Eur Spine J (2011) 20:79–86
DOI 10.1007/s00586-010-1521-8

ORIGINAL ARTICLE

The patient-specific functional scale is more responsive than the Roland Morris disability questionnaire when activity limitation is low

Amanda M. Hall · Chris G. Maher ·
Jane Latimer · Manuela L. Ferreira ·
Leonardo O. P. Costa

Received: 20 January 2010 / Revised: 4 June 2010 / Accepted: 3 July 2010 / Published online: 14 July 2010

Abstract The primary objective of this study was to determine which questionnaire, the Roland Morris disability questionnaire (RMDQ) or the patient-specific functional scale (PSFS), was better at detecting change in activity limitation in a large cohort of patients with low back pain undergoing rehabilitation. A secondary aim was to determine if the responsiveness of the questionnaires was influenced by the patient's level of activity limitation at baseline. Responsiveness statistics, including effect size statistics, Pearson's *r* correlations and receiver operative characteristic (ROC) curve analysis were used to determine ability to detect change in activity limitation on 831 patients with low back pain. Data were analysed at two time points; directly after treatment (termed short-term) and several weeks post-treatment (termed mid-term). The data were subsequently re-analysed on sub-sets of the full cohort according to the level of activity limitation from RMDQ baseline scores. In the total cohort we found that the PSFS was more responsive than the RMDQ; however, in the subgroup with high activity limitation this pattern was not observed. This is true for time points up to 6 months post-treatment. In conclusion, the RMDQ and PSFS both demonstrate good responsiveness according to the definitions given in previous guidelines. The PSFS is more responsive than the RMDQ for patients with low levels of activity limitation but not for patients with high levels of activity limitation.

Keywords Back pain · Disability · Activity limitation · Rehabilitation · Responsiveness

Introduction

A patient's level of a*ctivity limitation* is widely used as a primary indicator of successful treatment or rehabilitation for patients with spinal conditions. In order to determine the effects of a treatment on improving *activity limitation* for patients with low back pain, researchers and clinicians require appropriate tools for accurate assessment. Two of the most commonly used measures to assess activity limitation associated with low back pain are the Roland Morris disability questionnaire (RMDQ) [1] and the patient-specific functional scale (PSFS) [2]. Both been shown to provide reliable measures of activity limitation [1, 3]; however, there is some uncertainty regarding their responsiveness. Responsiveness is defined as the ability of an outcome measure to detect true change in a patient's health status over time [4]. Since recent research has been reported that treatments for low back pain produce small to moderate effects at best with little difference between treatment type [5, 6], it is crucial that instruments used to assess outcome are sensitive enough to detect small changes over time.

The literature defines two aspects of responsiveness: *internal responsiveness* and *external responsiveness*, each

A. M. Hall (✉) · J. Latimer · L. O. P. Costa
The George Institute for International Health,
Faculty of Medicine, The University of Sydney,
PO Box M201, Level 7, 341 George St,
Missenden Road, Sydney, NSW 2050, Australia
e-mail: amandahall@george.org.au

C. G. Maher
Musculoskeletal Division, The George Institute
for International Health, Faculty of Medicine,
The University of Sydney, PO Box M201,
Level 7, 341 George St, Missenden Road,
Sydney, NSW 2050, Australia

M. L. Ferreira
The George Institute for International Health,
The Faculty of Health Sciences,
The University of Sydney, Sydney, Australia

having their own definition and methods for assessment. Husted et al. [7] has defined *internal responsiveness* as the ability of a measure to change over a pre-specified time frame, and *external responsiveness* as the relationship between the change in a measure and the corresponding change in an external standard.

It may be argued that a patient-specific measure such as the PSFS would be more responsive than a disease-specific measure such as the RMDQ. Despite the fact that previous studies have investigated the relative responsiveness of the RMDQ and the PSFS [10–14], there is still uncertainty as to which questionnaire is more responsive. In the five studies that have examined *internal responsiveness*, four of the five studies concluded that the PSFS is superior to the RMDQ [8–11], while one found the RMDQ more responsive [12]. In the four studies [10, 12–14] assessing *external responsiveness* there was little agreement in the findings of these studies. Two [9, 10] found no difference between the two questionnaires with regard to responsiveness, one found that the PSFS [11] was more responsive and one found the RMDQ [12] more responsive.

