Benchmarking Routine Psychological Services: A Discussion of Challenges and Methods

Jaime Delgadillo, Dean McMillan, Chris Leach, Mike Lucock, Simon Gilbody and Nick Wood

Behavioural and Cognitive Psychotherapy / *FirstView* Article / October 2012, pp 1 - 15
DOI: 10.1017/S135246581200080X, Published online:

Behavioural and Cognitive Psychotherapy: page 1 of 15
doi:10.1017/S135246581200080X

Benchmarking Routine Psychological Services: A Discussion of Challenges and Methods

Jaime Delgadillo

Leeds Community Healthcare NHS Trust, UK

Dean McMillan

University of York, UK

Chris Leach and Mike Lucock

South West Yorkshire Partnership NHS Foundation Trust and University of Huddersfield, UK

Simon Gilbody

University of York, UK

Nick Wood

Leeds Community Healthcare NHS Trust, UK

Background: Policy developments in recent years have led to important changes in the level of access to evidence-based psychological treatments. Several methods have been used to investigate the effectiveness of these treatments in routine care, with different approaches to outcome definition and data analysis. **Aims:** To present a review of challenges and methods for the evaluation of evidence-based treatments delivered in routine mental healthcare. This is followed by a case example of a benchmarking method applied in primary care. **Method:** High, average and poor performance benchmarks were calculated through a meta-analysis of published data from services working under the *Improving Access to Psychological Therapies* (IAPT) Programme in England. Pre-post treatment effect sizes (ES) and confidence intervals were estimated to illustrate a benchmarking method enabling services to evaluate routine clinical outcomes. **Results:** High, average and poor performance ES for routine IAPT services were estimated to be 0.91, 0.73 and 0.46 for depression (using PHQ-9) and 1.02, 0.78 and 0.52

Reprint requests to Jaime Delgadillo, Leeds Community Healthcare NHS Trust - Primary Care Mental Health, The Reginald Centre, Second Floor, 263 Chapeltown Road, Leeds LS7 3EX, UK. E-mail: jaime.delgadillo@nhs.net

for anxiety (using GAD-7). Data from one specific IAPT service exemplify how to evaluate and contextualize routine clinical performance against these benchmarks. **Conclusions:** The main contribution of this report is to summarize key recommendations for the selection of an adequate set of psychometric measures, the operational definition of outcomes, and the statistical evaluation of clinical performance. A benchmarking method is also presented, which may enable a robust evaluation of clinical performance against national benchmarks. Some limitations concerned significant heterogeneity among data sources, and wide variations in ES and data completeness.

Keywords: Benchmarking, primary care, depression, anxiety.

Introduction

In the last decade, two important streams of research evidence have come together resulting in an important paradigm shift and transformation of mental healthcare. First, epidemiological studies have drawn attention to the high prevalence, health, societal and economic impact of depression and anxiety disorders (Das-Munshi et al., 2008; Kessler et al., 2003; McManus, Meltzer, Brugha, Bebbington and Jenkins, 2009). Second, psychotherapy research has progressively aligned itself to the evidence-based medicine movement, amassing empirical support for the efficacy of psychological interventions (Chambless et al., 1998). Consequently, the dissemination of empirically supported treatments (EST) has been advocated by clinical guidelines (National Institute for Health and Clinical Excellence, 2007a, b), which in turn have led to changes in the organization of psychological services.

An example of this movement can be found in the *Improving Access to Psychological Therapies* (IAPT) programme. This large-scale government funded initiative has been piloted and implemented in England since 2008, organizing EST for depression and anxiety in a series of progressively intensive stages of therapy – otherwise referred to as a stepped-care model (Richards and Suckling, 2009). There is, however, debate in the field about the "real world" effectiveness of EST in these routine services and outside of the closely controlled conditions of an efficacy study. The argument about external validity, for example, suggests that the results of clinical trials are not necessarily generalizable to ordinary healthcare populations that present a range of co-morbidities and demographic factors that render them distinct to research participants (Chambless and Ollendick, 2001; Franklin and DeRubeis, 2006). Others have argued that delivering effective EST requires close adherence to protocols used in clinical trials (Roth and Fonagy, 2004; Siev, Huppert and Chambless, 2009) and specific competences (Roth and Pilling, 2007). In routine practice, treatment fidelity is likely to vary from service to service, and this may account for differences in clinical outcomes (Glover, Webb and Evison, 2010).

Therefore, the large-scale dissemination of EST in routine practice raises some important questions. How do routine outcomes compare to those of efficacy studies? How do we measure outcomes in routine practice? How can we evaluate the outcomes of a routine service compared to other similar services? The main objective of this paper is to discuss some key challenges related to the measurement, definition and comparative analysis of clinical outcomes. A second objective is to present a case example of how to deal with some of these methodological challenges and evaluate the outcomes of a routine service against national benchmarks and clinical trials. References and analyses focus on the English IAPT

programme to contextualize the discussion and illustrate the application of benchmarking methods.

Measuring, defining and comparing clinical outcomes

Selecting outcome measurement tools

With the psychotherapy research literature describing over 100 different patient reported outcome measures, many with proven diagnostic validity and reliability (Wahl, Meyer, Löwe and Rose, 2010), specifying a set of measures for a routine service can be challenging. A second challenge is that routine services in primary care often see a heterogeneous group of patients with multiple diagnoses. This introduces a tension between selecting diagnosis-specific tools versus more general measures of improvement (e.g. overall psychological distress, functioning or quality of life), or idiosyncratic measures that are internally valid to the treatment of individual patients. Another challenge is the selection of a set of measures that are acceptable to patients in terms of their content and length. Some important considerations for the selection and application of outcome measures are summarized below, based on outcomes research and clinical guidelines (Blais et al., 2012; McAleavey, Nordberg, Kraus and Castonguay, 2012; Minami, Wampold, Serlin, Kircher and Brown, 2007; National Screening Committee, 2003; Newnham and Page, 2010; NICE, 2011; Wahl et al., 2010).

