Assessment of change in clinical evaluation

Middel, Lambertus Johannes

7

Conclusions and discussion

7.1 INTRODUCTION

When clinicians are interested in patient-assessed health as a relevant outcome of treatment, the choices that need to be made after consulting clinical epidemiologists or social scientists, often put them in a state of indecision. Many choices can be made on the basis of concepts belonging to the universe of 'quality of life' that is the most appropriate one in the context of the disease and treatment. Subsequently, after selection of the appropriate measures, it becomes apparent how complex and varied the methods used to assess change are. Regardless of this, choices still have to be made from the various methods to come to a valid interpretation of that change in terms of clinical relevance or clinical importance.
In order to study the assessment of treatment-related change in health-related functional status, we have used various methods to estimate the magnitude of change over time and evaluated criteria for what constitutes relevance from the patient's perspective. The first part of the discussion will focus on some problems of effect size interpretation. The other part of the discussion is focussed on the reliability and validity of the direct mode of assessing change in health-related functional status using so-called global or transition questions.

7.2 MAIN RESULTS AND CONSEQUENCES FOR METHODOLOGY OF ASSESSING CHANGE IN HRFS

In Chapter 3 we established that a Dutch version of a new HRFS instrument (the "Minnesota Living with Heart Failure-Questionnaire") measures what it purports to measure and is reliable and able to detect change over time; we decided to use it in the methodological part of the thesis. The MLHF-Q was used in a group of patients undergoing treatments with known physiological efficacy, but no evidence was available about improvement in HRFS. It was expected that change in HRFS would occur, and that this would be moderate. We evaluated two 'yardsticks' for the interpretation of change in scores for the researcher and the patient (Chapter 5). In the first place, we discovered that Cohen's yardstick for the interpretation of change magnitude is used inaccurately by researchers for a variety of effect size indices or so-called responsiveness measures (Chapter 4). Many researchers do not know how to interpret the magnitude of difference between mean scores in health-related functional status. We demonstrated that overestimation and underestimation of effect appeared in 20 to 50 percent of the estimated standardised response means

(SRM' s). Several researchers who use (for example) the standard deviation of the baseline score, or the standard deviation of the change in score as the denominator, refer to Cohen who merely used the pooled standard deviation to express mean differences in standard deviation units. Secondly, we showed that the patient yardstick (the external criterion), i.e. the judgement of what constitutes 'trivial', 'small', 'moderate' or 'large' change appeared to be in keeping with Cohen's thresholds for 'trivial', 'small', 'moderate' or 'large' change over time. To our knowledge, no longitudinal studies have been performed in which the concurrent or convergent validity of health state transition questionnaires was investigated. Moreover, the responsiveness of multi-item transition scales has rarely been assessed. In order to investigate the psychometric properties of the retrospective 'transition questionnaire', we used the MLHF-Q in two modes: repeated measured HRFS before and after treatment to assess change serially, and in a retrospective mode, with MLHF-Q items modified in transition questions. We showed that direct assessment of the amount of change measured by the instrument's items belonging to a dimension of HRFS is comparable with the corresponding dimension of serially assessed change measured by the same items.

7.2.1 Statistically significant but how to interpret the magnitude of change?

The problem of testing change over time with null-hypothesis goes together with the dilemma that with large samples, trivial change may be statistically significant. There are many approaches towards estimating the relevance of change scores in health status outcome measures, but even the apparent simplicity of standardising mean differences, may bring inaccurate estimation of effect size according to Cohen's thresholds. The choice of standard deviations of baseline scores and change scores (from stable subjects) etc. (Table 1.1, Chapter 1) in the denominator is in conflict with the thresholds Cohen provided for the interpretation of this index. In this thesis, the studies addressing the clinical efficacy of intrathecal baclofen infusion and the psychometric properties of the MLHF-Q (Chapters 2 and 3) used the SD of change scores in the denominator (following others uncritically in applying Cohen's rule of thumb for effect size interpretation). If we take Cohen's original work (1) as being valid, we will have to rectify interpretations of the meaning of the estimated magnitude according to the results from these analyses. In both studies, 40 Standardised Response Mean indices were interpreted according to Cohen's thresholds for pooled estimates of standard deviation (ESp) out of which 20 turned out to be overestimation of treatment-related effect. (Table 7.1). In another study Chapter 4) we analysed this problem using results from other researchers. This

