Published paper
Morley, S., Williams, A. and Hussain, S. (2008) *Estimating the clinical effectiveness of cognitive behavioural therapy in the clinic: Evaluation of a CBT informed pain management programme,* Pain, Volume 137 (3), 670 - 680.

Estimating the clinical effectiveness of cognitive behavioural therapy in the clinic: evaluation of a CBT informed pain management programme

Stephen Morley

University of Leeds and St James' University Hospital Leeds

Amanda Williams

University College London

Sumerra Hussain

University of Leeds

Correspondence

Stephen Morley, Institute of Health Sciences, University of Leeds, 101 Clarendon Road, Leeds, LS2 9LJ, UK. Tel: +44 113 343 2722; Fax: +44 113 243 3719; Email: s.j.morley@leeds.ac.uk

Abstract

Randomized controlled trials and meta-analyses provide evidence for the *efficacy* of cognitive-behaviourally informed treatment (CBT) programmes for chronic pain. The current study aims to provide *practice-based evidence* for the effectiveness of CBT in routine clinical settings. Over a 10 year period 1013 pain patients were accepted into a 4 week in-patient pain management programme. Data from more than 800 patients was available at pre-treatment and at one month post-treatment and for around 600 patients at pre-treatment and at 9 months follow-up. Measures reported in this analysis were pain experience and interference, psychological distress (depression and anxiety), self-efficacy, catastrophizing, and walking. Change from pre-treatment to post-treatment and follow-up was assessed with conventional statistical tests, the computation of effect sizes and by the reliable change index (RCI) and clinically significant change (CSC) methodology. These analyses provide evidence of statistical improvement at post-treatment and follow-up and the RCI/CSC methodology suggested that between 1 in 3 and 1 in 7 (depending on the outcome measure) achieved clinically significant gains. There was also evidence that a small percentage of patients (1% - 2%) reliably deteriorated during the period of treatment. The limitations in the inferences that can be drawn from this study and of the methodology are discussed. A case is made for the application of benchmarking methods using data from RCTs in order to more fully evaluate practice and to generate better quality practice based evidence.

1. Introduction

The practice of evidence based medicine is espoused by many governments and health care organisations. The aims of evidence based medicine include the allocating of effective treatment in a timely and efficient manner and, increasingly in socially regulated health care systems, managing access to scarce resources. The basis of evidence based medicine are high quality randomized controlled trials (RCTs) and integration of their results through systematic review and meta-analysis [41; 42]. Psychologically based treatments for chronic pain as a generic condition, or for specific diagnostic groups where pain is a significant symptom and problem e.g. rheumatoid and osteo-arthritis, have developed over 40 years; many RCTs and several meta-analyses attest to treatment efficacy under controlled conditions in comparison with both wait list and control treatments [44; 12; 27].

While RCTs and meta-analysis have been used to establish the *efficacy* of psychological treatments they do not necessarily provide evidence of *effectiveness:* whether the treatment provides a measurable beneficial effect when delivered to patients in other service contexts [4]. Generalising from RCTs to other contexts is problematic where there is less control over, for example, selection and recruitment of patients, implementation of treatment to a manualized standard, or checks on the adherence and delivery of the therapeutic protocol [55]. What is required is evidence from clinical practice, i.e. practice based evidence such as that described by Barkham and Mellor-Clark [4], for the *effectiveness* of treatment. These authors and others argue that a model of research into efficacious and effective treatment should be based on a cycle in which evidence based practice (EBM) informs clinical practice which then generates evidence and questions (practice based evidence) for testing under more controlled conditions of EBM.

The aim of this article is twofold: It seeks to generate evidence from routine clinical practice for the effectiveness of a cognitive behavioural pain management programme and in doing so it identifies issues concerning the way in which the majority of RCTs in this field measure and report outcomes. Most trials of psychological treatments use continuous measures and the subsequent analyses focus on differences between group means at the end of treatment (suitably controlled for pre-treatment differences and other confounds) that are evaluated with conventional inferential statistics (*P* values). Aside from the fact that *P* values are sensitive to sample size, the test statistics are not generally referenced to any external criterion. This methodology is not useful for evaluation of treatment where there is no control group, as is typical in clinical practice. In these situations, pre-post treatment differences may be computed but provide no information about the clinical significance of changes. In the current study we employed Reliable Change Index (RCI) / Clinically Significant Change (CSC) methodology [30; 29] to evaluate a 4 week in-patient CBT rehabilitation programme for chronic pain.

2. Methods

The data reported in this study were collected as part of routine procedure by the clinic. The data were available in an anonymised database. Permission to access and interrogate the database using an explicit protocol was given by the relevant UK NHS Trust research ethics committee.