- The selection of outcome measures from a pool of candidate tools should weigh up the following features: (a) research evidence of robust validity and reliability; (b) research evidence of acceptability to patients; (c) ease of administration and interpretation; (d) cost of training and implementation; (e) availability in several languages and formats to maximize accessibility; (f) the extent of published normative data to make comparative research more feasible.
- An adequate set of measures should balance: (a) diagnosis-specific measures that are matched to the service's target population and inclusion criteria; (b) "higher level" measures of functioning and/or quality of life; and (c) idiosyncratic measures that can capture individualized patient goals and outcomes. Such a combination of measures may render a good balance between reactivity (wide relevance to the treatment population) and specificity (e.g. precision of disorder-specific measures to the symptoms of individuals) (Minami et al., 2007). The number and length of measures should take into consideration the potential time onus and burden on clinicians and patients.
- Collecting these measures at repeated intervals can significantly maximize data completeness (Richards and Suckling, 2009) and enables their utilization in providing live feedback to patients and clinicians, which can in itself improve outcomes for some patients (Shimokawa, Lambert and Smart, 2010).
- Repeated exposure to cognitive tests (such as memory or intelligence tests) can influence performance through practice, and this is problematic when aiming to investigate the influence of a specific intervention over and above the influence of so-called "practice effects" (McCaffrey and Westervelt, 1995). In theory, it is possible that practice effects and/or fatigue may ensue in a process of repeated psychological symptom measurement (Larsen and Fredrickson, 1999). However, there is yet no convincing evidence that practice

or fatigue effects may undermine the validity of repeated outcome measurement outside of the specific area of neuropsychological testing.

IAPT is an example of a large-scale programme using a specific set of measures that meet many of the above criteria. It employs brief validated measures of depression and anxiety (PHQ-9 and GAD-7; Kroenke, Spitzer, Williams and Löwe, 2010), which are publicly available in several languages, and generally acceptable to patients in spite of criticisms by health professionals (Dowrick et al., 2009). They are used at repeated measurement points, alongside measures of functioning and adjustment, and several disorder-specific measures that are more sensitive to individual cases (IAPT National Programme Team, 2011).

Defining improvement

Selecting an adequate set of measures then poses the challenge of defining clinically meaningful improvement (and deterioration). Most diagnostic measures are interpreted based on normative cut-off scores that discriminate those respondents who are more or less likely to meet criteria for a specific condition. For example, a score $\geqslant$10 on the PHQ-9 measure used in the IAPT programme is considered to be indicative of major depression (Kroenke, Spitzer and Williams, 2001). Hence, several outcome evaluations have estimated depression improvement rates based on the numbers of patients with PHQ-9 scores $\geqslant$10 whose scores reduced below this cutpoint after treatment (Clark et al., 2009; Glover et al., 2010; Richards and Suckling, 2009). An important problem with this definition of improvement is that it is particularly susceptible to "Type I error", which occurs when cases with no meaningful symptom change are assumed to have improved. One explanation is that this method does not account for measurement error that may be due to inherent limitations in the precision of the questionnaire. Another related explanation is that cases with marginal symptom reductions are counted as "recovered", when such small reductions may possibly be due to natural fluctuations or regression to the mean.[1] A way to deal with this may be to use a multidimensional approach that would involve triangulating two or more correlated measures to enable a more holistic assessment of change (McAleavey et al., 2012). Clark et al. (2009), for example, proposed a multidimensional method referred to as "IAPT recovery". According to this method, patients scoring above the clinical cut-off in a depression (PHQ-9) or anxiety measure (GAD-7) at baseline will have "recovered" if their post-treatment scores are below the cut-offs for both measures. Although this is a stricter way to define outcomes, the possibility of Type I error still remains, particularly if small changes in both measures are due to regression to the mean.

A more robust way to minimize Type I error is to employ a reliable change index (RCI), which is a statistic used to evaluate whether observed changes on a scale are statistically reliable and not solely due to chance (Jacobson and Truax, 1991). The RCI is essentially a certain number of points on a scale above which observed changes are considered statistically reliable. For example, Richards and Borglin (2011) propose that a minimum reduction of 6 points in the PHQ-9 measure would be indicative of reliable improvement. Taken together,

[1]Regression to the mean is a phenomenon where extreme measures are statistically more likely to reduce closer to the mean at subsequent measurement points.

the RCI and the diagnostic cut-off score can help to define reliable and clinically significant improvement, which has been recommended as a robust method for assessing recovery in psychological interventions (Evans, Margison and Barkham, 1998; Jacobson and Truax, 1991; McMillan, Richards and Gilbody, 2010).

There are, however, some disadvantages to using a dichotomous definition of improvement (e.g. recovered vs. not recovered). In primary care, not all patients will necessarily meet criteria for a significant mental disorder; in fact, clinical guidelines advocate offering evidence-based treatments to people with mild or sub-threshold depression and anxiety (NICE, 2007a, b). Therefore, a dichotomous outcome definition does not account for symptom changes in patients with mild disorders. Although only limited symptom reductions can be expected in patients with sub-clinical symptoms, it is possible that such patients may deteriorate, and therefore rates of reliable deterioration should also be reported if dichotomous outcome definitions of recovery are adopted. Another limitation is that focusing only on "recovery" fails to recognize significant symptom changes for patients who may still meet criteria for a common mental disorder but may feel considerably better or less severely disabled.

