THIRD PRIZE

ASSESSING THE CLINICAL SIGNIFICANCE OF CHANGE SCORES RECORDED ON SUBJECTIVE OUTCOME MEASURES

Hugh Hurst, DC,[a] and Jennifer Bolton, PhD[b]

ABSTRACT

Background: To date, clinical trials have relied almost exclusively on the statistical significance of changes in scores from outcome measures in interpreting the effectiveness of treatment interventions. It is becoming increasingly important, however, to determine the clinical rather than statistical significance of these change scores.

Objective: To determine cutoff values for change scores that distinguish patients who have clinically improved from those who have not.

Method: Data were obtained from 165 back and 100 neck patients undergoing chiropractic treatment. Patients completed the Bournemouth Questionnaire (BQ) before treatment and the BQ and Patient's Global Impression of Change (PGIC) scale after treatment. Three statistical methods were applied to individual change scores on the BQ. These were (1) the Reliable Change Index (RCI); (2) the effect size (ES); and (3) the raw and percentage change scores. The PGIC scale was used as the "gold standard" of clinically significant change.

Results: The RCI, using the cutoff value of >1.96, appropriately identified clinical improvement in back patients but not in neck patients. An individual ES of approximately 0.5 had the highest sensitivity and specificity in distinguishing back and neck patients who had undergone clinically significant improvement from those who had not. In terms of raw score changes, percentage BQ change scores [(raw change score/baseline score) x 100] of 47% and 34% were identified as having the highest sensitivity and specificity in distinguishing clinically significant improvement from nonimprovement in back and neck patients, respectively.

Conclusion: This study provides a methodological framework for identifying clinically significant change in patients. This approach has important implications in providing clinically relevant information about the effect of a treatment intervention in an individual patient. (J Manipulative Physiol Ther 2004;27:26-35)

Key Indexing Terms: *Clinical Significance; Statistical Significance; Sensitivity; Specificity; Neck Pain; Back Pain*

INTRODUCTION

Evidence-based medicine advocates the application of findings from clinical trials in the treatment of individual patients. However, results from research studies are usually given as group mean values and the statistical significance of their differences. Data analyzed in this way give no indication of the proportion of patients in the group achieving a clinically important benefit from the treatment intervention. The information is therefore of limited clinical relevance, since there is no indication of the likelihood of a good response in a single patient. To counteract this, treatments are now being evaluated in terms of numbers needed to treat (NNT). NNT is an easily interpreted statistic informing the clinician of the number of patients that must be treated for a single patient to improve.[1,2] To calculate the NNT statistic, it is necessary to identify those patients in the group who have undergone a clinically important improvement.

Defining the proportion of patients who have clinically improved is problematic, however, when the outcome of interest is subjective and there are no directly measurable end points to indicate that the patient's condition has resolved. An example is in evaluating the effect of treatment in nonspecific back and neck pain where the outcomes of most interest are changes in patients' self-reported levels of pain and disability. In such cases, it is necessary to distinguish those individual change scores on pain and disability

[a]Private practice of chiropractic, Bristol, UK.
[b]Anglo-European College of Chiropractic, Bournemouth, England.
Submit requests for reprints to: Jennifer Bolton, PhD, Anglo-European College of Chiropractic, 13-15 Parkwood Road, Bournemouth BH5 2DF, England. (e-mail: *jbolton@aecc.ac.uk*).
Paper submitted September 5, 2003.
Copyright © 2004 by National University of Health Sciences.
0161-4754/$30.00
doi:10.1016/j.jmpt.2003.11.003

scales that represent clinically important change from those that do not.

There are now a number of methods available for identifying clinically important intraindividual change in subjective outcome measures.[3,4] These fall into 1 of 2 camps: the statistical or distribution-based methods on 1 hand and the global ratings or anchor-based methods on the other. The most common of the statistical methods are the effect size (ES) statistic and the Reliable Change Index (RCI), as well as simple change scores on the outcome measure itself.

The ES statistic is a method whereby mean differences between pretreatment and posttreatment scores can be standardized to quantify an intervention's effect in units of standard deviation (SD). It is therefore independent of measuring units and can be used to compare outcomes.[5] ES statistics are widely used to assess the magnitude of treatment-related changes over time and can be applied both to group data and to data recorded from a single patient.[4] Using threshold values put forward by Cohen[6] and Testa,[7] ES values for group mean changes and individual changes, respectively, can be interpreted as small, medium, or large treatment effects. The question remains, however, as to how effect sizes relate to patients' own perceptions of change in their condition and how effect sizes can be interpreted as clinically important effects. For example, thresholds for individual effect sizes in terms of clinically important change would enable patients to be identified as improved or not.

The RCI, originally proposed by Jacobsen et al[8] and later modified by Christensen and Mendoza,[9] is similar to the ES statistic in that it calculates mean differences between pretreatment and posttreatment scores but divides the difference by a standard error of measure that includes not only the SD of the measure but also its reliability coefficient. RCI values can be referenced to the normal distribution, and values that exceed 1.96 are unlikely ($P < .05$) unless an actual and reliable change has occurred.[3] Again, the question arises as to how this statistical method of arriving at a clinically important change compares with patients' own perceptions of a real and worthwhile change in their condition following treatment.

