Assessment of change in clinical evaluation

Middel, Lambertus Johannes

5

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?

**Berrie Middel, Msc*, Roy Stewart, Msc*,
Jelte Bouma, Ph.D.*, Eric van Sonderen, Ph.D.*,
Wim J.A. van den Heuvel, Ph.D*.**

* Northern Centre for Healthcare Research. School of Medicine, University of Groningen, The Netherlands

Keywords: Health status indicators; prospective studies; questionnaires; heart failure; treatment outcome; clinically relevant change; stratified effect size

Submitted

SUMMARY

Some clinical trials perform repeated measurement over time and estimate clinically relevant change in instrument's score with global ratings of perceived change or so-called transition questions.
The conceptual and methodological difficulties in estimating the magnitude of clinically relevant change over time in health related functional status (HRFS) are discussed. This paper investigates the concordance between the amount of serially assessed change with effect size estimates (the researcher's perspective) with global ratings of perceived change (the patient's perspective) is described.
A total of 217 patients who were scheduled for diagnostic examination were included, and the Minnesota Living with Heart Failure Questionnaire, extended with MOS-20 items, was assessed before and after medical intervention (Percutaneous Transluminal Coronary Angioplasty, Coronary Artery Bypass Grafting or pharmaco-therapy) . Global questions were applied to assess perceived change over time in for every item from domains of physical and emotional functioning and used as the external criterion of relevant change in the analysis of items. Global questions corresponding with overall change in these domains were used in the comparison of change in physical and emotional functioning scales. Two effect size indices were used: 1. ES (mean change/SDpooled) and 2. ES (mean change/SDchange). A method is described to calculate a value indicating the extent of discordance between the researcher's interpretation of magnitude of change and the external criterion (the patient's perspective).
Findings suggest effect size ES (mean change/SDpooled) was in keeping with the magnitude of change indicated by patient's judgement, or their category of subjective meaning, for all scales. Furthermore, in cases that the magnitude of change estimated with the SRM (mean change/SDchange) was not confirmed empirically by the external criterion ratings, the discordance could be interpreted as a trivial discordance.

5.1 INTRODUCTION

In publications on methods of assessment of change in health-related functional status (HRFS), the concept of responsiveness is used as either a psychometric quality of a measurement instrument or an indicator of the amount of change over time. The use of the term *responsiveness,* however, confuses the reader because the concept of responsiveness, used in papers addressing treatment-related health status change, can refer to a varying composite of aspects:

1. the ability to detect change over time [1-7] or the extent to which a measure is sensitive to *real* change in health-related functional status (HRFS) [8] ;
2. the sensitivity of a health status instrument by analogy with test performances in clinical practice (the ability of an instrument to detect the smallest change), or as a property of measures used to assess the effectiveness of medical interventions [5,6,9-11];
3. the ability to detect a clinically relevant or important change over time, according to an external criterion, to distinguish *between* improved and non-improved subjects [7,12-15];
4. the relative strength of correlation *between* the change in instrument score and an external criterion of perceived change or satisfaction with treatment [16,17].

There seems to be no unambiguous method to define and assess the concept of responsiveness in terms of measuring clinically relevant change in HRFS. Clinicians, for instance, use reference values (reference range) for clinical 'laboratory' health status indicators, such as blood sodium or erythrocyte sedimentation rate, as anchors for the degree of deviation from what can be valued as 'normal'. Reference values also give the opportunity to value changes after treatment as being trivial, or substantial and clinically relevant in the expected direction. In contrast, when HRFS is relevant in the treatment outcome evaluation, researchers do not have a 'population-based' reference range of values or common sense anchors for measures of e.g. physical functioning to value the outcome after treatment in terms of clinical relevance. In the absence of such a reference range or "golden standard", an estimate of clinically relevant change requires an external criterion to provide cut-off points or a reference range to discriminate between relevant and irrelevant change. One common method of interpretation is to compare health status score with a global subjective judgement of the direction and amount of change by clinician or patient [12,18,19] , often referred to as the external criterion. This subjective judgement is obtained by asking the extent to which deterioration or improvement has occurred since treatment, using a global question with verbal anchors ranging from a

dichotomous scale (e.g. improved vs. not improved) [20,21] to a 15-point scale ranging from -7='a very great deal worse' to +7 = 'a very great deal better' [22-25]. In other words, these verbal anchors can be used to estimate a relevant difference in an instrument's score over time. Thus, patients can be classified as having small but meaningful change in health status score if they state that they have changed 'a little' or 'somewhat' (sometimes defined as the minimal clinically important difference). Change scores represent moderate change if patients felt they had changed 'moderately' or 'a good deal'; scores represent large change if patients state that they have changed 'a great deal' or a 'very great deal' [22,23,26]. Mean differences can be standardized to quantify an intervention's effect in units of standard deviation, and allow comparison of the different outcomes of one intervention, independent of the measuring units. The resulting statistical measure is known as effect size index.