An alternative approach employed in many psychotherapy clinical trials and meta-analyses is to measure symptom changes using effect sizes (ES), a summary statistic that denotes the magnitude of changes observed in any given measure. ES account for all cases in a given sample, regardless of their diagnostic status at baseline, making it an inclusive measure for service level data. ES account for symptom changes in either direction: improvement, deterioration or no change. ES can also be easily interpreted using conventional definitions of small (around .20), moderate (.50) and large (.80) effects (Cohen, 1998). Furthermore, given their prominence in the research literature, ES enable wide comparisons across studies. In summary, each of these outcome definition methods has its particular advantages, and therefore a comprehensive investigation of outcomes could combine ES and rates of reliable and clinically significant improvement (RCSI).

Comparing and evaluating clinical outcomes

A number of methods have been used to investigate the effectiveness of evidence-based treatments applied in routine mental health care. One method is "benchmarking", which consists of the statistical comparison of routine clinical outcomes against those of clinical trials as high efficacy benchmarks, or against ES observed in control or "no treatment" groups as lower benchmarks for no significant treatment effects (Lueger and Barkham, 2010; Minami, Serlin, Wampold, Kircher and Brown, 2008; Weersing and Weisz, 2002). Minami et al. (2007, Minami, Serlin et al., 2008; Minami, Wampold et al., 2008) have considerably advanced this method by proposing standard procedures to calculate ES and to statistically compare these to published benchmarks. For example, using meta-analytic methods they estimated an ES of 0.15 for "no treatment" control groups of depressed patients, and argue that ES within a critical value (equivalent to 1/5 of a standard deviation) above this estimate may not be significantly different. Similarly, they estimated an efficacy benchmark of 0.79 based on clinical trials for depression; and reason that ES observed in routine practice within the critical value (1/5 of a standard deviation below or above) should be considered clinically comparable.

Benchmarking methods have been employed in published evaluations of IAPT services (Clark et al., 2009; Glover et al., 2010; Richards and Suckling, 2009). The two earlier studies reflect the effectiveness of pilot services termed "demonstration sites", which reported large clinical ES (between 1.06 and 1.38 for depression; 0.98 to 1.41 for anxiety) comparable to those of published randomized controlled trials of similar interventions. The Glover et al. (2010) report describes the clinical outcomes of the first year of implementation of the IAPT programme in 30 routine services termed "roll-out" sites, which suggests that clinical ES are largely variable across sites (ranging from as low as 0.38 and as high as 1.09). However, methods differ across these reports with regards to aggregation of data, calculation of ES, and definition of outcomes. Furthermore, outcomes are compared to those of clinical trials, but no systematic comparisons to lower benchmarks are made.

In the absence of a consistent benchmarking method for routine services, conventional league table ranking methods have been used as a means of evaluating the performance of services against one another. An example is found in the National Audit for Psychological Therapies report (Royal College of Psychiatrists, 2011). This report analyzed data for over 100 services that had available pre and post-treatment outcome measures. Services were ranked using percentiles and quartiles according to the observed rate of cases meeting criteria for recovery based on conventional cut-offs on several diagnostic measures. Although this is a common method used to rank services, it is limited by its lack of attention to measurement error, and it also ignores the possibility that outcomes may not necessarily be statistically different between services that are ranked as more or less effective. The actual recovery estimates for a service ranked as below average (e.g. quartile below 50%) may not be significantly different to that of an "above average service" (e.g. above 50% percentile), and apparent differences may be simply due to variations in sample size or data completeness. More importantly, such methods do not provide any insight into the factors that may account for differential performance between services. Differences in socio-economic status, for example, have been shown to be associated with outcomes in naturalistic treatment settings, with patients paying for private insurance showing modestly higher improvement (Blais et al., 2012). Other areas of healthcare have advanced sophisticated benchmarking methods that adjust for differences in case-mix, which may partly account for outcome differences between services (e.g. Orkin, 2010). To the best of our knowledge such methods have not yet been used in the evaluation of routine psychological services.

Case example of a benchmarking method

In the following section, we present an example of how to generate benchmark values against which a specific service's performance can be evaluated, using published data from the IAPT programme.

Outcome measures used in benchmark calculations

The measures used for the calculation of benchmark values were restricted to the PHQ-9 and GAD-7, since these measures are routinely and widely collected as part of a national dataset for patients accessing IAPT services in England, reasonably enabling broad comparisons for clinical outcomes aggregated at service level.

The PHQ-9 is a 9-item measure of depression based on the *Diagnostic and Statistical Manual* (DSM-IV) diagnostic criteria for major depressive disorder (Kroenke et al., 2001). Scores range from 0 to 27, with higher scores indicating greater severity of depression, and a score of 10 or above on this measure has been proposed as indicative of meeting criteria for major depression (Gilbody, Richards and Barkham, 2007; IAPT National Programme Team, 2011; Kroenke et al., 2001).

The GAD-7 is a 7-item measure originally developed as a screen for Generalized Anxiety Disorder (Spitzer, Kroenke, Williams and Löwe, 2006). The capacity of the GAD-7 to detect other anxiety disorders, including social phobia, post-traumatic stress disorder and panic disorder, has also been established (Kroenke, Spitzer, Williams, Monahan and Löwe, 2007). Scores range from 0 to 21, with higher scores indicating greater severity of symptoms; a score above 8 is indicative of an anxiety disorder.