2.1. Overview of sample and data collection

Between 1989 and 1998, 2041 patients attended a publicly funded (UK National Health Service) pain management service. The service is a specialist tertiary one that receives referrals from other pain clinics, general practice, orthopaedic and other specialist services from around the UK. Primary assessment was made by anaesthetists and clinical psychologists. The exclusion criteria were: inability to speak English, inability to climb stairs, current major psychiatric disorder (active psychosis, severe depression with high risk of suicidal attempt), suitable for further medical treatments following examination, pain duration less than one year, current opioid misuse (either chronic illegal opioid use or in a methadone maintenance programme). Patients were required to meet two of the following inclusion criteria: widespread disruption in non-work and/or work activity due to pain, habitual over-activity leading to increased pain, use of high levels pain medication with little reported benefit, high affective distress score, unnecessary use of medical aids, high levels of pain behaviour. Patients completed a package of assessments (see below) administered by assistant psychologists and others who were not part of the treatment team. During the period 1989-1998, several treatment options were trialled including 2 and 4 week programmes with in-patient and out-patient options. This article reports the outcomes for patients who attended a 4 week long in-patient programme and provided data at baseline (pre-treatment). Data were collected pre-treatment, one month post-treatment (post-treatment) and at 9 months follow-up (follow-up). There is variation in the number of data points available for each of the measures for two reasons: attrition at post-treatment and follow-up, and because over the 10 year period the programme occasionally changed measures, supplementing or replacing measures as they were improved. The sample sizes varied between 833 for pre-treatment/post-treatment comparisons and 527 for pre-treatment/follow-up comparisons.

2.2. Programme description

Williams et al (1999: p.60) [57] have previously described this programme as follows: "The unit was staffed by a consultant anaesthetist, two clinical psychologists, a physiotherapist, an occupational therapist, a senior nurse, and a secretary/administrator. The program was based on the work of Fordyce [20], Keefe and colleagues [23], [47], and Turk et al. [53] and incorporated operant and cognitive behavioural principles in all aspects. No other active treatments (such as nerve blocks or acupuncture) were offered once patients were accepted for the program. Treatment was carried out in hospital premises with hostel-type accommodation for in-patients who lived independently outside program hours and returned home at weekends. The in-patient programme was carried out over four weeks, four and a half days per week.

The program consisted of the following components, all supported by written materials: education concerning pain, disuse, drugs, and sleep; exercise routines for fitness, flexibility and muscle minimum strength, increasing gradually on a quota system; goal setting across all activities with quota increases and activity-rest scheduling (pacing); psychology sessions to improve problem solving, change maladaptive behaviours and to maintain those changes, with cognitive techniques to identify unrealistic and unhelpful thoughts and beliefs, and to challenge and change them; drug reduction applied to all pain-related drugs which had neither achieved analgesia nor improved function, with the usual aim of abstinence by discharge; applied relaxation;

relapse prevention and planning for crises; and sleep hygiene. Spouses and family members (where available) were involved in a small number of sessions."

2.3. Measures

The measures described in this section include those for which data was consistently collected in the 10 year period. During this period several other measures were trialled and either excluded or too few data were available for analysis. In order to obtain estimates of reliability and to set appropriate cut scores needed to determine RCI and CSC criteria, we carried out systematic searches of the literature for each measure. The results of these searches are briefly summarised for each measure. Computation of the RCI requires the standard deviation of the measure (obtainable from the sample) and an estimate of the reliability of the measure. Reliability data was obtained from published reports and wherever possible we used measures of internal consistency (Cronbach's α or the intraclass correlation coefficient). The measures of pain experience were single item measures for which internal consistency assessments are not possible: in these cases we used reliability based on stability (test-retest). The type of reliability coefficient and the values used in this study are shown in Table 1.

Jacobson's [29] determination of a clinically significant criterion employs statistical criteria to establish cut scores for continuous variables that essentially use the properties of the normal distribution. The criteria (*a*), (*b*) and (*c*) are defined as follows: (*a*) is achieved when the post-treatment (or follow-up) score lies outside of the range of the dysfunctional population, where the range is defined as extending 2 SD units beyond the mean for that population in the direction of a functioning population; (*b*) is achieved when the post-treatment (or follow-up) score lies within the range of the functioning population, where the range is defined as within 2 SD units of the mean of the functioning population; and (*c*) is defined as when the post-treatment (or follow-up) score is statistically more likely to be in the functional population – i.e. nearer to the mean of the functional population than the mean of the dysfunctional population. For practical reasons the sample of participants was regarded as the population. We then used the following rules to determine which criterion to apply for each measure. If distributional data were available from a suitable normative (functional) control sample we elected to use criterion (*c*). When no control sample was available we used criterion (*a*). In addition to Jacobson's criteria we used a normative criterion (*n*), determined by reference to established cut scores for the test. For example, the authors of the Hospital Anxiety and Depression Scale (HADS) [59] indicate that a score of 8 on each subscale may be regarded as a reasonable cut score for defining the presence of clinical levels of depression and anxiety. Details of the selected cut score are provided for each measure and a summary is shown in Table 1.

2.3.1. Pain experience (intensity, distress and interference)

Average pain intensity, pain distress and interference attributed to pain were measured on 0-100 (101 point) numerical rating scales (NRS) [32]. The literature search revealed several studies comparing the relative performance of scale types for these three constructs but information on test-retest reliability was sparse. The preferred internal consistency measure of reliability cannot be used on these measures as they are single item measures. Zautra et al. [58], reported 2-week test-retest correlations for the 101-point scale of 0.69 in a sample of women with fibromyalgia. Similar searches for distress and interference failed to produce clear guidance and we therefore adopted the 0.69 estimate of reliability.

Several recent studies have sought to establish criteria for judging clinically significant changes in pain intensity [17; 19; 8; 9; 18; 11; 25] in both acute and chronic pain and the value converges at about 30% change from the initial score. We therefore adopted this criterion for a clinically significant change for all three measures based on the NRS scales. Unlike the other CSC used in this study, the criterion for these three measures is ipsatized to the patient's baseline score and not set at a discrete score on the scale.

