Assessment of Validity, Reliability, Responsiveness and Bias of Three Commonly Used Patient-Reported Outcome Measures in Carpal Tunnel Syndrome

Hadeel Bakhsh(A,B,D,E,F), Inji Ibrahim(A,B,D,E,F), Wasim S. Khan(A,B,D,E,F), Peter Smitham(A,B,D,E,F), Nicholas Goddard(A,B,D,E,F)

University College London Institute of Orthopaedics and Musculoskeletal Science, Royal National Orthopaedic Hospital, Stanmore, Middlesex, London, UK

SUMMARY

Background. In recent years there has been an increase in the use of self-administered questionnaires to accurately assess intervention outcomes in hand surgery in order to determine the quality of healthcare. This prospective study aims to evaluate and assess the validity, reliability, responsiveness, and bias of a number of outcome measure for Carpal Tunnel Syndrome (CTS) including the disease-specific Boston questionnaires (BQ), and the region-specific Disability of Arm, Shoulder, and Hand (DASH) questionnaires and Manchester Modified Disabilities of the Arm, Shoulder and Hand (M2DASH) questionnaires, and comparing the results to Nerve Conduction Studies (NCS).

Material and methods. Forty-eight patients with clinical signs of CTS confirmed by NCS completed the BQ, DASH and M2DASH questionnaire at different time intervals peri-operatively. The scores were analysed to assess validity, reliability, responsiveness, and bias of the questionnaires. Validity analysis for the three questionnaires showed strong positive correlations and there was no age, gender, hand dominance, or side affected bias in the questionnaires.

Results. No significant correlation was obtained between the questionnaires and NCS. Significant results for responsiveness were noted in BQ symptom severity scale only.

Conclusions. 1. This study has shown that the BQ, DASH and M2DASH Questionnaires are valid and reliable outcome measures for CTS. 2. In terms of responsiveness, the DASH and M2DASH questionnaires are not as responsive as the BQ scores over the initial post-op recovery period. 3. We would therefore recommend that the Boston Questionnaire be used to assess early post-operative patient related outcome measures for Carpal Tunnel Syndrome.

Key words: carpal tunnel syndrome, Boston Questionnaire, DASH Questionnaire, validity, reliability, responsiveness, bias
BACKGROUND

Carpal Tunnel Syndrome (CTS) affects 345 new cases per 100,000 persons per year [1,2]. It results from a compression of the median nerve between the transverse carpal ligament, the flexor tendons and the carpal bones at the wrist level and is characterized by paresthesia, pain, and weakness in the forearm [3-7]. There are various conservative treatments for CTS and, if conservative treatments fail to provide symptomatic relief, surgical release of the compression is sought in 80% of all cases [4,8,9].

In recent years, it has become essential to accurately assess and monitor the intervention outcomes in hand surgery in order to determine the quality of healthcare and provide an evidence-based practice [10-12]. In order to evaluate the success of the surgical intervention for CTS, a variety of region-specific and disease-specific outcome measures have been developed and used over the years. Self-administered questionnaires are one of the most commonly used outcome measures. Katz et al (1994), found that measuring patients’ symptom severity using a self-administered questionnaire is four times more sensitive to clinical change. Moreover, measuring functional limitations caused by CTS using questionnaires has been found to be twice more sensitive to clinical progress than the usual objective tests. Accordingly, it is expected that the use of regional outcome measures specific to the hands will continue to increase [10,13].

Generic outcome measures have been shown to be less sensitive and responsive to clinical change over time in CTS. Therefore regional-specific and disease-specific questionnaires are the most widespread outcome measures used to evaluate the effect of CTS decompression surgery [11,14]. The most frequently used disease-specific questionnaire for CTS is the Boston Questionnaire (BQ), and the most frequently used region-specific questionnaire is the Disabilities of the Arm, Shoulder, and Hand score (DASH) [11]. Recently Khan et al. (2008) introduced the Manchester-modified version of the region-specific DASH score (M'DASH), which is validated as a reliable and responsive outcome measure for hand injuries only. Some clinicians consider Nerve Conduction Studies (NCS) a gold standard for diagnosis of CTS.

