn directly from the published studies or, where these were not available, we calculated them from the study data. RCSI data were available for 14 of the 31 studies in the quality review.

Selection of benchmarks

We selected the best available ES and RCSI benchmarks for each of the three clinical presentations – common mental health problems (CMHPs), panic disorder (PD), and bulimia nervosa (BN) – thereby yielding six condition-specific benchmarks. In doing so, however, we were mindful that there were no ideal benchmarks and that those selected were best approximations.

ES benchmarks

There was no body of efficacy literature pertaining to CMHPs that made a clear candidate for a benchmark. However, Minami, Wampold, Serlin, Kircher, and Brown (2007) derived measure-specific efficacy benchmarks for adult depression drawing from 11 studies. Notwithstanding that the spectrum of CMHPs is greater than depression, it is also likely that the high co-morbidity rates of depression with other mental health problems make it a plausible candidate. From the range of ES benchmarks reported by Minami and colleagues, we selected the ES for the BDI drawn from completer samples, which was 1.86. For PD, we derived the ES benchmark from Westen and Morrison's (2001) meta-analysis yielding an ES of 1.55 ($SD = 1.24$). This ES was virtually identical to the value (ES = 1.53) reported in a meta-analysis of a range of CBT treatment components for PD (Norton & Price, 2007).

RCSI benchmarks

Hunsley and Lee's (2007) review of effectiveness studies employed percentage rates of clinically significant improvement drawing from Westen and Morrison's (2001) meta-analytic review. However, due to an inconsistency in reporting between these two reports, we drew our benchmarks directly from Westen and Morrison as follows: depression, 54%; generalized anxiety disorder, 52%; and PD, 63%. Given the similarity between the rates for depression and generalized anxiety disorder, we selected the slightly more conservative index of 54% for CMHPs. As Westen and Morrison did not supply the improved/recovered benchmark for BN, we employed the meta-analysis of psychotherapy trials reported by Thompson-Brenner, Glass, and Westen (2003) that yielded a benchmark of 40.1% which we rounded to 40%.

Results

Scope of included studies

Of the 31 selected studies, 18 (58%) were conducted in the United Kingdom (UK), 3 in European countries, 8 in the United States of America (USA), 1 in Australia, and 1 in Canada. The majority of studies (29) were conducted since 2000. In terms of presenting

problems, the majority of studies (26) related to CMHPs, three to anxiety disorders, and one each for BN and psychoses/serious mental health problems. Summary information for these studies is presented in Table 1.

Quality appraisal of included studies

Table 2 lists the percentages of the Downs and Black criteria addressed by the included studies. The average level of quality appraisal criteria addressed by the practice-based studies was 65.5%. Levels of reporting were generally high (68.6%) with the lowest levels observed for internal validity – selection bias (17.4%). The mean quality rating score (final column, Table 1) was 21.0 ($SD = 3.46$) out of a possible maximum of 32 points.

Table 2. Percentages of Downs and Black criteria addressed by included studies

	Quality criteria				
	Reporting %	External validity %	Internal reliability %	Internal validity – confounding (selection bias) %	Overall %
Practice-based studies					
Serious mental health problems ($N = 1$)	27.3	63.6	60.0	0.0	34.4
CMHPs ($N = 26$)	68.9	86.7	62.3	17.7	69.1
PD ($N = 3$)	72.7	75.6	60.0	13.3	66.7
BN ($N = 1$)	90.9	72.7	60.0	40.0	75.0
Mean	68.6	84.4	61.9	17.4	65.5

Outcomes: Effect sizes

Common mental health problems

Of the 26 studies relating to CMHPs, 11 reported or provided figures for calculating ES data. Of these, 10 studies comprised data from UK NHS settings and 1 study comprised data from US managed health care settings. In order to be more assured of combining similar data, we excluded the US study (Minami *et al.*, 2008) from the meta-analysis and comment on this later.

We first considered ES data. Studies reported on multiple outcomes and measures. We carried out a meta-analysis based on studies addressing CMHPs and the outcome measured is general depressive symptomatology. The average ES (overall standardized mean difference, SMD) across these studies ($k = 10$) using a fixed effect model was 1.29 (95% CI = 1.26 – 1.33, $N = 10,842$) and ranged from SMD = 0.67 (95% CI = 0.14 – 1.21) to SMD = 1.52 (95% CI = 1.42 – 1.63). The individual and overall ESs, their associated 95% CIs, and the weighting of each study in terms of their contribution to the total sample, are presented in Figure 2. Mindful that the Stiles *et al.* (2008a,b) study contributed 50% of the data, we reran the analyses omitting this study to determine the contribution of this specific study and obtained an ES of 1.19 (95% CI = 1.15 – 1.24) suggesting that this specific study did not unduly influence the result. We also calculated the ES using a random effects model that yielded an ES of 1.14