2.3.2. Beck Depression Inventory

The original 21 item BDI was used. We adopted $\alpha = 0.81$ as the reliability coefficient from a review of the scale by [5]. A literature search revealed several studies comparing pain patients with and without depression as defined by DSM criteria. We identified one by [21] which used a sample of chronic pain patients with similar characteristics as the current sample. The data from Geisser at al. [20] was therefore used to establish a clinically significant change criterion (c) cut score of 19.68.

2.3.3. Hospital Anxiety and Depression Scale (HADS)

The HADS was developed as a measure of depression and anxiety to be applied in medical settings where patients are likely to have somatic symptoms attributable to a primary somatic illness rather than their mental health. The 14 items (7 each for depression and anxiety) therefore eschew reference to somatic manifestations of the target psychological states. The HADS has been widely used in many settings and has good internal consistency. Zigmond and Snaith [59] originally suggested that scores above 8 on each of the scales was indicative of clinical levels of anxiety and depression. Lowe et al. [37] have shown that this cut score appears to give optimal specificity and sensitivity.

2.3.4. Coping Skills Questionnaire Catastrophizing subscale (CSQ-Cat)

The CSQ is widely used to assess various aspects of coping. The original scale devised by Rosenstiel and Keefe [49] comprises 42 items and 7 subscales. Twenty years of research has established the critical role that catastrophizing plays in the maintenance of distress in chronic pain [50]. Reduction of catastrophizing is correlated with better emotional and behavioural outcomes in psychological treatment programmes [33; 6]. We therefore selected this subscale as a key indicator of cognitive change. The catastrophizing subscale consists of six items that address negative thoughts about pain as well as catastrophic thinking. Individuals are asked to rate the frequency of thoughts when they experience pain using a 7-point scale ranging from 0 (never) to 6 (always). The six items on the CSQ catastrophizing scale have been shown to have good internal reliability as well as a high degree of stability over time [35]. We selected the original estimate of $\alpha = 0.78$ [47] as the reliability coefficient and used Jacobson's (*c*) criterion to establish a clinical significant cut score.

2.3.5. Pain Self Efficacy Questionnaire (PSEQ)

The 10 item self-report PSEQ was developed as a pain specific assessment of Bandura's concept of self-efficacy; a person's beliefs about their ability to perform a particular behaviour in the face of "obstacles and aversive experiences" [3]. Examples of items include "I can still do many of the things I enjoy doing, such as hobbies or leisure activity, despite the pain". Respondents rate their belief in their ability to perform 10 activities on a seven-point scale ranging from 0 (not at all confident) to 6 (completely confident). Scores range from 0 to 60 with higher scores indicating stronger self-efficacy beliefs. The PSEQ has been shown to have good test-retest reliability and internal consistency [22; 1]. Asghari and Nicholas [1] report an internal consistency value $\alpha = 0.92$ which was used to assess the RCI. In the absence of a suitable control groups criterion (*a*) was used to set a clinically significant cut score.

2.3.5. 5-minute walk

The 5-minute walk test from a battery used to assess physical performance was selected as a behavioural outcome measure. The performance tests [26] were designed develop and normed by personnel involved in the pain management programme. The 5-minute walk test was derived from a longer 10-minute version and involves the patient walking up and down a corridor for the prescribed time. Patients were allowed to use the walls for support or to sit down where necessary and the number of meters walked was noted. Harding et al. [26] report excellent reliability for the 10-minute walk test (ICC = 0.94) and report that the 5-minute version is a reasonable alternative as both are highly correlated ($r = 0.99$). Jacobson's criterion (*a*) was used to set a clinically significant cut score.

2.4. Data Analysis

Possible bias attributable to attrition was checked by comparing the sub-sample available at each stage with drop-out sub-samples on demographic characteristics and the measure in question. Outcomes of the programme were evaluated by computing conventional inferential statistical tests and estimates of effect size, reliable change indices and the proportion of individuals achieving clinically significant change.

3. Results

3.1. Sample description

At pre-treatment 61.8% of the sample were female. Information on ethnic origin was available for 997 patients: the majority was white Caucasian (90.1%), with 3.8% Afro-Caribbean and 2.9% Asian. The average age of the sample was 45.7 years (SD = 11.7; range 18-84 years; median 46 years). Information on employment and social economic status was available for 994 individuals. Of these 6.1% were employed, a further 6.8% were employed but restricted by their pain, 27.7% were unemployed and this was largely attributed to pain, a substantial number were classed as permanently sick/disabled (38.9%); the remaining patients were retired (11.8%), homemakers (5.1%) or students (0.7%). With respect to occupational status 50.1% were classified as skilled manual/clerical, 25.9% as professional/managerial and 11.3% as unskilled manual. The majority of patients were receiving sickness or disability related income (70.4%) and for some (34.6%) this was the main source of income. Information regarding previous and current legal action was available for 956 patients. 20.1% were or had been involved in legal action; 74.2% of the sample was not involved in legal action.

The primary sites of pain were low back (62.7%), shoulder and upper limb (10.0%), lower limb (9.4%), neck (6.5%), abdomen (2.8%), head (3.0%), chest (2.2%), perineum, rectum and genital area (2.6%) and pelvis (1.5%). The mean time since onset was 113.2 months (SD = 108.3; range = 3-677; median 74 months). The majority were taking medication (95.6%).

3.2. Attrition

There were data on at least one measure pre-treatment for 1013 patients. Post-treatment data were available for between 720 and 833 patients depending on the measure, and at follow-up the range of sample sizes varied between 527 and 639. Information on reasons for attrition was not available. There were no systematic biases attributable to demographic characteristics or baseline values of the key measures.

