ORIGINAL ARTICLE

Outcome measures in chronic low back pain

Elaine F. Maughan · Jeremy S. Lewis

Received: 18 August 2009 / Revised: 18 January 2010 / Accepted: 21 February 2010 / Published online: 17 April 2010

Abstract The purpose of this prospective, single site cohort quasi-experimental study was to determine the responsiveness of the numerical rating scale (NRS), Roland–Morris disability questionnaire (RMDQ), Oswestry disability index (ODI), pain self-efficacy questionnaire (PSEQ) and the patient-specific functional scale (PSFS) in order to determine which would best measure clinically meaningful change in a chronic low back pain (LBP) population. Several patient-based outcome instruments are currently used to measure treatment effect in the chronic LBP population. However, there is a lack of consensus on what constitutes a "successful" outcome, how an important improvement/deterioration has been defined and which outcome measure(s) best captures the effectiveness of therapeutic interventions for the chronic LBP population. Sixty-three consecutive patients with chronic LBP referred to a back exercise and education class participated in this study; 48 of the 63 patients had complete data. Five questionnaires were administered initially and after the 5-week back class intervention. Also at 5 weeks, patients completed a global impression of change as a reflection of meaningful change in patient status. Score changes in the five different questionnaires were subjected to both distribution- and anchor-based methods: standard error of measurement (SEM) and receiver operating characteristic (ROC) curves to define clinical improvement. From these methods, the minimal clinically important difference (MCID) defined as the smallest difference that patients and clinicians perceive to be worthwhile is presented for each instrument. Based on the SEM, a point score change of 2.4 in the NRS, 5 in the RMDQ, 17 in the ODI, 11 on the PSEQ, and 1.4 on the PSFS corresponded to the MCID. Based on ROC curve analysis, a point score change of 4 points for both the NRS and RMDQ, 8 points for the ODI, 9 points for the PSEQ and 2 points for the PSFS corresponded to the MCID. The ROC analysis demonstrated that both the PSEQ and PSFS are responsive to clinically important change over time. The NRS was found to be least responsive. The exact value of the MCID is not a fixed value and is dependent on the assessment method used to calculate the score change. Based on ROC curve analysis the PSFS and PSEQ were more responsive than the other scales in measuring change in patients with chronic LBP following participation in a back class programme. However, due to the small sample size, the lack of observed worsening of symptoms over time, the single centre and intervention studied these results which need to be interpreted with caution.

Keywords Low back pain · Outcome measures · Minimally clinically important change · Responsiveness · Functional assessment · Clinical significance · Patient-reported outcomes

Introduction

Disability caused by low back pain (LBP) affects approximately one quarter of adults in any one year and is the

E. F. Maughan (✉)
Guy's and St Thomas' NHS Trust, London, UK
e-mail: Elaine.Maughan@nhs.net

J. S. Lewis
St George's NHS Healthcare Trust, London, UK

J. S. Lewis
St George's University of London, London, UK

J. S. Lewis
Therapy Department, Chelsea and Westminster Hospital
NHS Foundation Trust, London, UK

most common cause of physical disability in the working age population of the UK [1]. LBP that persists continuously or intermittently for longer than 3 months is deemed chronic. While not a disease, LBP is associated with substantial morbidity. People report that most if not all aspects of their lives are significantly affected by chronic pain [2]. In a survey conducted by Taylor [3], the impact of multiple areas of musculoskeletal pain, including LBP on health-related quality of life (HRQOL) was comparable to the HRQOL of patients with chronic liver disease prior to transplant and terminal cancer.

An historical review shows that there is no change in the pathology of LBP; however, our understanding and management has changed [1]. Restoration of normal function designed to address the patients' specific needs is considered a key outcome of physiotherapy for low back problems [4]. A systematic review, concluded that there is moderate evidence for clinical effectiveness of structured exercise programmes and that "back schools" reduce pain and improve function and return-to-work status, both in the short- and intermediate-term, compared with other treatments for recurrent and chronic LBP [5]. The UK-based National Institute for Health and Clinical Excellence (NICE) guidelines [6] for nonspecific LBP recommend a structured exercise programme tailored to the person: up to a maximum of 8 sessions over a period of up to 12 weeks.

Randomized control trials (RCTs) and the findings of systematic reviews of different treatment approaches for chronic LBP rarely show more than a small–moderate improvement in short-term outcomes from treatment [7]. Many treatment options exist and currently no one treatment has demonstrated superiority to the alternatives [8, 9]. One possible explanation for this is that chronic LBP is complex and multi-dimensional in nature [1]. The bio-psychosocial model of pain posits that the pain experience is a function of interacting combinations of patho-anatomical, neuro-physiological, physical and psychosocial factors which are different for each individual [8, 9]. As such the effects of treatment may be diluted in RCTs when applied to a heterogeneous group with diverse treatment needs [10].

Another possible explanation for small–moderate treatment effects seen in RCTs is the lack of consensus on what constitutes a "successful" outcome, how an important improvement has been defined and which specific outcome measures best measure this. This is difficult when the outcome of interest is subjective, and there are no definitive measurable end points to indicate when a patient is "better" [11]. RCTs have focused on the statistical significance of change in scores from outcome measures which reflects both the magnitude and variability of the treatment effects as well as the sample size. Statistical significance does not indicate the proportion of individuals in the group who achieved a clinical meaningful change from the treatment intervention [11]. Beaton et al. [12] carried out a qualitative study to explore the answer to the question "Are you better?" They reported that people varied in their definitions, both in terms of the type of change that they considered to be indicative of improvement and in the degree of importance of that change to them. This wide disparity between patients' expectation and perception of treatment further impairs our ability to definitively measure a "successful" outcome.