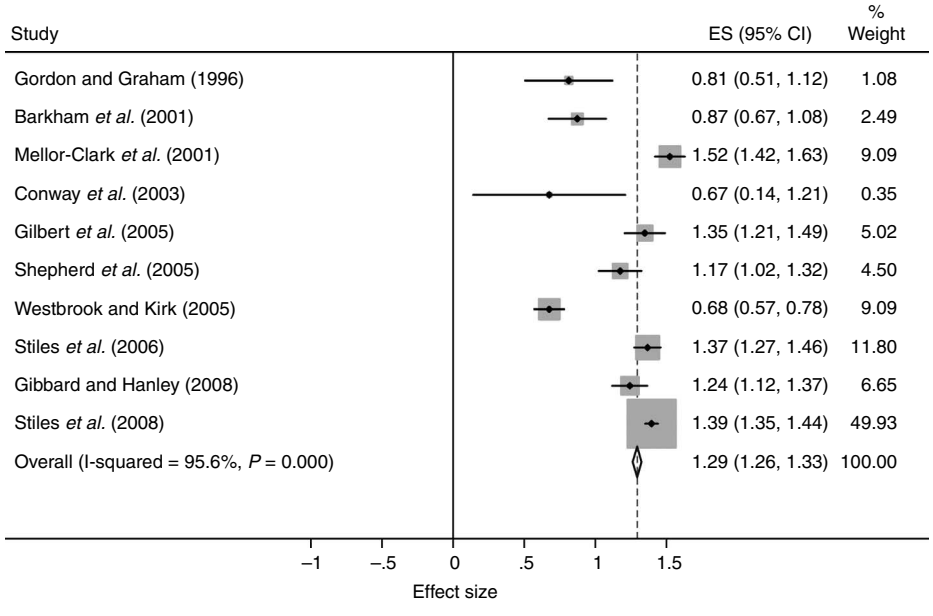

Figure 2. Pre- to post-therapy ESs for UK NHS practice-based studies of common mental health disorders.

(95% CI = 0.96 - 1.32). Recall that the comparator ES benchmark from efficacy studies was 1.86. We note that the I2 measure of heterogeneity is 'large' by conventional criteria.

Panic disorder

Two of the three studies relating to PD supplied ES data (Hahlweg, Fiegenbaum, Frank, Schroeder, & von Witzleben, 2001; Lincoln *et al.*, 2003). Five outcome categories were identified: agoraphobia, anxiety, depression, bodily symptoms, and general impairment. Results are considered separately for each of these outcome categories. All overall ESs were large, ranging from 0.79 to 1.08 with the highest overall ESs observed for bodily symptoms (SMD = 1.08, 95% CI = 0.95 - 1.22) and anxiety (SMD = 0.94, 95% CI = 0.82 - 1.07, $N = 575$). However, the ESs of both studies were associated with large confidence intervals. Because the number of studies did not meet the criterion of five, we do not present the analyses using a random effects model. Recall that the comparator ES benchmark from efficacy studies was 1.55.

Bulimia nervosa

The one study relating to BN (Thompson-Brenner & Westen, 2005) supplied ES data on the single-item Global Assessment Of Functioning ratings. The single ES was large (SMD = 1.60, 95% CI = 1.42 - 1.77, $N = 145$), indicating that clients experienced significant relief of symptoms, involving a shift from serious to mild in terms of severity. The number of studies did not meet the criterion for a random effects model and there was no appropriate benchmark available for ES data.

Outcomes: Reliable and clinically significant improvement

Common mental health problems

Fourteen of the twenty-six studies relating to CMHPs provided estimates of RCSI and the other associated categories of change. None of the studies relating to PD and BN supplied this data. The proportion of clients meeting the various categories of change for each study are shown in Table 3 and are grouped in chronological order within each type of service provision (i.e., primary, secondary, etc.). On average, rates of RCSI (i.e., recovered; Table 3) in primary care (56%) exceeded those for both primary–secondary (36%) and secondary (35%) services. The majority of the studies, across settings, equalled or surpassed the benchmark of 54% recovered.