To assess patients' own impressions of change, a global scale from "much better" through "no change" to "much worse" is commonly used.[5,10,11] Since patients themselves make a subjective judgement about the meaning of the change to them following treatment, this scale is often taken as the external criterion or "gold standard" of clinically important change.[11] This makes intuitive sense and underlies current debates on statistical versus clinical significance.[12] Hence, in clinical trials in which end points cannot be directly measured, for example in pain conditions, assessing patients' experiences and what makes a difference to them in terms of a worthwhile and meaningful improvement is pivotal. Moreover, it is worth noting that statistical significance of change scores is derived from outcome measures that again rely on patients' interpretations and subjective judgments colored by their experiences of their condition.

The study reported in this article uses a patient self-report global change questionnaire based on a 7-point numerical rating scale (NRS) to determine from the patients' own perspective the degree of change (improvement) following treatment. This change was judged for its clinical importance by asking patients just how noticeable the change was. Using this as the "gold standard" of clinically significant improvement, the objectives of the study were to determine the sensitivity and specificity of statistical methods of determining clinically significant improvement, namely: (1) the RCI; (2) the ES statistic; and (3) the outcome measure's raw score and percentage score changes. Deyo and Centor[13] highlighted the importance of a measure not only in its ability to detect a clinically important change when it has occurred but equally in its ability to detect when a clinically important change has not occurred. The issue is therefore not merely one of sensitivity to change but also the ability of a measure to distinguish between those patients who do improve and those who do not. All the statistical methods under test in this study were based on individual change scores before and after treatment recorded on the Bournemouth Questionnaire (BQ), a multidimensional outcome measure based on the biopsychosocial model of musculoskeletal pain and validated for use in back[14] and neck[15] pain patients.

Methods

Data Collection

Consecutive new patients attending a chiropractic practice in Bristol, England with an episode of neck or back pain were recruited to the study. Existing patients who had not attended the clinic in the previous 3 months or more and presented with a new episode of back or neck pain were also recruited in a consecutive manner to the study. All patients were over 16 years of age. Eligible patients were asked to complete a pretreatment questionnaire, after which they underwent treatment as usual. Following the fourth treatment visit, some 2 weeks later, patients were asked to complete a posttreatment questionnaire.

The questionnaires asked a number of questions, some of which do not contribute to the study reported here. The parts relevant to this study were the Bournemouth Questionnaire, completed both before treatment for the painful episode (baseline) and after the fourth visit (follow-up), and a 7-point numerical rating scale (NRS) on the Patient's Global Impression of Change (PGIC) at follow-up. The BQ consists of seven 11-point NRSs (0-10) covering different dimensions of the pain experience. The 7 subscales consist of pain intensity, disability in activities of daily living and in social activities, anxiety and depression, and fear-avoidance and locus of control behavior. The raw scores from each of

Since beginning treatment at this clinic, how would you describe the change (if any) in ACTIVITY LIMITATIONS, SYMPTOMS, EMOTIONS and OVERALL QUALITY OF LIFE, related to your painful condition? (tick ONE box).

No change (or condition has got worse)	☐	1
Almost the same, hardly any change at all	☐	2
A little better, but no noticeable change	☐	3
Somewhat better, but the change has not made any real difference	☐	4
Moderately better, and a slight but noticeable change	☐	5
Better, and a definite improvement that has made a real and worthwhile difference	☐	6
A great deal better, and a considerable improvement that has made all the difference	☐	7

Fig 1. *Patients' Global Impression of Change (PGIC) scale.*

Table 1. *Calculation of sensitivity, specificity, accuracy, and positive (LR+) and negative (LR−) likelihood ratios*

Scale change scores/cutoffs	Clinically significant change: Improvement (positive outcome)	Clinically significant change: Nonimprovement	Totals
≥ y (positive outcome)	a	b	a + b
< y	c	d	c + d
Totals	a + c	b + d	(a + b + c + d) = n

Values for score change ≥ y: sensitivity = a/(a + c); specificity = d/(b + d); accuracy = (a + d)/n.
n = number of observations. LR+ = sensitivity/(1-specificity); LR− = (1-sensitivity)/specificity.

the subscales are summed to give the total raw score (maximum score 70) on the BQ scale. This total BQ scale score was used in all calculations in this study. The psychometric properties of the BQ have been rigorously tested in both back[14] and neck[15] patients, and either the back or the neck BQ was administered to patients, as appropriate. The PGIC scale was modified to tease out from patients exactly what the change in their condition following treatment meant to them (Fig 1).