If we use an effect size index to assess the magnitude of treatment-related change over time, regardless of its outcome parameter and range of standardized values, we can give it meaning "with the 'straitjacket' provided by Cohen some thirty years ago". [27] The values used to classify effect sizes for mean differences as 'small', 'medium', and large' and the widely used thresholds of Cohen for effect size interpretation are: trivial effect (ES = <.20) , small effect (ES = ≥ .20 <.50) , medium effect (ES = ≥ .50 <.80) and large effect (ES = ≥ .80). The point open to discussion is: how are these effect size interpretations related to subjective ratings of magnitude of change with global questions? Our study's objective was to compare the classes of responsiveness as defined by Cohen with the self-report perception of the magnitude of change.

We used two perspectives from which the importance of a change over time in health status can be determined and which we shall discuss further in the coming paragraphs:

1. the researcher's perspective, calculating the score difference between two points in time and estimating the magnitude of change in multi-item dimensions of health-related quality of life with an effect size index, and
2. the patient's perspective, estimating change by single global or transition questions at post-test, asking directly how much the patient has experienced improvement or deterioration in HRFS since treatment.

5.1.1. The researcher's perspective

A within-group effect size index is generally the result of subtracting baseline scores from post-test scores (or vice versa) and dividing the mean difference by a standard deviation. There is, however, still no consensus on the most appropriate strategy for interpreting the standardized change score in health-related quality of life as

treatment outcome in medical intervention evaluation. Guyatt et al. [9] recommended the Responsiveness Index (RI) as the ratio of the average, treatment-related, change to the variability of scores in stable subjects as the most appropriate measure of responsiveness. We believe that a measure of change is not a function of stable patients and is inherently prone to overestimation or underestimation of the magnitude of change, because the numerator and denominator are based on different samples [13,18] Therefore, the within-group effect size was estimated using two methods:

1. Cohen [27,28] who introduced the effect size, calculated as the mean change in score divided by the pooled standard deviation of some repeatedly assessed outcome measure used in an experiment as follows:

$$ES = d' \quad \frac{\overline{X}_{treatment} - \overline{X}_{control}}{\sigma}$$

   With this estimate of effect size, after analysing a wide sampling of behavioural research, Cohen developed his rules of thumb for effect size interpretation. [29]
2. The Standardised Response mean (SRM), which is calculated by dividing the mean change of a serially assessed measure by the standard deviation of the change score (i.e. difference in score before and after medical intervention). In contrast with what seems to be widely assumed, it was not Cohen, but Liang et al. [30] who introduced this effect size to avoid confusion with the effect size index proposed by Cohen for correct use of his power tables.

In this study both ES and SRM are conceived as an estimate of the magnitude of treatment-related change in domains of HRFS. The values of effect size indices, derived from mean change scores in health status measures, vary from approximately –2 to +2 in similar study designs. With these two methods the mean change scores are standardized on three scales, covering a disease-specific and a generic measure of physical functioning and one disease-specific measure of emotional functioning.

5.1.2. The patient's perspective

As mentioned before, another method of estimating change is the clinician's judgement, or posing global questions to patients regarding how much they have improved or deteriorated since treatment. We consider the self-determined direction and magnitude of change as the best estimate of clinically relevant change. Therefore, this study has used the patient's perspective as the external criterion to estimate the

magnitude of perceived improvement in the domains of physical and emotional functioning.

5.2 PATIENTS AND METHODS

To ensure that change in health status occurred, we selected a group of patients undergoing a treatment with known efficacy, and used a disease-specific instrument with known sensitivity in detecting change over time. [31,32] The study sample was composed of patients who, after a diagnostic Coronary Angio Graphy (CAG), were scheduled for Percutaneous Transluminal Coronary Angioplasty (PTCA), for Coronary Artery Bypass Grafting (CABG), or were treated with a pharmaco-therapy. To assess change in physical and emotional functioningserially, we used items from a disease-specific health-status instrument (namely, the Minnesota Living with Heart Failure Questionnaire (MLHF-Q) [33,34] and from the MOS-20, a generic instrument. [35]

To assess change directly in both health domains, we modified all items into direct questions of perceived change (transitional questions), to be administered at post-treatment. Three items of the MOS-20, valued as most appropriate for the study sample, were used in the same format as a measure of physical functioning.