Table 3. Summary of recovery rates for practice-based studies

Study	Outcome measure	Recovered (%)	Reliably improved only[a] (%)	No reliable change (%)	Reliable deterioration (%)
Westen and Morrison (2001) benchmark		54			
UK primary care					
Mellor-Clark *et al.* (2008)	CORE-OM	58	17	23	2
Evans *et al.* (2003)	CORE-OM	64	17	–	–
Gilbert *et al.* (2005)	CORE-OM	52	18	29	1
Shepherd *et al.* (2005)	CORE-OM	45	21	31	3
Stiles *et al.* (2006)	CORE-OM	61	19	19	1
Gibbard and Hanley (2008)	CORE-OM	54[b]	14[b]	31	1
Stiles *et al.* (2008a,b)	CORE-OM	58	19	21	1
Mean[c]		56.0	17.9	25.7	1.5
UK primary–secondary care					
Lucock *et al.* 2003)	CORE-OM	42	18	39	3
Stiles *et al.* (2003)	CORE-OM	30	–	–	–
UK secondary care					
Barkham *et al.* (2001)	CORE-OM	39	15	40	6
Westbrook and Kirk (2005)	BDI	34	14	50	2
Westbrook and Kirk (2005)	BAI	32	18	47	3
Other					
Snell *et al.* (2001)	OQ-45	21	18	55	13
Conway *et al.* (2003)	BSI	7	27	66	0
Hansen and Lambert (2003)	OQ-45	14	20	57	8

Note. BAI, Beck Anxiety Inventory; BDI, Beck Depression Inventory; BSI, Brief Symptom Inventory; CORE-OM, Clinical Outcomes in Routine Evaluation – Outcome Measure; OQ-45, Outcome Questionnaire 45.

[a]In some studies, the numbers of patients making reliable change included those patients whose scores were below the clinical cut off.

[b]These specific percentages were calculated from the raw data provided by the authors.

[c]The sum of the mean percentages exceed 100% due to rounding of values for individual studies.

Bulimia nervosa

Thompson-Brenner and Westen (2005) did not compute estimates of RCSI rates but provided rates of patients significantly improved and recovered. Based on clinician ratings, 89.6% patients were significantly improved and 52.7% recovered in their eating

disorder symptoms indicating that the rates from the practice-based study exceeded the efficacy trial benchmark.

Outcomes and quality ratings

In order to test for associations between quality of studies and reported outcomes, we correlated quality ratings with both ES ($N = 14$) and RCSI ($N = 14$) rates. Total quality ratings were correlated $r(14) = .85$, $p < .004$ with ESs and $r(14) = .25$, $p = .40$ with RCSI estimates. Hence, higher ESs were associated with better quality studies.

Discussion

In contrast to the vast number of efficacy studies available for systematic reviews and meta-analyses, the number of practice-based studies we identified was small ($N = 31$) and similar to the number reported by Hunsley and Lee (2007) for effectiveness studies of adult disorders ($N = 21$). However, the emerging profile of the yield from practice-based studies can inform policy and practice when considered in conjunction with findings from other methodological approaches.

Of most importance to practitioners is evidence from practice-based research showing that the psychological treatment conducted in routine clinic settings is effective for a range problems experienced by clients. CMHPs yielded an uncontrolled ES of 1.29 and anxiety disorders yielded uncontrolled ESs within the range of 0.86–1.12. These ESs are consistent with practice based and effectiveness studies published subsequent to our electronic search. For example, Richards and Suckling (2009), drawing on routinely collected data from one of the UK's Improving Access to Psychological Therapies demonstration sites, reported uncontrolled ESs of 1.38 for depression as measured by the PHQ-9 (Patient Health Questionnaire) and 1.41 for anxiety as measured by the GAD-7 (Generalised Anxiety Disorder Assessment). Van Ingen, Freiheit, and Vye (2009) reported an uncontrolled ES (weighted) of 1.35 from a meta-analysis of 11 effectiveness studies of anxiety disorders derived from a variety of outcome measures (although the range of unweighted ESs ranged from 0.27 to 3.29).

The effect we found for common mental health disorders (ES = 1.29) indicates the average client following therapy would have been at the 90th percentile prior to therapy. For comparison, the ES benchmark value of 1.86 corresponds to the 97th percentile at pre-therapy. Using ES, calculated as the change score divided by the pre-treatment *SD*, for this comparison may systematically disadvantage practice-based studies, however. As noted in a report of a direct comparison of several efficacy and practice-based studies, the larger ESs in the efficacy studies appeared to be 'mainly attributable to the systematically smaller pre-treatment SD among the selected clients in the randomized trials as compared with the unselected clients in the practice-based studies. The effect is merely hidden when – as is often the case, even in otherwise sophisticated meta-analytic reviews – each study is compared in terms of its own measures, ignoring the possibility that selection procedures systematically narrowed the pre-treatment distribution of scores' (Barkham *et al.*, 2008, p. 412).

ESs varied across outcome categories in plausible ways. For example, the ES for depression in the Westbrook and Kirk (2005) study of anxiety and depression (ES = 0.68) was considerably lower than the mean. Whereas the majority of studies relating to CMHPs pertained to primary care service provision, Westbrook and

Kirk's study was of people presenting with more chronic conditions. The Barkham *et al.* (2001) study returned a similarly smaller ES (ES = 0.87) from secondary care settings (see Figure 1).