Data Analyses

Reliable change index. The individual RCI for each patient was calculated as the difference in raw scores on the BQ at baseline and follow-up, divided by the standard error of the differences between 2 test scores (S_{diff}) where $S_{diff} = \sqrt{2(SE)^2}$ and $SE = SD_b\sqrt{1\text{-}r}$, and where SD_b is the SD of the group baseline scores and *r* is the reliability coefficient (intraclass coefficient).[9] In this case, *r* is 0.95 and 0.65 for the back[14] and neck[15] BQ, respectively. Individual patients were each categorized as improved if the individual RCI exceeded 1.96.[16]

Effect size. The individual ES statistic was calculated based on the method of Kaziz et al[17] and adapted for use in individual patients.[4] To compute the individual's ES, the difference in a patient's scores at baseline and follow-up on the BQ was divided by the SD of the group baseline scores. Individual patients were each categorized as having undergone a small, moderate, or large change using the threshold values of 0.2, 0.6, and 1.0, respectively, as proposed by Testa.[7]

Raw change and percentage change scores. The absolute or raw change scores from the BQ were obtained by subtracting the follow-up from the baseline scores for each patient. Since positive outcomes were denoted by a reduction in scale scores, this resulted in a numerically positive change in most patients. The percentage change score was calculated as the raw change score divided by the baseline score (×100).[11,18]

Sensitivity and specificity of cutoff and score change values in identifying clinically significant change. The a priori definition of clinically significant improvement was the PGIC categories of either "a great deal better" or "better." Thus, patients scoring either 6 or 7 on the PGIC scale (Fig 1) were categorized as "improved." However, in a similar way to Farrar et al,[11] because this definition is arbitrary, sensitivity and specificity calculations were also carried out for "moderately better" or "better" (ie, patients scoring 5, 6, and 7) and for "a great deal better" only (ie, patients scoring 7). Sensitivity and specificity of cutoff values and score change values were computed for each of these categories of clinically significant change as shown in Table 1.[19] Values that gave the best balance between the highest sensitivity, the highest specificity, and the highest accuracy were selected as the most fitting in identifying clinically significant improvement in individual patients as defined by the PGIC scale. In addition, the likelihood ratios[20] of the cutoff values and change scores as determined by the best balance of sensitivity and specificity were calculated as shown in Table 1.

Table 2. *Categorization of clinically significant change (improvement) using three methods in back and neck pain patients*

	Back pain		Neck pain	
*PGIC scale**	%	(n)	%	(n)
1) No change or worse	6.1	(10)	2.0	(2)
2) Almost the same	2.4	(4)	2.0	(2)
3) A little better	7.3	(12)	5.0	(5)
4) Somewhat better	2.4	(4)	8.0	(8)
5) Moderately better	24.9	(41)	20.0	(20)
6) Better	39.4	(65)	37.0	(37)
7) A great deal better	17.6	(29)	26.0	(26)
RCI (Improvement) (>1.96 cutoff)	63.6	(105)	26.0	(26)
ES				
1) Small improvement (>0.2 cutoff)	80.0	(132)	79.0	(79)
2) Moderate improvement (>0.6 cutoff)	63.6	(105)	61.0	(61)
3) Large improvement (>1.0 cutoff)	51.5	(85)	44.0	(44)

Back pain patients (n = 165).
Neck pain patients (n = 100).
(n) = number of observations.
PGIC, Patients' Global Impression of Change; *RCI,* Reliable Change Index; *ES,* Effect Size.
*See Figure 1 for actual wording.

Results

One hundred sixty-five back and 100 neck pain patients were recruited to the study between November 2000 and September 2001. Of the total patient sample, approximately half were males (51%) and the mean age was 40.5 (±13.91 [SD]) years. There was no difference in either the gender ratio or age between back and neck patients. There was an approximately even split between acute and chronic cases being treated. In the back pain group, 58% reported that their current episode of pain had lasted less than 7 weeks, with a corresponding figure of 55% in the neck pain group. Most patients in the back pain group reported a history of the complaint (73%), while in the neck pain group, just over half (55%) reported similar episodes in the past. The mean period of time between completion of the baseline and follow-up questionnaires was 14.2 (±10.85) and 15.1 (±10.86) days in back and neck patients, respectively.

Table 2 and Figure 2 show the proportion of patients categorized as undergoing a clinically important improvement using the anchor-based method of the PGIC scale and 2 distribution methods, namely the RCI and the ES statistics. Using the PGIC scale, 17.6%, 57.0%, and 81.9% of back patients were categorized as clinically improved using cutoff scores on the PGIC of 7, ≥6, and ≥5, respectively. The proportions of neck patients were similar, with 26.0%, 63.0%, and 83.0% categorized as clinically improved using the same sliding scale of cutoff values. For the RCI method of categorizing patients, there was a notable difference in back and neck pain patients. In back patients, the proportion that had improved was 63.6%, whereas in neck patients, the proportion was just 26.0%. This difference may be the result of the significantly lower reliability coefficient of the neck BQ (0.65) compared with the back BQ (0.95). This results in reducing the size of individual RCI values and therefore the number of patients with RCI values meeting the criterion of >1.96 that defines reliable improvement. Using the ES on the other hand, which does not include the reliability coefficient of the measuring instrument, proportions were again comparable in neck and back patients. The less rigorous individual ES cutoff value of >0.2 gave similar proportions of improved patients (80.0% and 79.0% in back and neck patients, respectively) as the cutoff value of ≥5 on the PGIC scale, while the more rigorous individual ES cutoff value of >0.6 gave similar proportions of improved patients (63.0% and 61.0% in back and neck patients, respectively) as the cutoff value of ≥6 on the PGIC scale (the a priori definition of clinically significant improvement). As a cautionary note in interpreting these results, no indication is possible at this stage of data analysis as to whether or not the proportions of patients categorized as improved by these 3 methods are actually the same patients.