5.2.1. Patient selection

Patients were recruited from January to December 1998 from Groningen University Hospital, Martini Hospital Groningen, and Weezenlanden Hospital in Zwolle, in the Netherlands. Patients with other incapacitating diseases or cognitive impairments, aged 75 or older, or who did not speak Dutch were excluded. Ethical approval was obtained from each participating hospital's ethics committee. We prospectively administered the questionnaire at baseline and 6 weeks after the decision for non-invasive intervention, or 6 weeks after the day a PTCA /CABG-intervention was executed. Patients returned questionnaires at baseline accompanied by written informed consent. Returned questionnaires were routinely checked for completeness. If many questions or pages were not filled in, either a copy was sent with a kind request for completion or, in cases of only one question's omission, patients were interviewed by telephone. Because the questionnaire's completeness was monitored by a computer program both at baseline and follow-up, we effectively reduced the non-response on questions, and consequently, on scales.
We presumed that at baseline, i.e. prior to CAG, neither patients nor cardiologists had information about decisions concerning either intervention, and would thus not

affect the assessment of subjective health, most likely reducing the risk of 'floor and ceiling' effects. However, this control for potential bias resulted in logistic problems and, six months after the study began, we were forced to select patients waiting for outpatient treatment (PTCA) or hospital admission (CABG) as soon as they were scheduled on the waiting list.

5.2.2. Data Collection and measures

The Minnesota Living with Heart Failure Questionnaire (MLHF-Q) is a disease-specific instrument composed of 21 items and three scales measuring the following: the physical functioning dimension (8 items), the emotional functioning dimension (5 items) and the overall score on HRFS (21 items). Eight separate items, not assessing an underlying construct or dimension of HRFS, measure social and economic impairments which patients relate to their heart failure, and are part of the overall score. The original MLHF-Q items were phrased as follows: "Did your heart failure prevent you from living as you wanted during the last month by making your sleeping well at night difficult?"
The response options range from 'no' (score 0), very little (score 1) to very much (score 5). The total score ranges between 0 and 105, the physical dimension (sub-scale) between 0 and 40, the emotional dimension (sub-scale) between 0 and 25. To assess physical functioning, we extended the questionnaire with 3 items from the MOS-20. [35] These three items were: "Did your heart failure prevent you from living as you wanted during the last month by making it difficult for you to 1) bend, stoop or lift light objects 2) lift heavy objects, like moving a table and 3) run at a fast pace? "
Two methods of assessing change in health-related quality of life (HRQL) with multi-item scales were applied with the study data: HRQL-domains were serially measured with items from MLHF-Q and MOS-20, and consequently, the patient's perception or subjective significance of change was captured at follow-up of each of the MLHF-Q and MOS-20 items in terms of the extent of feeling improved, deteriorated or not changed.
These transition items are designed to elicit information regarding perceived change over time in specific aspects or in domains of the patient's heath status. For each item in the questionnaire (except the items on socio-economic impairments: 'hospitalisation and medical costs'), patients rated at post-test the degree to which they perceived that change had occurred, on that particular item, since baseline assessment.
Two global questions corresponded to the domains of physical functioning and emotional functioning of the serially assessed questionnaire. These items were

intended to capture change in the domains of HRFS as prospectively measured and were phrased as follows: "Since the last time I filled out the questionnaire (or: since my operation), my physical problems are ….. ", 2) "Since the last time I filled out the questionnaire (or: since my operation), my emotional problems are ….."
For every transition item and global question the patient was asked to circle the answer best describing his/her perception of the direction and magnitude of change at post-test on a seven-point Likert scale: 1) a great deal worse; 2) moderately worse; 3) a little worse; 4) no change; 5) a little better; 6) moderately better and 7) a great deal better.

5.3 DETERMINATION OF CONCORDANCE BETWEEN TWO INTERVALS OF MAGNITUDE OF CHANGE

Given that we established this study to evaluate assessment methods in health status, we selected patients undergoing treatment with known efficacy; there were only 20 patients indicating post-treatment deterioration. Therefore we excluded the calculated effect sizes of patients who deteriorated. Consequently, we analysed the concordance between magnitude of change according to Cohen and the external criterion as follows:

Cohen's thresholds Standardized Change score		**External Criterion (global question):** Since the last time you filled out the questionnaire (or since your operation), has there been any change in your physical/emotional problems related with your heart failure?
Trivial effect	= <.20	no change
Small effect	= ≥ .20 <.50	a little better
Medium effect	= ≥ .50 <.80	moderately better
Large effect	= ≥ .80	a great deal better

The effect sizes of the interval Cohen defined as 'trivial effect' vary between the values ES=0 and ES = .20. We have presumed that, for example, the subjective judgement 'there has been no change' matches the judgement of the researcher using Cohen's rules of thumb for effect size, in order to give meaning to the estimated magnitude of change within this interval. If we use the verbal anchor of the external