3.3. Conventional statistical tests and effect size estimates

Table 2 shows the mean values and standard deviations for each variable at each time point and the sample size at each point of measurement. The reported values for the pre-treatment data are based on the sample sizes used in the pre-treatment to post-treatment comparisons and these values were not significantly different to the values obtained when the reduced sample available for pre-treatment to follow-up comparisons were made. Conventional statistical tests were conducted with repeated measures ANOVA and t-tests. All pre-treatment to post-treatment and pre-treatment to follow-up comparisons involved large sample sizes ($n > 500$) and were significant with P levels beyond .001.

The magnitude of pre-treatment to post-treatment and pre-treatment to follow-up differences was also expressed as a series of effect sizes and their 95% Confidence Intervals (95%CI). Table 2 also displays the correlations for each measure across the occasions of testing and the corresponding pair-comparison t-test values. These data were used to compute estimates of effect size (d) and its variance ($\mathrm{var}(d)$) for correlated data [14]:

$$d = t_c\,[2(1-r)/n]^{½} \quad \text{and} \quad \mathrm{var}(d) = [2(1-r)/n] + [d^2/(2n-2)]$$

Where, t_c = the value of correlated t-test statistic; r = correlation, n = sample size.

The estimates of the effect sizes and their 95% confidence intervals for both pre-treatment to post-treatment and pre-treatment to follow-up comparisons are displayed in Figure 1. The range of values for the pre/post comparison, displayed as a filled circle, was 0.26 (pain intensity) to 0.73 (PSEQ) with a median value of 0.51. The corresponding values for the pre-treatment to follow-up comparisons, displayed as an open circle, were 0.27 (pain intensity), 0.62 (PSEQ) with a median value of 0.41. In every case the lower bound confidence limit for the effect size was greater than 0 indicating that the mean value of the sample is reliably greater than the point of no change.

3.4. Reliable Change Index (RCI) and Clinically Significant Change (CSC)

The RCI and Jacobson's CSC criterion (*a*) was computed for each measure using the available reported reliability coefficients cited in the Measures section (above) and the standard deviation of the sample at pre-treatment assessment. The combination of RCI and CSC criteria enables patients to be classified into a number of outcome groups: these are represented schematically in a 'tramline display' shown in Figure 2 [43]. The figure schematically represents the scores from a measure obtained at pre-treatment and post-treatment, in which a reduction in the score represents an improvement. The main diagonal (dotted line) represents scores that do not change during the period of measurement. The symmetrical parallel lines either side of the main diagonal represent the upper and lower boundaries of the RCI which, in this case have been set with a confidence interval (CI) of 95%. The figure also includes 8 points, labelled **i** to **vii** representing individuals who illustrate several possible combination of pre- and post-treatment scores corresponding to distinct outcomes. Data points (**ii** and **vii**) that fall within the confines of the RCI confidence intervals may therefore be regarded as representing ‘no change’ i.e. the magnitude of any change is explained by the error attributable to the imperfect reliability of the measure. Data points outside of the confidence intervals (all other points) can be considered as representing a change that is not likely to be accounted for by the unreliability of the measure, i.e. ‘genuine change’. In Figure 2 individuals **iii, iv** and **viii** show a reliable improvement from pre-treatment to post-treatment while individuals **i**, **v and vi** show a reliable deterioration at post-treatment. (We acknowledge that the effect of statistical regression [7] would be to rotate the line of no-change line and its confidence intervals away from the main diagonal. Minor adjustments to the RCI may be made by incorporating the expected regression effect into the computation of the RCI [28; 24]. Results of a recent large scale simulation study concluded that the traditional RCI computations are sufficient [2].)

In Figure 2 the CI has been set at the conventional 95%. This may be regarded as a somewhat conservative in that it sets a reasonably stringent criterion for defining change. The criterion may by relaxed by choosing another parameter e.g. 90%, the effect of this will be to contract the CI lines toward the main diagonal, as a consequence more individuals will be classified as changed. This aspect of selection of the CI magnitude is a function of investigator preference and the need to weigh the balance between Type 1 and Type 2 errors in defining change. The magnitude of the CI is also affected by the reliability of the test and the variance of the sample. Increasing the reliability of the test and decreasing the variance of test scores will both decrease the magnitude of the boundaries of the CI for a given criterion value (95% or 90%).

Figure 2 also illustrates the implications of imposing a defined criterion for assessing clinically significant change. In the figure the CSC cut scores have be superimposed as reference lines on both the pre-treatment (x-axis, vertical line) and post-treatment (y-axis, horizontal line) scales. These reference lines separate the sample into individuals who are above and below the CSC criterion at each occasion of measurement. In Figure 2 high scores on the scale indicate dysfunction (e.g. high levels of distress or interference in daily activities). Thus data points (**i**, **ii**, **iii** and **iv**) to the right of the pre-treatment cut score (vertical line) represent individuals who were above the defined clinical cut point at pre-treatment assessment, while data points **v** to **viii** represent individuals who were below the clinical criterion at pre-treatment. The horizontal line represents the cut-point superimposed on the post-treatment axis: points above the line represent patients who remain 'dysfunctional' at post-treatment and data points below represent patients who are below the criterion set for clinical significance. One important feature of the present data is that at pre-treatment a significant proportion of the current sample was below a CSC criterion for some of the measures. This situation arose because the selection criteria used did not require patients to meet these criteria for admission to the programme: for instance, high levels of distress (high scores on the BDI and HADS) were not a prerequisite for entry.