Responsiveness

Responsiveness refers to the ability of a measurement tool to detect real or important change over time when it has occurred (and equally when it has not occurred) in the concept being measured [13]. This has led to a search for the elusive "minimally clinically important difference" (MCID) in the scores of these measures, which ideally would identify when an individual or a group is "better" above and beyond change due to measurement error [14]. Jaeschke et al. [15], defined the MCID as the "the smallest difference in a score of a domain of interest that patients perceive to be beneficial".

Methods of exploring responsiveness can be classified either as those that measure change alone (distribution-based methods) or those that measure clinically meaningful change (anchor-based methods). Both anchor- and distribution-based methods have advantages and limitations, with neither superior to the other therefore a combination of both methods is preferred [16].

Distribution-based methods

The distribution-based method measures the statistical significance of the change scores in the measure. There is agreement that the standard error of measurement (SEM) is the best method of calculating statistically meaningful change [17]. The SEM accounts for the possibility that some of the change observed with a particular measure may be attributable to random error [14]. The SEM can then be used to calculate the smallest detectable change (SDC) for the score which reflects the smallest observed change in score that is above measurement error [18].

Anchor-based methods

The anchor-based method determines the smallest important difference in a measurement instrument that relates to a corresponding change in a reference measure of clinical/health status (the "anchor"). There is no gold standard "anchor" to assess true change of status. A global impression of change (GIC) instrument requires the respondent to compare post-treatment with pre-treatment status and judge whether meaningful change has taken place over the retest period. Patient global impression of change (PGIC) has been recommended by the initiative on methods, measurement, and pain assessment in clinical trials (IMMPACT) for use in chronic pain clinical trials as a core outcome measure of global improvement with treatment [19]. Lauridson et al. [20] demonstrated that the patients' global retrospect of treatment effect is robust in discriminating between those who have improved and those who remain unchanged. Receiver operating characteristics (ROC) analysis is an anchor-based method of examining a measure's responsiveness [18]. ROC curves can be used to determine the most accurate (highest specificity and sensitivity) cutoff for change scores. ROC curves are also used to rank the ability of competing measures to detect clinical change.

To achieve comprehensive multidimensional evaluation of outcome in LBP IMMPACT [19] propose six core outcome domains that should be considered: (1) pain, (2) physical functioning, (3) emotional functioning, (4) participant ratings of improvement and satisfaction with treatment, (5) symptoms and adverse events and (6) participant disposition. The NICE guidelines [6] for nonspecific LBP recommend that any intervention should have a high impact on patients' outcomes in particular pain, disability or psychological distress. This current study evaluated the following core domains; numerical rating scale (NRS), Roland–Morris disability questionnaire (RMDQ), Oswestry disability index (ODI), patient-specific functional scale (PSFS) and the pain self-efficacy questionnaire (PSEQ).

There are no agreed scientific grounds or empirical evidence to determine the optimum method of estimating the MCID [21]. Which instrument shows the best performance with respect to the individual patient versus a population view of the MCID can only be answered by a direct comparison between the instruments in a single study population. As such, the purpose of this study was twofold: first, to demonstrate the responsiveness of five different questionnaires PSFS, PSEQ, NRS, ODI and the RMDQ (using both anchor- and distribution-based methods) and second, to compare which of these tools best measures change in patients with chronic LBP following participation in a back class programme.

Methodology

Ethical review

This study was reviewed and approved by the ethical review board at the Lewisham Local Research Ethics Committee (reference number 07/Q0701/12) and the research and development centre for Greenwich, Lambeth, Lewisham and Southwark PCTs (reference number RDLAM 355).

Subjects/study population

Consecutive patients with LBP referred by their physiotherapist to the back class (at the Pulross centre in Brixton) were eligible for the study. The back class is run by a physiotherapist once a week for 5 weeks. The class included a practical education session in order to improve patients' management and understanding of their pain. Topics covered included; anatomy of the spine, mechanisms of chronic LBP, goal setting, posture, pacing activity, management of pain and returning to exercise. The class also included an exercise session to strengthen and stretch the main muscle groups with particular emphasis on the stomach, trunk and buttock muscles. Participants were taught how to exercise safely and pace their activity.

Inclusion criteria: Patients with LBP aged 18 and over, both males and females, LBP duration greater than 3 months with or without radiation to the lower limbs, not undergoing any other concurrent treatments for pain other than routine analgesia and sufficient level of spoken and written English language. Exclusion criteria included: Spinal surgery in the past 12 months, LBP as a result of new spinal fracture, infection, malignancy, inflammatory joint disease, clinical evidence of the need for specific interventions or further investigation, unstable neurological signs/symptoms, general health problems that prevented the patient from participating in an exercise programme, pregnancy and poor English comprehension. Potential participants were contacted by the researcher and received both verbal and written information pertaining to the essential elements of the study. Participation in the study was voluntary, and participating subjects signed informed consent documentation. Patients completed a booklet of five of outcome measures before the class and on completion of treatment. The following outcome measures were used.

Roland–Morris disability questionnaire (RMDQ)

The RMDQ consists of 24 statements about activity limitations due to back pain, e.g. walking, lying and self-care [22]. Patients were asked to answer yes or no to each statement. Each positive answer is worth one point with scores ranging from 0 (no disability) to 24 (severely disabled).

Oswestry disability index (ODI) version 2

The ODI is divided into ten sections to assess the level of pain and interference with several physical activities including; sleeping, self-care, sex life, social life and travelling ([23]; Medical Research Council 1989). Each question has a possible six responses which are scored from 0 to 5. Patients were asked to tick one response statement in each section that was most relevant to them. The score for each section was added and divided by the total possible score (fifty if all sections are completed), and the resulting score was multiplied by a hundred to yield a percentage score with 0% equivalent to no disability and 100% equivalent to a great deal of disability.

Numerical rating scale (NRS)

The NRS asked patients to rate their pain intensity on an 11-point scale where 0 indicates no pain and 10 indicates worst imaginable pain.

