ORIGINAL ARTICLES

THE BOURNEMOUTH QUESTIONNAIRE: A SHORT-FORM COMPREHENSIVE OUTCOME MEASURE. II. PSYCHOMETRIC PROPERTIES IN NECK PAIN PATIENTS

Jennifer E. Bolton, PhD, and B. Kim Humphreys, DC, PhD

ABSTRACT

Objective: To modify an existing outcome measure (Bournemouth Questionnaire [BQ]) for use in patients with nonspecific neck pain and test its psychometric properties.

Design: Prospective longitudinal study in which the questionnaire was administered on 3 occasions (pretreatment, retest, and posttreatment).

Setting: Anglo-European College of Chiropractic outpatient clinic and 8 field chiropractic practices.

Method: Seven core items relating to the biopsychosocial model of pain were included in the original questionnaire (back BQ). The wording of one of these items (disability in activities of daily living) was modified to include activities likely to be affected by neck pain. Testing of the neck BQ was carried out in 102 patients with nonspecific neck pain.

Results: The instrument demonstrated high internal consistency on 3 administrations (Cronbach's alpha = 0.87, 0.91, 0.92). All 7 items were retained on the basis that they each significantly contributed to the total score (item-corrected total score correlations >0.43) and to the instrument's responsiveness to clinical change (item change-corrected total change score correlations >0.42). The instrument was reliable in test-retest administrations in stable subjects (ICC = 0.65). The instrument demonstrated acceptable construct validity and longitudinal construct validity with established external measures. The treatment effect size of the instrument was found to be high (1.67).

Conclusion: The neck BQ covers the salient dimensions of the biopsychosocial model of pain, is quick and easy to complete, and has been shown to be reliable, valid, and responsive to clinically significant change in patients with nonspecific neck pain. Its use as an outcome measure in clinical trials and outcomes research is recommended. (J Manipulative Physiol Ther 2002;25:141-8)

Key Indexing Terms: *Neck Pain; Outcome Measure; Reliability; Validity; Responsiveness*

INTRODUCTION

Musculoskeletal conditions are extremely common and costly in terms of individual suffering, treatment, work absenteeism and compensation payments. Although much of the focus in the past has been on low back pain, there is now increasing recognition that neck pain, either nonspecific in origin or as a result of trauma, substantially contributes to these costs.[1] As with back pain, the underlying disease and cause in neck pain remain unclear, there is a significant risk of chronicity, and permanent disability and invalidity ensue for a small proportion of sufferers.[2,3] It is because of these and other similarities with back pain that when dealing with neck pain, an illness (biopsychosocial) rather than disease (medical) model is considered more appropriate.[4]

Therapies for neck pain are aimed at the relief of pain, stiffness, and disability and include exercise, mobilization, traction, acupuncture, and manipulation.[5,6] However, to date, there is not enough evidence in terms of both quantity and quality of clinical trials to make any strong recommen-

Anglo-European College of Chiropractic, Bournemouth BH5 2DF, England.

Submit reprint requests to: J.E. Bolton, PhD, MA, Anglo-European College of Chiropractic, 13-15 Parkwood Rd, Bournemouth BH5 2DF, England.

Paper submitted November 21, 2000; in revised form January 8, 2001.

0161-4754/2002/$35.00 + 0 **76/1/123333**
doi:10.1067/mmt.2002.123333

dations on treatment for neck pain.[4,7-9] Given the present climate of evidence-based medicine and practice, there is considerable scope for further studies into the effectiveness and efficacy of treatment interventions for nonspecific and traumatic neck pain.

Arguably, part of the reason for the paucity of good quality trials in neck pain is a lack of standardized outcome measures. The complexity of musculoskeletal pain as an illness means that there are several salient dimensions that can be measured, including physical impairment, pain intensity, and disability. There now seems to be some consensus, however, that although physical impairment and functional capacity measures such as muscle strength and endurance, and range of motion, might be useful as secondary outcomes, primary measures should center on those outcomes of direct relevance to the patient, including pain, disability, overall perceptions of improvement, and quality of life.[10,11]

In contrast to the relatively high number of condition-specific measures available for pain and disability in low back pain,[10] it remains somewhat surprising that there are very few measures for pain and disability in neck pain. To date, pain and disability indexes that have been specifically designed for use in neck patients include the Neck Disability Index (NDI),[12] the Northwick Park Neck Pain Questionnaire,[13] the Copenhagen Neck Functional Disability Scale,[14] and most recently, the Neck Pain and Disability Scale.[15] Of these measures, the NDI is the most commonly used in neck pain studies. All of these measures, however, concentrate on pain and disability and do not include other dimensions in the illness model, namely the affective and the cognitive.