The combination of RCI and CSC criteria enables patients to be classified into a number of outcome groups. As consequence an expanded range of outcomes is possible. For those patients above the CSC criterion at pre-treatment there are four possible outcomes (**i** to **iv** in Figure 2):

- **i.** reliable deterioration from their pre-treatment level
- **ii.** no change

iii. reliable improvement but not to such a degree that they may also be regarded as clinically improved

iv. clinical significant change i.e. reliable change and the post-treatment score exceeds the clinical criterion

For those patients below the CSC criterion at pre-treatment there are also four possible outcomes.

v. clinically significant deterioration i.e. reliable change and post-treatment score above the cut point

vi. reliable deterioration – but not to such a degree that the post-treatment score exceeds the cut point for clinical significance

vii. no change

viii. reliable improvement

There are thus eight possible outcomes when the sample is spread across the pre-treatment CSC cut score. To capture this aspect of the analysis the sample was divided into those above and below the defined CSC criteria at pre-treatment and the criteria for reliable and clinically significant change applied to each subset of the data. The summary statistics, frequencies and rounded percentages for the complete sample in each of the categories **i** to **viii**, are shown in Table 3.

As noted, the method used to set CSC criterion meant that there were no pre-treatment observations below the criterion for the three pain related measures (intensity, distress and interference). The majority of individuals were categorized as unchanged for these measures with a significant minority achieving a clinically significant change of more than 30% reduction in their scores from baseline. The response rate for achieving a clinically significant change ranged from about 1 in 8 for pain intensity to 1 in 4 for pain interference. It is also notable that no patients fell into the reliable change category (***ii***). This is attributable to a statistical quirk in that the degree of change necessary for a reliable change exceeded the value required to meet the CSC criteria (see Table 1). This position would change with increased reliability of the measure, currently set at 0.69 for all measures, or by altering the CSC criteria.

The CSC criterion for the three scales assessing distress assigned a significant proportion of the sample below the cut point at pre-treatment (BDI 55%; HADS anxiety 22%; HADS depression 30%). When only those above the CSC criterion are considered, the response rate for achieving a CSC ranged from about 1 in 3 (33%, BDI) to 1 in 4 (25%, HADS depression) to just over 1 in 5 for the HADS anxiety scale (18%). A further smaller number achieved a reliable improvement that did not reach clinical significance (BDI, 5%; HADS anxiety 12%; HADS depression 9%). The remaining measures, CSQ, PSEQ and 5 minute walk test, all used Jacobson's criterion (*a)* which set the CSC criterion at the extreme tail of the clinical sample. As a consequence relatively few patients are assigned below the CSC criterion at pre-treatment. The response rates for those above the CSC criterion ranged from about 17% (PSEQ) to 4% (CSQ catastrophizing) and 6% (5 minute walk test). Relatively more patients achieved a reliable change (CSQ catastrophizing = 15%; PSEQ = 33%; 5 minute walk = 34%).

When taken as a whole, individuals who were below the CSC criteria for the pre-treatment measures remained below the criteria at post-treatment. The majority of scores were essentially unchanged but a few individuals made some reliable gains. A few showed reliable deterioration but did not exceed the CSC criterion, and a small number did show what might be regarded as clinical deterioration. There was also evidence that a small number of patients above the CSC criteria at pre-treatment showed deterioration at post treatment. A similar pattern of responding was shown at follow-up.

4. Discussion

The main aim of this study was to establish evidence for the effectiveness of CBT for the management of chronic pain in a routine clinical setting and to explore different methodologies for assessing outcome. The conventional statistical criteria used in RCTs provide evidence of significant change. However, the large sample size ensures observed changes are highly significant and interpretation on the basis of *P* values alone is spurious. Similarly, the observed effect sizes would generally be regarded as ranging from the upper end of *small* (.3 to .49) to *medium* (.5 to .8) magnitude [10]. While these observations support the interpretation that change occurred in this clinical sample, neither statistic can be used to support claims that any change was above and beyond that produced by the measurement error inherent in the scales, or that the changes were clinically important.

RCI/CSC methodology has a number of advantages: the RCI sets criteria for determining whether the magnitude of observed change is spurious (attributable to measurement error) and the CSC sets criteria for determining the clinical meaning of the observed change. Using RCI/CSC evidence emerged that for measures of pain, emotional distress and self efficacy between one third and one fifth of patients achieved clinically significant outcome. A considerably smaller number (6%, or 1 in 17) achieved a clinically significant change on a measure of behavioural activity, the 5 minute walk test. A caveat in interpreting these outcome figures is that the data analysis was performed per protocol rather than as intention-to-treat. This may therefore over-estimate the degree of change in the sample; although our missing data analyses gave no reason to suspect that participants with incomplete data sets were substantially different.

Where data are dichotomised e.g. improved/unimproved it is possible to estimate how many patients must be treated in order for one to benefit (Number Needed to Treat – NNT). NNT computations require data from treated and control arms of an RCT [42]. The NNT statistic is generally not used to express outcomes for psychological treatments as most trials employ continuous outcome measures that are expressed as effect sizes in meta-analyses [44; 13; 27]. In two exceptions, for pain and gastrointestinal symptoms [36] and children with headache [16], NNT values between 2 and 3 have been reported. In the current study the values corresponding to the number achieving a clinically significant change might be regarded as pseudo-NNT (given the absence of a control group). Thus if 25% achieved a CSC i.e. 1 in 4 responded, gives the crude NNT estimate of 4.

