Assessing the clinical significance of change scores following carpal tunnel surgery

Rouin Amirfeyz · Alanna Pentlow · Julian Foote · Ian Leslie

Received: 22 July 2007 / Revised: 14 September 2007 / Accepted: 17 September 2007 / Published online: 31 October 2007
© Springer-Verlag 2007

Abstract This article presents a prospective longitudinal study to determine the cut-off values for change scores of DASH, Levine, and Kamath questionnaires to distinguish clinical improvement following carpal tunnel surgery. Fifty-four patients (40 female, 14 male), with positive nerve conduction studies, were prospectively followed up. Three questionnaires (DASH, Levine, and Kamath) were posted to patients at four and two weeks prior to their operation and then six weeks following surgery. A patient global impression of change (PGIC) score was completed for patients to rate the overall change in their symptoms. According to the PGIC, 93% of patients improved. The cut-off values for raw change scores that best define clinically significant improvement following carpal tunnel release were 20.9 for DASH, 0.47 for Levine, and 1.97 for the Kamath questionnaire. This study provides a methodological framework for identifying clinically significant changes following treatment. A questionnaire follow-up of patients is now possible using the data provided.

Résumé Etude prospective longitudinale dans le but de déterminer les valeurs seuils, les scores DASH, de LEVINE et de KAMATH correspondant à l'amélioration clinique

R. Amirfeyz · A. Pentlow · J. Foote
Trauma & Orthopaedics, Avon Orthopaedic Centre,
Bristol, UK

I. Leslie
Trauma & Orthopaedics,
Bristol Royal Infirmary,
Bristol, UK

R. Amirfeyz (✉)
75 The Praedium, Chapter Walk,
Bristol BS6 6WB, UK
e-mail: ramirfeyz@yahoo.com

après chirurgie du canal carpien. 54 patients (40 femmes et 14 hommes), ayant un électromyogramme positif ont été suivis de manière prospective. 3 questionnaires (DASH, LEVINE et KAMATH) ont été envoyés aux patients à 4 et 2 semaines avant leur intervention et à 6 semaines après l'intervention. Un score d'impression globale de changement subjectif (PGIC) a été rempli par les patients pour mesurer les changements survenus dans leur symptomatologie. 93% des patients ont été améliorés selon le score PGIC. Les valeurs seuils de changement brut des scores correspondant le mieux à l'amélioration clinique suivant la chirurgie de libération du canal carpien étaient 20.9 pour le DASH, 0.47 pour le LEVINE et 1.97 pour le KAMATH. Cette étude fournit une base méthodologique pour identifier les améliorations cliniques significative après traitement. Le suivi des patients par questionnaire est maintenant possible par l'utilisation des données fournies.

Introduction

It is important to be able to assess accurately and have evidence to prove the effectiveness of clinical interventions. Evidence-based medicine applies the results of clinical trials to the treatment of individual patients. Results from research are usually given as a group mean and the statistical significance of their difference [11]. There is some debate about the relevancy of a model that bases treatment of an individual on the results of a group of patients [17]. More relevant to the patient and the clinician is the proportion of patients who undergo a particular treatment intervention and achieve a clinically significant improvement. This knowledge will provide an individual patient with information regarding the likelihood that they will

benefit from the procedure [4]. However, defining a clinically significant change can be difficult, particularly where the outcome measure is subjective, for example, with pain [23].

Carpal tunnel release is a common operation in orthopaedic surgery. The success of the procedure is determined by a decrease in the severity of symptoms and an increase in function. Where success is assessed by the operating surgeon it is subject to observer bias [16]. To overcome this, self-administered questionnaires, which assess both physical function and severity of symptoms can be used before and after a treatment intervention to look for change. Such questionnaires have been shown to be more sensitive to clinical change than objective neuro-physiological testing [5, 9]; however, despite attempts to quantify clinically important change there is little consensus in the literature of how to determine the magnitude of change in a self-administered questionnaire that is of clinical importance [25].

There are two main types of methods for identifying clinically important intra-individual changes in subjective outcome measures [7, 19].

The first type consists of anchor-based methods, where an external judgement of meaningful change is made by a patient or expert. The most common of this type is the patient global impression of change (PGIC) score. Here, the patient ranks their change following an intervention on a scale from 1 to 7, with 1 representing “no change” and 7 representing “a great deal better”. As patients are making a subjective judgement about the meaning of change, to them this scale is taken as being the “gold standard” of clinically important change [25]. The *a priori* definition of clinically significant change suggests that PGIC values of 6 or more correlate best with actual change [24].