criterion to determine the interval in which the effect size index should lie, the magnitude of change, according to the researcher's effect size interpretation, will deviate from the patient's interpretation. The concordance between ES and global question, used as external criterion, will never be perfect if we make rigid comparisons. For example: in a sub-group considering themselves unchanged after treatment, the estimated magnitude of change in score was reflected by an ES= 0.25 which, according to Cohen, was evaluated as a small improvement by the investigator (small effect: ES = 0.20 – 0.50). This ES of 0.25 is presumed to be out of range with the interval that, theoretically, should correspond with the retrospective 'no change' rating (ES = 0 – 0.20). If we want to validate the researcher's interpretation of the amount of prospective change using the external criterion as an anchor point, we need an indication of the extent of discordance between two interpretations. Thus, how serious is the ES=0.25 deviation from the 'no change' interval's upper limit, in this case ES=0.20? To find an answer we calculated a value by means of which every effect size index can be evaluated within the intervals determined by the external criterion anchor points. We considered the effect size of a treatment in three situations in which the estimated magnitude of change (ES) is: 1. concordant with the external criterion interval; 2. discordant with the external criterion in terms of overestimation (the ES represents, according to Cohen's thresholds, a larger magnitude of change than indicated by the patient's judgement), and 3. discordant with the external criterion interval in terms of underestimation (the ES reflects a smaller magnitude of change according to the assumed interval corresponding with the patient's judgement).

The value by which we express the concordance between the researcher's interpretation of the estimated magnitude of change, Cohen's thresholds and the magnitude of perceived change (according to the external criterion), has the advantage of being easily interpreted, and its range is from 0 to 1. With the interpretation of the values between this minimum and maximum we must take into account in which of the three aforementioned situations has the comparison between effect size and external criterion has been made.

In the event that an effect size lies within the interval concordant with the external criterion, or represents a surplus of the effect size, the maximum value is +1, whereas the maximum is –1 when the effect size reflects a lower magnitude of change than determined by the external criterion. When the effect size is concordant with the external criterion, a value of zero means that the ES coincides with the lower limit of the interval, and the value 1.0 means that it coincides with the upper limit.

In addition, we used 0.50 as the range midpoint with its class boundaries of 0.40 and 0.60 and, in the case of discordance, we signified the calculated value as follows: 0 – 0.20 = 'poor discordance'; > 0.20 – < 0.40 'small discordance'; 0.40 - < 0.60 'fairly

large discordance'; 0.60 - < 0.80 'large discordance'; and 0.80 – 1.00 'very large discordance'.
In the situation of overestimation or underestimation, the values receive a positive or negative sign respectively, and can be interpreted as the extent of the surplus or shortfall of effect size as determined by the external criterion.

5.3.1. Effect size concordant, according to the external criterion

Change in a serially assessed physical functioning scale was interpreted as small (ES = 0.26) in the group of patients considering themselves physically 'a little better' (see: table 5.3-b). We presumed that the upper and lower limits of this retrospective judgement of improvement in physical functioning corresponds with the effect size interval of what is assumed by Cohen to be a 'small effect', ES = .20 - .50, (range .30). We determined the value, expressing 'the extent of concordance' as follows: the distance between the operative ES (0.26) and the interval's lower limit was determined (0.26 minus 0.20 = 0.06) and divided by the interval's range (0.06/0.30=0.20).
When the magnitude of change valued by the researcher, with Cohen's rule of thumb, is concordant with the amount of change valued by the patient's judgement, the indicator has a minimum value of 0 with a maximum of 1. The value is 0 when identical with the interval's lower limit, and 1 when identical with the interval's upper limit. In this example the value of 0.20 indicates that we can interpret it as the proportion it occupies from the interval's range in the direction of the lower limit.

5.3.2 Effect size discordant, overestimation according to the external criterion

It will occasionally occur that the estimated magnitude of change does not correspond with the external criterion. The interpretation of the magnitude of change, according to Cohen, can indicate a larger effect than was expected to correspond within the judgement of the patient. For example: an ES= 0.75 was found with the group of patients that considered their improvement as 'a little better'. If we presume that this judgement corresponds with effect sizes ranging between 0.20 – 0.50, we will conclude that an ES = 0.75 is an overestimation in relation to the external criterion by crossing the threshold of ES = 0.50. To get an estimate of the seriousness of this deviation we cannot calculate the indicator in the open-ended interval for large effect (ES $\geq$ 0.80 standard deviation units). A maximum value ($\geq$.80) of the studied effect size is necessary to estimate the extent of concordance with the external criterion. Therefore, we fixed the maximum of

standardized change over time at the 1.26 SD we detected in our sample, and calculated the range between this maximum and the upper limit of the interval as determined by the external criterion (1.26 – 0.50 = 0.76). The difference between the operative effect size and the upper limit of the interval corresponding with the external criterion, 0.25 (0.75 – 0.50), is divided by the range of the interval, resulting in a value of 0.33 (0.25/0.76). According to our rule of thumb we would value the discordance with the external criterion as small.

5.3.3. Effect size discordant, underestimation according to the external criterion

Suppose an ES = 0.73 was found in relation to the external criterion 'a great deal better' which, according to our assumption, is considered relating an underestimation to the lower bound of the interval corresponding with the external criterion, in this case ES=0.80.
We calculated the range between the maximum value of ES in our sample and the interval's lower limit as determined by the external criterion (-1.26 – 0.80 = -2.06). The difference between the operative effect size and the lower limit of the interval corresponding with the external criterion is 0.07 (0.80 – 0.73); divided by the interval's range, this gives a value of – 0.03 (0.07/-2.06). We consider this a trivial discordance with the interval determined by the external criterion.