Method

Outcome benchmarks were estimated using data published by Glover et al. (2010), reporting ES for 30 "roll-out" sites that applied the IAPT programme in routine practice during the first year of implementation in England. Our strategy was to generate high, medium and low performance estimates based on the observed outcomes in a national sample of services. We then estimated clinical outcomes using contemporaneous data from one routine IAPT site in Leeds, and used this as a case example to illustrate how to evaluate one service against national benchmarks.

The calculations of benchmark values were based on cases in the datasets that: (a) had received at least two appointments, the first being a pre-treatment suitability assessment contact and the second being a treatment session; (b) had pre-treatment measures available that could be taken as a baseline; and (c) had been discharged from treatment. Using intent-to-treat principles, this included patients who may have dropped out in addition to those who completed treatment; the last observed measure was taken as a post-treatment rating. Pre-post treatment uncontrolled ES and confidence intervals were calculated using the formulas proposed by Minami, Serlin et al. (2008, pp. 517–520); this is a comparable measure to Cohen's *d* (Cohen, 1998), computed for repeated measures and weighted by sample size. An important innovation introduced by Minami and colleagues is the calculation of 95% confidence intervals with reference to "critical values" that are equivalent to 1/5 of a standard deviation. These authors reason that when comparing clinical effect sizes against efficacy benchmarks, statistically significant differences may not always be clinically important. The "critical value" factored into their confidence interval calculations ensures that comparisons between routine clinical outcomes and performance benchmarks should be statistically and clinically significant. Following these authors, standardized ES were aggregated using conventional meta-analytic methods and forest plots to calculate measure-specific estimates for all of the 30 IAPT roll-out sites into a summary figure for depression (PHQ-9) and another for anxiety (GAD-7). We calculated "average" benchmarks based on the mean ES for all 30 sites. Next, we calculated "high" and "low" performance benchmarks based on pooled ES for the four highest performing sites and four poorest performing sites. We selected four sites because this number represented the bottom eighth and the top eighth of the 30 sites. PHQ-9 and GAD-7 ES for the Leeds site were then statistically compared to the benchmark estimates.

Table 1. Leeds IAPT site data

Demographics $n = 2891$		
Mean age (*SD*)	37.9 (13.1)	
Female	64.3%	
White British	88.9%	
In employment	62.9%	
Off work due to ill health	10.6%	
Receiving state financial support	16.9%	
Referred by a GP	80.2%	
Scoring in the clinical range in PHQ-9 measure (depression)	73.4%	
Scoring in the clinical range in GAD-7 measure (anxiety)	80.2%	
Prescribed psychotropic medication	45.3%	
Mean therapy contacts (*SD*)	5.7 (3.9)	
Outcome data	Pre-treatment	Post-treatment
PHQ-9 mean (*SD*)	14.1 (6.4)	8.9 (6.9)
GAD-7 mean (*SD*)	12.6 (5.3)	7.9 (5.9)

Notes: Based on all discharged cases with $\geqslant 2$ contacts and available pre/post treatment data; proportions could only be estimated for the subset of valid data records. GP = general medical practitioner

As recommended in the preceding discussion, we calculated reliable and clinically significant improvement rates (RCSI; Jacobson and Truax, 1991) as an additional performance indicator. This could only be estimated for the Leeds site, since we only had access to case-level data for this service, and there is no reference to RCSI in Glover et al. (2010). Following Richards and Borglin (2011), our definition of RCSI for depression required a patient with a PHQ-9 $\geqslant 10$ at baseline to have (a) reduced this score by at least 6 points and (b) to have a post-treatment score < 10. For GAD-7, a score $\geqslant 8$ was taken as a cut-off for a diagnosis of an anxiety disorder, and RCSI required (a) a reduction of at least 5 points and (b) a post-treatment score < 8.

Results

A brief contextual description of the Leeds site is presented first, followed by a comparative analysis of these data against national benchmark values. The Leeds IAPT site offers evidence-based psychological therapy organized in a stepped-care model, following national treatment guidelines (NICE 2007a, b, 2011). This service received referrals for 14,453 people between the March 2009–2010 data collection period, 5783 of whom accessed therapy following assessment. A total of 2891 cases that were discharged from the service at the time of data collection had both pre and post-therapy outcome measures available, and were included in data analysis (77% data completeness). Treatment options at the time of this study included cognitive behavioural therapy (CBT), computerized CBT, brief guided self-help, and psycho-education groups. Approximately 86% of patients were offered low intensity (step 2 in the stepped-care model) treatment options with a mean duration of 4.8 ($SD = 2.64$) sessions. Approximately 13% were offered high intensity CBT (step 3) with a mean duration of 9.9 ($SD = 6.72$) sessions, and less than 1% were referred to long-term psychotherapy in secondary care (step 4). Table 1 describes basic demographic characteristics for the Leeds dataset, which

are largely comparable to the data reported in the benchmark report cited above in terms of gender, ethnic mix, employment status and diagnostic characteristics.

Forest plots of depression (PHQ-9) and anxiety (GAD-7) ES are presented in Figures 1 and 2, which represent the clinical performance of 30 IAPT roll-out sites and the Leeds site for comparison. These figures also overlay estimated high, average and low performance benchmarks. Considerable variation in sample size and ES across sites was evident, as well as varying degrees of precision in these calculations, denoted by the width of confidence intervals (which include critical values weighted by sample size). Amongst the 30 roll-out sites, the Leeds site's ES were ranked in 8th place, or first quartile (25.8%) taking $n = 31$ as denominator. The aggregated data estimates based on which ES were calculated can be found in the Glover et al. (2010) report and are therefore not presented in this paper. The relevant estimates for Leeds PCMHS are in Table 1.