In contrast to the ES comparisons, the majority of practice-based studies met the benchmark of 54% recovered, based on RCSI criteria (Table 3). In primary care studies, the proportion of patients classified as recovered averaged 56%. It should be noted that some studies included clients scoring below the clinical cut off point at pre-therapy, which has the effect of reducing the numbers of patients who could achieve RCSI; clients who began below the clinical cut point pre-therapy could not, by definition, move from the clinical to the non-clinical population. Inclusion of these studies yielded a more conservative improvement rate for practice-based studies. As a further caution, RCSI rates are sensitive to investigators' arbitrary choice of some parameters; for example, small differences in estimated reliability of the measure lead to large differences in RCSI thresholds (Barkham *et al.*, 2008).

Rated quality of the practice-based data set

The practice-based studies sampled were characterized by low levels of internal validity according to the Downs and Black criteria (see Table 2). Within these practice-based studies, treatment was not determined by protocol or manual and we are not able to make therapy modality-specific generalizations. None of the studies had a control group, partly because of ethical and practical constraints, but more specifically because the absence of a control group was a criterion for inclusion in this sample. These limitations underline why practice-based results be seen as complementing those from efficacy trials, which have different limitations (Barkham & Margison, 2007).

Of the 31 studies, 17 did not provide pre–post therapy data for the calculation of ESs and an overlapping group of 17 studies did not provide data for calculation of rates of recovery, so they could not be benchmarked. All of these studies relied on pre–post therapy data, thereby restricting any generalizations to treatment completers. Collecting session-by-session data would help to alleviate this problem by permitting estimates of progress to the point clients stopped attending.

Although the Downs and Black checklist had been designed to evaluate both randomized and non-randomized trials it did not prove to be responsive to the design features of practice-based research. Some features seen only in terms of deficit (e.g., lack of control group, heterogeneous nature of patient population, non-manualized treatments) define the practice-based paradigm's aims and conceptual underpinnings. To address this disparity, we revised and adapted the checklist (see Appendix B) to render it applicable for practice-based research. It is encouraging that the majority of studies reported high levels of external validity and we would further recommend the technique of presenting a data flowchart akin to a CONSORT diagram whereby it is possible to (a) locate the sample studied as representative of the population referred and (b) to ensure comprehensiveness in the study sample by means of, for example, including all consecutively referred clients for a specified time period.

Finding that higher quality studies reported higher ESs (i.e., quality ratings were correlated positively and significantly with ESs) suggests that better designed studies that report data adequately and comprehensively are better able to detect therapeutic and clinical effects in routine practice. This interpretation supports the case for methodological rigour in the field of practice-based evidence. Alternatively, the correlation could reflect more tightly managed sampling in the higher quality studies,

leading to smaller pre-treatment SDs and consequently larger ESs. We tested this by correlating the quality ratings with the pre-treatment SDs and found a null correlation $r(14) = .26$, $p = .38$. Hence, in the current sample of studies, higher ESs were not associated with restricted sampling.

Other limitations

Uncontrolled ES estimates, of the kind we reported, confound effects of the passage of time, effects of common factors and genuine placebo effects (Westen & Morrison, 2001). More generally, small sample size and associated large confidence intervals were a common feature of many studies included in this review. In terms of the clinical presentations, the benchmarks did not perfectly match the selected studies in terms of populations and/or measures used. For example, 9 of the 10 studies included in the common mental health ES meta-analysis used measures of general symptoms for clients suffering from CMHPs while our comparison used the symptom specific ES benchmark for depressed patients. On the treatment side, it is questionable whether tightly controlled protocol-driven treatments delivered by therapists trained specifically for the purposes of a specific study are the same when transported to routine settings even if they have the same name.

Conclusion

When patients receive treatment as normally delivered within routine practice they experience significant relief of symptoms. When RCSI rates are benchmarked against results from efficacy trials, there is evidence to suggest that practice-based studies do meet the standard required. However, when ESs are used, outcomes fall short of those achieved in trials. Hence, the results of benchmarking against efficacy trials depend to some extent on the outcome index used. ESs appear more biased against practice-based studies because of the impact of restricted variance in trials, but both indexes are problematic. Although it is indicated that differences in outcomes between efficacy trials and practice-based research are partly attributable to methodological issues, substantive factors are also likely to contribute. Therefore, in addition to attending to methodological issues, further work is required to understand the relative contribution of these factors. Future research should seek greater comprehensiveness (and thereby representativeness) of data samples and an acceptable level of reporting of data – both of which are essential if treatments as practised are to be usefully evaluated.