Further analyses of these data using 2 × 2 tables (Table 1), however, does categorize patients using both the PGIC scale and either the RCI or the ES method, and the accuracy provides a measure of agreement of categorization of patients as "improved" and "not improved" between the 2 methods. The sensitivity, specificity, and accuracy of the RCI and the ES of categorizing individual patients as "improved" against the 3 cutoff values of the PGIC scale as the "gold standard" are shown in Table 3. For the RCI, the best balance between high sensitivity and high specificity in back patients was achieved using the a priori definition of clinical improvement on the PGIC scale (cutoff ≥6). In contrast, in neck patients, the best balance was achieved for the more rigorous cutoff value of 7 on the PGIC scale, although even here the sensitivity of the RCI was not that high. This is as expected, given the relatively low reliability of the neck BQ and the resultant small proportion of neck patients categorized as improved using the RCI method (Table 2 and Fig 2).

Using the ES as a method of identifying individual patients who have improved or not, the best balance between high sensitivity and high specificity was achieved in back and neck patients using the cutoff value of ±6 on the PGIC as the "gold standard" of clinical improvement (the a priori definition) (Table 3). Table 4 expands the cutoff values for the ES method and shows the actual individual ES cutoff with the best balance between high sensitivity and high specificity in identifying patients who have improved using the a priori definition of improvement. Cutoff individual ES values of >0.4 in back patients and >0.5 in neck patients were shown to be the best in distinguishing patients who had improved from those who had not (Table 4). Moreover, these cutoff values gave 72% and 80% agreement between

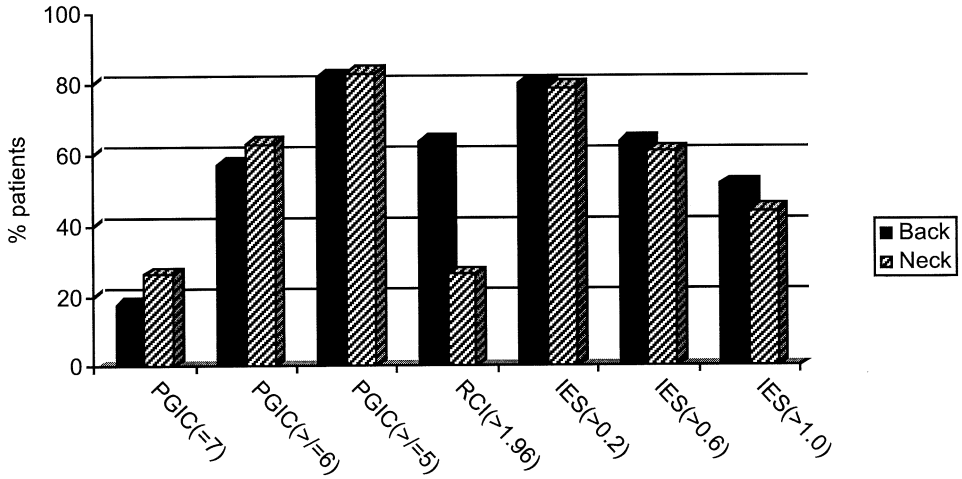

Fig 2. *Categorization of patients as clinically improved. Patients with back pain (n = 165) and neck pain (n = 100) were categorized as clinically improved using the PGIC (Patients' Global Impression of Change) scale, the RCI (Reliable Change Index) statistic, and the IES (Individual Effect Size) statistic. Cutoff values are given in parentheses.*

Table 3. *Sensitivity, specificity, and accuracy of the RCI and ES in identifying clinically significant change (improvement)*

	RCI		ES					
	>1.96		>0.2		>0.6		>1.0	
Cutoffs Patients	Back	Neck	Back	Neck	Back	Neck	Back	Neck
PGIC								
Cutoff = 7								
Sensitivity (%)	25.7	61.5	22.2	32.9	25.7	41.0	29.4	45.5
Specificity (%)	96.7	86.5	100	100	96.7	97.4	94.1	89.3
Accuracy (%)	51.5	80.0	37.6	47.0	51.5	63.0	61.2	70.0
Cutoff ≥ 6								
Sensitivity (%)	72.4	92.3	65.9	70.9	72.4	83.6	77.7	90.9
Specificity (%)	70.0	47.3	78.8	66.7	70.0	69.2	65.0	58.9
Accuracy (%)	71.5	59.0	68.5	70.0	71.5	78.0	71.5	73.0
Cutoff ≥ 5								
Sensitivity (%)	93.3	100	90.2	91.1	93.3	96.7	94.1	100
Specificity (5)	38.3	23.0	51.5	47.6	38.3	38.5	31.3	30.4
Accuracy (%)	73.3	43.0	82.4	82.0	73.3	74.0	63.6	61.0

Defined using 3 cutoff points of the PGIC scale in back (n = 165) and neck (n = 100) pain patients.
PGIC, Patients' Global Impression of Change; *RCI,* Reliable Change Index; *ES,* Individual Effect Size.

the individual ES and the PGIC in the categorization of patients as improved and not improved in back and neck pain patients, respectively. Calculation of the positive likelihood ratios (Table 4) showed that patients who had improved were approximately 3 times as likely to have scores above these cutoff values than patients who had not improved.