Pain self-efficacy questionnaire (PSEQ)

Patients were asked to rate how confident they were at that time despite the presence of their pain in performing ten activities listed by selecting a number on a 7 point scale, where 0 equals "not at all confident" and 6 equals "completely confident" [24]. Scores on the PSEQ may range from 0 to 60, with higher scores indicating stronger self-efficacy beliefs.

Patient-specific functional scale (PSFS)

Patients were asked by their clinician to identify up to three activities that they had difficulty with or were unable to perform as a result of their back pain and to rate these activities on an 11-point scale from 0: unable to perform activity to 10 able to fully perform the activity at same level before back pain [25]. At follow-up, patients were allowed to access their original scores and were invited to rescore each activity according to their current perception of their performance.

Patient global impression of change (PGIC)

On completion of the back class each patient completed a global impression of change to assess whether the patient was better, about the same, or worse. PGIC was measured on a 7-point scale where 1 = completely recovered, 2 = much improved, 3 = slightly improved, 4 = no change, 5 = slightly worse, 6 = much worse and 7 = vastly worsened. In this current investigation, patients whose mean score was greater than 2 which corresponded to the categories much improved and completely better were considered to have improved, scores between 3, 4 and 5 were grouped as unchanged and scores 6 and 7 (worse than ever) were grouped as deteriorated.

Statistical analyses

The data were analysed on both Microsoft Office Excel 2003 and the Statistical Package for Social Sciences (SPSS) version 14, Chicago, USA, 2005. The change scores between baseline and on completion of the class were calculated for each outcome.

Two methods were used to quantify responsiveness: a distribution-based method expressed by the standard error of measurement (SEM) and an anchor-based method by ROC curve analysis.

Distribution-based method

The patients whose status remained stable were used to determine the SEM which allowed the calculation of smallest detectable change in each measure.

Standard error of measurement (SEM)

The SEM indicates the precision of the outcome measure [26]. The $\mathrm{SEM} = \mathrm{SD}\sqrt{(1 - r)}$ where $\mathrm{SD} =$ standard deviation of the scores at baseline, $r =$ reliability coefficient calculated by the intra-class correlation coefficient (ICC). The ICC is a reliability parameter that relates the measurement error to the variability between subjects [27] calculated by dividing the inter-individual variation by the total variation (inter-individual variation plus the intra-individual variation). The smallest detectable change (SDC) was then calculated by the formula $\mathrm{SDC} = 1.96 \times \sqrt{2} \times \mathrm{SEM}$ [18]. The SDC at the 95% CI is equal to 1.96 (z value for the 95% CI, two-tailed) multiplied by $\sqrt{2}$ to adjust for the error associated with taking two measurements (baseline and follow-up). Since only unchanged patients were assessed, patients with a score less than or equal to the SDC have a 95% chance that no real change has occurred [28].

Anchor-based method

Receiver operating characteristics (ROC) curve analysis assessed the ability of each questionnaire to distinguish patients who had and had not changed according to an external criterion.

Receiver operating characteristic (ROC) curve

Deyo and Centor [29] suggested that an instrument's ability to correctly identify a clinically important change could be evaluated like a diagnostic test in terms of sensitivity and specificity. A ROC curve is produced by plotting the sensitivity (the number of patients correctly identified as improved by the questionnaire divided by the number of patients as improved according to the GIC) against the specificity (number of patients correctly identified as unchanged by the questionnaire divided by the number of unchanged patients according to the GIC) were calculated. For each possible cutoff value the sensitivity (y-axis) was plotted against 1 − specificity (x-axis) to generate a ROC curve [28].

An instrument that can discriminate well between two groups of patients would have a plot where sensitivity sharply increases, while 1 − specificity remains low [29].

The area under the curve (AUC) can be interpreted as the probability of correctly identifying the improved patients from the non-improved patients. The area ranges from 0.5 (no accuracy in distinguishing improved from non-improved) to 1.0 (perfect accuracy). The greater the total area under the ROC curve indicates the instrument's accuracy. From each ROC curve the MCID can be estimated [30]. This is identified as the cutoff value that gives the best balance between the highest sensitivity and the highest specificity, i.e. the lowest overall misclassification in the ROC analysis [31]. This is represented pictorially as the point nearest the upper left-hand corner of the graph.

Results

Over a one year period, a total of 63 consecutive patients who were referred to the exercise and advice programme, who fulfilled the inclusion criteria and did not fulfil the exclusion criteria consented to participate. Fifteen subjects did not complete the study leaving 48 subjects. Eleven subjects did not complete the back class due to a variety of reasons such as; personal and family sickness as well as childcare issues. Four subjects failed to complete the outcome measures on completion of the class. The study population characteristics are shown in Table 1. At follow-up, 23 patients (48%) were classified as having improved based on the PGIC and 25 patients (48%) were classified as having remained stable (no change). No patients were classified as having worsened.

For each outcome measure, the means and standard deviations at baseline and 5 weeks are presented as a group average, improved and unchanged patients in Table 2. To determine whether the baseline scores for those patients whose status remained stable differed from the baseline scores for those who reported an improvement, Mann–Whitney nonparametric tests were performed for each instrument. It appears that those who did not improve had worse baseline scores for all measures; this reached significance ($P < 0.05$) in baseline scores for NRS, RMDQ and PSEQ. At 5-week follow-up, there were significant differences in all outcomes between those who had improved and those whose status remained unchanged.

Table 1 Population characteristics

	Baseline	Improved ($n = 23$)	No change ($n = 25$)
Age			
Range (years)	25–78	25–76	28–78
Mean age (years)	52	55	50
Gender			
Male (%)	16 (33)	6 (26)	10 (40)
Female (%)	32 (67)	17 (74)	15 (60)
Working status			
Retired	17	10	7
Employed	22	11	11
Unemployed	9	2	7
Duration of LBP			
Range	6 months–20 years	1–15 years	6 months–20 years
Mean (years)	6	5	6.5
Classes attended	Completed all five sessions (%)	74	72
	Completed four of the five sessions (%)	26	28

Intra-class correlation coefficient (ICC)

The intra-class correlation coefficient (ICC) for each instrument was calculated by dividing inter-individual variation by the total variation (inter-individual variation plus the intra-individual variation) produced using SPSS version 14.0, 2005. The ICC ranges from 0 (no agreement) to 1 (perfect agreement) above 0.75 is considered good reliability [32]. The ICC values calculated in this study were 0.92 for the NRS, 0.9 for RMDQ, and the ICC for both PSEQ and PSFS were 0.92 and 0.91, respectively.