As a result, there is a need for a self-report measure that incorporates the affective and cognitive aspects of neck pain, in addition to pain severity and disability. Although there are a number of established psychological measures covering the affective and cognitive domains,[16] they have neither been specifically designed for use nor tested in patients with neck pain. Moreover, in a recent study[17] it has been argued that although a number of instruments exist that can be used to measure the salient dimensions in the illness model, to use them all is impractical, particularly in studies evaluating the effectiveness of treatment interventions in which patient outcomes are documented in the busy, routine clinical setting.

As a result of the need for documenting outcomes in patients with back pain in the clinic setting, the Bournemouth Questionnaire (BQ) was recently developed and tested.[17] The BQ is a comprehensive, valid, and reliable outcome measure reflecting the multidimesionality of the musculoskeletal illness model, but at the same time short and practical for repeated use in both clinic- and research-based settings. The aims of this study were therefore to (1) modify the back BQ for use in patients with nonspecific neck pain, and (2) to test its validity, reliability, and responsiveness.

Methods

Questionnaire Development and Testing

Item selection, wording, and scaling. Construction and testing of the original BQ for use in patients with back pain are described by Bolton and Breen.[17] The BQ was devised from the salient dimensions of the biopsychosocial (or illness) model first described by Waddell[18] for back pain. On the basis of the assumption that neck pain, like back pain, is explained by an illness model of musculoskeletal pain, the same 7 core items that make up the back BQ were selected for the neck BQ. These core items were not altered apart from some minor modifications to the wording of the disability in activities of daily living (ADL) scale (scale 2, see Appendix). Thus the activities "walking," "climbing stairs," and "getting in/out of bed/chair" included in the back BQ were omitted and replaced by the activities "lifting," "reading," and "driving."[12,14] No other changes were made to the original questionnaire, including the scaling responses for each of the items. A copy of the neck BQ is shown in the Appendix.

Reliability, validity, and responsiveness. These psychometric properties were evaluated in the same way as described for the back BQ,[17] on the basis of the methodologic frameworks outlined by Kirshner and Guyatt[19] and Streiner and Norman.[20] In brief, any measure that evaluates longitudinal change over time (outcome measure) must tap areas relevant and responsive to change in the condition under test (item selection), show stable intrasubject variation with insignificant variation between stable replicate measures (test-retest reliability), display a strong relationship between change scores and change scores in established measures over time (external longitudinal construct validity), and have the power to detect clinically important differences over time (responsiveness).[19]

Data Collection

The study was multicenter in that 8 field chiropractic practices and 1 teaching outpatient chiropractic clinic recruited patients to the study. New patients or patients with a new complaint of neck pain were asked to complete a battery of questionnaires on their first visit before seeing the practitioner (pretreatment questionnaires). This battery of questionnaires consisted of the neck BQ, the NDI,[12] the Copenhagen Neck Functional Disability Scale (NFDS),[14] and the SF36 (a generic health status measure).[21] The NDI and NFDS have been shown to be valid and reliable in patients with neck pain, although, as far as we are aware, the responsiveness of these instruments has not been reported. The SF36 is a widely used and validated short form of the US Medical Outcomes Survey questionnaire.[21,22] It contains 8 subscales, each assessing a dimension of function or

health status. Unlike the NDI and NFDS, it produces 8 separate scores rather than a single total score.

On leaving the clinic the same day, patients were asked to complete a second neck BQ to evaluate reliability (test-retest). To minimize memorization of responses to the initial questionnaire, the order of the 7 core items on the retest questionnaire was scrambled. Only those patients who reported that their condition remained unchanged from that when completing the pretreatment questionnaire by use of an 11-point global improvement scale were included in the test-retest reliability analysis. Patients completed, by mail, the same battery of questionnaires (posttreatment) as that administered before starting their treatment, approximately 4 to 6 weeks later.

Data Analysis

All data were analyzed in the same way as that reported for the psychometric testing of the back BQ.[17] In brief, the following data analyses were used:

Homogeneity and reliability. To determine the homogeneity of the 7 core items so that they could be summed to give a total score, Cronbach's alpha coefficient and item-corrected total correlation coefficients were used.[20] Homogeneity is considered acceptable when Cronbach's alpha exceeds approximately 0.7 but is not higher than approximately 0.9, and item score correlations (Pearson's correlation coefficient, *r*) with corrected total scores are not less than 0.2.[20]

Test-retest reliability was investigated using the intraclass correlation coefficient (ICC), calculated from a repeated measures 2-way analysis of variance table.[20] ICC values ascend from 0 to 1 on a continuum of increasing strength of agreement.