5.4 RESULTS

Of the 398 candidates screened for inclusion in this study, 139 (34.9%) did not return the first questionnaire. Questionnaires were received from the remaining 259 patients. We could not test the probability of systematic differences between non-respondents and the study sample because information was inaccessible without written informed consent from the patients not returning the first questionnaire.
Forty-two patients (16.2%) dropped out before the post-test assessment. The reasons for not responding at post-test were because the patient died (n= 7), had no heart failure (n=9), refused further participation (n=9), was too ill at post-test (n=3), had moved (n=3) or did not react at all (n=11). To ensure that the patients who left at post-test did not deviate systematically from the study group, their characteristics at the time they returned the first questionnaire were compared with the baseline characteristics of those who completed the post-test questionnaire. Except for education level (the study sample had a statistically significant higher education), the demographic characteristics of the two groups were similar. This comparison also

showed no statistically significant differences in mean scores on baseline health-status scales.

Analyses were based on 217 subjects (83.8%) who filled in the questionnaires at both baseline and post-test.
The mean age of the patients was 60.6 (SD ±9.43) with a range of 25 to 75. Sixty-one (28%) were female and 156 (72%) male. Men were more likely to have a partner, live with someone, have a higher education, and be employed. Five percent, 44%, 21% and 27%, respectively, had a self-reported NYHA-class I to IV at baseline. At follow-up, 64 (29%) had undergone a CABG, 71 (33%) a PTCA, and 82 (38%) were being treated with pharmaco-therapy.
Additional sample characteristics are presented in Table 5.1

Table 5.1 *Sociodemographic Characteristics*

		N (%)
Marital status	Married	170 (78)
	Cohabiting	15 (7)
	Partner, not cohabiting	2 (1)
	Unmarried	6 (3)
	Divorced	7 (3)
	Widow/Widower	16 (7)
Living situation	Alone	28 (13)
	With others	185 (85)
Education[1]	Grade 6	44 (20)
	Technical School (grades 7-9)	61 (28)
	Junior High School (grades 7-9)	34 (16)
	Junior High School incl. vocational education	33 (15)
	High School/A-levels	6 (3)
	College (4 yr.)	22 (10)
	University (5 yr.+)	7 (3)
Employment status	Employed	57 (26**)**
	Unemployed	147 (68)

1. These categories are used by the Dutch National Institute for Statistics (CBS) to classify education level

5.5 ANALYSIS

5.5.1. Item-analysis

Every questionnaire item was linked to a global question addressing the same health status aspect, and for 23 items the change scores were standardized and broken down by the item-related global question rating. Thus, regardless of an item's domain we calculated 4,798 response combinations out of a total of 4,991 (217 x 23), representing missing data of less than 4%. In table 5.2 we show the relationship between the researcher's judgement of magnitude of change and that of the patient, for every repeatedly measured item (except those such as 'being restricted by costs of healthcare' which were not suitable to ask for improvement after treatment). The stratified SRM (for every global question rating, the mean change score was divided by the standard deviation of the observed change) does not differ significantly from the ES (for every global question rating, the mean change score was divided by the pooled standard deviation of baseline and post-test scores). The magnitude of change estimated with both SRM and ES (interpreted according to Cohen) is not in concordance with the interval determined by the rating 'a great deal better' but, considering the calculated value, represents a trivial deviation. When the effect sizes have values in concordance with the external criterion, the calculated value shows a tendency towards the interval's upper limit. Although it seems that Cohen's thresholds of magnitude of change over time appear to confirm the patient's judgement of the extent of improvement, this approach has a certain weakness since we analysed item response combinations, while clinicians are concerned with estimated magnitude of treatment effects in patients.

Table 5.2: *Estimation of magnitude of change on items with the Standardized Response Mean and Effect Size index, broken down by corresponding values of item-related external criterion of perceived magnitude of change.*

Global question/ Corresponding External criterion	Effect size Interval	Number of response combinations	SRM	Within corresp. interval	Value	Effect size	Within corresp. interval	Value
No change	0 – 0.20	3314	0.15	yes	0.75	0.14	yes	0.70
A little better	0.20 – 0.50	513	0.47	yes	0.90	0.41	yes	0.70
Moderately better	0.50 – 0.80	521	0.72	yes	0.73	0.63	yes	0.43
A great deal better	0.80 – max.	450	0.77	no	0.01	0.79	no	0.01