secondary analysis of data from other studies revealed that 20% of the estimated effect sizes did not fall in the same magnitude of change category according to the Cohen's thresholds.

Table 7.1 *Comparison of forty Standardised Response Means calibrated into Cohen's pooled effect size index (ESp) from Chapters 2 and 3 in this thesis.*

	Effect Size (ESp)			
	Trivial	Small	Moderate	Large
SRM	0 - <.20	≥.20 - <.50	≥.50 - <.80	≥ .80
Trivial	2			
Small	3	4		
Moderate		9	8	
Large			8	6

Thus, the SRM interpretation of effect magnitude with the thresholds Cohen with the ESp calculated on the same data (transformation of the same mean change over time into units of pooled standard deviation may result in dramatic differences (50% of the SRM indices are overestimated). Unfortunately, we still have no algorithm for effect size indices calculated with the standard deviation from baseline scores or from change scores in stable subjects according to an external criterion. Furthermore, even in a situation where we are able to reliably interpret effect size, we cannot differentiate between a 'large' and 'very large' effect since the cut-off point for large has a theoretical range from ES > .80 to infinite. However, Hopkins' [2] Likert-scale approach is able to give meaning to the extension of the scale to the level above large for Cohen's effect size statistic: ES = 0 - < .20 trivial effect; ES = ≥ .20 - < .60 small effect; ES = ≥.60 < 1.20 moderate effect; ES ≥1.20 - < 2.0 large effect; ES ≥ 2.0 -

< 4.0 very large and ES ≥ 4.0 - ∞ is considered to be 'nearly perfect'. In addition to thresholds for effect magnitudes, Hopkins elaborated Cohen's thresholds for correlation coefficients, relative risks and odds ratio. Despite this promising attempt to proceed with a more complete scale of effect magnitude, further research will need to provide empirical evidence for the external validity of this new rule of thumb for effect size interpretation irrespective of health status measure and research designs. Ever since Jacob Cohen wrote his well-known book [1], the effect size has been a problematic parameter in clinical evaluation, and several promising alternatives (for example, the "Reliable Change Index"), have been developed [3], improved and criticised [4-8]. In future studies statistical computer programmes may be able to give the researcher additional information on some treatment effect indices (notwithstanding the fact that no consensus exists on a method for signifying the magnitude of change within and between experimental and control groups that is meaningful in particular treatment contexts). Nevertheless, implementing effect sizes standard in the representation of statistical results may require researchers to change long-held patterns of behaviour.

7.2.2. Concordance between the researcher's interpretation of effect size and the patient's perception.

With global rating scales, respondents describe their state of health by answering just one question. An example of this would be 'would you describe your health as very good, good, fair, poor or very poor? Global, single-item measures of perceived overall health have been shown to be reliable and valid [9-11] and able to predict both change in functional status and mortality [12,13]. Furthermore, some studies have shown that the correlation between single global judgements of health with multiple-item dimensions (scales) of health status is not perfect [9,14,15] while other results indicate fair and good relationships [16-19]. One of the research questions in this study was to determine the concordance between the patient's perceived magnitude of change in a domain of health-related functional status (the external criterion), and the magnitude of change as estimated by the researcher using effect size estimates. The patient's perception of magnitude of change was assessed at item level and at scale or domain level. At item level, perceived magnitude was assessed with the instrument's items transformed into transition items. At scale level, for each domain (physical and emotional function) a single global question was put that covered the content of change in the corresponding domain. Change in these domains of HRFS was assessed with a repeatedly measured multi-item scale. Assessing the meaningfulness of changes in longitudinally assessed HRFS scores might have been hampered by the weak reliability and validity of single global questions that measure