Validity. Face validity and construct validity were not subjectively judged in this study because these analyses had already been undertaken in the back BQ.[17] Because the original questionnaire was essentially unchanged (apart from minor modifications to the wording of 1 of the 7 core items), it was considered unnecessary to repeat these procedures for the neck BQ.

External construct validity was tested by calculating the correlation (Pearson's *r*) between the scores of the items in the BQ with those of their counterpart external measures.[20] Similarly, the longitudinal construct validity[19] of the BQ was tested by calculating the correlation (r) of the within-subject longitudinal changes in item scores with those of established measures.

Responsiveness. In contrast to validity and reliability, responsiveness is often neglected in the psychometric testing of an instrument.[23] For an outcome measure, it is essential that it is able to detect clinical change over time. In this study, only those patients who reported that they had improved on the posttreatment questionnaire by use of an 11-point global improvement scale were included in testing responsiveness of the neck BQ. Patients' self-report of improvement was considered a clinically significant change for the purposes of data analyses in this study.

Internal responsiveness of each of the 7 core items of the questionnaire was investigated by determining the strength of the correlation (*r*) between the change scores for each item and the corrected total change score. This correlation should be 0.3 or greater to ensure that each item contributes significantly to the overall responsiveness of the instrument.[24]

Comparing the ability to detect clinically significant change between the BQ and established measures tested external responsiveness. There are several methods of calculating external responsiveness.[23] In this study, 2 methods were used, both of which give the treatment effect size for an outcome measure. In the first case, the effect size was calculated according to the method of Kaziz et al,[25] in which the mean change in scores is divided by the standard deviation of the baseline scores. In the second, the effect size was calculated according to the method of Cohen,[26] in which the mean score change is divided by the standard deviation of the change scores.

Results

Subjects

One-hundred two patients with nonspecific neck pain as the main presenting complaint were recruited to the study. The mean age of the sample was 45.4 (SD 14.81) years, and 64 (62.7%) of the patients were female. Twenty-one of these patients (20.6%) suffered from neck pain alone. Most patients in the sample (79.4%), however, complained of neck pain with associated symptoms, including the shoulders and upper limbs, low back, and headache. Around half the sample (45.1%) reported that their current episode of neck pain had lasted more than 7 weeks, with almost two thirds (62.7%) reporting that they had suffered from similar episodes in the past. Of the 77 patients who were in paid employment, 45 (58.4%) reported that they had not taken time off work because of their neck complaint. The length of work absence varied from 1 day to 1 month in most of the remaining subjects, with 5 patients reporting that they had taken more than 1 month off work.

Homogeneity of Items

Cronbach's alpha coefficient was approximately 0.9 for all 3 administrations of the questionnaire (ie, pretreatment, retest, and posttreatment) (Table 1). These results support the hypothesis that the instrument taps different dimensions of the same attribute and that, as a consequence, all 7 items can be summed to give a total score. Again, for all 7 items at each of the 3 administrations, the item-corrected total correlation coefficients (Table 1) were well above the 0.2 cut-off advocated by Streiner and Norman.[20] This demonstrates that all items contributed significantly to the total score and therefore should be retained in the questionnaire.

Table 1. *Internal consistency of the Bournemouth Questionnaire*

Item	Item-corrected total correlations Pearson's *r*							Cronbach's alpha (Total score)
	1	2	3	4	5	6	7	
Pretreatment	0.66	0.77	0.72	0.72	0.53	0.65	0.49	0.87
	(99)	(99)	(98)	(98)	(98)	(97)	(98)	(99)
Retest	0.70	0.83	0.84	0.71	0.69	0.78	0.43	0.91
	(90)	(89)	(90)	(90)	(90)	(90)	(90)	(90)
Posttreatment	0.80	0.79	0.70	0.80	0.62	0.80	0.79	0.92
	(71)	(72)	(72)	(72)	(72)	(72)	(71)	(72)

Number of observations in parentheses.

Table 2. *Test-retest reliability of Bournemouth Questionnaire*

Item	ICC	LOA	
		Upper limit	Lower limit
1	0.60	1.56	0.65 (45)
2	0.52	2.56	1.31 (45)
3	0.50	2.69	1.44 (45)
4	0.59	2.54	1.42 (45)
5	0.63	1.96	0.90 (45)
6	0.53	2.49	1.12 (44)
7	0.63	1.07	−0.22 (45)
Total	0.65	11.49	5.96 (45)

Number of observations in parentheses.