The self-administered BQ was developed by Levine et al (1993) and is considered the ‘gold standard’ for CTS [12,15-18]. The DASH questionnaire is an outcome measure specifically designed for upper-extremity injuries and disorders. The DASH has been validated as a reliable and responsive outcome measure for upper limb pathologies including CTS [12,19-22]. Some authors argue that the DASH questionnaire is not as sensitive to clinical change as disease-specific questionnaires [21,22], whereas others have demonstrated that the DASH is similar to, or even better than, disease-specific measures [12,19]. A study by Dowrick (2006) showed that although the DASH is region-specific outcome measure of the upper-extremity, it also measures disability in lower-extremity injuries [23]. Khan et al (2008) introduced a shorter and modified version of the DASH, the Manchester-modified Disability of Arm, Shoulder, and Hand (M'DASH), which is more specific to upper-limb injury than the original DASH [24].

This research aims to investigate the validity, reliability, responsiveness, and bias of the BQ, DASH and M'DASH questionnaire for CTS compared to NCS data.

MATERIAL AND METHODS

Design & Sample Size

We conducted a prospective longitudinal study of patients who attended the outpatient orthopaedic clinic and day surgery unit at the Royal Free Hospital, Hampstead, UK from May to July 2011 inclusive. Patients were enrolled in the study and asked to complete the BQ, DASH and M'DASH questionnaires while waiting for their appointment, or before their CTS decompression surgery. All patients had the same peri-operative management under the care of a single surgeon (NG).

Inclusion & Exclusion Criteria

Patients with a confirmed diagnosis of CTS using NCS who were scheduled for primary open carpal tunnel decompression, and who completed at least two sets of questionnaires were included in the study. Patients were excluded from the study if they were under 18 years old, pregnant or unable to complete the questionnaire due to language or cognitive difficulties. These inclusion and exclusion criteria are consistent with previous studies [11,24,26,27]. A total of 48 patients met the criteria and were included in the study.

Procedure

Forty-eight patients completed the first set of questionnaires in the clinic or at the day surgery unit before undergoing CTS decompression. Patients were then sent a further two sets of questionnaires spaced out over the next six to eight weeks. The questionnaires were used to determine the validity and bias of the questionnaires. Responsiveness was assessed in patients who completed the questionnaire pre-oper-
tively and post-operatively. Reliability was assessed in patients who completed the two sets of questionnaires post-operatively.

**Statistical Analysis**

All data analysis was conducted by SPSS version 19 [28]. A $p$ value of $<0.05$ was taken as statistically significant. The questionnaire scores were calculated as specified in the questionnaires.

Construct validity was expressed in correlation coefficients [29]. Internal consistency was assessed by calculating Cronbach’s alpha for each of the questionnaires [30]. To establish test-retest reliability, the scores were correlated for the first and the second time point using the interclass correlation coefficient (ICC) [30]. For group comparisons, a value of 0.70 or higher is considered acceptable. For internal-responsiveness, the scores preoperatively and postoperatively were assessed for how sensitive they were to clinical change [31]. This was investigated by calculating effect size (ES) [mean change/ (SD of the baseline scores)], and the standard response mean (SRM) [(postoperative mean – preoperative mean)/SD of the difference] [31]. ES values $<0.50$, $0.50$-$0.80$, and $>0.80$ were considered to represent small, moderate, and large degrees of responsiveness, respectively [27,31], while SRM values of 0.2 indicating small, 0.5 indicating moderate, and greater than 0.8 large change [12,22]. A paired $t$-test was used to compare the results of the BQ, DASH and M'DASH questionnaires. For external-responsiveness, the changes of scores preoperatively and postoperatively were compared. The ES and SRM values were compared using a paired $t$-test. Bias in the scores was investigated by performing correlation analysis with age, and linear regression analysis for assumed independent variables including age, gender, hand dominance, and the side injured [29]. Evidence that a questionnaire is not correlated with different variables would give a low correlation, less than 0.30.