Tables 5, 6, 7, and 8 show the results of the sensitivity and specificity of raw change and percentage change scores on the back BQ and neck BQ in identifying patients who have clinically improved, defined using the 3 categories of improvement on the PGIC scale. As the category of improvement defined by the PGIC scale becomes less rigorous, so the change scores required to identify improvement are reduced. Thus, raw change scores of ±23, ±14, and ±9 (Table 5) and percentage change scores of ±64, ±47, and ±31% (Table 6) on the BQ in back pain patients identified clinical improvement defined by PGIC cutoff scores of 7, ±6, and ±5, respectively. Corresponding values in neck patients were ±18, ±9, and ±6 (Table 7) and ±58, ±34, and ±24% (Table 8), respectively. These data suggest that the BQ is a more responsive instrument to change in neck patients compared with back patients. Thus, taking the a priori definition of the PGCI cutoff of ±6 to denote clinically significant improvement, raw change scores of ±14 and ±9 and percentage change scores of ±47% and ±34% on the back and neck BQ, respectively, were selected as best distinguishing patients who have improved from those who have not. These cutoff scores were associated with levels of accuracy between 73% and 80%, indicating good agreement between change scores on the BQ and PGIC scores in

Table 4. *Sensitivity, specificity, accuracy, and likelihood ratios of expanded ES in identifying clinically significant change (improvement)*

Back pain					
ES cutoffs	>0.2	>0.3	>0.4	>0.5	>0.6
Sensitivity (%)	65.9	68.0	70.8	70.9	72.4
Specificity (%)	78.8	77.5	75.6	70.9	70.0
Accuracy (%)	68.5	70.3	72.1	70.9	71.5
LR+	3.1	3.0	2.9	2.4	2.4
LR−	0.43	0.41	0.39	0.41	0.39
Neck pain					
ES cutoffs	>0.2	>0.3	>0.4	>0.5	>0.6
Sensitivity (%)	70.9	73.3	79.4	83.8	83.6
Specificity (%)	66.7	68.0	71.9	74.3	69.2
Accuracy (%)	70.0	72.0	77.0	80.0	78.0
LR+	2.1	2.3	2.8	3.3	2.7
LR−	0.44	0.39	0.29	0.22	0.24

Defined using the PGIC scale (cutoff ≥ 6) in back (n = 165) and neck (n = 100) pain patients.
PGIC, Patients' Global Impression of Change; *ES,* Effect Size; *LR+*, Positive Likelihood Ratio; *LR−*, Negative Likelihood Ratio.

Table 5. *Sensitivity, specificity, accuracy, and likelihood ratios of raw change scores of the BQ in identifying clinically significant change (improvement)*

PGIC cutoff = 7								
Raw change scores	≥5	≥10	≥20	≥21	≥22	≥23	≥24	≥30
Sensitivity (%)	100	93.1	72.4	72.4	72.4	72.4	62.1	37.9
Specificity (%)	29.4	45.6	70.6	73.5	75.0	77.9	79.4	87.5
Accuracy (%)	41.8	53.9	70.9	73.3	74.6	77.0	76.4	78.8
LR+	1.4	1.7	2.5	2.7	2.9	3.3	3.0	3.0
LR−	0.00	0.15	0.39	0.38	0.37	0.35	0.48	0.71
PGIC cutoff ≥ 6								
Raw change scores	≥5	≥10	≥13	≥14	≥15	≥16	≥20	≥30
Sensitivity (%)	90.4	78.7	76.6	74.5	70.2	66.0	54.6	28.7
Specificity (%)	43.7	62.0	69.0	71.8	73.2	74.7	85.9	98.6
Accuracy (%)	70.3	71.5	73.3	73.3	71.5	69.7	67.9	58.8
LR+	1.6	2.1	2.5	2.6	2.6	2.6	3.9	20.5
LR−	0.22	0.34	0.34	0.36	0.41	0.46	0.53	0.72
PGIC cutoff ≥ 5								
Raw change scores	≥5	≥6	≥7	≥8	≥9	≥10	≥20	≥30
Sensitivity (%)	84.4	80.7	78.5	74.1	72.6	70.4	43.7	20.0
Specificity (%)	63.3	63.3	66.74	66.7	76.7	80.0	93.3	96.7
Accuracy (%)	80.6	77.6	76.4	72.7	73.3	72.1	52.7	33.9
LR+	2.3	2.2	2.4	2.2	3.1	3.5	6.5	6.1
LR−	0.25	0.31	0.32	0.39	0.36	0.37	0.60	0.83

Defined using 3 cutoff points of the PGIC scale in back pain patients (n = 165).
BQ, Bournemouth Questionnaire; *PGIC,* Patients' Global Impression of Change, *LR+*, Positive Likelihood Ratio; *LR−*, Negative Likelihood Ratio.

identifying improved and nonimproved patients. They are also associated with a positive likelihood ratio of between 2.6 and 3.4 (Tables 5, 6, 7, and 8), indicating that patients who have improved are approximately 3 times more likely to have a change score equal to or above the cutoff score than a patient who has not improved. The corresponding range of negative likelihood ratios of between 0.23 and 0.36 (Tables 5, 6, 7, and 8) indicates that a patient who has improved is approximately one third as likely to have a change score below the cutoff score than a patient who has not improved.