The second type of method is distribution based and quantifies clinically meaningful groups and individual changes based on statistical parameters. One example of this is the effect size statistic. This gives an indication of the magnitude of the effect of treatment, in either groups or individuals, and can be used to calculate the sensitivity of self-administered questionnaires to detect clinically significant changes [15].

Another distribution method statistic is the reliable change index (RCI) devised by Jacobsen et al. [12]. RCI scores are used to determine whether an individual has improved sufficiently and if the change is not likely to be due to simple measurement unreliability. RCI values can be referenced to the normal distribution and a value >1.96 is unlikely unless an actual and reliable change has occurred.

In this study, three outcome questionnaires, DASH, Levine and Kamath, were used to evaluate the success of carpal tunnel surgery. The aim was to compare the sensitivity of the questionnaires and to establish cut-off values of (pre-op to post-op) change scores which best define a

clinically significant improvement by comparing them to the gold standard PGIC scale. If a patient could be defined as clinically improved by using a self-administered questionnaire then such questionnaires could be used as a form of postoperative follow-up and may be able to reduce the number of outpatient clinic visits and reduce costs.

Methods

Fifty-four patients who were listed for carpal tunnel surgery were prospectively followed up at a general orthopaedic unit in Bristol from May 2005 until February 2006. Three different questionnaires—DASH, Levine (function and symptoms), and Kamath—were posted to patients four weeks prior to their operation date.

The DASH (disabilities of the arm, shoulder and hand) questionnaire is a 30-part questionnaire designed to evaluate disabilities and symptoms in one or more upper limb disorders [10]. Studies of reliability have shown the DASH questionnaire to be both valid and reliable in assessing carpal tunnel syndrome [2, 6].

The Boston questionnaire by Levine et al. is a well-recognised, validated, disease-specific questionnaire comprising two parts: one assess function and the other severity of symptoms. Some studies have found it to be more sensitive than DASH [1] whilst others show comparable results [5].

The final questionnaire designed by Kamath et al. [13] is based on the Boston questionnaire. It consists of nine questions with a yes or no response and has been shown to have an 85% sensitivity in assessing patients for carpal tunnel syndrome.

Patients were asked to complete and return all of the questionnaires. Questionnaire completion was repeated two weeks later to check for intraobserver error. This time interval was chosen as sufficiently long to prevent patients remembering previous answers but short enough to prevent significant changes in symptom severity.

Surgery to decompress the carpal tunnel was then performed under either local or general anaesthetic. Six weeks post surgery, patients were asked to complete the same set of three questionnaires to assess for change in scores. A PGIC score was also completed for patients to rate the overall change in their symptoms since treatment.

Data analysis

The raw change scores for each of the questionnaires were calculated by subtracting the post-op score from the pre-op score. The percentage change score was also calculated by dividing the raw change score by the baseline score ($\times 100$).

The effect size statistic was calculated for individual raw change scores as the individual change score divided by the SD of the group baseline scores using the method of Kazis [14]. For individual effect size, 0.2, 0.6, and 1.0 are small, moderate, and substantial changes, respectively [22].

Correlation between the results obtained four weeks and two weeks presurgery was measured using the Pearson's correlation coefficient. Reliable change index scores were calculated for each patient by dividing the raw change score by $SDb \times \sqrt{2(1-r)}$, where SDb is the standard deviation of baseline scores and r is the reliability coefficient calculated using the Pearson's coefficient.

Sensitivity and specificity of cut-off values in identifying clinically significant change were calculated. Scores of 5 and above, 6 and above, and 7 on the PGIC were used in calculating the sensitivity and specificity of cut-off values for each of the questionnaires. 2×2 tables were created to categorise patients using both the PGIC and effect size or RCI methods as improved or not improved. From these tables the sensitivity, specificity, and accuracy were calculated. Cut-off values of effect size and RCI scores that gave the best balance between high sensitivity and specificity and the highest accuracy were chosen as the most fitting in identifying cut-off values for clinically significant change in individual patients as defined by the "gold standard" PGIC.

The raw change score that produced this effect size was calculated by identifying the effect size that most accurately defined clinically important change, thus giving the cut-off value for raw change score that equates with clinically significant change.

In a similar way, by comparing categories of PGIC which showed clinically important change with percentage change scores and calculating the sensitivity and specificity which best defined improvement, cut-off values for percentage change scores were obtained.