Estimates of heterogeneity of ES for roll-out sites were calculated using the Q test (Cochran, 1954) and the related I^2 statistic to provide a measure of the percentage of variation across sites (Higgins, Thompson, Deeks and Altman, 2003). For depression (PHQ-9), Q (29) = 313.5, $I^2 = 91\%$, $p < .001$; for anxiety (GAD-7), Q (29) = 369.8, $I^2 = 92\%$, $p < .001$. All were highly significant ($p < .001$) with I^2 much higher than the 75% value generally taken as an index of high variation, indicating large heterogeneity between sites.

Figure 3 provides a graphical summary of key performance indicators, using the Leeds site data to illustrate the benchmarking of routine clinical outcomes. Vertical lines in this graph represent the continuum of ES for PHQ-9, and GAD-7, with overlaid high, average and low performance benchmarks for national services. The Leeds site's ES for depression and anxiety (PHQ-9 = 0.81 / GAD-7 = 0.90) were significantly larger than low (0.46 / 0.52) and average (0.73 / 0.78) benchmarks, but significantly smaller than high level benchmarks (0.91 / 1.02). This was determined based on the null hypothesis that if the benchmark estimate is contained within the confidence intervals of the ES for the routine service, there is no significant difference between estimates (following Minami, Serlin et al., 2008). Furthermore, the Leeds site's ES were significantly different to the "no treatment" lower level benchmark (0.15 ES) proposed by Minami, Serlin et al. (2008), but not statistically different to the efficacy benchmark from clinical trials (0.79 ES). Finally, reliable and clinically significant improvement rates (RCSI) for the Leeds site were 42.2% for depression and 43.5% for anxiety. These estimates were closely comparable to depression (41%) and anxiety (40%) RCSI estimates reported by Richards and Borglin (2011) for the IAPT demonstration site.

Discussion

This paper presents a considered review of challenges and methods for the selection of an adequate set of psychometric measures, the operational definition of outcomes, and the statistical comparison and evaluation of clinical performance. This study used historical outcome estimates available in the public domain; the intention was to exemplify a robust method that could be employed to analyze more current datasets such as those generated routinely by the English IAPT programme.

There are some limitations concerning the benchmarking method presented in this paper. Data quality and completeness appeared to vary considerably across IAPT roll-out sites, with a wide range of completion rates ranging from 21.4% to 100% (Glover et al., 2010). Data completion rates will likely lead to biases in the outcome data, as noted by Clark et al.