There may be several potential reasons for the different findings with regard to responsiveness. First, all studies used relatively small samples which may adversely impact the precision of the estimate of responsiveness [13] and second, the studies enrolled patients with quite different levels of activity limitation with RMDQ mean scores ranging from 5.7 to 12.0 (out of 24). It is plausible that the RMDQ may be less responsive in people with minimal activity limitation as the items tend to reflect very high levels of activity limitation (e.g. “I stay in bed most of the time because of my back pain”). However, the question of whether the degree of activity limitation affects the relative responsiveness of these instruments has not been systematically investigated.

The primary aims of this study are to (a) compare the responsiveness of the RMDQ and the PSFS in a large cohort of low back pain patients and (b) determine if the relative responsiveness of these measures depends on the degree of activity limitation. We also aimed to determine if the responsiveness of these questionnaires was influenced by the duration of follow-up.

Methods

Design

We obtained data from four randomized controlled trials [14–17] studying patients with low back pain that were registered with Australia New Zealand Clinical Trials Registry and accessible to the author. In all trials PSFS and RMDQ were administered by an investigator who was blinded to treatment allocation and follow-up rates were above 85% (Table 1). In all trials, activity limitation was measured by the RMDQ and the PSFS at baseline, directly after treatment which we termed ‘short-term follow-up’ and at a follow-up time-point between 3 and 6 months from baseline which we termed ‘mid-term follow-up’.

The RMDQ consists of 24 yes/no items regarding activities of daily living that may be affected by back pain. Each item that is answered “yes” is scored one point with scores ranging from 0 representing “no disability” to 24 representing “extremely severe disability” [1]. The PSFS differs to the RMDQ in that it asks the person to nominate three important activities they are not able to do, or are having difficulty performing, because of their back pain. Then, each activity is scored on a Likert scale ranging from 0 (unable to perform the activity) to 10 (able to perform the activity at pre-injury level). The scores are then summed and averaged yielding a total score out of 10 [18].

Participants/data source

All studies recruited subjects with low back pain residing in New South Wales, Australia or Auckland, New Zealand from primary and tertiary care settings as well as from community volunteers. Individual studies’ inclusion criteria differed by duration of pain symptoms, which ranged from 0 to 6 weeks duration (terms acute pain) [15], 6–12 weeks duration (termed sub-acute pain) [17] and pain persisting for more than 3 months in duration (termed chronic pain) [14, 19] these definitions are recommended by the Cochrane Back Review Group [20]. Characteristics of the included trials including trial design, type of intervention and participant’s characteristics of age, gender, pain level and activity limitation level at baseline are presented in Table 1. Each trial has been rated for methodological quality and statistical reporting using the 11-item PEDro scale [21] The items are: (1) eligibility criteria and source; (2) random allocation; (3) concealed allocation; (4) baseline comparability; blinding of (5) subjects, (6) therapists, and (7) assessors; (8) adequate follow-up; (9) intention-to-treat analysis; (10) between-group statistical comparisons; and (11) point measures and measures of variability reported. The last ten items are used to calculate the total PEDro score, with this score determined simply by the number of items met (item 1 is not included in the total PEDro score as it relates to generalizability rather than internal validity or statistical reporting). Each trial is evaluated by two independent raters and, when there is disagreement between the raters for any item, arbitration by a third rater if necessary. Reliability is moderate for consensus ratings of individual items of the PEDro scale (kappa values range from 0.50 to 0.79) and also moderate for the total PEDro score [intraclass