Standard error of measurement (SEM) and the smallest detectable change (SDC)

For each instrument, the corresponding ICC was used to calculate the SEM. The SEM was then used to indicate the smallest detectable change (SDC). Based on the values of the SEM, the SDC for the NRS is 2.4 points which out of a maximum eleven points equates to a 22% score change. The SDC for the RMDQ was 4.9 points which out of a maximum 24 points equates to a 21% score change. The SDC for the ODI was 16.7 points, which equates to a 17% score change. 10.9 points on the PSEQ of a maximum 60 points equates to an 18% score change and 1.4 on the PSFS of a maximum 11 points equates to a 13% score change.

Table 2 Descriptive data of each outcome measure at baseline and 5 weeks later post back class

	Baseline			*P* values*	5 weeks			*P* values*
	Average	Improved	No change		Average	Improved	No change	
NRS								
Mean (SD)	5 (2.6)	5 (2.7)	6 (2.3)	0.053*	4 (2.3)	3 (2.0)	5 (2.2)	0.007*
Range	0–10	0–9	0–10		0–9	0–7	0–9	
RMDQ								
Mean (SD)	11 (6.1)	9 (6.1)	14 (5.4)	0.01*	9 (6.3)	6 (4.9)	12 (6.1)	0.001*
Range	1–21	1–21	4–22		0–23	0–17	1–23	
ODI								
Mean (SD)	29 (20)	24 (18.2)	35 (20.2)	0.06	29 (1.9)	20 (14.9)	38 (18.4)	0.001*
Range	4–78	4–68	4–78		4–71	4–56	4–71	
PSEQ								
Mean (SD)	38 (14.8)	42 (14.6)	34 (13.8)	0.028	43 (15)	51 (11.9)	35 (13.6)	0.00*
Range	9–60	15–60	9–57		13–60	28–60	13–60	
PSFS								
Mean (SD)	4 (1.6)	4 (1.4)	4 (1.8)		6 (2.1)	7 (1.4)	5 (2.1)	0.00*
Range	1–7	2–6	1–7	0.264	1–10	3–10	1–8.3	

RMDQ Roland–Morris disability questionnaire (score 0–24); *ODI* Oswestry disability index score (0–100); *NRS* numerical rating scale (score 0–10); *PSEQ* pain self-efficacy questionnaire (score 0–60); *PSFS* patient-specific functional scale score (0–10)

* Significant ($P < 0.05$) values based on the Mann–Whitney *U* test

ROC curves

Figure 1 compares the ROC curve produced for each instrument. The AUC for the PSEQ is 0.73 and PSFS is 0.86 which indicates that these measures are responsive to change over time. The AUC of the RMDQ and ODI were 0.64 and 0.67, respectively. The NRS was least responsive with AUC at 0.5 which indicates no accuracy for detecting change.

The MCID is identified as the optimal cutoff change with the best balance between the highest sensitivity and the highest specificity, i.e. the lowest overall misclassification in the ROC analysis [31]. MCID was 4 points for both the NRS and RMDQ, 8 points for the ODI, 9 points for the PSEQ and 2 points for the PSFS.

A summary of results (Table 3) demonstrate that the MCID score change for each outcome to determine those who have changed and those who have remained stable are dependent on the distribution- or anchor-based method used. Based on the values of the ROC curve analysis the PSEQ and PSFS appears to be more consistent and responsive to change than the other scales in measuring change in patients with chronic LBP following participation in a back class programme.

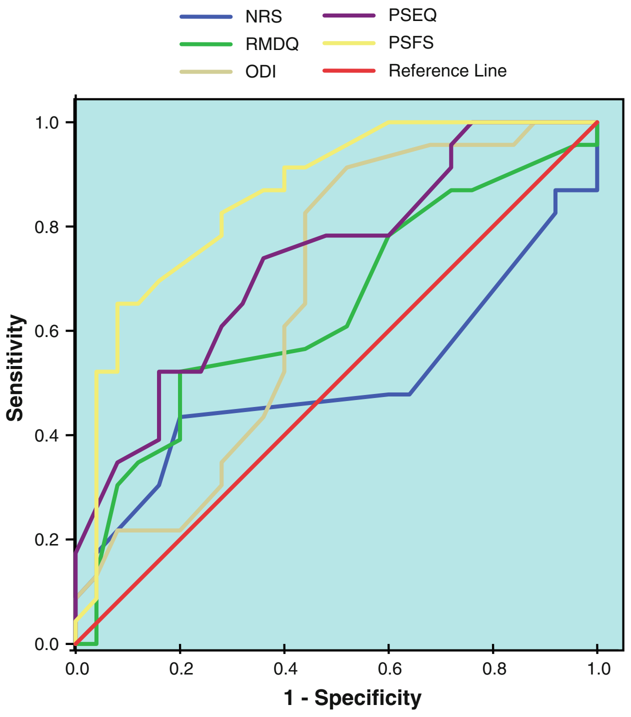

Fig. 1 ROC curves comparing NRS, RMDQ, ODI, PSEQ and PSFS

Discussion

Following a synthesis of the literature, the recent NICE guidelines made three definitive recommendations for the management of low back pain, which include: acupuncture, manual therapy and exercise classes [6]. In addition, in the Department of Health “High quality care for all” document [33], it was stated that outcome measurements are to be