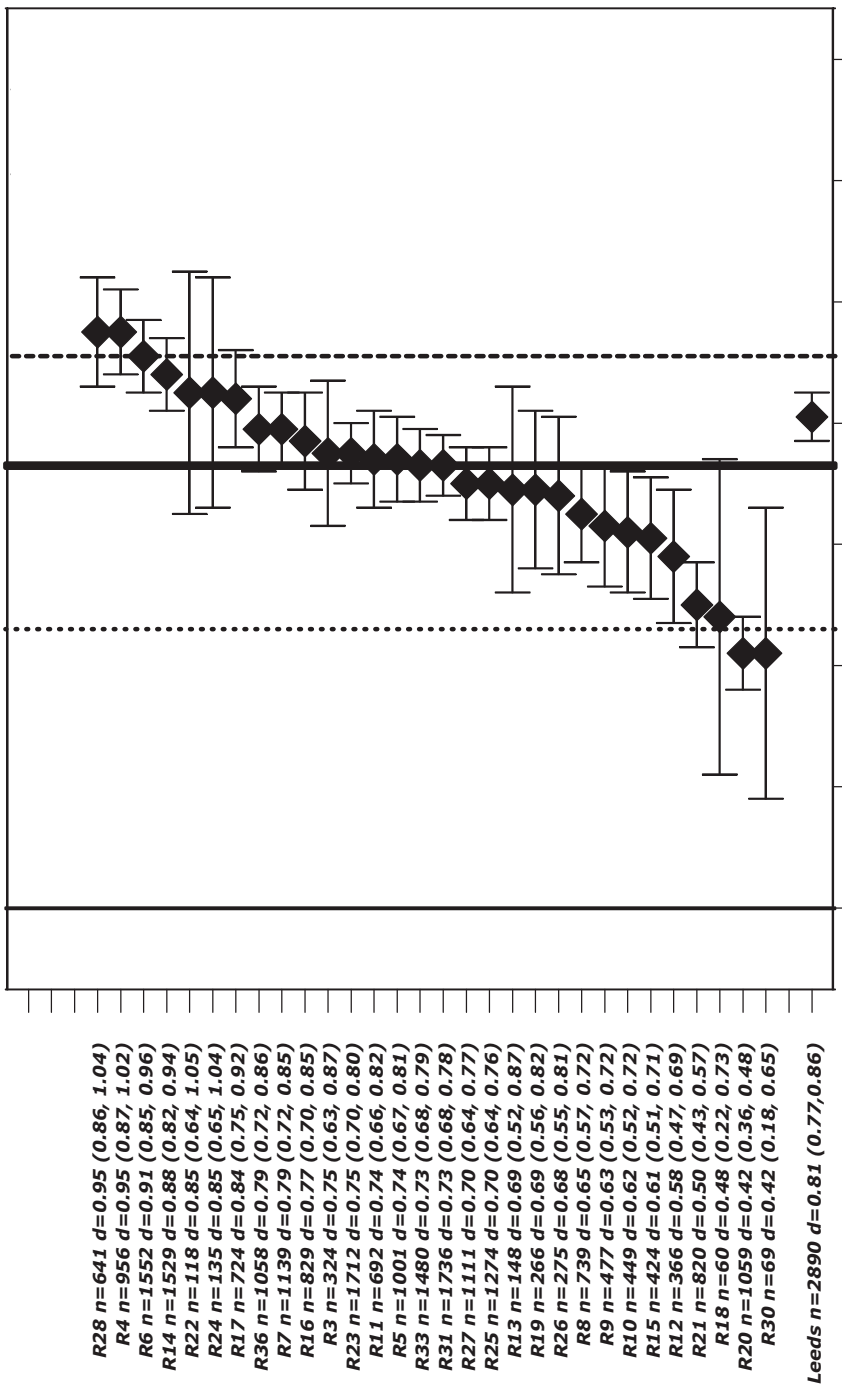
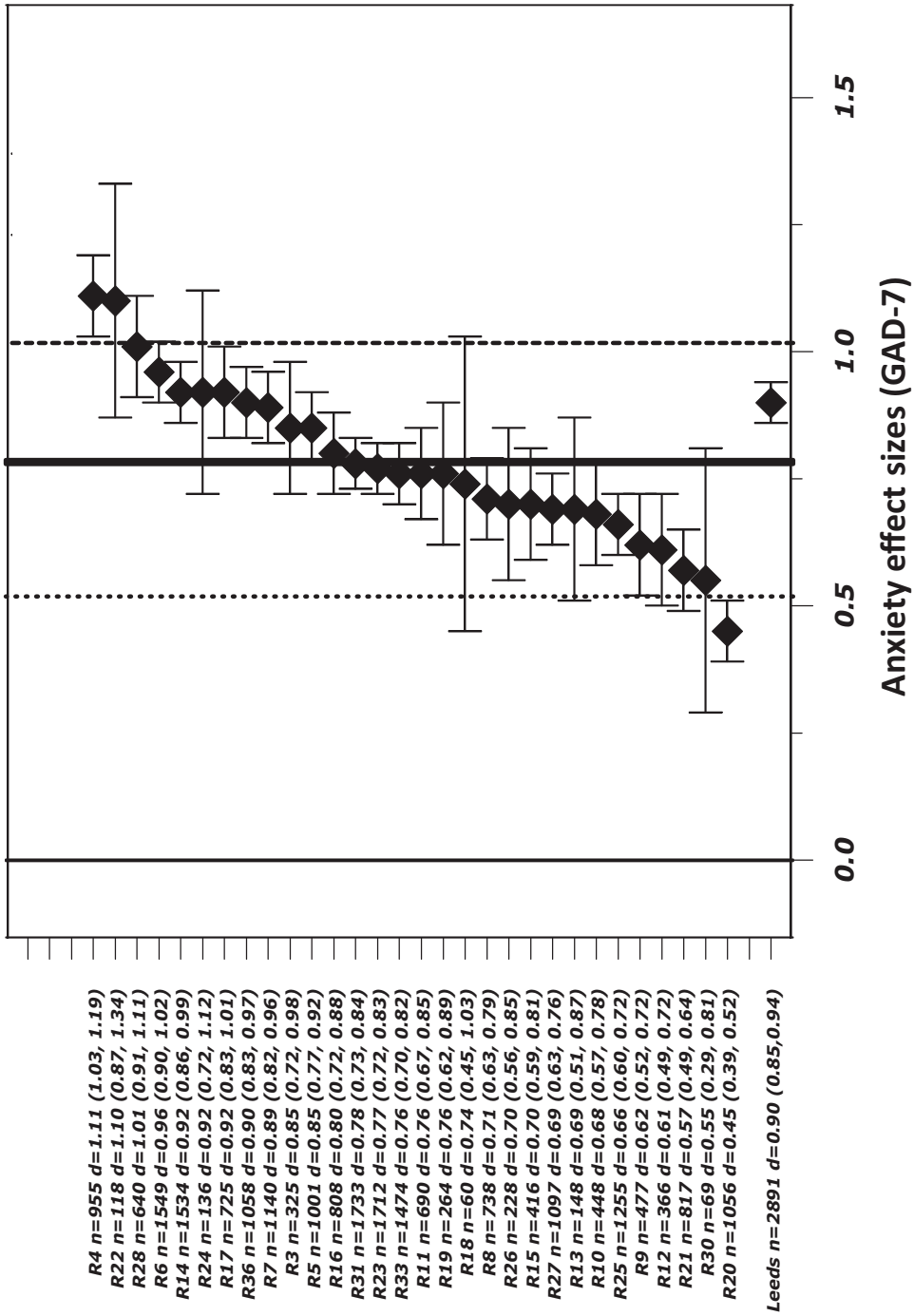