Table 1 Characteristics of included RCTs

Characteristic	Hancock [15]	Pengel [17]	Ferreira [14]	Costa [28]
Source of subjects	GP referral	Health Care Professional Referral Community Advertisement Hospital wait list	Physiotherapist referral Teaching hospital	Physiotherapist referral Teaching hospital
N (randomized)	240	260	240	154
Follow-up (%)	98	89	88	94
Pain duration (weeks)	0–6	6–12	>12	>12
Age mean (SD)	40.7 (15.6)	49.9 (15.8)	53.6 (15.0)	53.7 (12.8)
Women (%)	44	48	69	60
Pain intensity[a] [mean (SD)]	6.5 (1.7)	5.4 (2.0)	6.3 (2)	6.7 (2.1)
RMDQ-24[b] [mean (SD)]	13.1 (5.4)	8.4 (5.0)	13.5 (5.5)	13.3 (5.0)
PSFS[c] [mean (SD)]	3.9 (1.8)	3.8 (1.9)	3.6 (1.4)	3.3 (1.8)
Trial design	Factorial	Factorial	3 arm trial	2 arm trial
Intervention type	1. NSAIDs 2. Placebo NSAIDs 3. SMT 4. Placebo SMT	1. Ex. and advice 2. Sham Ex. and advice 3. Ex. and Sham advice 4. Sham Ex. and Sham advice	1. Spinal manip. therapy 2. Motor control Ex. 3. General Ex.	1. Motor control Ex. 2. Placebo
Intervention duration (weeks)	4[d] (max)	6	8	8
Number of sessions	12 (max)	12	12	12
Trial quality[e]	9/10	9/10	8/10	9/10

SMT spinal manipulative therapy, *Ex.* exercise

[a] The 0–10 NRS measures the participant's level of pain on a scale of 0–10 where 0 is "no pain" and 10 is "worst pain possible"

[b] The RMDQ consists of 24 yes/no items with scores ranging from 0 representing "no disability" to 24 representing "extremely severe disability

[c] The PSFS asks the participant to nominate 3 important activities they are not able to do or are having difficulty performing because of their back pain. Each activity is scored on a numerical rating scale from 0 (unable to perform the activity) to 10 (able to perform the activity at pre-injury level). The scores are then summed and averaged yielding a total score out of 10

[d] Duration of treatment was until patient was recovered up to a maximum of 4 weeks

[e] Trial quality was rated using the PEDro scale which is a 11-point scale used to assess risk of bias in randomized controlled trials [21]

correlation coefficient (type 1, 1) 0.68]. The PEDro score for each trial is listed in Table 1.

Data analysis

We measured both *internal* and *external responsiveness* of the questionnaires. All analyses were conducted using SPSS 17.0 software. *Internal responsiveness* was calculated using effect size (ES) statistics, which relates the magnitude of change of the patients to some measure of variation in the population. While there have been different measures of variation proposed, we chose to express ES in terms of two measures of variation (a) variation at baseline and (b) variation of the change scores, and to compare the results; ES(a) = mean change/SD of baseline score and ES(b) = mean change/SD of change score also referred to as the standardized response mean (SRM) [7]. For both methods, we calculated ES using paired *t* tests with 84% confidence intervals for both RMDQ and PSFS. We chose 84% confidence intervals because non-overlapping 84% confidence intervals are equivalent to a *Z* test of means at the 0.05 level [11, 22].

External responsiveness requires comparing the responsiveness on the questionnaire being studied to an external criterion of health status. While there is no universally recommended gold standard for an external criterion of health status, we decided to use the global perceived effect scale (GPE) because it was common to all included trials and has been previously used for this purpose [9, 11]. The GPE used in all trials is an 11-point scale which ranges from −5 being "vastly worse" to +5 being "completely recovered" and 0 being "unchanged" [23].

The first method we used to calculate *external responsiveness* treated the GPE as a continuous outcome measure and used the Pearson's *r* correlation test to measure the correlations between the change scores (baseline to follow-up) on both the PSFS and RMDQ with the follow-up GPE score. To determine if the correlations were significantly different, we compared the *r* values using Cohen's test of paired correlations [24].This method provides information regarding how the change scores of the measure we evaluated and the scores on the external criterion varied together.

The second measure of *external responsiveness* relied upon receiver operating characteristics (ROC) curve. The area under the ROC curve describes the ability of a questionnaire to distinguish patients who have and have not changed according to an external criterion [13]. We continued to use the GPE as the external criterion and dichotomized it into those who changed (a score of 4 or more on the GPE, *improved*) and those who did not change (a score of 3 or less on GPE, *not improved*). We also conducted a sensitivity analysis by conducting new analyses using a stricter cut-off (a score of 5 indicating *improved* and a score of 4 or less as *not improved*) and a less strict cut-off (a score of 3 or greater indicating *improved* and a score of 2 or less as *not improved*). To determine if the areas under the ROC curves for the RMDQ and PSFS outcomes were statistically different we used the HONG KONG ROC Statistical Program [25] to calculate a Delong statistic for each pair of RMDQ and PSFS ROC curves [26].