the transitional state of health. However, we showed that at item level as well as at scale level, the external criterion appeared to be in keeping with Cohen's thresholds for 'trivial', 'small, 'moderate' and 'large' effects. Furthermore we compared our results with data from Osoba et al. [16] (Chapter 5) who used an identical transition scale for the external criterion but a different effect size index (mean change scores divided by the standard deviation of baseline scores). The concordance between longitudinal effect magnitude and the transition ratings of "moderately better" and "very much better" in the physical functioning domain was not perfect (see Table 7.2).

Table 7.2 *Stratified effect sizes ($\bar{X}1 - \bar{X}2$/Sdbaseline) of change over time in domains of health-related functional status*

			physical functioning			emotional functioning			social functioning			global functioning	
Corresponding Effect size interval		ES	within corresp. Interval	value	ES	within corresp. interval	value	ES	within corresp. interval	value	ES	within corresp. interval	value
No change	0 – 0.20	0.09	y	0.45	0.35	n	0.16	0.07	y	0.35	0.06	y	0.30
A little better	0.20 – 0.50	0.09	n	-0.08	0.43	y	0.77	0.22	y	0.07	0.51	n	0.02
Moderately better	0.50 – 0.80	0.16	n	-0.21	0.84	n	0.13	0.26	n	-0.15	0.73	y	0.77
A great deal better	0.80 – max (1.11)	0.38	n	-0.22	1.11	y	1.00	0.81	y	0.03	0.86	y	0.19

Source: Osoba et al. [16]

Different effect size indices may yield different outcomes. In addition, varying numbers of global ratings of a transition question makes comparison with results from other studies inconsistent and weak. The different distances between ratings and the necessity of collapsing or merging an anchor point to allow comparison can lead to differences in the relationship to the magnitude of standardised change over time. Another threat of concordance between external criterion and amount of change over time is that the composite of aspects belonging to (for example) the instrument's domain of physical functioning does not correspond with the set of aspects in the patient's mind by when he or she is asked " has there been any change in your physical problems"?

7.2.3 Reliability and convergent validity of transition scales

Eliciting the direction and magnitude of change in evaluative studies by directly asking "how have you been feeling since the bypass operation?", (as clinicians frequently do when they see patients after treatment) has both been criticised as well as considered to be a reliable and valid approach in evaluation of treatment. One confounding factor that may affect the reliability and validity of direct transition questions is known as 'recall bias'. It is assumed that because of this recall bias effect, patients are not considered able to make accurate and reliable estimates of their health status, either before treatment or at another point in the history of their illness. [20]

Acting on Coughlin's conclusions, [21] we minimised the recall bias by taking the shortest possible time span between the first questionnaire and follow up to reduce errors in recollection. We also selected interventions such as PTCA or CABG for this part of the thesis since the significance, vividness and meaningfulness of these events contribute to the accuracy of recall. The second source of error that may occur is the present health status influencing the patient's perception of the direction and magnitude of treatment-related change over time. [20,22,23] By choosing treatment with a known efficacy in a study aimed at comparing repeated measurement of HRFS with transition questions at follow-up, this 'present state bias' was assumed not to be a significant confounding factor. Correlation between present state questions and concordant transition questions seem 'logical' in a sample of patients who underwent treatment with known efficacy. Consequently, it was expected that a perceived improvement in, for example, 'climbing stairs' should correlate with no limitations in climbing stairs after PTCA or CABG when these treatments are aimed at improving the physical condition of climbing stairs at baseline.