Table 3. *External construct validity of items of the Bournemouth Questionnaire*

Item	Correlation with	Pearson's *r*	
		Pretreatment	Posttreatment
1	NFDS (usual pain intensity scale)	0.58 (99)	0.83 (70)
2	SF36 (physical functioning scale)	−0.43 (98)	−0.59 (71)
	NFDS	0.52 (98)	0.45 (70)
	NDI	0.39 (98)	0.65 (71)
3	SF36 (social functioning scale)	−0.43 (96)	−0.45 (69)
	NFDS	0.62 (96)	0.47 (70)
	NDI	0.41 (97)	0.71 (71)
4	SF36 (role-emotional scale)	−0.37 (94)	−0.44 (68)
5	SF36 (mental health scale)	−0.44 (95)	−0.55 (69)
6	SF36 (question 8)	−0.48 (94)	−0.47 (69)
7	SF36 (general health scale)	−0.14 (96)	−0.20 (70)

Number of observations in parentheses.

Table 4. *External construct validity of the Bournemouth Questionnaire*

Correlation of BQ with:	Pearson's *r*	
	NDI	NFDS
Pretreatment	0.51 (98)	0.63 (98)
Posttreatment	0.71 (71)	0.48 (70)

Number of observations in parentheses.

Test-Retest Reliability

In stable subjects who reported no change in their condition between the pretreatment and retest administrations of the BQ, the ICC values for each of the 7 items ranged from 0.50 to 0.63 (Table 2). The ICC of the total score of the BQ was 0.65, indicating moderate agreement between total scores in these patients. With these data, the limits of agreement of the BQ total score (equal to the absolute mean difference between test-retest observations ± 2 standard error of the differences) were 6.0 and 11.5 (Table 2). This indicates that total score changes of the neck BQ greater than 12 are indicative of real change over and above the variability of the instrument in stable subjects. For individual scales of the BQ, the limits of agreement values are also given in Table 2. The upper limits ranged from 1.07 for scale 7 to 2.69 for scale 3. Hence, change scores of approximately 1 to 3, depending on the scale, are indicative of real change, taking into account the reliability of individual scales.

Validity

Tables 3 and 4 show the data testing the external construct validity of the BQ. Each of the 7 items of the BQ was tested either against whole instruments or against individual scales of these instruments as considered most appropriate to the construct of the BQ item under test. For example, item 1 of the BQ relating to pain intensity was tested against the usual pain intensity scale of the NFDS. There was some difficulty in identifying a suitable scale for item 6 of the BQ (fear-avoidance behavior) among the 3 external measures. As a result, a single question from the SF36 judged as measuring a similar attribute to this item was used as the external measure (SF36-question 8: "In the past 4 weeks, how much did your pain interfere with your normal work? (both work outside the home and housework)"). Construct validity was tested for both pretreatment and posttreatment administrations of the questionnaire. As shown in Table 3, apart from item 7 of the BQ, there was overall good correlation between item scores of the BQ and those of their counterpart measures. The negative sign on some of these correlations merely indicates that the 2 scales being com-

Table 5. *External longitudinal construct validity of items of the Bournemouth Questionnaire*

Item	Correlation with	Pearson's *r*
1	NFDS (usual pain intensity scale)	0.49 (69)
2	SF36 (physical functioning scale)	−0.39 (69)
	NFDS	0.47 (69)
	NDI	0.43 (70)
3	SF36 (social functioning scale)	−0.09 (66)
	NFDS	0.55 (68)
	NDI	0.48 (70)
4	SF36 (role-emotional scale)	−0.30 (64)
5	SF36 (mental health scale)	−0.31 (66)
6	SF36 (question 8)	−0.41 (66)
7	SF36 (general health scale)	−0.08 (68)

Number of observations in parentheses.

Table 6. *External longitudinal construct validity of the Bournemouth Questionnaire*

	Pearson's *r*	
Correlation of BQ with:	NDI	NFDS
	0.50 (70)	0.44 (68)

Number of observations in parentheses.

Table 7. *Responsiveness (internal longitudinal construct validity) of items of the Bournemouth Questionnaire*

Change score item-corrected total change score correlations							
	Pearson's *r*						
Item	1	2	3	4	5	6	7
	0.70 (70)	0.82 (71)	0.73 (71)	0.73 (71)	0.42 (71)	0.64 (71)	0.66 (70)

Number of observations in parentheses.