Effect of disability level and length of follow-up on responsiveness:

The responsiveness analyses were run on the full dataset and then on subsets of the data defined according to activity limitation at baseline. Median baseline RMDQ scores were used to divide the cohort into low and high activity limitation subgroups. Analyses were run for short and mid-term follow-up. The short-term time-point was measured directly after treatment and was slightly different for each study (4 weeks [15], 6 weeks [17], 8 weeks [14, 16]). The 'mid-term' time-point was measured several weeks post-treatment and varied amongst studies (12 weeks [14–16], 26 weeks [14, 16]).

Results

Participants

All participants included in this study were recruited from New South Wales, Australia or Auckland, New Zealand. The mean (SD) pain scores at baseline were 6.2 (2.0) on a 0–10 pain scale. Disease-specific activity limitation measured with the RMDQ was 11.9 (5.6) on the 0–24 scale. Patient-specific activity limitation measured with the PSFS was 3.8 (1.7) on the 0–10 scale. Descriptive information on each included trial is presented in Table 1. Descriptive information on the sample included for analysis in this study is presented Table 2.

Internal responsiveness (Tables 3, 4)

The RMDQ and PSFS each demonstrated good internal responsiveness indicated by effect sizes that were above the cut-off of 0.8 recommended by Husted et al. [7]. This demonstrates that both questionnaires are good at detecting change in activity limitation in patients with low back pain. At both short- and mid-term follow-up, the PSFS was more responsive as evidenced by non-overlapping 84% CI.

When we investigated if these questionnaires demonstrate the same responsiveness in people with different levels of activity limitation, we found that the PSFS was more responsive than the RMDQ in patients presenting with "low activity limitation" at baseline with all four comparisons statistically significant. For patients with "high activity limitation" at baseline, the RMDQ was more responsive with the difference in effect sizes statistically significant for two of the four comparisons.

Table 2 Characteristics of study participants at baseline

Characteristic	Full LBP cohort	Low activity limitation[a]	High activity limitation[b]
n	831	453	378
Age (years)	49.7 (15.9)	50.4 (16.9)	48.9 (14.6)
Gender			
Female (%)	56	54	57
Pain intensity			
0–10 NRS	6.2 (2.0)	5.5 (1.9)	7.0 (1.8)
Activity limitation			
RMDQ-24 item	11.9 (5.6)	7.6 (3.2)	17.1 (2.9)
PSFS	3.8 (1.7)	4.3 (1.8)	3.1 (1.5)

Mean scores and Standard deviations are reported

[a] Classified by a baseline RMDQ score of 0–12

[b] Classified by a baseline RMDQ score of 13–24

External responsiveness (Tables 3, 4)

Pearson correlations demonstrated that both the RMDQ and PSFS correlated well with the GPE indicating similarities in their ability to detect when a meaningful change has occurred to the patient. The correlations were statistically significantly different for both comparisons. Both questionnaires were above cut-offs for acceptable area under the curve (AUC > 0.70) and are therefore able to distinguish between patients who improve and those who do not according to the GPE. The primary ROC analysis revealed that the PSFS demonstrated a greater AUC than the RMDQ; however, this result was not consistently observed with the sensitivity analyses. While there was a trend for the PSFS to be superior for "low activity limitation" and the RMDQ for "high activity limitation" there was no consistent evidence from the statistical comparison of the paired AUC values.