Discussion

In this study, 3 statistical methods derived from different computations of change scores on the BQ were investigated for their ability to distinguish patients who had undergone a clinically significant change from those who had not. The a priori definition of clinically significant improvement was a score of 6 or more on a 7-point NRS based on patients' global impression of change in their condition following treatment. This equated to feeling better or much better and a noticeable, worthwhile, and meaningful change. This an-

Table 6. *Sensitivity, specificity, accuracy, and likelihood ratios of percentage change scores of the BQ in identifying clinically significant change (improvement)*

PGIC cutoff = 7										
Percentage change scores	≥20	≥30	≥40	≥50	≥60	≥63	≥64	≥65	≥66	
Sensitivity (%)	100	100	93.1	89.7	79.3	72.4	72.4	72.4	65.2	
Specificity (%)	30.4	34.1	42.2	52.6	67.4	70.3	71.1	71.1	72.6	
Accuracy (%)	42.7	45.5	50.9	58.5	69.1	70.3	70.9	70.9	70.9	
LR+	1.4	1.5	1.6	1.9	2.4	2.4	2.5	2.5	2.4	
LR−	0.00	0.00	0.16	0.20	0.31	0.39	0.39	0.39	0.48	
PGIC cutoff ≥ 6										
Percentage change scores	≥20	≥30	≥40	≥46	≥47	≥48	≥49	≥50	≥60	
Sensitivity (%)	91.5	88.3	85.1	80.9	79.8	78.7	77.7	76.6	62.8	
Specificity (%)	47.1	50.0	64.3	72.9	74.3	74.3	74.3	74.3	88.6	
Accuracy (%)	72.1	71.5	75.8	77.0	77.0	76.4	75.8	75.2	73.3	
LR+	1.7	1.8	2.4	3.0	3.1	3.1	3.0	3.0	5.5	
LR−	0.18	0.23	0.23	0.26	0.27	0.29	0.30	0.31	0.42	
PGIC cutoff ≥ 5										
Percentage change scores	≥20	≥30	≥31	≥32	≥33	≥34	≥35	≥40	≥50	≥60
Sensitivity (%)	83.7	80.74	79.3	77.8	77.8	77.8	77.0	73.3	63.7	49.6
Specificity (%)	65.5	69.0	75.9	79.3	79.3	79.3	79.3	79.3	86.2	100
Accuracy (%)	80.0	78.2	78.2	77.6	77.6	77.6	77.0	73.9	67.3	58.2
LR+	2.4	2.6	3.3	3.8	3.8	3.8	3.7	3.5	4.6	∞
LR−	0.25	0.28	0.27	0.28	0.28	0.28	0.29	0.37	0.42	0.5

Defined using 3 cutoff points of the PGIC scale in back pain patients (n = 165).
BQ, Bournemouth Questionnaire; *PGIC,* Patients' Global Impression of Change; *LR+,* Positive Likelihood Ratio; *LR−,* Negative Likelihood Ratio.

Table 7. *Sensitivity, specificity, accuracy, and likelihood ratios of raw change scores of the BQ in identifying clinically significant change (improvement)*

PGIC cutoff = 7										
Percentage change scores	≥5	≥10	≥15	≥16	≥17	≥18	≥19	≥20	≥30	≥40
Sensitivity (%)	100	92.3	76.9	73.1	73.1	73.1	69.2	69.2	30.8	11.5
Specificity (%)	39.2	60.8	67.6	71.6	75.7	77.0	79.7	83.8	96.0	100
Accuracy (%)	55.0	69.0	70.0	72.0	75.0	76.0	77.0	80.0	80.0	77.0
LR+	1.7	2.4	2.4	2.6	3.0	3.2	3.4	4.3	7.7	∞
LR−	0.00	0.12	0.34	0.38	0.36	0.35	0.39	0.37	0.72	0.89
PGIC cutoff ≥ 6										
Percentage change scores	≥5	≥6	≥7	≥8	≥9	≥10	≥20	≥30	≥40	
Sensitivity (%)	87.3	85.7	85.7	81.0	77.8	73.0	44.4	17.5	4.8	
Specificity (%)	56.8	62.2	70.3	73.0	75.7	81.1	94.6	100	100	
Accuracy (%)	76.0	77.0	80.0	78.0	77.0	76.0	63.0	48.0	39.0	
LR+	2.0	2.3	2.9	3.0	3.2	3.9	8.2	∞	∞	
LR−	0.22	0.23	0.20	0.26	0.29	0.33	0.59	0.83	0.95	
PGIC cutoff ≥ 5										
Percentage change scores	≥3	≥4	≥5	≥6	≥7	≥10	≥20	≥30	≥40	
Sensitivity (%)	86.8	84.3	80.7	79.5	75.9	62.7	36.1	13.3	3.6	
Specificity (%)	58.8	70.6	76.5	88.2	88.2	94.1	100	100	100	
Accuracy (%)	82.0	82.0	80.0	81.0	78.0	68.0	47.0	28.0	20.0	
LR+	2.1	2.9	3.4	6.7	6.4	10.6	∞	∞	∞	
LR−	0.23	0.22	0.25	0.23	0.27	0.40	0.64	0.87	0.96	

Defined using 3 cutoff points of the PGIC scale in neck pain patients (n = 100).
BQ, Bournemouth Questionnaire; *PGIC,* Patients' Global Impression of Change; *LR+,* Positive Likelihood Ratio; *LR−,* Negative Likelihood Ratio.

chor-based method has been used in many other studies to determine clinically significant change.[11,21-23] In the absence of a true gold standard, asking patients themselves what constitutes a meaningful change to them, with all the attendant internal and external factors that might influence such judgment, seems intuitively the best that can be done when investigating issues of clinically important change.