After applying the method of Asseltine et al, [23] it was concluded that there were no differences in responsiveness (the Standardised Response Mean) between longitudinal change scores and transition scores in the domains of emotional and physical function. The SRM of the MLHF-Q physical function scale was 0.56 for change scores and 0.53 for transition scale scores whereas the SRM s of the emotional function scale were 0.31 and 0.30 respectively. It was hypothesised that if a distinction between invasive and non-invasive treatment and between improvement and stability in angina pectoris were made, differences in magnitude of effect would be found. Invasive PTCA/CABG treatments were expected to produce more change whereas non-invasive treatment was expected to produce very little change in HRFS over time. Strikingly, the physical scale's SRM s in the invasive treatment groups ranged between .73 and .78 (improved group .82 and .89 respectively) and in the non-invasive group, the SRM s ranged from. 29 to .12 (stable group .35 and .28). Although smaller in magnitude, the SRM indices of the emotional functioning scale showed similar results. These outcomes will be published after this thesis [24].

Table 7.3 *Responsiveness (SRM) of the different measuring methods for groups of patients*

	Treatment		Angina Pectoris [a]		
Measure	Invasive (N=135)	Non inv. (N=82)	Improved (N=87)	Stable (N=121)	Overall (N=217)
Physical scale					
SCS[b]	. 73 (. 09)[d]	. 29 (. 10)	. 82 (.11)	. 35 (. 09)	. 56 (. 07)
URS[c]	. 78 (. 09)	. 12 (. 09)	. 89 (.11)	. 28 (. 08)	. 53 (. 07)
Emotional scale					
SCS[b]	. 39 (. 09)	. 17 (. 11)	. 48 (. 09)	. 14 (. 09)	. 31 (. 07)
URS[c]	. 36 (. 08)	. 18 (. 12)	. 51 (. 11)	. 13 (. 09)	. 30 (. 07)

[a] NYHA classification; [b] Serial change scores; [c] Unweighted retrospective scores; [d] Values between brackets are standard errors.

When improved and stable groups were broken down by type of treatment the responsiveness indices (SRM s) of the improved CABG/PTCA and stable patients ranged from .95 to 1,00 and from .48 to 54, respectively. SRM s of improved and stable patients treated with pharmaceuticals ranged from. 44 to. 50 and from. 04 to. 18, respectively.

7.3 RECOMMENDATIONS FOR PRACTICE AND RESEARCH

So long as no consensus reached on standards for evaluating, using and interpreting effect size estimates of treatment related change in clinical research, there is an important need to develop uniform and widely accepted criteria to give meaning to the size of an effect. This lack of precision is not only relevant when evaluating treatment-related change within and between groups, but, even more important in the estimation of power in the planning phase of a trial. Standardisation of effect size interpretation needs reference ranges of health-related functional status assessed with population surveys. Furthermore, longitudinal research is needed to discriminate between changes in HRFS over time in a sample drawn from the general population, with change in a sub-sample of chronically ill patients. In other words, with knowledge about a reference range of an indicator of health-related functional status in the general population, we can recognise that there are differences. Furthermore, with a longitudinally assessed estimate of autonomous change in the same sample, we will be able to better understand the meaningfulness of treatment-related effects. In studies on the measurement of health-related quality of life and HRFS, effect sizes (ES) have been used as surrogates for clinically relevant change when change over time in outcome was substantial. However, ES do not provide a complete understanding of the meaningfulness of the observed change. Patients have to perceive a change in the performance of daily activities in order to rate the direction and degree of change; moreover, even when this perceived change is small in magnitude, it may still be perceived as a significant one by the patient. According to Osoba, [19] the significance of change as perceived by the subject 'should be of paramount consideration' in future attempts to define the meaningfulness of change in HRFS or health-related quality of life. The development of multi-item transition measures may cover change in the relevant underlying domain more representatively. Therefore, we suggest that measures that assess more concrete aspects of the patient's HRFS will provide greater accordance between serial and transition measures of change.