Table 3 Internal and external responsiveness of RMDQ and PSFS at short-term follow-up

	n	Effect size (a) (84% CI)	Effect size (b) (84% CI)	Correlations of the change scores with GPE at discharge	*P**	AUC (GPE cut-off for improvement = 3 or better)	*P***	AUC (GPE cut-off for improvement = 4 or better)	*P***	AUC (GPE cut-off for improvement = 5 or better)	*P***
LBP											
RMDQ	831	1.01 (0.96–1.06)	0.92 (0.87–0.97)	0.57	0.03	0.85	0.48	0.81	0.001	0.83	0.01
PSFS	831	1.70 (1.63–1.78)	1.11 (1.07–1.16)	0.61		0.85		0.85		0.87	
Low activity limitation											
RMDQ	453	1.08 (0.99–1.17)	0.81 (0.74–0.87)	0.59	1.0	0.81	0.34	0.80	0.06	0.81	0.01
PSFS	453	1.55 (1.46–1.65)	1.11 (1.04–1.17)	0.60		0.82		0.84		0.87	
High activity limitation											
RMDQ	378	2.84 (2.67–3.01)	1.20 (1.12–1.27)	0.70	0.01	0.92	0.01	0.88	0.37	0.90	0.09
PSFS	378	2.16 (2.02–2.30)	1.12 (1.06–1.21)	0.65		0.88		0.88		0.87	

Short-term follow-up refers to time-point at discharge from treatment: acute = 4 weeks, sub-acute = 6 weeks, chronic = 8 weeks

Effect size (a): *Standardized effect size* defined as the difference between the mean baseline scores and follow-up scores divided by the standard deviation of baseline scores [7]

Effect size (b): *Standardized response mean* defined as the difference between the mean baseline scores and follow-up scores divided by the standard deviation of the change score [7]

The effect sizes are significantly different if the 84% CI between the RMDQ and PSFS do not overlap [11]. Since the CI do not overlap for each cohort listed in this table, the effect sizes are significantly different for the PSFS and RMDQ

* *P* values of two-tailed test for paired Pearson's *r* correlations Cohen and Cohen [22, 29]

** *P* values are from DeLong's test of paired AUC values [24]

Table 4 Internal and external responsiveness of RMDQ and PSFS at mid-term follow-up

	n	Effect size (a) (84% CI)	Effect size (b) (84% CI)	Correlations of the change scores with GPE at discharge	*P**	AUC (GPE cut-off for improvement = 3 or better)	*P***	AUC (GPE cut-off for improvement = 4 or better)	*P***	AUC (GPE cut-off for improvement = 5 or better)	*P***
LBP											
RMDQ	820	1.04 (0.98–0.09)	0.88 (0.83–0.93)	0.59	0.00	0.83	0.77	0.83	0.05	0.84	0.26
PSFS	820	1.80 (1.72–1.88)	1.11 (1.06–1.16)	0.65		0.86		0.86		0.85	
Low activity limitation											
RMDQ	452	1.13 (1.03–1.22)	0.767 (0.7–0.83)	0.63	0.20	0.83	0.43	0.81	0.37	0.77	0.28
PSFS	452	1.74 (1.65–1.84)	1.81 (1.11–1.25)	0.61		0.84		0.81		0.76	
High activity limitation											
RMDQ	368	2.97 (2.72–3.10)	1.13 (1.05–1.20)	0.72	0.04	0.89	0.18	0.92	0.17	0.93	0.16
PSFS	368	2.16 (2.00–2.31)	1.04 (0.97–1.12)	0.70		0.87		0.90		0.91	

Mid-term follow-up refers to time-point 6–18 weeks after discharge from treatment: acute = 12 weeks, sub-acute = 12 weeks, chronic = 26 weeks

Effect size (a): *Standardized effect size* defined as the difference between the mean baseline scores and follow-up scores divided by the standard deviation of baseline scores [7]

Effect size (b): *Standardized response mean* defined as the difference between the mean baseline scores and follow-up scores divided by the standard deviation of the change score [7]

The effect sizes are significantly different if the 84% CI between the RMDQ and PSFS do not overlap [11, 30]. Since the CI do not overlap for each cohort listed in this table, the effect sizes are significantly different for the PSFS and RMDQ

* *P* values of two-tailed test for paired Pearson’s *r* correlations Cohen and Cohen [22, 29]

** *P* values are from DeLong’s test of paired AUC values [26]

Discussion

This study including a cohort of 831 patients with low back pain is the largest study to date to investigate the responsiveness of the PSFS and RMDQ in a head to head comparison. In the total cohort we found that the PSFS was clearly more responsive than the RMDQ; however, in the subgroup with high activity limitation this pattern was not observed and there was some evidence that the RMDQ was superior to the PSFS. For the patients with low activity limitation the PSFS was consistently the more responsive measure. To our knowledge this is the first study to find evidence that the level of activity limitation may influence the rankings of responsiveness of these two activity limitation measures.