Table 8. *Effect sizes of the BQ, NDI and NFDS*

	Δ	SD*	SD†	Effect size*	Effect size†
BQ (69)	22.8	13.66	15.80	1.67	1.43
NDI (69)	5.7	7.09	6.78	0.80	0.83
NFDS (67)	3.7	6.27	5.82	0.59	0.63

Number of observations in parentheses.
Δ Mean change scores.
*Standard Deviation of baseline scores.
†Standard Deviation of change scores.
Effect size* Kaziz et al. (1989).
Effect size† Cohen (1977).

pared were scoring in opposite directions. All of these correlations, apart from those for item 7, were statistically significant ($P < .001$). Similarly, there were statistically significant correlations ($P < .001$) between the total scores of the BQ and the NDI and NFDS at both pretreatment and posttreatment administrations (Table 4). These data are broadly favorable to the external construct validity of the individual items and of the total score of the neck BQ.

Tables 5 and 6 show the external longitudinal validity of the individual items of the BQ and the total scores using the same external measures as for external validity testing (Tables 3 and 4). Correlations for change scores between the BQ and the counterpart external measures were overall comparable (although in some cases slightly lower) with those in external validity testing. With the exception of item 3 and the social functioning scale of the SF36 and, as before, item 7 and the general health scale of the SF36, all correlations were statistically significant ($P < .001$). Again, these data are broadly favorable to the longitudinal construct validity of the individual items (Table 5) and of the total score (Table 6) of the neck BQ.

Responsiveness

Table 7 shows the internal longitudinal construct validity, or responsiveness, of each of the 7 items in the questionnaire. All of the change score item-corrected total change score correlation coefficients were above the 0.3 cutoff advocated by Stratford et al.[24] This indicates that all of the items in the neck BQ are responsive to clinically significant change (as determined by self-perceived global improvement) and that each contributes significantly to the change in the total score. On the basis of these results, none of the items is redundant to the total score in terms of the ability of the neck BQ to detect clinically significant change.

Effect sizes for the neck BQ calculated according to the method of Kaziz et al[25] and Cohen[26] are shown in Table 8. As can be seen, relative to both the NDI and the NFDS, the effect size of the neck BQ was large. As a consequence, a much smaller sample of patients would be required to demonstrate clinically significant change at a statistically significant level. From Table 8, it is clear that it is the greater mean change in scores between pretreatment and posttreatment administrations of the BQ that accounts for this larger effect size. The reason for this is illustrated by transforming the raw scores of each of the 3 measures (ie, BQ, NDI, and NFDS) to percentage of the total score, and directly comparing pretreatment and posttreatment values. For the BQ, NDI, and NFDS, the mean pretreatment percentage scores were 50.8% (SD 19.5%; n = 69), 28.7% (SD 14.18%; n = 70), and 35.0% (SD 20.91%; n = 69), respectively. Similarly, the mean posttreatment scores for the BQ, NDI, and NFDS were 18.4% (SD 16.91%; n = 70), 17.59% (SD 13.50%; n = 69), and 22.8% (SD 18.19%; n = 68), respectively. Hence, although the posttreatment scores for all 3 measures were comparable in percentage terms, the mean pretreatment percentage score for the BQ was substantially greater than that of either the NDI or the NFDS.

Discussion

It is surprising, given the prevalence of neck pain and its impact on the individual and society,[27] that there are not more instruments that have been developed and tested for use in trials evaluating treatment interventions in this condition. The few that do exist are all geared toward the pain severity and disability dimensions of the condition and as such do not take a wider view of neck pain based on a biopsychosocial model of pain. That neck pain, like back pain, is more likely explained by a biopsychosocial model than a medical one, is now entirely in keeping with our present understanding of musculoskeletal pain disorders and the current moves away from passive treatment to active rehabilitation in the management of nonspecific neck pain.[5,6,28]

As a consequence, there is a need for an outcome measure that comprehensively incorporates the salient dimensions of the biopsychosocial pain model for use in neck studies. At the same time, such an outcome measure must be practical for use, not only in the research setting, but also in the busy routine clinic setting if it is to be used to evaluate both the efficacy and the effectiveness of treatment interventions.[29] These same considerations were behind the development and testing of a new, short-form comprehensive outcome measure for use in patients with back pain.[17] The back BQ contains 7 core items: (1) pain intensity; (2) disability in ADL and (3) in social activities; the emotional dimensions of (4) anxiety and (5) depression; and the cognitive aspects of (6) fear-avoidance behavior and (7) pain locus of control. Mindful of the similarities between back and neck pain, and the need for a comprehensive yet short outcome measure for use in neck pain patients based on the biopsychosocial model, this study was formulated to modify the back BQ and then test its psychometric properties in patients with nonspecific neck pain.