However, when a patient rates a reduction in (for example) difficulty in climbing stairs, as 'large', it does not necessarily imply that a patient will view this subjectively significant change as being important. Future areas of research aimed at quantification of meaningful change in HRFS should also include the importance patients assign to that change, even if it is experienced as being small. One piece of research has produced examples that seem promising extensions of transition questions. In this approach, the respondent rates the direction and the degree of perceived change by a assigning a value that has meaning to the respondent for the

experienced change, as well as by rating the degree of importance the respondent assigns to perceived change. In evaluation of treatment-related change in clinical trials, the importance assigned to the small improvement in one item of a domain of HRFS may outweigh a moderate deterioration in another item belonging to the same domain.

Finally, the following are key issues in the debate on methods for estimating clinically important change: Significance of treatment effects: significance to whom [25] who is to say what is important? [26] and "ask patients what they want "[27-29] have increasingly become apparent. To give clinically relevant meaning to change scores gained on two different points in time using HRFS instruments, several investigators suggest that the current approaches could be improved by taking more explicit account of patients' perceptions and expectations. A new paradigm is incorporating individual patient perspectives, expectations and preferences with respect to the effects of (innovative) treatments in the outcome measures. With scoring systems based on individualised measures such as the so-called Goal Attainment Scale (GAS) or Patient Specific Index (PCI), each patient essentially receives his or her 'own instrument' and these instruments seem to show an improved sensitivity to change in health-related functional status when compared with conventional methods. [30,30-34,34-37]

This thesis is aimed at supporting clinicians, health professionals, investigators and administrators in the understanding and critical evaluation of the psychometric properties of health status measures and methods in estimating and interpreting change in patient-assessed health outcomes. Health professionals increasingly stress that in the realisation of effective care and expected outcome of planned change in the process of care delivery, patients' preferences are essential sources of information. The operationalisation of the patient's perception of the severity of limitation in domains of health-related functioning, or operationalisation of individual preference or weighted relevance of items of health-related functional status measures is still in its infancy. However, for health administrators and decision-makers, investigation into the validity of patient-specific HRFS instruments used to evaluate the outcomes of innovative treatment and care, standardisation of methods is required. HRFS instruments cannot be used in the evaluation of treatment and care without a valid way of ascertaining what change in measured difference scores means.

References

1. Cohen J. Statistical power analysis for the behavioural sciences. revised edition. New York: Academic Press; 1977.
2. Hopkins WG. A new view of statistics: Effect Magnitudes.1997;http://sportsci.org/resource/stats/effectmag.html:
3. Jacobson NS, Truax P. Clinical Significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology 1991;59:12-9.
4. Speer DC. Clinically significant change: Jacobson and Truax (1991) revisited [published erratum appears in J Consult Clin Psychol 1993 Feb;61(1):27]. J.Consult.Clin.Psychol. 1992;60(3):402-8.
5. Hageman WJJM, Arrindell WA. A further refinement of the reliable change (RC) index by improving the pre-post difference score: introducing RC_{ID}. Behav.Res.Ther. 1993;31(7):693-700.
6. Hageman WJ, Arrindell WA. Establishing clinically significant change: increment of precision and the distinction between individual and group level of analysis [see comments]. Behav.Res.Ther. 1999;37(12):1169-93.
7. Hageman WJ, Arrindell WA. Clinically significant and practical! Enhancing precision does make a difference. Reply to McGlinchey and Jacobson, Hsu, and Speer. Behav.Res.Ther. 1999;37:1219-33.
8. Maassen GH. Kelley's formula as a basis for the assessment of reliable change. Psychometrika 2000;65(2):187-97.
9. Kempen GIJM, Miedema I, van den Bos GAM, Ormel J. Relationship of domain-specific measures of health to perceived overall health among older subjects. J Clin Epidemiol 1998;51(1):11-8.
10. Cunny KA, Perri M. Single-item vs. multiple-item measures of health-related quality of life. Psychol Rep 1991;69:127-30.
11. Gough IR, Furnival CM, Schilder W, Grove W. Assessment of the quality of life of patients with advanced cancer. European Journal of Clinical Oncology 1983;19:1161-5.
12. Idler EL, Kasl SV. Self-ratings of health: Do they also predict change in functional ability? Journal of Gerontol Soc Sci 1995;1995(50):S344-S353
13. Idler EL, Kasl SV. Health perceptions and survival: Do global evaluations of health status really predict mortality? Journal of Gerontol Soc Sci 1991;46:S55-S65
14. Kempen GIJM. The MOS Short-Form General Health Survey: single item vs. multiple measures of health-related quality of life; some nuances. Psychol Rep 1992;70:608-10.
15. Ziebland S. Jenkinson C, editors.Measuring health and medical outcomes. London: UCL Press; 1999;Measuring changes in health status.
16. 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.
17. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. Transition questions to assess outcome in rheumatoid arthritis. British Journal of Rheumatology 1993;32:807-11.