Table 3 Summary of results

	Distribution		Anchor (ROC analysis)	
	SEM	SDC	MCID	AUC
NRS	0.86	2.4	4	0.5
RMDQ	1.78	4.9	3.5	0.64
ODI	6.06	16.7	7.5	0.67
PSEQ	3.95	10.9	8.5	0.73
PSFS	0.5	1.4	2.3	0.86

RMDQ Roland–Morris disability questionnaire (score 0–24); *ODI* Oswestry disability index score (0–100); *NRS* numerical rating scale (score 0–10); *PSEQ* pain self-efficacy questionnaire (score 0–60); *PSFS* patient-specific functional scale score (0–10); *SEM* standard error of measurement; *SDC* smallest detectable change; *MCID* minimal clinical important difference; *AUC* area under the curve

used to determine the quality of a service. In order to understand the impact of LBP on the patient's life, clinicians require reliable, valid and responsive measurement tools that accurately assess function and monitor change over time [31].

In this current investigation, score changes in the five different questionnaires were subjected to both distribution- and anchor-based methods: standard error of measurement (SEM) and receiver operating characteristic (ROC) curves to define clinical improvement. From the anchor-based methods, a score change of 4 points for the NRS, 3.5 for the RMDQ, 7.5 for the ODI, 8.5 for the PSEQ and 2.3 for the PSFS was defined as the smallest difference that patients perceive to be worthwhile. This compares to a score change of 2.4 for the NRS, 4.9 for the RMDQ, 16.7 for the ODI, 10.9 for the PSEQ, and 1.4 for the PSFS allowing one to have 95% confidence that the observed change is real change above measurement error derived from distribution-based methods. To combine these methods to be confident that clinically important change has occurred beyond measurement error the SDC should be less than the MCID [18]. In this current investigation, the MCID was greater than the SDC for the NRS and PSFS which suggests a score change of 4 points on the NRS and 2 points on the PSFS are associated with clinically important change and scores less than this are associated with measurement error. Based on ROC curve analysis, the PSFS and PSEQ were more responsive than the NRS, RMDQ and ODI to detect clinically important change in patients with chronic LBP.

In this current investigation, based on the SEM, the SDC for the NRS is 2.4 points, 5 points for the RMDQ and 17 points for the ODI. These results are comparable to an expert panel's consensus on clinical interpretation proposed for the NRS, RMDQ, and ODI: 2 for NRS, 5 for RMDQ and 10 for ODI as the SDC [21]. In clinical psychology, 1.96 × $\sqrt{2}$ × SEM has been defined as the threshold for classifying important improvement, whereas a change of 1 SEM in a health quality of life questionnaire for patients with chronic heart and respiratory disease was determined as the minimum threshold of change by Wyrich et al. [34]. Childs et al. [35] calculated the SDC of the NRS in patients with LBP at the 95% CI as 1.96 × SEM and concluded that a 2-point change on the NRS is necessary to exceed the bounds of statistical error and to be considered clinically meaningful. Had the formula in this current investigation been used which calculated the SDC as 1.96 × $\sqrt{2}$ × SEM, a 2.8-point change on the NRS would have been considered necessary. Previously Stratford et al. [36] and Davidson and Keating [26] calculated the SDC as 5 points for RMDQ and 10 points for ODI, respectively, by the formula 1.65 × $\sqrt{2}$ × SEM to correspond with the 90% CI. Clinicians need to be aware that disagreement exists regarding how many SEMs an individual must change in order for that change to confidently exceed the bounds of measurement error to be considered significant [37].

In this current investigation, the AUC for the NRS was 0.51 indicating that the NRS did no better than chance alone to discriminate for improvement. This contrasts to the results reported by Childs et al. [35] who found the AUC to be 0.72 at 1 week and 0.92 at 4 weeks. Farrar et al. [38] reported a similar area under the curve (0.87) for patients who were classified as very much improved. Salaffi et al. [39] also reported an area under the curve of 0.89 for the NRS. Farrar's high degree of interrelationship between change in pain intensity and the PGIC was not found in this current investigation. The value of the area under the curve in this study may be due to small sample size. Although not explored in this study, there are perhaps other components of the patient's response (as opposed to pain intensity reported by the NRS), such as psychological and psychosocial factors which influenced their perception of overall improvement. Turk [8] reported that relief from chronic pain is rarely achieved by current treatments. Even patients who report they have improved are not pain-free. Baldwin et al. [40] reported that the main determinant of return to work in this population is not a reduction in the intensity of pain but how well the individual is able to adapt to the pain. The 2,724 patients in the study by Farrar et al. [38] were involved in a placebo-controlled trial of pregabalin treatment for chronic pain, whereas in this current investigation the intervention was a back class, where improvement may have been perceived as a better self-management and understanding of pain rather than a reduction in pain intensity.

Based on ROC curve analysis in this study, the AUC of the RMDQ and ODI were 0.64 and 0.67, respectively. However, Stratford et al. [36] and Beurskens et al. [30] previously calculated the AUC for RMDQ as 0.84 and 0.79, respectively. Beurskens et al. [30] in the same study also reported the area under the curve for the ODI as 0.78. In this study, the PSFS (AUC 0.86) and PSEQ (AUC 0.73) proved to be the more responsive than RMDQ, ODI and NRS. Beurskens et al. [41] also concluded that the PSFS was more sensitive to change but less specific when compared to the RMDQ and pain. Pengel et al. [42] reported that the PSFS was the most responsive outcome measure when compared to the RMDQ and physical impairment measures in patients with subacute (between 6 weeks and 3 months duration) LBP. While both the RMDQ and ODI are an important measure of overall disability, they may conceal improvements in specific activities that are relevant to the patient [14, 43]. Research conducted by Hudak et al. [44], investigating the relationship between patient satisfaction and treatment outcome, indicated that this could be facilitated by developing strategies to elicit the patients' most important reason for undergoing treatment. The advantage of the PSFS is that only changes in activities experienced as most relevant by the patient and that have the potential to improve are assessed [41].