Acknowledgements

We thank Dave Saxon for his support in using the Stata software, Dr Graham Paley for acting as an independent rater for the quality of a sample of review studies, and Dr Terry Hanley and Isabel Gibbard for enabling additional analyses of their data.

References

(* denotes inclusion in the review; †denotes inclusion in meta-analyses)

*Baker, R., Baker, E., Allen, H., Golding, E., Baker, R., Thomas, P., *et al.* (2002). A naturalistic longitudinal evaluation of counselling in primary care. *Counselling Psychology Quarterly*, *15*, 359–373.

Barkham, M., & Margison, F. (2007). Practice-based evidence as a complement to evidence-based practice: From dichotomy to chiasmus. In C. Freeman & M. Power (Eds.), *Handbook of evidence-based psychotherapies* (pp. 443–476). Chichester: Wiley.

*Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., *et al.* (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*, 184–196.

Barkham, M., Stiles, W. B., Connell, J., Twigg, E., Leach, C., Lucock, M., *et al.* (2008). Effects of psychological therapies in randomized trials and practice-based studies. *British Journal of Clinical Psychology*, *47*, 397–415. doi:10.1348/014466508X311713

*Booth, H., Cushway, D., & Newnes, C. (1997). Counselling in general practice: Clients' perceptions of significant events and outcome. *Counselling Psychology Quarterly*, *10*, 175–187.

*Borkovec, T. D., Echemendia, R. J., Ragusea, S. A., & Ruiz, M. (2001). The Pennsylvania practice research network and future possibilities for clinically meaningful and scientifically rigorous psychotherapy effectiveness research. *Clinical Psychology: Science and Practice*, *8*, 155–167.

*Brown, G. S., & Jones, E. R. (2005). Implementation of a feedback system in a managed care environment: What are patients teaching us? *Journal of Clinical Psychology*, *61*, 187–198.

Clark, D. M., Fairburn, C. G., & Wessly, S. (2008). Psychological treatment outcomes in routine NHS services: A commentary on Stiles *et al.* (2007). *Psychological Medicine*, *38*, 629–634.

*†Conway, S., Audin, K., Barkham, M., Mellor-Clark, J., & Russell, S. (2003). Practice-based evidence for a brief time-intensive multi-modal therapy guided by group-analytic principles and method. *Group Analysis*, *36*, 413–435.

Department of Health (1996). *NHS psychotherapy services in England: Review of strategic policy.* London: Her Majesty's Stationary Office.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiological Community Health*, *52*, 377–384.

Eisen, S. V., & Dickey, B. (1998). Mental health outcome assessment: The new agenda. *Psychotherapy: Theory, Research, Practice, Training*, *33*, 181–189.

*Evans, C., Connell, J., Barkham, M., Marshall, C., & Mellor-Clark, J. (2003). Practice-based evidence: Benchmarking NHS primary care counselling services at national and local levels. *Clinical Psychology and Psychotherapy*, *10*, 374–388.

Farhall, J., & Cotton, S. (2002). Implementing psychological treatment for symptoms of psychosis in an area mental health service: The response of patients, therapists and managers. *Journal of Mental Health*, *11*, 511–522.

*†Gibbard, I., & Hanley, T. (2008). A five year evaluation of the effectiveness of person centred counselling in routine clinical practice in primary care. *Counselling and Psychotherapy Research*, *8*, 215–222.

*†Gilbert, N., Barkham, M., Richards, A., & Cameron, I. (2005). The effectiveness of a primary care mental health service delivering brief psychological interventions: A benchmarking study using the CORE System. *Primary Care Mental Health*, *3*, 241–251.

*†Gordon, K., & Graham, C. (1996). The impact of primary care counselling on psychiatric symptoms. *Journal of Mental Health*, *5*, 515–523.

*Hahlweg, K., Fiegenbaum, W., Frank, M., Schroeder, B., & von Witzleben, I. (2001). Short- and long-term effectiveness of an empirically supported treatment for agoraphobia. *Journal of Consulting and Clinical Psychology*, *69*, 375–382.

*Hansen, N. B, & Lambert, M. J. (2003). An evaluation of the dose-response relationship in naturalistic treatment settings using survival analysis. *Mental Health Services Research*, *5*, 1–12.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

*Hirsch, C., Jolley, S., & Williams, R. (2000). A study of outcome in a clinical psychology service and preliminary evaluation of cognitive-behavioural therapy in real practice. *Journal of Mental Health*, *9*, 537–549.

*Houghton, S., & Saxon, D. (2007). An evaluation of large group CBT psycho-education for anxiety disorders delivered in routine practice. *Patient Education and Counselling*, *68*, 107–110.

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

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*, 300–307.