**RESULTS**

The mean age of patients was 60 (SD 14.46) and ranged from 36-87. There were 37 female and 11 male patients.

**Validity**

Patients scores at presentation were used to assess construct and criterion validity, by using Spearman’s correlation coefficient for BQ (subscales: BQ-functional & BQ-symptom severity), DASH, M'DASH scores and NCS results (sensory action potential (SAP), distance motor latency (DML), and motor conduction velocity (MCV)). The questionnaire scores appeared to have excellent correlation with each other as shown in Table 1. There was, however, no significant correlation between the three questionnaires and the NCS components.

**Reliability**

Internal consistency tests showed that the questionnaires were internally consistent with a high Cronbach’s alpha as shown in Table 2. Test-retest reliability results are shown in Table 3, and all results were significant ($p<0.001$), suggesting good test-retest reproducibility.

**Responsiveness**

Standard Response Mean (SRM), Effect Size (ES), and the mean change of scores were calculated for all of the questionnaires. The results in Table 3 show that with the exception of the BQ, the other questionnaires were internally consistent with a high Cronbach’s alpha as shown in Table 2. Test-retest reliability results are shown in Table 3, and all results were significant ($p<0.001$), suggesting good test-retest reproducibility.

---

**Table 1. Assessment of the BQ, DASH and M2DASH questionnaires, and NCS Correlations for Construct & Criterion Validity**

<table>
<thead>
<tr>
<th></th>
<th>BQ-F</th>
<th>BQ-SS</th>
<th>DASH</th>
<th>M'DASH</th>
<th>MCV</th>
<th>SAP</th>
<th>DML</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BQ-F</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=48)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BQ-SS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=48)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>0.773</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DASH</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=48)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>0.867</td>
<td>0.770</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>M'DASH</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=48)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>0.828</td>
<td>0.809</td>
<td>0.948</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MCV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=43)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>-0.151</td>
<td>-0.019</td>
<td>-0.148</td>
<td>-0.147</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.334</td>
<td>0.944</td>
<td>0.341</td>
<td>0.746</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SAP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=43)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>-0.125</td>
<td>-0.042</td>
<td>-0.125</td>
<td>-0.093</td>
<td>0.630</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.426</td>
<td>0.791</td>
<td>0.424</td>
<td>0.554</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DML</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=43)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>0.052</td>
<td>-0.159</td>
<td>0.025</td>
<td>-0.025</td>
<td>-0.553</td>
<td>-0.660</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.739</td>
<td>0.310</td>
<td>0.872</td>
<td>0.874</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.01 level (2-tailed). Bold numbers indicate significance (BQ-F: Boston Questionnaire- Functional Scale; BQ-SS: Boston Questionnaire- Symptom Severity Scale; DASH: Disability of Arm, Shoulder, and Hand scores; M'DASH: Manchester-Modified Disability of Arm, Shoulder, and Hand scores; MCV: Motor Conduction Velocity; SAP: Sensory Action Potential; DML: Distance Motor Latency)
naires did not show significant results ($p>0.05$), with a very small effect size recorded for the M 2DASH ($ES=-0.056$). The BQ subscales provided moderate and large SRM and ES, and only BQ-SS showed significant results indicating that the scale is sensitive to clinical change ($p=0.000$).

Bias

Linear Regression analysis showed that the scores of all three questionnaires were not associated with age, gender, affected side or dominant hand, as demonstrated by $p$-values of more than 0.05. Hence, they are unbiased outcome measures. Additionally, correlation analysis, using Pearson’s coefficient, confirmed no bias with age in all three questionnaires; BQ subscales (BQ-F $r=0.069$, BQ-SS $r=-0.156$), DASH ($r=0.086$), M’DASH ($r=0.076$).