18. Bjorner JB, Kristensen TS. Multi-item scales for measuring global self-rated health. Investigating of construct validity using structural equation models. Research On Aging 1999;21(3,May):417-39.
19. Osoba D. Interpreting the meaningfulness of change in health-related quality of life scores: lessons from studies in adults. Int.J.Cancer 1999;12:132-7.
20. Streiner DL; Norman GR. Health measurement scales. A practical guide to their development and use. second edition. Oxford: Oxford University Press; 1995.
21. Coughlin SS. Recall bias in epidemiologic studies. J.Cinical Epidemiology 1990;43(1):87-91.
22. 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.
23. Aseltine RH, Carlson KJ, Fowler FJ, Barry MJ. Comparing prospective and retrospective measures of treatment outcomes. Medical Care 1995;33(uppl.):AS67-AS76
24. Goudriaan H, Middel B, Van Duijn MAJ, Stewart RE, van den Heuvel WJA. Responsivesness of serially assessed change after treatment in Health Related Functional Status compared with identical retrospectively assessed transition scales, weigthed transition scales and patient speciftc indices. submitted for publication 2001;
25. Mitchell PH. The significance of treatment effects: significance to whom? Medical Care 1995;33(4):AS280-AS285
26. Lachs MS. The more things change... Journal of Clinical Epidemiology 1993;46(10):1091-2.
27. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis and Rheumatism 1985;28(5):542-7.
28. 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.
29. Wright JG, Rudicel S, Feinstein AR. Ask Patients what they want. Evaluation of individual complaints before total hip replacement. J Bone Joint Surg 1994;76-B(2):229-34.
30. 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.
31. Tugwell P, Bombardier C, Buchanan WW, Goldsmith C, Grace E, Bennett KJ, Williams HJ, Egger M, Alarcon GS, Guttadauria M, et al. Methotrexate in rheumatoid arthritis. Impact on quality of life assessed by traditional standard-item and individualized patient preference health status questionnaires. Arch Intern Med 1990;150:59-62.
32. 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.

33. MacKenzie RC, Charlson ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. Arch Intern Med 1986;146:1325-9.
34. Rockwood K, Stolee P, Fox RA. Use of goal attainment scaling in measuring clinically important change in the frail elderly [see comments]. J.Clin.Epidemiol. 1993;46(10):1113-8.
35. Rockwood K, Joyce B, Stolee P. Use of goal attainment scaling in measuring clinically important change in cognitive rehabilitation patients. J Clin Epidemiol 1997;50(5):581-8.
36. Gordon JE, Powell C, Rockwood K. Goal attainment scaling as a measure of clinically important change in nursing-home patients. Age and Ageing 1999;28:275-81.
Gordon J, Rockwood K, Powell C. Assessing patients' views of clinical changes [letter]. JAMA 2000;283(14):1824-5.