A limitation of the current study concerns the range of measures used in this setting. We examined nine outcomes they were not spread equally across possible domains of measurement [44; 45; 51]. Two measures captured pain experience and three sampled emotional distress. The measure of interference with daily activity was limited to single numerical rating scale. The programme did, at various times, use the Sickness Impact Profile and the Medical Outcomes Study (SF-36), but neither of these measures was used extensively enough to generate a significant sample size. Only one measure of physical capacity was analysed (5 minute walk test) and the degree to which this measure can be generalised to other behavioural activities is debatable. The remaining two measures, the CSQ and PSEQ, might arguably be regarded as process rather than outcome measures in that they capture cognitive variables generally deemed to be relevant in the successful implementation of CBT. The evidence that a significant proportion of the patients experienced a reliable change in these measures is in accord with the intended impact of the treatment but this cannot be taken as prima facie evidence that cognitive change was the essential causal process in the treatment programme [6; 54]. Medication and health care data was collected but the format of the data was not reducible to scaled measures. The present data underscore the requirement to establish robust measures for routine clinical use to achieve a data set that captures the critical domains [51; 15].

RCI/CSC methodology offers advantages over conventional inferential statistics in evaluating programmes but there are several problems associated with the method as illustrated within this study. First the results obtained by this method are a function of the estimates of reliability and sample variance. For example, the impact of the reliability on the proportion of individuals achieving a reliable change is evidenced in the contrast between the pain measures (r = 0.69) and the 5 minute walk (r = .94) – see Table 3. The lower (test-retest) reliability for the pain measures reflects the fact that pain is not stable over time [31]. As a consequence few participants achieved a reliable change for the pain measures but it also reflects the fact that a larger shift in the mean level of pain is required in order to be able to detect that a reliable change has occurred. Second, the availability and choice of normative data and reference samples influences the proportions achieving a clinically significant change. We approached this problem in a pragmatic manner in selecting norms from various sources but the availability of large normative data sets for chronic pain patients over a range of measures would be a considerable asset [46]. This approach led to variation in the criteria used to assess clinically significant change and circumvented detailed consideration of the definition of clinically significant change in a chronic population. RCI/CSC methodology uses the normative statistical properties of the scales to determine the cut points [34]. Exceptions to this were the choice of external norms for the HADS and pain measures, but there is a difference between these measures in the way the criteria have been established. The HADS cut score represents the best point for discriminating between depression/non-depression when an externally defined diagnostic ‘gold standard’ is used [37]. In contrast, the criterion for the pain was derived from patients’ global impression of change as much improved or better at the *end* of pharmacological trials [19]. These criteria are arguably not fully appropriate for assessing the outcomes in the current setting. One approach that appears promising is to construct criteria on the basis of patient defined outcomes [52; 48]. Robinson et al. [48] and Thorne and Morley (in preparation) both reported that chronic pain patients estimated that they would regard a change of greater than 50% as a successful or acceptable outcome for a range of outcome measures.

The RCI/CSC methodology explicitly separated patients above and below CSC cut points pre-treatment, a feature which has not generally been reported in RCTs of psychological treatments for chronic pain. The separation of the sample pre-treatment was particularly noticeable for the measures of emotional adjustment (BDI, HADS). This feature both indicates the heterogeneity within the chronic pain population and suggests that unless poor emotional adjustment is included in the selection criterion, a significant proportion of patients will show little change on these measures because they are already ‘below threshold’. At the other end of the spectrum, RCI/CSC methodology also provides a method for identifying a small proportion of patients who deteriorate during the course of treatment. This phenomenon is recognised within the general psychotherapy literature [40] but has barely been investigated in pain management. Conventional inferential statistical analyses do not consider either the threshold or the deterioration aspects of the current data.

RCI/CSC methodology is able to rule out measurement error as a plausible explanation for change and to determine whether individuals may be regarded as clinically improved, but this method cannot answer questions relating to the cause of change. It remains unclear whether observed changes were the result of specific treatment effects, non-specific treatment effects, or the mere passage of time. However, for the current sample one might infer that the change was not entirely attributable to time. An RCT [56] conducted within this setting using the same treatment protocol indicated the superiority of treatment against a 12 week wait list control and a generalisation study [57] showed relatively small advantages to randomised over nonrandomised patients treated at INPUT. It is possible to develop benchmarks of treatment outcome for application to routine clinical programmes. Benchmarks may be used to determine whether the magnitude of the average change in the clinic corresponds to the average change in RCTs. Minami et al. [38; 39] have described a strategy and developed criteria for psychotherapy efficacy in depression. Data from RCTs were used to generate benchmark effect sizes for pre-treatment – post-treatment comparisons from the treatment and control arms of the trials. These benchmarks controls for the passage of time, non-specific and specific treatment effects. The effect size estimates generated in the current study need to be compared against the corresponding benchmarks developed from CBT trials in chronic pain. Such comparison would facilitate the evaluation of the programme effectiveness overall but the effect size statistic does not directly relate to clinical effectiveness, for which benchmarks based on categorical outcomes which include those of clinical significance are required.