Results

Of the 54 patients who were asked to complete the sets of questionnaires, 43 returned a full set and were included in the analysis. Of these patients 37 were female, 17 were male, and the mean age of the patients was 55 years.

The mean raw change scores (and standard deviation) for each of the questionnaires were 1.1(22.7), 0.7 (0.8), 12.4 (0.94), and 1.8 (1.97) for the Levine symptoms, Levine function, DASH, and Kamath questionnaires, respectively.

Pearson's coefficient of reliability for each questionnaire was calculated using the scores obtained four and two weeks preoperatively (Table 1). Values closest to 1 show the best correlation between scores and therefore the least intraobserver error.

Table 1 Pearson's coefficient of reliability

Pearson coefficient (r)	
DASH	0.88
Levin symptom	0.78
Levin function	0.78
Kamath	0.72

Table 2 shows improvement with effect size and RCI statistics. Using the cut-off values for the PGIC of ≥ 5 , ≥ 6 , and 7, the percentage of patients classing themselves as improved was 93% (40 patients), 67.4% (29 patients), and 46.5% (20 patients), respectively. For the RCI method the percentage of patients classified as improved varied between questionnaires with 46.5% (20 patients) for the DASH, 69.8% (30 patients) for the Levine symptoms, 39.5% (17 patients) for the Levine function, and 44.2% (19 patients) for the Kamath questionnaire. Using the effect size statistic, the percentage of patients who had improved gradually decreased with the three different cut-off values for small, moderate, and large improvement.

In order to ascertain whether the patients who showed improvement on the PGIC were the same individuals that had shown improvement on the RCI and effect size score 2×2 tables were used for both the RCI and the effect size. The accuracy provides a measure of agreement of categorization of patients "improved" or "not improved" between the two methods. The sensitivity, specificity, and accuracy of RCI and effect size of categorising individual patients as improved against the three cut-off values of the PGIC are shown in Table 3.

For the RCI the best balance between high sensitivity and high specificity was achieved using a PGIC value of ≥ 6 for the DASH, Levine function, and Kamath questionnaires and a value of ≥ 5 for the Levine symptom questionnaire.

Using the effect size the best balance between high sensitivity and high specificity was found using a PGIC value of ≥ 6 for the DASH, Kamath, and Levine function questionnaires and a value of ≥ 5 for the Levine symptom questionnaire.

The cut-off values for the effect size method were expanded to calculate the exact effect size which has the

Table 2 Categorising percentage of patients as improved using effect size and RCI for the DASH, Levine symptoms, Levine function, and Kamath questionnaires

	DASH	Levine symptoms	Levine function	Kamath
RCI improvement >1.96	46.5%	69.8%	39.5%	44.2%
Effect size				
Small improvement >0.2	67.4%	90.7%	72.1%	74.4%
Moderate improvement >0.6	51.2%	83.7%	58.1%	60.5%
Large improvement >1.0	44.2%	67.4%	37.2%	44.2%

Table 3 Sensitivity, specificity, and accuracy of the effect size in identifying clinically significant change

PGIC cut-offs	>0.2				>0.6				>1			
	DASH	Levine symptoms	Levine function	Kamath	DASH	Levine symptoms	Levine function	Kamath	DASH	Levine symptoms	Levine function	Kamath
Cutoff=7												
Sensitivity	65.0%	95.0%	75.0%	75.0%	55.0%	85.0%	65.0%	60.0%	55.0%	70.0%	35.0%	50.0%
Specificity	30.4%	13.0%	30.4%	26.1%	52.2%	17.4%	47.8%	39.1%	65.2%	34.8%	60.9%	60.9%
Accuracy	46.5%	51.2%	51.2%	48.8%	53.5%	48.8%	55.8%	48.8%	60.5%	51.2%	48.8%	55.8%
Cutoff ≥6												
Sensitivity	72.4%	93.1%	82.8%	79.3%	62.1%	86.2%	72.4%	58.6%	58.6%	72.4%	48.3%	51.7%
Specificity	42.9%	14.3%	50.0%	35.7%	71.4%	21.4%	71.4%	35.7%	85.7%	42.9%	85.7%	71.4%
Accuracy	62.8%	67.4%	72.1%	65.1%	65.1%	65.1%	72.1%	51.2%	67.4%	62.8%	60.5%	58.1%
Cutoff ≥5												
Sensitivity	70.0%	90.0%	72.5%	75.0%	52.5%	82.5%	60.0%	60.0%	47.5%	67.5%	40.0%	45.0%
Specificity	66.7%	0.0%	33.3%	33.3%	66.7%	0.0%	66.7%	33.3%	100.0%	33.3%	100.0%	66.7%
Accuracy	69.8%	83.7%	69.8%	72.1%	53.5%	76.7%	60.5%	58.1%	51.2%	65.1%	44.2%	46.5%