5.5.2. Scale analysis

To evaluate the concordance between the magnitude of change in domains of health-related functional status and an external criterion, we used the standardized change scores of scales and a single global question intended to correspond with the repeated measures of physical and emotional functioning. Tables' 5.3a to 5.3c present mean scores across global ratings of perceived change in functioning. Mean scores increase as the rating of global perceived magnitude of change increases, confirming the outcome of other studies [22,23,36,37]. Similarly, within each of the four categories of degree of improvement as perceived by the patient, the SRM reflect systematically more change than the ES, whereas the differences between these indices are very small with regard to the 3-item scale of physical functioning, and the domain of emotional functioning. Both effect size indices, as estimates of the magnitude of prospective change, indicate that the 3-item scale was less responsive than the MLHF-Q physical functioning scale, regardless of the perceived improvement rating (tables 5.3a and 5.3b). Overall effect sizes in the sample (see: *total* in tables 5.3a and 5.3b) indicated the same difference between the 3-item scale from a generic instrument (MOS-20)('small': SRM = 0.44 en ES = 0.42) and disease-specific scales ('moderate': SRM = 0.59 en ES =0.58), a consistent result in other studies [38-41]. Change in the emotional functioning domain seemed less relevant for this group of patients, given that 79% declared no change after treatment. Furthermore, the overall effect sizes of this scale are, according to Cohen, small.

The magnitude of change estimated with effect sizes ES (mean change/SDpooled) was, according to our rule of thumb, in keeping with the magnitude of change indicated by the patient's judgement, or their category of subjective meaning, for all scales. Furthermore, in cases that the magnitude of change estimated with the SRM (mean change/SDchange) was not confirmed by the external criterion ratings (Tables 5.3a and 5.3c), the discordance was trivial.

Table 5.3 a *Stratified effect sizes of change over time in the disease-specific physical functioning dimension (8 items)*

Global question/ External criterion	Corresponding Effect size interval	N	Mean change score	SRM	Within corresp. interval	Value	ES	Within corresp. interval	Value
No change	0 – 0.20	71	1.99	0.25	n	0.05	0.20	y	1.0
A little better	0.20 – 0.50	35	4.18	0.52	n	0.03	0.43	y	0.77
Moderately better	0.50 – 0.80	44	6.96	0.87	n	0.15	0.71	y	0.70
A great deal better	0.80 – max. (1.26)	45	7.11	0.89	y	0.20	0.72	n	- 0.04
Total		195	4.60	0.59			0.58		

Table 5.3 b *Stratified effect sizes of change over time in the physical functioning dimension (3 items)*

Global question/ External criterion	Corresponding Effect size interval	N	Mean change score	SRM	Within corresp. interval	Value	ES	Within corresp. interval	Value
No change	0 – 0.20	70	0.53	0.11	y	0.55	0.11	y	0.55
A little better	0.20 – 0.50	34	1.17	0.26	y	0.20	0.25	y	0.17
Moderately better	0.50 – 0.80	42	3.00	0.67	y	0.57	0.64	y	0.47
A great deal better	0.80 – max.(1.26)	45	3.82	0.85	y	0.11	0.81	y	0.02
Total		191	1.80	0.44			0.42		

Table 5.3 c *Stratified effect sizes of change over time in the disease-specific emotional functioning dimension (5 items)*

Global question/ External criterion	Corresponding Effect size interval	N	Mean change score	SRM	Within corresp. interval	Value	ES	Within corresp. interval	Value
No change	0 – 0.20	159	1.13	0.21	n	0.01	0.20	y	1.00
A little better	0.20 – 0.50	17	2.71	0.53	n	0.04	0.48	y	0.93
Moderately better	0.50 – 0.80	11	3.16	0.59	y	0.30	0.56	y	0.20
A great deal better	0.80 – max. (1.26)	15	6.80	1.26	y	1.00	1.21	y	1.00
total		202	1.68	0.35			0.32		

5.6 DISCUSSION

The values used to classify effect sizes for difference between means as small, medium, or large "was arbitrary but seamed reasonable", as Cohen [27] stated when he stressed that investigators should render their own judgement on the matter. Many researchers evaluating change in health-related quality of life measures, using effect size as an indicator of an instrument's responsiveness or to estimate magnitude of change over time, seem to have adopted Cohen's thresholds with the same rigidity that ' $\alpha = .05$' has been adopted. [42,43] Some researchers pose that an effect size is synonymous with the importance of change over time, without questioning who determines what should be considered trivial or important whether modified by terms such as 'minimal' or not [37] and the dependence of effect size interpretation on the perspective of the interpreter. [12]

To signify the importance of change in this study, we used an anchor-based approach, [44] asking patients to judge treatment-related magnitude of change over time by responding to global or 'transition questions' intended to be analogous with the instruments' domains of health.