Acknowledgements

We thank André Lynam-Smith for his extensive help in data cleaning and preparation; and Guy's and St Thomas' Hospital NHS Trust for permission to access the data. Sumerra Hussain was funded by the West Yorkshire NHS Confederation. The authors declare no conflicts of interest.

References

[1] Asghari A, Nicholas MK. Pain self-efficacy beliefs and pain behaviour. A prospective study. Pain 2001;94(1):85-100.

[2] Atkins DC, Bedics JD, McGlinchey JB, Beauchaine TP. Assessing clinical significance: Does it matter which method we use? J Consult Clin Psychol 2005;73(5):982-989.

[3] Bandura A. Self-efficacy: The exercise of control. New York: W H Freeman/ Times Books/ Henry Holt & Co, 1997.

[4] Barkham M, Mellor-Clark J. Bridging evidence-based practice and practice-based evidence: Developing a rigorous and relevant knowledge for the psychological therapies. Clin Psychol Psychother 2003;10(6):319-327.

[5] Beck AT, Steer RA, Garbin MG. Psychometric properties of the Beck Depression Inventory: Twenty five years of evaluation. Clin Psychol Review 1988;8(1):77-100.

[6] Burns JW, Kubilus A, Bruehl S, Harden RN, Lofland K. Do changes in cognitive factors influence outcome following multidisciplinary treatment for chronic pain? A cross-lagged panel analysis. J Consult Clin Psychol 2003;71(1):81-91.

[7] Campbell DT, Kenny D. A primer on regression artefacts. New York: Guilford, 1999.

[8] Cepeda MS, Africano JM, Polo R, Alcala R, Carr DB. Agreement between percentage pain reductions calculated from numeric rating scores of pain intensity and those reported by patients with acute or cancer pain. Pain 2003;106(3):439-442.

[9] Cepeda MS, Africano JM, Polo R, Alcala R, Carr DB. What decline in pain intensity is meaningful to patients with acute pain? Pain 2003;105(1-2):151-157.

[10] Cohen J. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press, 1977.

[11] Cucchiaro G, Farrar JT, Guite JW, Li Y. What postoperative outcomes matter to pediatric patients? Anesthesia Analgesia 2006;102(5):1376-1382.

[12] Dixon D, Pollard B, Johnston M. What does the chronic pain grade questionnaire measure? Pain 2007;130(3):249-253.

[13] Dixon KE, Keefe FJ, Scipio CD, Perri LM, Abernethy AP. Psychological interventions for arthritis pain management in adults: A meta-analysis. Health Psychol 2007;26(1):241-250.

[14] Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. Psychol Methods 1996;1(2):170-177.

[15] Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, Kerns RD, Stucki G, Allen RR, Bellamy N, Carr DB, Chandler J, Cowan P, Dionne R, Galer BS, Hertz S, Jadad AR, Kramer LD, Manning DC, Martin S, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robbins W, Robinson JP, Rothman M, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Wernicke J, Witter J. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113(1-2):9-19.

[16] Eccleston C, Morley S, Williams ACdeC, Yorke L, Mastroyannopoulou K. Systematic review and meta-analysis of randomised controlled trials of

psychological therapy for chronic pain in children and adolescents. Pain 2002;99(1-2):157-165.

[17] Farrar JT. What is clinically meaningful: outcome measures in pain clinical trials. Clin J Pain 2000;16(2 Suppl):S106-112.

[18] Farrar JT, Berlin JA, Strom BL. Clinically important changes in acute pain outcome measures: a validation study. J Pain Symptom Management 2003;25(5):406-411.

[19] Farrar JT, Young JP, Jr., LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001;94(2):149-158.

[20] Fordyce WE. Behavioral methods for chronic pain and illness. St Louis: Mosby, 1976.

[21] Geisser ME, Roth RS, Robinson ME. Assessing depression among persons with chronic pain using the center for epidemiological studies-depression scale and the beck depression inventory: A comparative analysis. Clin J Pain 1997;13(2):163-170.

[22] Gibson L, Strong J. The reliability and validity of a measure of perceived functional capacity for work in chronic back pain. J Occup Rehabil 1996;6(3):159-175.

[23] Gil KM, Ross SL, Keefe FJ. Behavioral treatment of chronic pain: four pain management protocols. In: RD France, KDD Krishnan editors. Chronic Pain. Washington: American Psychiatric Association, 1988.

[24] Hageman WJJM, Arrindell WA. A further refinement of the reliable change (RC) index by improving the pre-post difference score: Introducing RCID. Behav Res Ther 1993;31(7):693-700.

[25] Hanley MA, Jensen MP, Ehde DM, Robinson LR, Cardenas DD, Turner JA, Smith DG. Clinically significant change in pain intensity ratings in persons with spinal cord injury or amputation. Clin J Pain 2006;22(1):25-31.

[26] Harding VR, Williams AC, Richardson PH, Nicholas MK, Jackson JL, Richardson IH, Pither CE. The development of a battery of measures for assessing physical functioning of chronic pain patients. Pain 1994;58(3):367-375.

[27] Hoffman BM, Papas RK, Chatkoff DK, Kerns RD. Meta-Analysis of Psychological Interventions for Chronic Low Back Pain. Health Psychol 2007;26(1):1-9.

[28] Hsu LM. Reliable changes in psychotherapy: Taking into account regression toward the mean. Behavioral Assessment 1989;11(4):459-467.

[29] Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. J Consult Clin Psychol 1999;67(3):300-307.

[30] Jacobson NS, Truax P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59(1):12-19.