Asghari and Nicholas [45] revealed that self-efficacy is one of the most significant factors influencing treatment outcome for patients with chronic pain. Nicholas [46] supports the use of the PSEQ as both a screening instrument to determine patients' confidence in performing normal activities despite pain and as an evaluating tool to measure outcomes after treatment. Sensitivity of the PSEQ to treatment effects were reported by Williams et al. [47]. To date, no studies have used anchor-based methods to examine criteria from clinically important changes on the PSEQ. However, ROC analysis demonstrated that the PSEQ is responsive to clinically important change over time.

From the ROC curve analysis, MCID calculated in this study was 4 points for both the NRS and RMDQ, 8 points for the ODI, 9 points for the PSEQ and 2 points for the PSFS. Beurskens et al. [30] previously reported the MCID for the PSFS between 1.8 and 2.4 points and the MCID for the ODI as between 4 and 6 points. Stratford et al. [36] reported that the MCID for the RMDQ was 5 points. The area under the curve and the MCID of the PSEQ has not been reported previously.

One methodological challenge in responsiveness analyses is the lack of a gold standard for the construct of clinical change [31]. Patient global impression of change (PGIC) has been recommended by IMMPACT [19] for use in chronic pain clinical trials as a core outcome measure of global improvement with treatment and has been extensively used as a comparison in much of the literature regarding responsiveness. Norman et al. [48], however, questions the reliability of a single item global change scale as the standard for evaluating a multi-item tool. Patients' ability to recall their previous health state is questionable. Davidson and Keating [26] and Schmitt and Di Fabio [49] found retrospective judgment of change to be influenced by the respondents' current status on the day the instrument is completed rather than their previous health status. The measurements are generally administered at the same point in time so that errors in both the global impression of change and the outcome measure are likely to be correlated. However, Von Korff et al. [50] stated that recall of chronic pain in terms of average intensity and interference with activities has acceptable levels of validity for up to a 3-month recall period. Ultimately patients decide whether a treatment is beneficial and the PGIC provides the single best measure of the significance from the patients' perspective [16].

Jaeschke et al.'s [15] full definition of the MCID is: "the smallest difference in a score of a domain of interest that patients perceive to be beneficial and would mandate, in the absence of troublesome side effects and excessive costs, a change in the patients' management". Taking this definition into consideration, the MCID greatly depends on the type of anchor and the anchor's definition of important change which in its very nature is arbitrary [51]. In agreement with other authors [35–38] *much improved*, was set as the standard to reflect minimally important improvement in the current investigation. Little research has been carried out on the importance of change, for example if a patient indicates that they are slightly improved, it is a *minimal* change, but it is unknown how *important* this change is to the patient. Some authors use *slight improvement*, while others use *much better* to be the minimally important improvement as measured by the anchor [51]. A qualitative study conducted by Yelland and Schluter [52] on a 110 patients with chronic LBP undergoing treatment reported a wide disparity between the minimal reductions in pain and disability that makes a treatment worthwhile to an individual with LBP and what percentage reductions in pain and disability they desired. In their study, patient's self-reported outcome also depended to some extent on meeting their pre-treatment expectations. A good illustration of this from the current investigation is subject 2 who listed walking and standing more than 10 min as the first two activities he had difficulty with while the third activity listed was running. The back class did improve his duration of standing and walking; running was unchanged. He subjectively reported no overall improvement. This appears contradictory; however, it may be that to be able to walk and stand longer would be considered worthwhile but being able to run is ultimately what this patient desired.

The NICE guideline [6] recommends a structured exercise programme tailored to the person: up to a maximum of 8 sessions over a period of up to 12 weeks. However, uncertainty exists as to whether a back class over a 5-week period may substantially reduce the pain, disability and work absence associated with chronic LBP. With respect to this, the immediate aim of the back class was to facilitate patients to overcome limitations in function and to return safely in a graded manner to normal activity levels, with the ultimate goal of improving function and social participation [1]. Although the treatment offered may not lead to a complete resolution of symptoms, the intervention may be still worthwhile to an individual with LBP with regard to better self-management and understanding of their pain [8].

Outcome measures that allow patients to generate their own item content are a useful step in capturing how the disorder has affected them and subsequently how treatment may influence this. The findings of the current investigation are in agreement with Walsh et al. [53] who demonstrated that as the PSFS reflects the patient's personal objectives it was more responsive to the effect of an intervention than an outcome measure that has less relevance to the individual patient. Studies investigating the PSFS have examined its responsiveness in comparison to a global rating of change score [25], to the subscales of the short form-36 [43], to the neck disability index [54] and the RMDQ [42]. Each study found the PSFS more sensitive to change over time compared with the other self-report measures.

Stratford et al. [36] demonstrated that the magnitude of MCID is dependent on baseline scores, and hence the MCID falls between 1 and 2 points for patients with low initial scores and between 7 and 8 points with high initial scores. Jordan et al. [55] compared methods used to derive MCID and defined clinical improvement as a 30% reduction from baseline RMDQ and back pain rated better on a global rating scale. Farrar et al. [38] also examined the relationship between baseline scores and the MCID for the NRS and suggested a reduction of 30% from baseline score to indicate a clinically important difference. IMMPACT [56] benchmark a 10–20% reduction in the NRS to reflect minimally important changes and reductions of $\geq$30% to reflect at least moderate clinically important differences. In this current investigation, the baseline scores for NRS, PSEQ and RMDQ were significantly higher between improved and unchanged subgroups in baseline. Due to the small sample size the effect of different baseline measures was not assessed in this current investigation, but should be considered for future research.

This study presents both the statistical and clinical significance of change in scores from outcome measures of pain, disability and pain self-efficacy to interpret treatment effect. However, due to the small sample size from one centre with one intervention, this result needs to be interpreted with caution.