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

*Kates, N., Crustolo, A.-M., Farrar, S., & Nikolaou, L. (2002). Counsellors in primary care: Benefits and lessons learned. *Canadian Journal of Psychiatry*, *47*, 857–862.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Layard, R. (2006). The case for psychological treatment centres. *British Medical Journal*, *332*, 1030–1032.

*Lincoln, T. M., Rief, W., Hahlweg, K., Frank, M., von Witzleben, I., Schroeder, B., *et al.* (2003). Effectiveness of an empirically supported treatment for social phobia in the field. *Behaviour Research and Therapy*, *41*, 1251–1269.

*Lucock, M., Leach, C., Iveson, S., Lynch, K., Horsefield, C., & Hall, P. (2003). A systematic approach to practice-based evidence in a psychological therapies service. *Clinical Psychology and Psychotherapy*, *10*, 389–399.

McEvoy, P. M., & Nathan, P. (2007). Effectiveness of cognitive behavior therapy for diagnosticay heterogeneous groups: A benchmarking study. *Journal of Consulting and Clinical Psychology*, *75*, 344–350.

*†Mellor-Clark, J., Connell, J., Barkham, M., & Cummins, P. (2001). Counselling outcomes in primary health care: A CORE system data profile. *European Journal of Psychotherapy, Counselling and Health*, *4*, 65–86.

*Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S., & 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*, 116–124.

Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (2007). Benchmarks for psychotherapy efficacy in adult major depression. *Journal of Consulting and Clinical Psychology*, *75*, 232–243.

Mullin, T., Barkham, M., Mothersole, G., Bewick, B. M., & Kinder, A. (2006). Recovery and improvement benchmarks for counselling and the psychological therapies in routine primary care. *Counselling and Psychotherapy Research*, *6*, 68–80.

*Nettleton, B., Cooksey, E., Mordue, A., Dorward, I., Ferguson, J., Johnston, J., *et al.* (2000). Counselling: Filling a gap in general practice. *Patient Education and Counseling*, *41*, 197–207.

Norton, P. J., & Price, E. C. (2007). A meta-analytic review of adult cognitive-behavioral treatment outcome across the anxiety disorders. *Journal of Nervous and Mental Disease*, *195*, 521–531.

Richards, D. A., & Suckling, R. (2009). Improving Access to Psychological Therapies (IAPT): Phase IV prospective cohort study. *British Journal of Clinical Psychology*, *48*(4), 377–396. doi:10.1348/014466509X405178

Roth, A. D., & Parry, G. (1997). The implications of psychotherapy research for clinical practice and service development: Lessons and limitations. *Journal of Mental Health*, *6*, 367–380.

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crit-Christoph, P., Hazelrigg, P., *et al.* (1997). Evidence that therapy works in clincally representative conditions. *Journal of Consulting and Clinical Psychology*, *65*, 355–365.

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*, 512–529.

*†Shepherd, M., Ashworth, M., Evans, C., Robinson, S. I., Rendall, M., & Ward, S. (2005). What factors are associated with improvement after brief psychological interventions in primary care? Issues arising from using routine outcome measurement to inform clinical practice. *Counselling and Psychotherapy Research*, *5*, 273–280.

*Smith, H. B., Sexton, T. L., & Bradley, L. J. (2005). The practice research network: Research into practice, practice into research. *Counselling and Psychotherapy Research*, *5*, 285–290.

*Snell, M. N., Mallinckrodt, B., Hill, R. D., & Lambert, M. J. (2001). Predicting counseling center clients' response to counseling: A 1-year follow-up. *Journal of Counseling Psychology*, *48*, 463–473.

StataCorp (2001). *Statistical software: Release 7.0*. College Station, TX: Stata Corporation.

*†Stiles, W. B., Barkham, M., Mellor-Clark, J., & Connell, J. (2006). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary-care routine practice: Replication in a larger sample. *Psychological Medicine*, *36*, 555–556.

*†Stiles, W. B., Barkham, M., Mellor-Clark, J., & Connell, J. (2008a). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary-care routine practice: Replication in a larger sample. *Psychological Medicine*, *38*, 677–688.

Stiles, W. B., Barkham, M., Mellor-Clark, J., & Connell, J. (2008b). Routine psychological treatment and the Dodo verdict: A rejoinder to Clark *et al.* (2008). *Psychological Merdicine*, *38*, 905–910.

*†Stiles, W. B., Leach, C., Barkham, M., Lucock, M., Iveson, S., Shapiro, D. A., *et al.* (2003). Early sudden gains in psychotherapy under routine clinic conditions: Practice-based evidence. *Journal of Consulting and Clinical Psychology*, *71*, 14–21.