This study identified from 70% to 80% agreement in categorizing patients as improved or not improved between asking patients directly on a PGIC scale and indirectly using cutoff values with high sensitivity and specificity on outcome measures. Since both methods rely on patients' own subjective judgements about change in their condition, this is reassuring. Many agreement studies rule out agreement

Table 8. *Sensitivity, specificity, accuracy, and likelihood ratios of percentage change scores of the BQ in identifying clinically significant change (improvement)*

PGIC cutoff = 7											
Percentage change scores:	≥20	≥30	≥40	≥50	≥56	≥57	≥58	≥59	≥60		
Sensitivity (%)	96.2	96.2	92.3	92.3	84.6	80.8	76.9	73.1	73.1		
Specificity (%)	33.8	44.6	54.1	66.2	74.3	75.7	77.0	77.0	77.0		
Accuracy (%)	50.0	58.0	64.0	73.0	77.0	77.0	77.0	76.0	76.0		
LR+	1.5	1.7	2.0	2.7	3.3	3.3	3.3	3.2	3.2		
LR−	0.11	0.085	0.14	0.17	0.21	0.25	0.3	0.35	0.35		
PGIC cutoff ≥ 6											
Percentage change scores:	≥20	≥30	≥33	≥34	≥35	≥36	≥37	≥40	≥50	≥60	
Sensitivity (%)	85.7	84.1	84.1	82.5	82.5	81.0	79.4	77.8	69.8	50.8	
Specificity (%)	46.0	64.9	73.0	75.7	75.7	75.7	75.7	75.7	86.5	89.2	
Accuracy (%)	71.0	77.0	80.0	80.0	80.0	79.0	78.0	77.0	76.0	65.0	
LR+	1.6	2.4	3.1	3.4	3.4	3.3	3.3	3.2	5.2	4.7	
LR−	0.31	0.25	0.22	0.23	0.23	0.25	0.27	0.29	0.35	0.55	
PGIC cutoff ≥ 5											
Percentage change scores:	≥20	≥23	≥24	≥25	≥26	≥27	≥28	≥30	≥40	≥50	≥60
Sensitivity (%)	81.9	79.5	77.1	77.1	75.9	75.9	75.9	75.9	68.7	59.0	43.4
Specificity (%)	64.7	70.6	76.5	76.5	76.5	76.5	82.4	82.4	94.1	100	100
Accuracy (%)	79.0	78.0	77.0	77.0	76.0	76.0	77.0	77.0	73.0	66.0	53.0
LR+	2.3	2.7	3.3	3.3	3.2	3.2	4.3	4.3	11.6	∞	∞
LR−	0.28	0.29	0.30	0.30	0.32	0.32	0.29	0.29	0.33	0.41	0.57

Defined using 3 cutoff points of the PGIC scale in neck pain patients (n = 100).
BQ, Bournemouth Questionnaire; *PGIC,* Patients' Global Impression of Change; *LR+,* Positive Likelihood Ratio; *LR−,* Negative Likelihood Ratio.

that occurs by chance by using the κ statistic in data analyses instead of simple percent agreement. However, in this case, since the data were not recorded as binary variables, the κ statistic was not considered to be an appropriate method of analysis.

One of the 3 statistical methods used to categorize patients as improved and not improved, the RCI, gave anomalous results both in identifying the proportion of neck patients in the sample who improved and in calculations involving the PGIC scale. Neither of these findings was apparent when the RCI was used in back patients. The reliability coefficient of the neck BQ was relatively low, and this may have resulted in an overrigorous threshold for identifying patients who improved. Caution is therefore indicated when identifying clinically important improvement using the RCI for outcome measures in which reliability is moderate to poor.

The results of the sensitivity and specificity analyses showed that the second statistical method used in this study, the individual ES statistic, can be used to distinguish patients who improve from those who do not using the a priori definition of clinically important improvement from the PGIC scale. The findings of this study show that clinically significant improvement is indicated for individual back patients with an ES statistic of 0.4 or more and individual neck patients with an ES statistic of 0.5 or more. The similarity of these 2 values suggests that an overall individual ES cutoff of 0.5 for both types of patients rather than the exact values would be more convenient for use in a clinical setting and in the design of clinical trials.