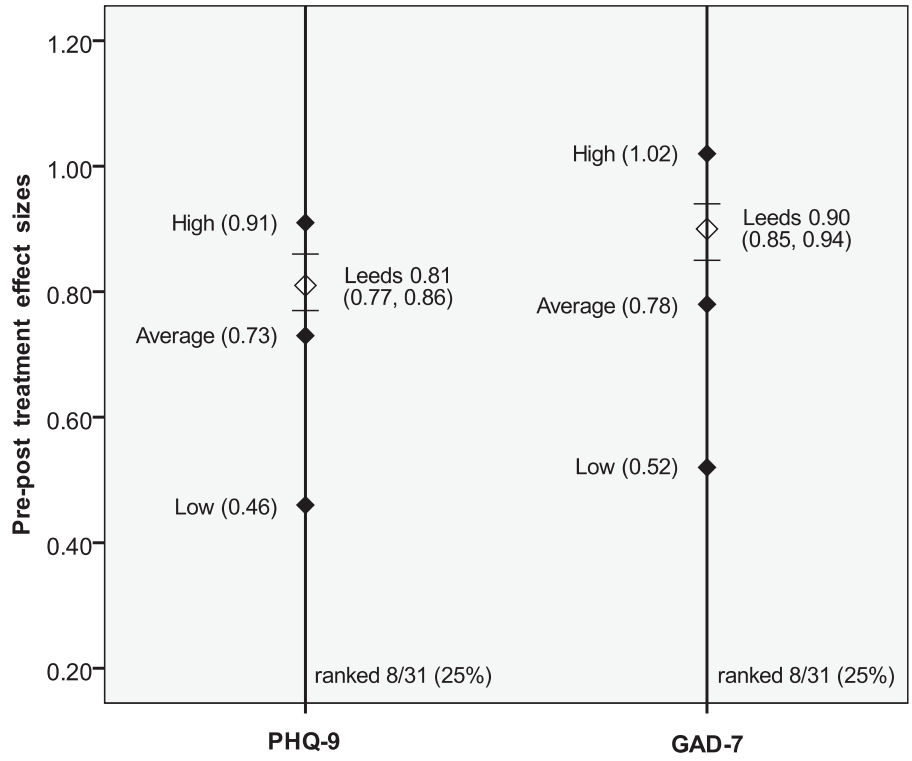
Figure 1. Forest plot of PHQ-9 effect sizes (and 95% CI) for IAPT roll-out sites (R), and Leeds site

Notes: Solid line = average benchmark (0.73); dotted lines = low (0.46) and high (0.91) performance benchmarks



Notes: Solid line = average benchmark (0.78); dotted lines = low (0.52) and high (1.02) performance benchmarks

Figure 2. Forest plot of GAD-7 effect sizes (and 95% CI) for IAPT roll-out sites (R), and Leeds site



Valid *n*	2890	2891
Data completeness	77%	77%
n cases >= cutoff	2125	2320
RCSI	42.2%	43.5%

Figure 3. Benchmarking of key performance indicators for Leeds IAPT site

(2009). The very high degree of heterogeneity between sites in ES (over 90%) should also be considered and researched. This may be due to the fact that sites in this historical dataset were at very different stages of development and implementation, and may also relate to completion rates mentioned above. However, the reasons for this high degree of variation will be complex and we have merely highlighted some factors likely to account for this.

It is important that the performance of a service is not judged simply in terms of its placement relative to the different benchmark values. Although mediation analyses were not possible to present in this benchmarking report since case-level data were not available to do this, it is noted that other authors have already described factors associated with differential outcomes in IAPT services. For example, Gyani, Shafran, Layard and Clark (2011) argue that variability in the performance of IAPT roll-out sites was related to differences in baseline severity, proportions of patients stepped up to high intensity treatment, greater length of treatment, and delivery of NICE recommended disorder-specific treatments for patients in

high intensity therapy. In this context, the relative placement of the Leeds site in amongst the benchmark values may be partly explained by its higher mean number of treatment sessions and higher degree of treatment fidelity[2] at step 3 compared to the "average" IAPT site reported in Glover et al. (2010).

Furthermore, as previously described, it is possible that the demographic and diagnostic mix varies across services, and this may account for some differences in clinical effect. Benchmarking methods that adjust for case-mix (e.g. accounting for baseline severity, socio-economic status, chronic health and comorbidity) may be developed as these data emerge in the future. Future benchmarking developments could also compare the relative ES estimates obtained with disorder-specific measures versus those obtained with more general distress measures. Clearly all of these factors should be considered when evaluating the effectiveness of a service and one of the concerns of benchmarking is that it could be used in a simplistic way, ignoring contextual, diagnostic and population factors such as those described above.

References

Blais, M. A., Sinclair, S. J., Baity, M. R., Worth, J., Weiss, A. P., Ball, L. A., et al. (2012). Measuring outcomes in adult outpatient psychiatry. *Clinical Psychology and Psychotherapy*, *19*, 203–213.

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Christoph, P., et al. (1998). Update on empirically validated therapies, II. *The Clinical Psychologist*, *51*, 3–16.

Chambless, D. L. and Ollendick, T. H. (2001). Empirically supported psychological interventions: controversies and evidence. *Annual Review of Psychology*, *52*, 685–716.

Clark, D. M., Layard, R., Smithies, R., Richards, D. A., Suckling, R. and Wright, B. (2009). Improving access to psychological therapy: initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy*, *47*, 910–920.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129.

Cohen, J. (1998). *Statistical Power Analysis for the Behavioural Sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.

Das-Munshi, J., Goldberg, D., Bebbington, P. E., Bhugra, D. K., Brugha, T. S. and Dewey, M. E. (2008). Public health significance of mixed anxiety and depression: beyond current classification. *British Journal of Psychiatry*, *192*, 171–177.

Dowrick, C., Leydon, G. M., McBride, A., Howe, A., Burgess, H., Clarke, P., et al. (2009). Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. *British Medical Journal*, *338*, b663.

Evans, C., Margison, F. and Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence-based Mental Health*, *1*, 70–72.

Franklin, M. E. and DeRubeis, R. J. (2006). Are efficacious laboratory-validated treatments readily transportable to clinical practice? In J. C. Norcross, L. E. Beutler and R. F. Levant (Eds.), *Evidence-Based Practices in Mental Health: debate and dialogue on fundamental questions* (pp. 375–383). Washington DC: American Psychological Association.