Thompson-Brenner, H., Glass, S., & Westen, D. (2003). A multidimensional meta-analysis of psychotherapy for bulimia nervosa. *Clinical Psychology: Science and Practice*, *10*, 269–287.

*Thompson-Brenner, H., & Westen, D. (2005). A naturalistic study of psychotherapy for bulimia nervosa, part 1: Comorbidity and therapeutic outcome. *Journal of Nervous and Mental Disease*, *193*, 573–584.

Van Ingen, D. J., Freiheit, S. R., & Vye, C. S. (2009). From the lab to the clinic: Effectiveness of cognitive-behavioral treatments for anxiety disorders. *Professional Psychology: Research and Practice*, *40*, 69–74.

*Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology*, *73*, 914–923.

*†Westbrook, D., & Kirk, J. (2005). The clinical effectiveness of cognitive behaviour therapy: Outcome for a large sample of adults treated in routine practice. *Behaviour Research and Therapy*, *43*, 1243–1261.

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

*Wolgast, B. M., Lambert, M. J., & Puschner, B. (2003). The dose-response relationship at a college counseling center: Implications for setting session limits. *Journal of College Student Psychotherapy*, *18*, 15–29.

Received 31 July 2008; revised version received 27 July 2009

Appendix A

Search history

(1) Practice-based evidence.mp. [mp = title, abstract, heading word, table of contents, key concepts]
(2) Routine Practice.mp. [mp = title, abstract, heading word, table of contents, key concepts]
(3) Benchmarking.mp. [mp = title, abstract, heading word, table of contents, key concepts]
(4) Transportability.mp. [mp = title, abstract, heading word, table of contents, key concepts]
(5) Transferability.mp. [mp = title, abstract, heading word, table of contents, key concepts]
(6) Clinical$ representat$.mp. [mp = title, abstract, heading word, table of contents, key concepts]
(7) Evidence based practice.mp. [mp = title, abstract, heading word, table of contents, key concepts]
(8) (External valid$ adj findings).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(9) (Applicab$ adj findings).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(10) (Applicab$ adj intervention$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(11) (Empiric$ support$ adj treatment$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(12) (Empiric$ support$ adj intervention$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(13) Clinical$ Effective$.mp. [mp = title, abstract, heading word, table of contents, key concepts]
(14) (Dissem$ adj treatment$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(15) (Dissem$ adj intervention$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(16) (Clinical Practice adj intervention$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(17) (Clinical Practice adj treatment$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(18) (Service deliv$ adj intervention$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(19) (Service deliv$ adj treatment$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(20) (Clinical$ effective$ adj2 evaluat$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(21) (Service deliv$ adj evaluat$).mp. [mp = title, abstract, heading word, table of contents, key concepts]
(22) 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21
(23) limit 22 to (human and English language)

Appendix B

Downs and Black rating sheet: Adapted version

ID

Reporting Yes = 1 No = 0 Unable to determine = 0

1	Is the hypothesis/aim/objectives of the study clearly described	☐	
2	Are the main outcomes to be measured clearly described in the introduction or methods section	☐	If the main outcomes are first mentioned in the results section, the question should be answered No
3	Are the characteristics of the clients included in the study clearly described	☐	Inclusion and/or exclusion criteria should be given. Emphasis on inclusion and exclusion criteria, other characteristics are age/gender/morbidity
4	Are the interventions/treatments of interest clearly described?	☐	Treatments and placebo (where relevant) that are to be compared should be clearly described
5	Are the distributions of principal confounders in each group of clients to be compared (or within a single group) clearly described?	☐	A list of principal confounders is provided. Morbidity, co-morbidity, age, gender, previous history. Good qual will include adjustment regression or matching
6	Are the main findings of the study clearly described?	☐	Simple outcome data (including denominators and numerators) should be reported for all major findings so that the reader can check the major analyses and conclusions. This question does not cover statistical testes which are considered below
7	Does the study provide estimates of the random variability in the data for the main outcomes?	☐	In non normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation, or confidence intervals should be reported. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes
8	Have all the important adverse events that may be a consequence of the intervention/treatment been reported?	☐	This should be answered yes if the study demonstrates that there was a comprehensive attempt to measure adverse events (A list of adverse events is provided). E.g. early discontinuation of therapy
9	Have the characteristics of clients lost to follow-up been described?	☐	This should be answered yes where there were no losses to follow-up or where losses to follow-up were so small that findings would be unaffected by their inclusion. This should be answered no where a study does not report the number of patients lost to follow-up. Follow − up = post − therapy, or loss from study at baseline

Appendix B. (*Continued*)