The amount of change estimated with the ES in this study showed, in contrast to the SRM, the highest concordance between the researcher's interpretation according to Cohen and the patient's rating of perceived change, used as external criterion. Additionally, neither of the two estimates of amount of change, i.e. ES and SRM, deviated from the interval determined by the rating from the external criterion with regard to the 3-item physical functioning scale. Our data suggest that the difference in the number of items in the disease specific (5 items) and generic (3 items) physical functioning scale may be related to the extent to which the magnitude of change over time judgements are consistent with the external criterion. Given this study's design, this could not be verified. Despite the fact that the SRM was less concordant with the interval associated with perceptible change in the same domain, as rated by the patient, in comparison with the ES, we can conclude that Cohen's thresholds 'small', 'medium', and 'large' appear to be in keeping with the external criterion. It is possible that the terminology used in the global rating of change in physical functioning covered the content of the 3-item scale from a generic instrument more precisely than the content of the disease-specific items in this study sample.

In our approach we have, in order to achieve a conveniently arranged and comprehensible questionnaire, abandoned external criteria with 15 anchor points. Although it is known that raw change scores derived from repeated measurement of health-related quality of life increase with the amount of change as perceived by the patient, this phenomenon is no guarantee that effect sizes will fall in Cohen's range as determined by the external criterion. Osoba et al. [37] used ES (mean change/SD

baseline) in comparison with the same rating scale method as this study does. They concluded: "Cohen's estimates appear to be confirmed empirically in our direct study of the degree of change experienced by women who received chemotherapy for breast cancer." We applied our approach to their data and concluded that the magnitude of change in the domains of emotional functioning, social functioning, and global functioning were consistent with the external criterion. In contrast, two underestimates of effect were found in physical functioning determined by the ratings 'moderately better' and 'a great deal better'.

In our attempt to confirm Cohen's estimates empirically with data from studies with a 15-point global rating scale, [23] the only replicated result relates to the phenomenon that means of change scores consistently increase with the perceived magnitude of change. However, this concordance between mean raw scores and the external criterion was demonstrated only after merging seven global ratings into three. For example: 1)'almost the same', 'hardly any better at all', 2)'a little better' and 3) 'somewhat better' represent small improvement; 4)"a good deal better", 5) "moderately better" represent moderate improvement and 6)"a great deal" and 7) "a very great deal better" a large improvement. Reducing the original rating scale could cause a fallacy in the comparison of effect size and external criterion. The distances between the merged ratings probably differ from those between the anchor points on a seven-point rating scale, which can lead to differences in the relation with magnitude of standardized change assessed over time.

Myriad effect size indices have been developed, from which the researcher can choose, and no universal criteria exist to interpret this statistic [45,46]. In this study the criterion validity of the interpretation by Cohen's thresholds was evaluated, and results were compared with other studies. Future research is needed to clarify effect size interpretation, using other effect size indices or methods to assess functional status, e.g. weighed or unweighed scores in patient specific health status measures. [41,47-49]

REFERENCES

1. Stockler MR, Osoba D, Goodwin P, Corey P, Tannock IF. Responsiveness to change in health-related quality of life in a randomized clinical trial: A comparison of the Prostate Cancer Specific Quality Of Life Instrument (PROSQOLI) with analogous scales from the EORTC QLQ-C30 and a Trial Specific Module. J Clin Epidemiol 1998;51(2):137-45.
2. Murawski MM, Miederhoff PA. On the generalizability of statistical expressions of health related quality of life instrument responsiveness: a data synthesis. Quality of Life Research 1998;7:11-22.
3. Sneeuw KCA, Aaronson NK, Sprangers MAG, Detmar SB, Wever LDV, Schornagel JH. Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. J Clin Epidemiol 1998;51(7):617-31.
4. Taylor R, Kirby B, Burdon D, Caves R. The assessment of recovery in patients after myocardial infarction using three generic quality-of-life measures. J Cardiopulmonary Rehabil 1998;18:139-44.
5. Wiebe S, Rose K, Derry P, McLachlan R. Outcome assessment in epilepsy: comparative responsiveness of quality of life and psychosocial instruments. Epilepsia 1997;38(4):430-8.
6. Russel MGVM, Pastoor CJ, Brandon S, Rijken J, Engels LGJB, Van der Heijde DMFM, and et al. Validation of the dutch translation of the Inflammatory Bowel Disease Questionnaire (IBDQ): A health related quality of life questionnaire in inflammatory bowel disease. Digestion 1997;58282-8.
7. Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. J.Clinical Epidemiology 1991;44(4/5):417-21.
8. Parkerson GR, Willke RJ, Hays RD. An International Comparison of the reliability and responsiveness of the Duke Health Profile for measuring health-related quality of life of patients treated with Alprostadil for erectile dysfunction. Medical Care 1999;37(1):56-67.
9. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal Chron.Dis. 1987;40(2):171-8.
10. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control.Clin.Trials. 1991;12(4 Suppl):142S-58S.
11. Katz JN, Gelberman RH, Wright EA, Lew RA , Liang MH. Responsiveness of Self-Reported and Objective Measures of Disease Severity in Carpal Tunnel Syndrome. Medical Care 1994;32(11):1127-33.
12. Van der Windt DAWM, Van der Heijden GJMG, De Winter AF, Koes BW, Deville W, Bouter LM. The responsiveness of the Shoulder Disability Questionnaire. Ann Rheum Dis 1998;57:82-7.
13. 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.

14. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an anology to diagnostic test performance. J.Chronic Disease 1986;39 11):897-906.
15. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. The Journal of Rheumatology 1993;20(3):429-36.
16. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J.Clin.Epidemiol. 1995;48(11):1369-78.
17. Katz JN, Punnett L, Simmons BP, Fossel AH, Keller RB. Workers' Compensation Recipients with Carpal Tunnel Syndrome: The Validity of Self-Reported Health Measures. American Journal of Public Health 1996;86(1):52-6.
18. Norman G. Issues in the use of change scores in randomized trials. J.Clin.Epidemiology 1989;42(11):1097-105.
19. Deyo RA, Patrick DL. The significance of treatment effects: The clinical perspective. Medical Care 1995;33(4):AS286-AS291
20. Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Hanna B. The MACTAR patient preference disability questionnaire- An individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. Journal of Rheumatology 1987;14(3):446-51.
21. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the Rehabilitation Activities Profile and the Barthel Index. Journal of Cliical Epidemiology 1996;49(1):39-44.
22. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimally clinically important difference. Controlled Clinical Trials 1989;10:407-15.
23. Juniper EF, Gyuatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. Journal of Clinical Epidemiology 1994;47(1):81-7.
24. Wyrich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Medical Care 1999;37(5):469-78.
25. Wyrich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J.Clin.Epidemiol. 1999;52(9):861-73.
26. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50(8):869-79.
27. Cohen J. Statistical power analysis for the behavioural sciences. revised edition ed. New York: Academic Press; 1977.
28. Cohen J. A Power Primer. Psychological Bulletin 1992;112(1):155-9.
29. Lipsey MW. Design sensitivity. Statistical power for experimental research. SAGE Publications,London.; 1990.
30. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Medical Care 1990;28(7):632-42.

31. Guyatt GH. Measurement of health-related quality of life in heart failure. JACC 1993;22(4 (Supplement A)):185A-91A.
32. Guyatt GH. Measurement of health-related quality of life in heart failure. Special Issue: Heart disease: The psychological challenge. The Irish Journal of Psychology 1994;15(1):148-63.
33. Rector TS, Tschumperlin LK, Kubo SH, Bank AJ, Francis GS, McDonald KM, Keeler CA, Silver MA. Use of the Living With Heart Failure questionnaire to ascertain patients' perspectives on improvement in quality of life versus risk of drug-induced death. J.Card.Fail. 1995;1(3):201-6.
34. Rector TS, Cohn JN. Assessment of patient outcome with the Minnesota Living with heart Failure questionnaire: Reliability and validity during a randomized, double blind, placebo-controlled trial of pimobendan. American Heart Journal 1992;October,124(4):1017-25.
35. Stewart AL, Hays RD, Ware JE. The MOS Short-form General Health Survey: Reliability and validity in a patient population. Medical Care 1988;26:724-35.
36. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: Reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1996;50(1):79-93.
37. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. Journal of Clinical Oncology 1998;16(1):139-44.
38. Bessette L, Sangha O, Kuntz KM, Keller RB, Lew RA, Fossel AH, Katz JN. Comparative responsiveness of generic versus disease-specific and weighted versus unweighted health status measures in carpal tunnel syndrome. Medical Care 1998;36(4):491-502.
39. Stadnyk K, Calder J, Rockwood K. Testing the measurement properties of the Short Form-36 Health Survey in a frail elderly population. J Clin Epidemiol 1998;51(10):827-35.
40. Vaile JH, Mathers M, Ramos-Remus C, Russel AS. Generic health instruments do not comprehensively capture patient perceived improvements in patients with carpal tunnel syndrome. The Journal of Rheumatology 1999;26(5):1163-6.
41. Wright JG and Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50(3):239-46.
42. Thompson B. If statistical significance tests are broken/misused, what practices should supplement or replace them? Theory & Psychology 1999;9(2):165-81.
43. Cohen J. The earth is round ($p<.05$). American Psychologist 1994;49(12):997-1003.
44. Lydick E, Epstein RS. Interpretation of quality of life changes. Quality of Life Research 1993;2:221-6.
45. Kraemer HC. Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. Psychoneuroendocrinology 1992;17(6):527-36.
46. Sechrest L, Yeaton WH. Magnitudes of experimental effects in social science research. Evaluation Review 1982;6(5):579-600.
47. Wright JG, Young NL. The patient-specific index:Asking patients what they want. The Journal of Bone and Joint Surgery 1997;79-A(7):974-83.
48. MacKenzie RC, Charlson ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. Arch Intern Med 1986;146:1325-9.

49. Browne JP, O'Boyle CA, McGee HM, McDonald NJ, Joyce CRB. Development of a direct weighting proced ure for quality of life domains. Quality of Life Research 1997;6:301-9.