The study has shown that the third statistical method under test can also be used to distinguish patients who have improved from those who have not. Raw change scores of 14 or more and percentage change scores of 47% or more were best associated with the a priori definition of clinical change in back pain patients. Corresponding cutoff values in neck pain patients were lower at 9 or more for raw change scores and 34% or more for percentage change scores. Using a similar definition of clinically important improvement, Farrar et al[11] showed that a percentage change score of approximately 30% on an 11-point pain intensity NRS best distinguished chronic pain patients who had improved from those who had not. In an accompanying study to this one, using the BQ in a different sample of neck pain patients and using the RCI (but without the correction factor proposed by Christensen and Mendoza[9]) to identify clinically improved patients, corresponding cutoff values were raw score changes of 13 or more and percentage change scores of 33% (Bolton, submitted for publication). The similarity of the cutoff percentage change score value in both studies suggests this might be more appropriate as a clinical tool in identifying patients who have improved. Moreover, percentage change score is a standardized measure that is more easily interpretable, particularly when different outcome measures with different scales are in use. Farrar et al[11] concluded that in studies in which there is high variability in baseline pain levels, the relationship between percentage change and clinical improvement will be more consistent than the relationship between raw change and clinical improvement.

This article provides a methodological framework for interpreting statistical computations from outcome measures in terms of their clinical significance. In essence, it treats these computations as diagnostic tests in determining the presence or absence of a clinically significant change. There is a considerable amount of potential bias in the evaluation of diagnostic tests[24] and a strength of this study was that it avoided selection bias by recruiting patients in a consecutive manner. However, the study only looks at scores from 1 outcome measure in a limited patient group and change that occurs over a relatively short period of time. Moreover, the modified PGIC scale has not been tested for reliability or validity, nor has it been shown to be a valid external criterion for clinically significant change, even though we used it as such. In an area where there is an array of methods to define minimal important difference (anchor-based and statistical), more work is required to identify just what does constitute a clinically important difference, so that it can be used with confidence as a valid external criterion in future studies. Further work is also required into other outcome measures and other conditions. In particular, the reliability of the cutoff values reported in this study should be investigated by repeating the work in different samples of patients. In conditions such as back and neck pain, which are notoriously unpredictable and heterogeneous, issues of reliability are of paramount importance when cutoff values are being proposed for use in other settings. It is also the case that since this study's design did not include a control group, no conclusions have been drawn on the cause of the improvement observed in these patients and therefore the effect of the treatment intervention.

Conclusion

This study presents a number of threshold values on statistical computations from change scores that best identify patients undergoing clinically significant change from those who have not. This work is based, however, on the PGIC as an external criterion of clinically significant change, and while this may be both conceptually reasonable and clinically relevant, it remains to be seen whether or not this is a valid assumption. By identifying proportions of patients who have undergone clinically important change, calculations can be made of the NNT and thus facilitate the application of group results from clinical trials to an individual patient. This transition from research setting to clinical setting underpins the principles of the practice of evidence-based health care.

Acknowledgments

With thanks to Ms Luci Rowe and Ms Christine Kite.

References

1. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomized trials. Br Med J 1998;316:690-3.
2. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine. London: Churchill Livingstone; 2000. p. 105-53.
3. Turk DC. Statistical significance and clinical significance are not synonyms! Clin J Pain 2000;16:185-7.
4. Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. J Eval Clin Pract 2000;6:39-49.
5. Middel B, Stewart R, Bouma J, van Sonderen E, van den Heuvel W. How to validate clinically important change in health-related functional status. Is the magnitude of the effect size consistently related to magnitude of change as indicated by a global question rating? J Eval Clin Pract 2001;7:399-410.
6. Cohen J. Statistical power analysis for the behavioural sciences. New York: Academic Press; 1977.
7. Testa M. Interpreting quality of life clinical trial data for use in the clinical practice of antihypertensive therapy. J Hypertens Suppl 1987;5:S9-S13.
8. Jacobson NS, Follette WG, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. Behav Ther 1984;15:336-52.
9. Christensen L, Mendoza J. A method of assessing change in a single subject: an alteration of the RC index. Behav Ther 1986;17:305-8.
10. Wyrwich K, Nienaber N, Tierney W, Wolinsky F. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 1999;37:469-78.
11. Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole M. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical rating scale. Pain 2001;94: 149-58.
12. Rowbotham MC. What is a 'clinically meaningful' reduction in pain? Pain 2001;94:131-2.
13. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39:897-906.
14. 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.
15. Bolton JE, Humphreys BK. The Bournemouth Questionnaire: a short-form comprehensive outcome measure. I. Psychometric properties in neck pain patients. J Manipulative Physiol Ther 2001;25:141-8.
16. Turk DC, Okifuji A, Sinclair JD, Starz TW. Interdisciplinary treatment for fibromyalgia syndrome: clinical and statistical significance. Arthritis Care Res 1998;11:186-95.
17. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27(Suppl 3): S178-89.
18. Little DG, MacDonald D. The use of the percentage change in Oswestry Disability Index score as an outcome measure in lumbar spinal surgery. Spine 1994;19:2139-43.
19. Farrar JT, Portenoy RK, Berlin JA, Kinman JL, Strom BL. Defining the clinically important difference in pain outcome measures. Pain 2000;88:287-94.
20. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine. London: Churchill Livingstone; 2000. p. 67-93.

21. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407-15.
22. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994;47:81-7.
23. Beurskens AJHM, de Vet HCW, Koke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71-6.
24. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.