[31] Jensen MP. Questionnaire Validation: A brief guide for readers of the research literature. Clin J Pain 2003;19(6):345-352.

[32] Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. Pain 1986;27(1):117-126.

[33] Jensen MP, Turner JA, Romano JM. Changes in beliefs, catastrophizing, and coping are associated with improvement in multidisciplinary pain treatment. J Consult Clin Psychol 2001;69(4):655-662.

[34] Kazdin AE. The meanings and measurement of clinical significance. J Consult Clin Psychol 1999;67(3):332-339.

[35] Keefe FJ, Brown GK, Wallston KA, Caldwell DS. Coping with rheumatoid arthritis pain: catastrophizing as a maladaptive strategy. Pain 1989;37(1):51-56.

[36] Lackner JM, Morley S, Dowzer C, Mesmer C, Hamilton S. Psychological treatments for irritable bowel syndrome: A systematic review and meta-analysis. J Consult Clin Psychol 2004;72(6):1100-1113.

[37] Lowe B, Spitzer RL, Grafe K, Kroenke K, Quenter A, Zipfel S, Buchholz C, Witte S, Herzog W. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. J Affect Disord 2004;78(2):131-140.

[38] Minami T, Serlin R, Wampold B, Kircher J, Brown G. Using clinical trials to benchmark effects produced in clinical practice. Quality and Quantity 2006;DOI 10.1007/s11135-006-9057-z.

[39] Minami T, Wampold BE, Serlin RC, Kircher JC, Brown GS. Benchmarks for Psychotherapy Efficacy in Adult Major Depression. J Consult Clin Psychol 2007;75(2):232-243.

[40] Mohr DC. Negative Outcome in Psychotherapy: A Critical Review. Clinical Psychology: Science and Practice 1995;2(1):1-27.

[41] Moore A, Edwards J, Barden J, McQuay H. Bandolier's Little Book of Pain. Oxford: Oxford University Press, 2003.

[42] Moore A, McQuay H. Bandolier's Little Book of Making Sense of the Medical Evidence. Oxford: Oxford University Press, 2006.

[43] Morley S. Trial design in psychological treatments: What can we tell patients? In: RA Moore, HJ McQuay editors. Systematic reviews in and meta-analyses in pain: Lessons from the past leading to pathways for the future. Seattle: IASP Press, in press.

[44] Morley S, Eccleston C, Williams ACdeC. Systematic review and meta-analysis of randomized controlled trials of cognitive behaviour therapy and behaviour therapy for chronic pain in adults, excluding headache. Pain 1999;80(1-2):1-13.

[45] Morley S, Williams ACdeC. Conducting and evaluating treatment outcome studies. In: RJ Gatchel, DC Turk editors. Psychosocial factors in pain. New York: Guilford, 2002. pp. 52-68.

[46] Nicholas MK, Asghari A, Blyth FM. What do the numbers mean? Normative data in chronic pain measures. Pain 2008; 134(1-2): 51-58.

[47] Philips HC. The psychological management of chronic pain: A treatment manual. New York: Springer, 1988.

[48] Robinson ME, Brown JL, George SZ, Edwards PS, Atchison JW, Hirsh AT, Waxenberg LB, Wittmer V, Fillingim RB. Multidimensional Success Criteria and

Expectations for Treatment of Chronic Pain: The Patient Perspective. Pain Medicine 2005;6(5):336-345.

[49] Rosensteil AK, Keefe FJ. The use of coping strategies in chronic low back pain. Pain 1983;17:33-44.

[50] Sullivan MJL, Thorn BE, Haythornthwaite JA, Keefe FJ, Martin M, Bradley LA, Lefebvre JC. Theoretical perspectives on the relationship between catastrophizing and pain. Clin J Pain 2001;17:52-64.

[51] Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N, Carr DB, Cleeland C, Dionne R, Farrar JT, Galer BS. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. Pain 2003;106(3):337-345.

[52] Turk DC, Dworkin RH, Revicki D, Harding G, Burke LB, Cella D, Cleeland CS, Cowan P, Farrar JT, Hertz S, Max MB, Rappaport BA. Identifying important outcome domains for chronic pain clinical trials: An IMMPACT survey of people with pain. Pain;In Press, Corrected Proof.

[53] Turk DC, Meichenbaum D, Genest M. Pain and behavioral medicine: a cognitive-behavioral perspective. New York: Guilford Press, 1983.

[54] Turner JA, Holtzman S, Mancl L. Mediators, moderators, and predictors of therapeutic change in cognitive-behavioral therapy for chronic pain. Pain 2007;127(3):276-286.

[55] Westen D, Novotny CM, Thompson-Brenner H. The Empirical Status of Empirically Supported Psychotherapies: Assumptions, Findings, and Reporting in Controlled Clinical Trials. Psychol Bull 2004;130(4):631-663.

[56] Williams A, Richardson P, Nicholas M, Pither C, Fernandes J. Inpatient vs. outpatient pain management: Results of a randomised controlled trial. Pain 1996;66(1):13-22.

[57] Williams ACdC, Nicholas MK, Richardson PH, Pither CE, Fernandes J. Generalizing from a controlled trial: The effects of patient preference versus randomization on the outcome of inpatient versus outpatient chronic pain management. Pain 1999;83(1):57-65.

[58] Zautra AJ, Johnson LM, Davis MC. Positive affect as a source of resilience for women in chronic pain. J Consult Clin Psychol 2005;73(2):212-220.

[59] Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361-370.

List of Tables