Limitations

Several limitations of this current investigation must be noted. First, the sample size was small from one centre, with one intervention. Subjects for this investigation only included those who were appropriate to attend the back class, consented to this pilot study, attended at least four out of five back classes and fully completed the outcomes measures pre- and post-intervention. Consequently the results should not be generalised to all individuals with chronic LBP. Second, no individuals in the current investigation reported a worsening of symptoms. As such, no information was available to establish reasonable cutoffs for determining how the instruments respond to a worsening of symptoms. Third, there was not a true test–retest period; the SDC was calculated for those patients who reported no change with the intervention; therefore, the results may be different if the distribution-based methods were assessed prior to any intervention for all patients. In addition, the period between administering the initial questionnaires and collecting the follow-up dated was only 5 weeks, and the findings may have been different if this follow-up period was of a longer duration. Finally, there was no control group, and there is no certainty that the changes observed related to the intervention of the natural history of the condition.

Indications for future research

This investigation was a pilot study. Future research is needed to demonstrate whether the methods used in this study will produce comparable results when applied to a larger population.

The aim of this study was to compare the ability of the questionnaires to measure change and not to evaluate the back class which served as the construct for change. However, the premise of the back class is to allow patients with chronic LBP to regain control of their pain by the re-introduction of physical fitness, mobility and previously valued activities using the principles of graded exposure which may bias the sensitivity of the outcome measures used in this study. Future research on these outcome measures should not only include larger samples and additional treatments, such as manual therapy and/or acupuncture, which have also been recommended by NICE [6].

The effect of different baseline measures was not assessed in this current investigation, but consideration of the amount of change for people with higher or lower baseline measures should be investigated in future studies. Future work should also evaluate whether an MCID for improvement is the same as for worsening of symptoms.

When measuring treatment effects in chronic LBP, both clinician and patient would benefit from an understanding of the change in common outcome measures that correspond to a meaningful improvement or worsening symptoms from a patient's perspective [14].

There is previous work on using different anchors to measure clinically important change, e.g. a priori judgment or percentage change scores; further comparisons of these methods used to calculate the MCID should be considered. The answer to the question "are you better?" could mean a complete resolution of symptoms for some, a state of acceptance to their condition or a redefinition of what being better would be like. Further qualitative work in this area is needed in terms of defining the best way to establish an anchor for the type of recovery measured with the outcome tools that are used.

Conclusion

The LBP has a significant impact on the quality of life of those affected. A well-developed and responsive outcome measure provides beneficial information to determine real change and evidence of treatment effectiveness [57]. In clinical practice, outcome measures are increasingly used as screening instruments, but there is little evidence to suggest that their use substantially changes patient management [58].One possible reason is that clinicians are sceptical of the psychometric properties of measures and have a poor understanding about the meaning of score changes. This study presented score changes in five different questionnaires to define clinical improvement. From the anchor-based methods, a score change of 4 points for the NRS, 3.5 for the RMDQ, 7.5 for the ODI, 8.5 for the PSEQ and 2.3 for the PSFS was defined as the smallest difference that patients perceive to be worthwhile. This compares to a score change of 2.4 for the NRS, 4.9 for the RMDQ, 16.7 for the ODI, 10.9 for the PSEQ, and 1.4 for the PSFS derived from distribution-based methods which ensures that the observed change is real and beyond measurement error. Evidence of robust measurement properties and a clear understanding of score change in patient-based measures are essential for clinicians to use these data to target the most appropriate treatment, to monitor the subsequent effects of treatment and thereby improve treatment outcomes for patients with LBP.

References

1. Waddell G (2006) Preventing incapacity in people with musculoskeletal disorders. Br Med Bull 77–78:55–69
2. Turk DC, Dworkin RH, Revichi D et al (2008) Identifying important outcome domains for chronic pain clinical trials: an IMMPACT survey of people with pain. Pain 137:276–285
3. Taylor W (2005) Musculoskeletal pain in adult New Zealand population: prevalence and impact. N Z Med J 118:1221
4. Lewis J, Hewitt J, Billington L et al (2005) A randomized control trial comparing two physiotherapy interventions for chronic low back pain. Spine 30:711–721
5. Heymans M, van Tulder M, Esmail R et al (2005) Back schools for nonspecific low back pain: a systematic review within the framework of the Cochrane Collaboration Back Review Group. Spine 30:2153–2163
6. NICE clinical guideline 88 (May 2009) Low back pain. Early management of persistent non-specific low back pain
7. Underwood M, Morton V, Farrin A (2007) Do baseline characteristics predict response to treatment for low back pain? Secondary analysis of the UK BEAM dataset. Rheumatology 46(8):1297–1302
8. Turk DC (2002) Clinical effectiveness and cost-effectiveness of treatments for patients with chronic pain. Clin J Pain 18:355–365
9. O'Sullivan P (2005) Diagnosis and classification of chronic low back pain disorders: maladaptive movement and control impairments as underlying mechanism. Man Ther 10:116–121
10. Wand BM, O'Connell NE (2008) Chronic non-specific low back pain—sub-groups or a single mechanism? BMC Musculoskelet Disord 9:11
11. Hurst H, Bolton J (2004) Assessing the clinical significance of change scores recorded on subjective outcome measures. J Manipulative Physiol Ther 27:26–35
12. Beaton D, Tarasuk V, Katz J et al (2001) Are you better? A qualitative study into the meaning of being better and its implications for health status measurement. Arthritis Rheumatol 42(supplement):S274
13. Terwee C (2003) On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 12:349–362
14. Hägg O, Fritzell P, Nordwall A (2003) The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 12:12–20
15. Jaeschke R, Singer J, Guyatt G (1989) Measurement of health status. Ascertaining the minimal clinical important difference. Control Clin Trials 10:407–415
16. Crosby R, Kolotkin R, Williams G (2003) Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 56:395–407
17. Fritz J, Irrgang J (2001) A comparison of a modified Oswestry low back pain disability questionnaire and the Quebec back pain disability scale. Phys Ther 81:776–788
18. Terwee C (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34–42
19. Dworkin RH, Turk DC, Farrar JT (2005) Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 113:9–19
20. Lauridsen H, Hartvigsen J, Korsholm L et al (2007) Choice of external criteria in back pain research: Does it matter? Recommendations based on analysis of responsiveness. Pain 131:112–120
21. Ostelo R, Deyo R, Stratford P et al (2008) Interpreting change scores for pain and functional status in low back pain. Spine 33:90–94
22. Roland M, Morris R (1983) A study of the natural history of back pain. Part 1: development of a reliable and sensitive measure of disability in low back pain. Spine 8:141–144
23. Fairbank J, Pynsent P (2000) The Oswestry disablility index. Spine 25:2940–2953