Reporting Yes = 1 No = 0 Unable to determine = 0			
10	Have actual probability values been reported (e.g. 0.035 rather than <0.05) for the main outcomes except where the probability value is less than 0.01	☐	
11	Have sufficient data been provided to enable calculation of outcomes such as pre–post ESs, estimates of reliable a nd clinically significant change	☐	If data are provided to enable calculation of any one of these outcomes score the question yes
External valitity/clinical representativeness Yes = 0 No = 0 Unable to determine = 0			
12	(a) Were the clients asked to participate in the study representative of the entire population from which they were recruited	☐	The study must identify the source population for clients and describe how the patients were selected. Clients would be representative if they comprised the entire source population, an unselected sample of consecutive clients, or a random sample. Random sampling is only feasible where a list of all members of the relevant population exists. Where a study does not report the proportion of the source population from which the patients are derived the question should be answered as unable to determine
	(b) Were clients referred through usual clinic routes	☐	
13	Were those clients who were prepared to participate representative of the entire population from which they were recruited?	☐	The proportion of those asked who agreed should be stated. Validation that the sample was representative would included demonstrating that the distribution of the main confounding factors was the same in the study sample and the source population
14	(a) Were client heterogeneous in personal characteristics	☐	
	(b) Were clients heterogeneous in terms of presenting problems	☐	
15	(a) Were the staff, places, facilities where the patients were treated representative of the treatment the majority of patients receive?	☐	For the question to be answered yes the study should demonstrate that the intervention was representative of that in use in the source population
	(b) Was the treatment conducted in a non university setting	☐	The question should be answered no if, for example, the intervention was undertaken in a specialist centre unrepresentative of the hospitals most of the source population would attend
	(c) Was implementation of treatment monitored (R)	☐	

Appendix B. (*Continued*)

External valitity/clinical representativeness Yes = 0 No = 0 Unable to determine = 0		
16	Were therapists experienced, professionals with regular caseloads	☐
17	Were therapists free to use a wide variety of procedures in treatment and not just limited to one treatment procedure	☐
18 (R)	Were therapists trained immediately before the study and in the specific treatment being studied	☐

Internal reliability Yes = 1 No = 0 Unable to determine = 0		
19	If any of the results of the study were based on 'data dredging' was this made clear	☐ Any analysis that had not been planned at the outset of the study should be clearly indicated. If no retrospective unplanned subgroup analysis were reported, then answer yes
20	Were the statistical tests used to assess the main outcomes appropriate	☐ The statistical techniques used must be appropriate to the data. For example, non parametric methods should be used for small sample sizes. Where little statistical analysis has been undertaken, but where there is no evidence of bias, the question should be answered yes. If the distribution of the data (normal or not) is not described it must be assumed that the estimates used were appropriate and the question should be answered yes
21	Was the compliance with the intervention/s/treatments reliable?	☐ Where there was non compliance with the allocated the question should be answered no
22	Were the main outcome measures used accurate (valid and reliable)	☐ For studies where the outcome measures are clearly described, the question should be answered yes. For studies which refer to other work or that demonstrates the outcome measures are accurate, the question should be answered yes
23	Do the analyses adjust for different lengths of follow-up of patients in different treatment groups?	☐ Where no comparison group score 0. Where lengths of follow-up the same score 1

Appendix B. (*Continued*)

Internal reliability confounding (selection) bias Yes = 1 No = 0 Unable to determine = 0			
24	Were the clients in different intervention/treatment groups recruited from the same population	☐	For example, clients for all comparison groups should be selected from the same source population. The question should be answered unable le to determine where there is no information concerning the source of patients included in the study. Where no comparison group score 0
25	Were the clients in different intervention/treatment groups recruited over the same period of time?	☐	For a study which does not specify the time period over which clients were recruited, the question should be answered unable to determine. Where no comparison group score 0
26	Was there adequate adjustment for confounding in the analysis from which the main findings were drawn	☐	This question should be answered no if the main conclusions of the study were based on analyses of treatment rather than intention to treat; the distribution of known confounders was not described; or the distribution of confounders differed between the treatment groups but was not taken into account in the analyses. if the effect of the main confounders was not investigated or confounding was demonstrated but no adjustment was made in the final analyses, the question should be answered no
27	Were losses of clients to follow-up taken into account?	☐	If the numbers of clients lost to follow-up are not reported, the question should be answered as unable to determine. If the proportion of lost to follow-up was too small to affect the main findings, the question should be answered yes
28	Did the study have sufficient power to detect a clinically important effect where the probability value for a difference being due to chance is less than 5%	☐	Sample sizes have been calculated to detect a difference of x and y%. Has power analysis been performed

	Size of smallest intervention group	
A	$<N1$	0
B	$N1$–$N2$	1
C	$N3$–$N4$	2
D	$N5$–$N6$	3
E	$N7$–$N8$	4
F	$N8+$	5