In this cohort, both questionnaires met the established criteria for good internal and external responsiveness suggested by the current clinimetric guidelines [13]. This finding held true regardless of how (internal vs. external*)* or when (short-term and mid-term follow-up) responsiveness was measured. These results indicate that both questionnaires are good at detecting change in activity limitation in patients with low back pain.

Level of activity limitation clearly affected the relative responsiveness of these two measures. The PSFS is clearly more responsive for people with “low activity limitation” (0–12 on the RMDQ) at baseline but not for people with “high activity limitation” (13–24 on the RMDQ) at baseline. This result may explain why some studies showed greater responsiveness for the PSFS and others the RMDQ. The three previous studies that reported the PSFS to be more responsive used populations in which the mean RMDQ at baseline was low (5.7–8.7) [9–11]. The single study reporting the RMDQ as more responsive used a population with a mean RMDQ of 12.0 [12].

Implications for treatment providers and researchers

The results of this study have several important *implications* for treatment providers and researchers. First, the magnitude of the effect size has been categorized into small ($ES < 0.2$), medium ($ES = 0.5$) and large ($ES \geq 0.8$) to help facilitate decisions regarding choice of outcome measure [7]. The ES of both RMDQ and PSFS are above 0.8 and considered large for the complete cohort of LBP patients at follow-up time points up to 6 months post-treatment. Second, there has been increasing research investigating the large number of people in the community with low back pain who are not seeking care but may benefit from community-based treatments. Given the evidence that this group of people report lower levels of activity limitation (up to five points less on the RMDQ) [27], the PSFS would be preferable for assessing outcome.

Strengths

There are a number of strengths in this study. First, there have been various methods to measure responsiveness in previous trials; the statistical methods used in this study included two measures for both internal and external responsiveness which allowed comparison of the results for consistency. A second strength is the size and clinical characteristics of the sample which allowed us to investigate responsiveness as a function of baseline disability level which has not been done in previous studies. Lastly, responsiveness was calculated using the data from four high-quality randomized controlled trials including patients from different sources and receiving different treatments, allowing us to generalize our results to a wider population base than any other study in this area.

Limitations

The authors recognize that one of the limitations of this study may be the choice of the GPE for the external criterion of health status. While the GPE is commonly used as an external criterion, there is no gold-standard measure of true change. Thus, our estimates of external responsiveness are only as accurate as the extent to which the external criterion actually reflects true change in patient’s overall health status.

Conclusion

The RMDQ and PSFS demonstrate good internal and external responsiveness with both instruments being above the required cut-off points recommended in previous guidelines. This is true for time points up to 6 months post-treatment. The PSFS is more responsive than the RMDQ for patients with low levels of activity limitation but not for patients with high levels of activity limitation.

References

1. Morris R (1983) A study of the natural-history of back pain.1. Development of a reliable and sensitive measure of disability in low-back pain. Spine 8:141–144
2. Stratford P, Gill C, Westaway M, Binkley J (1995) Assessing disability and change on individual patients: a report of a patient specific measure. Physiother Can 47:258–263
3. Stratford PW, Binkley J, Solomon P, Gill C, Finch E (1994) Assessing change over time in patients with low back pain. Phys Ther 74:528–533