highest sensitivity and specificity and so best distinguishes between patients who have and have not improved. Cut-off individual effect size values of >0.9 for the DASH, >0.2 for the Levine symptom, >0.5 for the Levine function, and >1.0 for the Kamath questionnaire were the most distinguishing. As individual effect size is raw change score divided by SD group baseline scores, the cut-off values for raw change scores which best distinguish patients who have improved can be calculated. The cut-off values for raw change scores that best define clinically significant improvement were 20.9 for DASH, 0.16 for Levine symptoms, 0.47 for Levine function, and 1.97 for the Kamath questionnaire.

The results of sensitivity and specificity of percentage change scores in identifying patients as clinically improved, defined using the PGIC scale cut-offs of ≥5, ≥6, and 7, were calculated.

Percentage change scores which best denote clinically significant change were 10% using the PGIC category of ≥5 for all of the questionnaires except the Levine function questionnaire where a percentage change score of 20% with the PGIC category of ≥6 gave the highest sensitivity and specificity.

Discussion

Carpal tunnel syndrome is the most common reason for elective referral in hand surgery [21]. Surgical release is generally successful but in the climate of evidence-based medicine the importance of reliably monitoring the effectiveness of treatment is well recognised. There is currently no gold standard for measuring the effectiveness of outcomes following carpal tunnel release [18]; thus, asking patients themselves what constitutes a meaningful change

is perhaps the best way of assessing clinically important change. In busy, overbooked outpatient clinics using outcome questionnaires could provide an easier and cheaper way to help follow-up patients and highlight those who have failed to improve. The data from such questionnaires can provide valuable information for clinical, audit, and research purposes.

Previous studies have shown such questionnaires to be reliable, reproducible, and responsive to clinical change [5, 9, 10, 13, 16]; however, there have been no studies to date which demonstrate what score change is needed between pre-op and post-op questionnaire to equate to clinical improvement, therefore making the scores from questionnaires difficult to interpret clinically. Two recent reviews of the literature have compared various available questionnaires in relation to carpal tunnel syndrome and, whilst none of the available ones have been shown to be perfect [20], the Levine questionnaire was favoured for this particular upper limb problem [3].

In this study three statistical methods were used to analyse change scores for three commonly used outcome questionnaires in carpal tunnel syndrome. The ability of each of the questionnaires to distinguish patients who had clinically improved from those who had not was assessed, and cut-off values for change scores which showed improvement were established.

All of the questionnaires showed good test–retest reliability (Pearson’s reliability coefficient $r > 0.72$) with the DASH questionnaire being the most reliable ($r = 0.88$).

Comparing the level of agreement between patients that had improved and those which had not using the “gold standard”, PGIC (taking the a priori definition of clinically significant improvement as being a score of 6 or more), and the RCI or effect size statistics derived from the question-

naires identified which questionnaires were the most sensitive to clinical change. The DASH, Levine symptom, and Levine function questionnaires showed similar correlations of 60–70% agreement in categorising patients between PGIC and RCI. The Kamath questionnaire only showed a 58.1% agreement. The Kamath questionnaire also performed worst when comparing PGIC to the effect size statistic (using a value of 0.6 for moderate improvement) with an agreement of 51.2% compared with 65.1% for the DASH and Levine symptoms questionnaires and 72.1% for the Levine function questionnaire.

This study found that although the raw change scores that correlated with the a priori definition of clinically significant improvement varied between questionnaires, the DASH questionnaire required a much bigger change score (20.1) compared to the others. The percentage change score providing the best agreement was 20% for all the questionnaires except the Kamath where 10% gave a better correlation.

The major limitation of this study was that the patients were only followed-up for six weeks whereas previous research has shown clinical improvement to peak at six months post-op [8]. If patients are not followed-up until their response to surgery is at its greatest, some patients may be deemed as not having improved although they improve later, adversely affecting outcome results.

This study provides a methodological framework for interpreting the results of three outcome questionnaires in assessing their clinical significance. The study only looked at outcomes for a limited patient group of 43 patients; thus, further work is required to investigate the reliability of the values reported here by repeating the investigation in further groups of patients.