Basing neck pain on a biopsychosocial model in the same way as back pain, and given the generic nature of the 7 core items in the back BQ, very little modification was made to the original questionnaire. Only 1 of the 7 items, namely disability in ADL was changed to exclude those activities likely to be affected by back pain and replace them with activities likely to be affected by neck pain. Because of these small changes, we now advocate the use of a generic BQ that can be used in *all* painful musculoskeletal complaints, including shoulder and extremity pain. In this generic BQ, the wording of the item on disability in ADL is phrased "How has your complaint interfered with your daily activities (housework, washing, dressing, lifting, reading, driving, climbing stairs, getting in/out of bed/chair, sleeping)?" This encompasses those activities, 1 or more of which is likely to be affected by each of the painful nonspecific musculoskeletal conditions. Apart from the small change in the wording of the disability in ADL scale of the neck BQ, no change was made to the response scaling for the questionnaire items. The 11-point NRS has previously been shown to be a responsive scale, as well as one that is relatively easy for patients to complete.[30,31]

To test for redundancy of items, the item-corrected total correlations and item change score-corrected total change score correlations were determined. The results of this study show that, in both cases, each of the 7 items contributes significantly to the total score of the BQ and that, as such, there are no redundant items in the neck BQ. Moreover, the results of the internal consistency tests (Cronbach's alpha) showed that the neck BQ is a homogenous instrument tapping different aspects of the same attribute (ie, the neck pain experience). This is further evidence that neck pain is more likely explained by a biopsychosocial model than a medical one.

The test-retest results showed that the neck BQ is a reliable instrument, and that in stable subjects there is moderate agreement in consecutive administrations of the questionnaire. From these data, it has been possible to show that a change score in excess of 12 points (out of a total of 70), or approximately 17%, is indicative of a real change ("signal") over and above the variability ("noise") of the measuring instrument. This is an important point that raises the matter of clinical change versus statistical change, and the fact that the two are not necessarily synonymous.[32]

Because no "gold standard" exists, testing the validity of instruments such as the BQ is difficult. The best that can be done is to use established measures that purport to measure similar constructs as the instrument under test. In the case of the neck BQ, this was even more difficult than usual because of the paucity of established measures specifically designed for use in neck patients. As a result, we chose to use the 2 most frequently used neck disability measures (the NDI and the NFDS), even though they only measure pain and disability. Both the NDI[12,33] and the NFDS[14] have undergone psychometric testing in neck patients. In addition, we used the generic health status measure, the SF36. This has been validated[21] and widely used in different populations, including patients with back pain.[22] Because the SF36 produces 8 separate scale scores rather than a single index, only the individual scales of the SF36 were used as external criteria for testing.

The total score of the neck BQ correlated significantly with the total scores of the NDI and the NFDS, both in terms of absolute scores (external construct validity) and the change scores over time (external longitudinal construct validity). When testing the external validity of the individual items of the neck BQ, we trawled the established measures and selected those that appeared to most closely match the attribute under test in each item. In one case, item 6 (fear-avoidance behavior), we were unable to find an appropriate scale or measure to use as an external criterion and we therefore used the scores from an individual question of the SF36 (question 8). Apart from item 7 (pain locus of control), there was moderate to strong (and in all cases

statistically significant) correlation with the chosen external measures, supporting the external construct validity and external longitudinal construct validity of individual items of the neck BQ. The poor (and statistically insignificant) correlation between item 7 and the general health scale of the SF36 was most likely due to the fact that the external scale does not adequately reflect the pain locus of control construct. It was however, the best fitting scale we could find. Moreover, the correlation between item 3 and the social functioning scale of the SF36 was low and not statistically significant when testing external longitudinal construct validity, but not when testing external construct validity. We are unable, at this time, to offer an explanation for this seemingly spurious finding.