4. Guyatt G, Walter S, Shannon H, Cook D, Jaeschke R, Heddle N (1995) Basic statistics for clinican. 4. Correlation and regression. Can Med Assoc J 152:497–504
5. Chou R, Huffman LH (2007) Nonpharmacologic therapies for acute and chronic low back pain: a review of the evidence for an American pain Society/American college of physicians clinical practice guideline. Ann Intern Med 147:492–504
6. Machado LAC, Kamper SJ, Herbert RD, Maher CG, McAuley JH (2009) Analgesic effects of treatments for non-specific low back pain: a meta-analysis of placebo-controlled randomized trials. Rheumatology 48:520–527. doi:10.1093/rheumatology/ken470
7. Husted JA, Cook RJ, Farewell VT, Gladman DD (2000) Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 53:459–468
8. Beurskens AJ, de Vet HC, Koke AJ, Lindeman E, van der Heijden GJ, Regtop W et al (1999) A patient specific approach for measuring functional status in low back pain. J Manip Physiol Ther 22:144–148
9. Costa LOP, Maher CG, Latimer J, Ferreira PH, Ferreira ML, Pozzi GC et al (2008) Clinimetric testing of three self-report outcome measures for low back pain patients in Brazil. Spine 33:2459–2463
10. Frost H, Lamb S, Stewart-Brown S (2008) Responsiveness of a patient specific outcome measure compared with Oswestry disability index v2.1 and Roland Morris disability questionnaire for patients with subacute and chronic low back pain. Spine 33:2450–2457
11. Pengel LHM, Refshauge KM, Maher CG (2004) Responsiveness of pain, disability, and physical impairment outcomes in patients with low back pain. Spine 29:879–883
12. Beurskens AJ, de Vet HC, Koke AJ (1996) Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 65:71–76
13. Terwee CB, Bot SDM, de Boer MR, van der Windt D, Knol DL, Dekker J et al (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34–42
14. Ferreira ML, Ferreira PH, Latimer J, Herbert RD, Hodges PW, Jennings MD et al (2007) Comparison of general exercise, motor control exercise and spinal manipulative therapy for chronic low back pain: a randomized trial. Pain 131:31–37
15. Hancock MJ, Maher CG, Latimer J, McLachlan AJ, Cooper CW, Day RO et al (2007) Assessment of diclofenac or spinal manipulative therapy, or both, in addition to recommended first-line treatment for acute low back pain: a randomised controlled trial. Lancet 370:1638–1643
16. Maher C, Latimer J, Hodges P, Refshauge K, Moseley L, Herbert RD et al (2005) The effect of motor control exercise versus placebo in patients with chronic low back pain [ACTRN012605000262606]. BMC Musculoskelet Disord 6(ARTN 54):1–8. doi:10.1186/1471-2474-6-54
17. Pengel LHM, Refshauge KM, Maher CG, Nicholas MK, Herbert RD, McNair P (2007) Physiotherapist-directed exercise, advice, or both for subacute low back pain—a randomized trial. Ann Intern Med 146:787–796
18. Stratford PW, Binkley JM, Riddle DL (2000) Development and initial validation of the back pain functional scale. Spine 25:2095–2102
19. Costa LDM, Henschke N, Maher CG, Refshauge KM, Herbert RD, McAuley JH et al (2007) Prognosis of chronic low back pain: design of an inception cohort study. BMC Musculoskelet Disord 8(ARTN 11):1–4
20. Higgins J, Green S (2008) Cochrane handbook for systematic reviews of interventions version 5.0.0. The Cochrane Collaboration
21. Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M (2003) Reliability of the PEDro scale for rating quality of randomized controlled trials. Phys Ther 83:713–721
22. Payton ME, Miller AE, Raun WR (2000) Testing statistical hypotheses using standard error bars and confidence intervals. Commun Soil Sci Plant Anal 31:547–551
23. Kamper SJ, Maher CG, Mackay G (2009) Global rating of change scales: a review of strengths and weaknesses and considerations for design. J Man Manip Ther 17 (in press)
24. Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences. L. Erlbaum Associates, Hillsdale
25. Cheng A. The HONG KONG ROC program. http://department.obg.cuhk.edu.hk/researchsupport/statmenu.asp
26. Delong ER, Delong DM, Clarkepearson DI (1988) Comparing the areas under two or more correlated receiver operating characteristic curves—a nonparametric approach. Biometrics 44:837–845
27. Ferreira ML, Machado G, Latimer J, Maher C, Ferreira PH, Smeets RJ (2009) Factors defining care-seeking in low back pain—a meta-analysis of population based surveys. Eur J Pain (in press)
28. Costa LOP, Maher CG, Latimer J, Hodges PW, Herbert RD, Refshauge KM et al (2009) Motor control exercise for chronic low back pain: a randomized placebo-controlled trial. Phys Ther 89:1275–1286
29. Cohen J (1988) Statistical power analysis for the behavioural sciences. L. Erlbaum Associates, Hillsdale
30. Tyron W (2001) Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an intergrated alternative method of conducting null hypothesis statistical tests. Psychol Methods 6:371–386