References

1. Amadio PC, Silverstein MD, Ilstrup DM, Schleck CD, Jensen LM (1996) Outcome assessment for carpal tunnel surgery: the relative responsiveness of generic, arthritis-specific, disease-specific, and physical examination measures. *J Hand Surg [Am]* 21:338–346
2. Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier C (2001) Measuring the whole or the parts? Validity, reliability and responsiveness of the disability of the arm, shoulder and hand outcome measure in different regions of the upper extremity. *J Hand Ther* 14:128–146
3. Changulani M, Okonkwo U, Keswani T, Kalairajah Y (2007) Outcome evaluation measures for wrist and hand—which one to choose? *Int Orthop* (in press). doi:10.1007/s00264-007-0368-z
4. Farrar JT, Portenoy RK, Berlin JA, Kinman JL, Strom BL (2000) Defining the clinically important difference in pain outcome measures. *Pain* 88:287–294
5. Greenslade JR, Mehta RL, Belward P, Warwick DJ (2004) Dash and Boston responsiveness of an outcome questionnaire? *J Hand Surg [Br]* 29:159–164
6. Gummesson C, Atroshi I, Ekdahl C (2004) The quality of reporting and outcome measures in randomized clinical trials related to upper-extremity disorders. *J Hand Surg [Am]* 29:727–734
7. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS (1998) Interpreting treatment effects in randomised trials. *BMJ* 316:690–693
8. Guyette TM, Wilgis EF (2004) Timing of improvement after carpal tunnel release. *J Surg Orthop Adv* 13:206–209
9. Heybeli N, Kutluhan S, Demirci S, Kerman M, Mumcu EF (2002) Assessment of outcome of carpal tunnel syndrome: a comparison of electrophysiological findings and a self-administered questionnaire. *J Hand Surg [Br]* 27:259–264
10. Hudak P, Amadio P, Bombardier C (1996) Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). *Am J Ind Med* 29:602–608
11. Hurst H, Bolton J (2004) Assessing the clinical significance of change scores recorded on subjective outcome measures. *J Manipulative Physiol Ther* 27:26–35
12. Jacobson NS, Follette WG, Revenstorf D (1984) Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behav Ther* 15:336–352
13. Kamath V, Stothard J (2003) A clinical questionnaire for the diagnosis of carpal tunnel syndrome. *J Hand Surg [Br]* 28:455–459
14. Kazis LE, Anderson JJ, Meenan RF (1989) Effect sizes for interpreting changes in health status. *Med Care* 27:S178–S189
15. Kirshner B, Guyatt G (1985) A methodological framework for assessing health indices. *J Chronic Dis* 38:27–36
16. Levine DW, Simmons BP, Koris MJ, Daltroy LH, Hohl GG, Fossel AH, Katz JN (1993) A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. *J Bone and Joint Surg [Am]* 75:1585–1592
17. Miles A, Charlton BG, Bentley P, Polychronis A, Grey J, Price N (2000) New perspectives in the evidence-based healthcare debate. *J Eval Clin Prac* 6:77–84
18. Rempel D, Evanoff B, Amadio PC (1998) Consensus criteria for the classification of carpal tunnel syndrome in epidemiologic studies. *Am J Public Health* 88:1447–1451
19. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB (2000) Evidence-based medicine how to practice and teach EBM, 2nd edn. Churchill Livingstone, London, pp 105–153
20. Sambandam SN, Priyanka P, Gul A, Ilango B (2007) Critical analysis of outcome measures used in the assessment of carpal tunnel syndrome. *Int Orthop* (in press). doi:10.1007/s00264-007-0344-7
21. Stevens JC, Sun S, Beard CM, O'Fallon WM, Kurland LT (1988) Carpal tunnel syndrome in Rochester, Minnesota, 1961 to 1980. *Neurology* 38:134–138
22. Testa M (1987) Interpreting quality of life clinical trial data for use in the clinical practice of antihypertensive therapy. *J Hypertens* 5 (suppl):S9–S13
23. Turk DC (2000) Statistical significance and clinical significance are not synonyms! *Clin J Pain* 16:185–187
24. Turk DC, Okifuji A, Sinclair JD, Starz TW (1998) Interdisciplinary treatment for fibromyalgia syndrome: clinical and statistical significance. *Arthritis Care Res* 11:186–195
25. Wyrwich KW, Wolinsky FD (2000) Identifying meaningful intra-individual change standards for health-related quality of life measures. *J Eval Clin Prac* 6:39–49