In contrast to reliability and validity, responsiveness (the ability to detect clinically significant change) is an often-neglected psychometric property of a measure. Considering that the ability to detect clinical change is an essential property of an evaluative measure, this is a serious shortcoming. Although there are several ways of estimating the responsiveness of an instrument,[23] arguably the most common approach is determination of the instrument's treatment effect size. As far as we are aware, there are no published data on the effect size of either the NDI or the NFDS. The data from this study demonstrate that the effect size of the neck BQ is large, and considerably greater than that of the NDI and the NFDS. This result has important implications for those conducting clinical trials and outcomes research in neck patients. One of the problems in clinical trials, even multicenter trials, is recruitment of patients. Use of an outcome measure with a large effect size substantially reduces the sample size needed to establish a clinically significant difference as statistically significant. We suggest that further work be done in this area to investigate whether or not the large effect size difference between the BQ and neck disability measures reported in this study is replicated in other patient populations. The data also reveal that the primary reason for this difference in effect size is almost certainly the higher (percentage) values of the pretreatment BQ scores of these patients when compared with those measured with other instruments.

This study has limitations. The patients recruited to the study were convenience samples from a chiropractic college teaching clinic and chiropractors' field practices. As such, these patients may not be representative of all patients with neck pain who present to chiropractors, nor may they be representative of other ambulatory patients with neck pain. The psychometric properties of the neck BQ should therefore be tested in other populations, including patients who suffer neck pain as a result of a traumatic injury. It is important to remember that an outcome measure validated in one patient group is not necessarily valid in another in which the patient characteristics, particularly levels of disability and chronicity of the complaint, may be different. No attempt was made in this study to distinguish between patients with acute and chronic neck pain. Finally, for the purposes of testing the responsiveness of the instrument, distinction was necessary between patients who had undergone clinically significant change and those who had not. Clinically significant change remains a debatable issue, which in the absence of consensus, we have defined as the self-report of patients' perceptions of improvement in their condition.

Conclusion

As a result of a lack of outcome measures specifically designed and developed for use in neck patients, and in particular ones based on neck pain as an illness, we have developed a short-form, comprehensive neck outcome measure. The neck BQ covers the salient dimensions of the biopsychosocial model of pain, is quick and easy to complete, and has been shown to be reliable, valid and responsive to clinically significant change in nonspecific neck pain patients. We therefore recommend the neck BQ as an outcome measure in clinical trials and in outcomes research for evaluating the efficacy and effectiveness of treatment interventions for nonspecific neck pain.

Acknowledgments

We would like to thank staff at the AECC Teaching Clinic and the chiropractors in private practice for their essential contribution in recruiting patients to this study.

REFERENCES

1. Borghouts JAJ, Koes BW, Vondeling H, Bouter LM. Cost-of-illness of neck pain in The Netherlands in 1996. Pain 1999;80:629-36.
2. Bovim G, Schrader H, Sand T. Neck pain in the general population. Spine 1994;19:1307-9.
3. Bogduk N. The neck. Bailliere's Clin Rheumatol 1999;13: 261-85.
4. Kjellman GV, Skargren EI, Oberg BE. A critical analysis of randomized clinical trials on neck pain and treatment efficacy: a review of the literature. Scand J Rehab Med 1999;31:139-52.
5. Jordan A, Bendix T, Nielsen H, Hansen FR, Host D, Winkel A. Intensive training, physiotherapy, or manipulation for patients with chronic neck pain. Spine 1998;23:311-9.
6. Murphy DR. Protocols for the management of patients with cervical spine syndromes. In: Murphy DR, editor. Conservative management of cervical spine syndromes. New York: McGraw Hill; 2000. p. 691-700.
7. Koes BW, Assendelft WJ, van der Heijden GJ, Bouter LM, Knipschild PG. Spinal manipulation and mobilisation for back and neck pain: a blinded review. BMJ 1991;303:1298-303.
8. van der Heijden GJ, Beurskens AJ, Koes BW, Assendelft WJ, de Vet HC, Bouter LM. The efficacy of traction for back and neck pain: a systematic, blinded review of randomized clinical trial methods. Phys Ther 1995;75:93-104.
9. Aker PD, Gross AR, Goldsmith CH, Peloso P. Conservative management of mechanical neck pain: systematic overview and meta-analysis. BMJ 1996;313:1291-96.
10. Bolton JE. Evaluation of treatment in back pain patients: clinical outcome measures. Eur J Chiropractic 1994;42:29-40.