24. Nicholas M (1989) Self-efficacy and chronic pain. Paper presented at the annual conference of the British Psychological Society, St. Andrews
25. Stratford P, Gill C, Westaway M et al (1995) Assessing disability and change on individual patients: a report of a patient specific measure. Physiother Can 47:258–263
26. Davidson M, Keating J (2002) A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 82:8–24
27. De Vet H, Terwee C, Knol D, Bouter L (2006) When to use agreement versus reliability measures. J Clin Epidemiol 59:1033–1039
28. Ostelo R, de Vet H (2005) Clinically important outcomes in low back pain. Best Pract Res Clin Rheumatol 19:593–607
29. Deyo R, Centor R (1986) Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 39:897–906
30. Beurskens A, de Vet H, Koke A (1996) Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 65:71–76
31. Grotle M, Brox JL, Vallestad N (2004) Functional status and disability questionnaires: what do they assess? A systematic review of back-specific outcome questionnaires. Spine 30:130–140
32. Brouwer S, Kuijer W, Dijkstra P, Goeken L et al (2003) Reliability and stability of the Roland Morris questionnaire: intraclass correlation and limits of agreement. Disabil Rehabil 26:162–165
33. Department of Health (2008) High quality care for all—NHS next stage review final report, section 4. Quality at the heart of everything we do, Crown Copyright, The Stationery Office, p 47. http://www.tsoshop.co.uk
34. Wyrwich K, Tierney W, Wolinsky F (1999) Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 52:861–873
35. Childs J, Riva S, Fritz J (2005) Responsiveness of the numeric pain rating scale in patients with low back pain. Spine 30:1331–1334
36. Stratford P, Binkley J, Riddle D, Guyatt G (1998) Sensitivity to change of the Roland-Morris back pain questionnaire: Part 1. Phys Ther 78:1186–1196
37. Wyrwich K (2004) Minimal important difference thresholds and the standard error of measurement: is there a connection? J Biopharm Stat 14:97–110
38. Farrar JT, Young J, LaMoreaux L et al (2001) Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 94:149–158
39. Salaffi F, Stancati A, Silvestri C, Ciapetti A, Grassi W (2004) Minimal critical important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. Eur J Pain 8:165–172
40. Baldwin ML, Butler RJ, Johnson WG et al (2007) Self-reported severity measures as predictors of return to work outcomes in occupational back pain. J Occup Rehabil 17:68
41. Beurskens A, de Vet H, Koke A (1999) A patient specific approach for measuring functional status in low back pain. J Manipulative Physiol Ther 22:144–148
42. Pengel L, Refshauge K, Maher C (2004) Responsiveness of pain, disability, and physical impairment outcomes in patients with low back pain. Spine 29:879–883
43. Chatman A, Hyams S, Neel J et al (1997) The patient specific functional scale: measurement properties in patients with knee dysfunction. Phys Ther 77:820–829
44. Hudak P, Wright J (2004) The characteristics of patient satisfaction measures. Spine 25:3167–3317
45. Asghari A, Nicholas M (2001) Pain self-efficacy beliefs and pain behaviour. A prospective study. Pain 94:85–100
46. Nicholas M (2007) The Pain self efficacy questionnaire: taking pain into account. Eur J Pain 11:153–163
47. Williams A, Richardson P, Nicholas M, Pither C, Harding V, Ralphs J (1996) Inpatient versus outpatient pain management results of a chronic pain trial. Pain 66:13–22
48. Norman G, Stratford P, Regehr G (1997) Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 50:869–879
49. Schmitt J, Di Fabio R (2005) The value of prospective and retrospective global change criterion measures. Arch Phys Med Rehabil 86:2270–2276
50. Von Korff M, Jensen P, Karoli P (2000) Assessing global pain severity by self-report in clinical and health services research. Spine 25:3140–3151
51. De Vet H (2007) Reproducibility and responsiveness of evaluative outcome measures. Int J Technol Assess Health Care 17:479–487
52. Yelland J, Schluter P (2006) Defining worthwhile and desired responses to treatment of chronic low back pain. Pain Med 7:38–45
53. Walsh D, Kelly S, Johnson P et al (2003) Performance problems of patients with chronic low back pain and the measurement of patient-centered outcome. Spine 29:87–93
54. Westaway M, Stratford P, Binkley J (1998) The patient specific functional scale: validation of its use in persons with neck dysfunction. J Sports Phys Ther 27:331–338
55. Jordan K, Dunn K, Lewis M et al (2006) A minimal clinically important difference was derived for the Roland-Morris disability questionnaire for low back pain. J Clin Epidemiol 59:45–52
56. Dworkin R, Turk D, Wyrwrich K, Beaton D et al (2008) Interpreting the clinical importance of treatment outcomes in chronic pain trials: IMMPACT recommendations. J Pain 9:105–121
57. Haywood K (2006) Patient reported outcome I: Measuring what matters in musculoskeletal care. Musculoskeletal Care 4:187–203
58. Greenhalgh J, Long A, Flynn (2005) The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? Soc Sci Med 60:833–843