**DISCUSSION**

The use of self-administered outcome measures is increasing and becoming more popular in hand surgery to monitor the effectiveness of treatment, justify intervention and cost implications and build a foundation of evidence-based health care [10,29, 30]. Self-administered questionnaires are more efficient than objective measures in assessing the overall function and disability of the hand from the patients’ own perspective [15,16]. Consequently, questionnaires must be evaluated thoroughly before they can be widely used in clinical practice.

This study shows that all three questionnaires have construct validity and criterion validity. There is a strong correlation between the three questionnaires suggesting similar construct, indicating that they measure the same concepts. The negative correlation between the different components of the NCS of the median nerve shows no significance in correlation with any of the three questionnaire scores. Previous studies assessing the relationship between NCS and self-administered questionnaires for CTS demonstrated similar results. Schrijver et al (2005) and Chan et al (2005) found no association between BQ & NCS preoperatively or postoperatively, and they concluded that NCS is more important in diagnosing CTS and distinguishing it from other pathologies that masquerade CTS e.g. radiculopathy or motor neuron disease rather than measuring the outcome of the surgery [32,33]. It thus appears that NCS and self-administered outcome measures are different independent measures with different constructs, and clinicians and researchers interested in CTS interventional outcomes should assess both measures. This conclusion is supported by other related studies [17,32-35].

This study adds further support to the reliability of the BQ, DASH and M’DASH questionnaires by confirming them as reliable outcome measures for CTS. This is demonstrated by the high internal consistency in the comparisons. The high value of Cronbach’s alpha in the DASH questionnaire is an indication of possible redundancy in the questions. High ICC indicates that the three questionnaires have test-retest reliability.
Although Kotsis et al (2005) and Gummesson et al (2003) had previously showed that the DASH is a responsive outcome measure for CTS when assessing patients pre-operatively and six months after CTS decompression [27,36], our study showed that the DASH and the M’DASH had a low negative effect size (ES) and standard response mean (SRM) indicating low sensitivity to change when compared to the BQ. In our study, sending further questionnaires eight weeks post-operatively may not have given the patients sufficient time to show improvement in their functional abilities. Most previous studies that aim to assess the outcome of CTS or other hand pathologies via questionnaires, allowed a period of 3-6 months after surgery for healing and recovery before sending the final questionnaire [27,29,36]. Mallick et al (2007), however, demonstrated that two-weeks after CTS surgery is an adequate period to show improvement in the questionnaire scores [18]. Different questionnaires may be more responsive at different phases of recovery [22]. Patients who undergo CTS decompression surgery may experience immediate symptomatic relief and only delayed functional improvement. In this study, the BQ with its eleven questions on symptom severity was able to detect clinical change better than the M’DASH, which contains only four questions that address symptom severity. The M’DASH may, however, be more appropriate in evaluating the physical function once patients are fully recovered.

This study confirmed that there is no age, gender, hand dominance, or side affected bias present in any of the three questionnaires evaluated. Furthermore, when all three questionnaires were investigated for correlation with age, they demonstrated very low correlation coefficients (r<0.30) proving there is no bias in the questionnaire scores when correlated with age.

For future studies, it would be useful if patient demographics were recorded to study whether such factors would influence patient responses to the questionnaires i.e. work, other pathological conditions, ethnicity, and medication for symptomatic relief. Also, as the majority of the questions target physical disability, it would be worth assessing the validity of the questions by correlating them with common objective outcome measures such as grip strength and two-point discrimination.

**CONCLUSION**

1. This study has shown that the BQ, DASH and M’DASH Questionnaires are valid and reliable outcome measures for CTS.
2. In terms of responsiveness, the DASH and M’DASH questionnaires are not as responsive as the BQ scores over the initial post-op recovery period.
3. We would therefore recommend that the Boston Questionnaire be used to assess early post-operative patient related outcome measures for Carpal Tunnel Syndrome.

**ACKNOWLEDGMENTS**
The authors are grateful to the Ministry of Higher Education of Saudi Arabia for granting HB a scholarship and the opportunity to continue educational development.
30. Terwee C B, Bot S D M, de Boer M R. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;34-42.