11. Bolton JE. Future directions for outcomes research in back pain. Eur J Chiropractic 1997;45:57-64.
12. Vernon H, Mior S. The neck disability index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14: 409-15.
13. Leak AM, Cooper J, Dyer S, Williams KA, Turner-Stokes L, Frank AO. The Northwick Park neck disability questionnaire, devised to measure neck pain and disability. Br J Rheumatol 1994;33:469-74.
14. Jordan A, Manniche C, Mosdal C, Hindsberger C. The Copenhagen neck functional disability scale: a study of reliability and validity. J Manipulative Physiol Ther 1998;21:520-7.
15. Wheeler AH, Goolkasian P, Baird AC, Darden BV. Development of the neck pain and disability scale. Spine 1999;24: 1290-4.
16. Bolton JE. The psychosocial profile of the chronic back patient. In: Stude D, editor. Spinal rehabilitation. Stamford, CT: Appleton & Lange; 1999: p. 231-43.
17. Bolton JE, Breen AC. The Bournemouth Questionnaire: a short-form, comprehensive outcome measure. I. Psychometric properties in back pain patients. J Manipulative Physiol Ther 1999;22:503-10.
18. Waddell G. A new clinical model for the treatment of low-back pain. Spine 1987;12:632-44.
19. Kirshner B, Guyatt GA. A methodological framework for assessing health indices. J Chron Dis 1985;38:27-36.
20. Streiner DL, Norman GR. Health measurement scales, 2nd ed. Oxford: Oxford University Press; 1995.
21. Brazier JE, Harper R, Jones NMB, O'Cathain A, Thomas KJ, Usherwood T, Westlake L. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ 1992;305:60-164.
22. Ware JE, Sherbourse CD. The MOS 36-item short form health survey (SF-36). Med Care 1992;30:473-81.
23. Bolton JE. On the responsiveness of evaluative measures. Eur J Chiropractic 1997;45:5-8.
24. Stratford P, Solomon P, Binkley J, Finch E, Gill C. Sensitivity of sickness impact profile items to measure change over time in a low-back pain patient group. Spine 1993;18:1723-7.
25. Kaziz LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178-89.
26. Cohen J. Statistical power analysis for the behavioural sciences. New York: Academic Press; 1977.
27. Ariens GAM, Borghouts JAJ, Koes BW. Neck pain. In: Crombie IK, Croft PR, Linton SJ, LeResche L, Von Korff M, editors. Epidemiology of pain. Seattle: IASP Press; 1999. p. 235.
28. Jordan A, Ostergaard K. Rehabilitation of neck/shoulder patients in primary health care clinics. J Manipulative Physiol Ther 1996;19:32-5.
29. Pittler MH, White AR. Efficacy and effectiveness. Focus Alt Comp Ther 1999;4:109-10.
30. Bolton JE, Wilkinson RC. Responsiveness of pain scales: a comparison of three pain intensity measures in chiropractic patients. J Manipulative Physiol Ther 1998;21:1-7.
31. Jensen MP, Miller L, Fischer LD. Assessment of pain during medical procedures: a comparison of three scales. Clin J Pain 1998;14:343-9.
32. Turk DC. Statistical significance and clinical significance are not synonyms! Editorial. Clin J Pain 2000;16:185-7.
33. Hains F, Waalen J, Mior S. Psychometric properties of the neck disability index. J Manipulative Physiol Ther 1998;21: 75-80.

APPENDIX

Global dimensions of the Neck Bournemouth Questionnaire

The following scales have been designed to find out about your neck pain and how it is affecting you. Please answer ALL the scales by circling ONE number on EACH scale that best describes how you feel:

1. Over the past week, on average how would you rate your neck pain?

No pain — Worst pain possible

0 1 2 3 4 5 6 7 8 9 10

2. Over the past week, how much has your neck pain interfered with your daily activities (housework, washing, dressing, lifting, reading, driving)?

No interference — Unable to carry out activities

0 1 2 3 4 5 6 7 8 9 10

3. Over the past week, how much has your neck pain interfered with your ability to take part in recreational, social, and family activities?

No interference — Unable to carry out activities

0 1 2 3 4 5 6 7 8 9 10

4. Over the past week, how anxious (tense, uptight, irritable, difficulty in concentrating/relaxing) have you been feeling?

Not at all anxious — Extremely anxious

0 1 2 3 4 5 6 7 8 9 10

5. Over the past week, how depressed (down-in-the-dumps, sad, in low spirits, pessimistic, unhappy) have you been feeling?

Not at all depressed — Extremely depressed

0 1 2 3 4 5 6 7 8 9 10

6. Over the past week, how have you felt your work (both inside and outside the home) has affected (or would affect) your neck pain?

Have made it no worse — Have made it much worse

0 1 2 3 4 5 6 7 8 9 10

7. Over the past week, how much have you been able to control (reduce/help) your neck pain on your own?

Completely control it — No control whatsoever

0 1 2 3 4 5 6 7 8 9 10