[2]At the time of this data collection, Step 3 interventions in the Leeds site comprised exclusively of CBT, whereas Glover and colleagues estimated that the "average" IAPT roll-out site offered 57.8% CBT and 50.1% Counselling at step 3.

Gilbody, S., Richards, D. and Barkham, M. (2007). Diagnosing depression in primary care using self-completed instruments: UK validation of PHQ–9 and CORE–OM. *British Journal of General Practice*, *57*, 650–652.

Glover, G., Webb, M. and Evison, F. (2010). *Improving Access to Psychological Therapies: a review of the progress made by sites in the first rollout year*. Stockton on Tees: North East Public Health Observatory.

Gyani, A., Shafran, R., Layard, R. and Clark, D. M. (2011). *Enhancing Recovery Rates in IAPT Services: lessons from analysis of the Year One data*. London: University of Reading, London School of Economics and Kings College London.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J. and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560.

IAPT National Programme Team (2011). *The IAPT Data Handbook: guidance on recording and monitoring outcomes to support local evidence-based practice*. Version 2.0. London: National IAPT Programme Team.

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

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., et al. (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association*, *289*, 3095–3105.

Kroenke, K., Spitzer, R. L. and Williams, J. B. W. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*, 606–613.

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O. and Löwe, B. (2007). Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine*, *146*, 317–325.

Kroenke, K., Spitzer, R. L., Williams, J. B. W. and Löwe, B. (2010). The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *General Hospital Psychiatry*, *32*, 345–359.

Larsen, R. J. and Fredrickson, B. L. (1999). Measurement issues in emotion research. In D. Kahneman, E. Diener and N. Shwarz (Eds.), *Well-being: the foundations of hedonic psychology* (pp.40–60). New York: Russell Sage Foundation.

Lueger, R. J. and Barkham, M. (2010). Using benchmarks and benchmarking to improve quality of practice and services. In M. Barkham, G. E. Hardy and J. Mellor-Clark (Eds.), *Developing and Delivering Practice-Based Evidence*. Chichester: Wiley.

McAleavey, A. A., Nordberg, S. S., Kraus, D. and Castonguay, L. G. (2012). Errors in treatment outcome monitoring: implications for real-world psychotherapy. *Canadian Psychology*, *53*, 105–114.

McCaffrey, R. J. and Westervelt, H. J. (1995). Issues associated with repeated neuropsychological assessments. *Neuropsychology Review*, *5*, 203–221.

McManus, S., Meltzer, H., Brugha, T., Bebbington, P. and Jenkins, R. (2009). *Adult Psychiatric Morbidity in England, 2007: results of a household survey*. Retrieved September 7, 2010 from http://www.ic.nhs.uk/pubs/psychiatricmorbidity07

McMillan, D., Richards, D. and Gilbody, S. (2010). Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. *Journal of Affective Disorders*, *127*, 122–129.

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

Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C. and Brown, G. S. (2008). Using clinical trials to benchmark effects produced in clinical practice. *Quality and Quantity*, *42*, 513–525.

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S. and 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.

National Institute for Health and Clinical Excellence (2007a). *Anxiety (amended): management of anxiety (panic disorder, with or without agoraphobia, and generalized anxiety disorder) in adults in primary, secondary and community care*. London: NICE.

National Institute for Health and Clinical Excellence (2007b). *Depression (amended): management of depression in primary and secondary care*. London: NICE.

National Institute for Health and Clinical Excellence (2011). *Common Mental Health Disorders: identification and pathways to care*. London: National Collaborating Centre for Mental Health.

National Screening Committee (2003). *The UK National Screening Committee's Criteria for Appraising the Viability, Effectiveness and Appropriateness of a Screening Programme*. London: HMSO.

Newnham, E. A. and Page, A. C. (2010). Bridging the gap between best evidence and best practice in mental health. *Clinical Psychology Review*, *30*, 127–142.

Orkin, F. (2010). Risk stratification, risk adjustment, and other risks. *Anesthesiology*, *113*, 1001–1003.

Richards, D. A. and Suckling, R. (2009). Improving access to psychological therapies: Phase IV prospective cohort study. *British Journal of Clinical Psychology*, *48*, 377–396.

Richards, D. A. and Borglin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: two year prospective cohort study. *Journal of Affective Disorders*, *133*, 51–60.

Roth, A. and Fonagy, P. (2004). *What Works for Whom? A critical review of psychotherapy research (2nd edn).* New York: Guilford Press.

Roth, A. D. and Pilling, S. (2007). *The Competences Required to Deliver Effective Cognitive and Behavioural Therapy for People with Depression and with Anxiety Disorders*. London: Department of Health. Retrieved February 10, 2012 from http://www.ucl.ac.uk/clinical-psychology/CORE/CBT_Framework.htm#Map

Royal College of Psychiatrists (2011). *National Audit of Psychological Therapies for Anxiety and Depression, National Report 2011*.

Shimokawa, K., Lambert, M. J. and Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology*, *78*, 298–311.

Siev, J., Huppert, J. and Chambless, D. L. (2009). The dodo bird, treatment technique, and disseminating empirically supported treatments. *The Behavior Therapist*, *32*, 69–75.

Spitzer, R., Kroenke, K., Williams, J. B. W. and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *66*, 1092–1097.

Wahl, I., Meyer, B., Löwe, B. and Rose, M. (2010). Measurement of patient reported outcomes in psychotherapy research. *Journal of Psychosomatic Research*, *68*, 676.

Weersing, V. R. and Weisz, J. R. (2002). Community clinic treatment of depressed youth: benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology*, *70*, 299–310.