Assessment of change in clinical evaluation

Middel, Lambertus Johannes

1.

Introduction

1.1 GENERAL INTRODUCTION

Chronic diseases such as rheumatism, spasticity and asthma are irreversible: clinicians and other health professionals can only minimise their patients' symptoms and improve their ability to function in day-to-day life. Physiologic measures are used to assess the severity of the disease. These objectives or laboratory tests can also be used as indicators of the course of the disease in the context of the treatment. In cardiology, for example, clinical measures such as left ventricular ejection fraction (LVEF), rate pressure product (i.e. heart rate × blood pressure), VO2 Max. and so on provide tools for classifying the severity of heart disease, and are also used in the assessment of improvement or deterioration in what these tools measure. A major disadvantage of these cardiac measures is that they do not necessarily reflect the patient's well being, health-related quality of life, or the ability to carry out his or her normal activities[1].

Although extending survival with minimal impairment is the primary goal of treatment, there is a growing recognition that the treatment should address other important goals as well, since for some chronic diseases, improvement of health-related quality of life (HRQL) or health-related functional status (HRFS) may be more important. In clinical studies, however, quality of life outcomes have turned out to be a 'kaleidoscopic' concept since no consensus exists with regard to the meaning of the concept in either the research community or the clinical community. Furthermore, the operationalization of the concept of (health-related) quality of life is heavily dependent on the disciplinary perspective in outcome assessment. This lack of consensus has given rise to the development of a myriad of measures involving different components whose conceptual dimensions vary. [2] .Therefore, instruments labelled as quality of life measures "may appear as health status, physical functioning, emotional functioning, perceived health status, symptoms, mood, need satisfaction, well being, and, often, several of these at the same time". [3] During the last 10 to 15 years, there has been an exponential increase in the development and use of instruments to measure the outcomes of medical interventions from the patient's perspective. A family of more than 150 instruments were identified in 75 studies; [4] in 1996, Spilker et al. catalogued nearly 215 measures in their second edition of "Quality of Life and Pharmacoeconomics in Clinical Trials" [5]. Since there is no consensus on the theoretical construct of quality of life, [3,6-9] the universe of domains belonging to this concept (and therefore the ongoing discussion on the selection of items by which it is operationalized), we prefer concepts such as health-related functional status. Functional status reflects the ability to perform the tasks of daily life in physical, emotional and social domains. There is also a growing agreement on

the components of these constructs and the validity of their measurement; for example, by validating these self-report measures with evidence-based measures. [10-12] By using the term health-related functional status (HRFS) in this thesis, we implicitly assume that a change in health status or functioning is indirectly related to the patient's subjective experience of quality of life.

For clinicians or other health professionals who feel the need to measure HRFS as an outcome in clinical trials, it is essential to know that the choice of available health status instruments is related to the methodological debate on the psychometric properties of instruments (in contrast to clinical outcomes such as physiologic measures). Consequently, this choice is also associated with methodological issues relating to the interpretation of outcome in terms of treatment-related change over time or differences between treated and control groups. Because improving patients' functional status has become a central therapeutic goal for many diseases, it is important that both clinicians and researchers develop a common understanding of 1) what HRFS concepts mean; 2) which measure is likely to be the most appropriate one in the context of the disease and aim of the study; 3) the methods to assess treatment-related change, and 4) the methods by which a valid interpretation of the magnitude of that change in terms of clinical relevance or clinical importance can be achieved.

1.2 OBJECTIVES AND MAIN RESEARCH QUESTIONS OF THIS THESIS

Health status measures have become an important part of clinical research in the evaluation of treatment efficacy. Furthermore, there is a need to assess treatment efficacy with evidence-based HRFS measures. When new instruments (e.g. the 'Minnesota Living with Heart Failure Questionnaire') are presented to the clinical and scientific community, [13] *reliability* and *validity* are traditionally the most important features of the instruments that are evaluated. An instrument is reliable if it gives the same result on repeated assessments of stable subjects whose circumstances have not changed (test-retest reliabilty), and when the test yields more or less the same results when administered by different observers (interobserver reliability). The validity of a measure refers to whether that measure does indeed measure the conceptually defined property (for example, perceived physical health). In testing the validity of a new physiological measure, there is often a golden standard or criterion measure available for comparison. In contrast with physiological measures, there is no gold standard for a functional status instrument against which to measure its validity. Therefore, the validity of physical health status can be investigated by a number of

different procedures by which the same construct is assessed (for example: self-report vs. performance-based tests); when a similar result occurs, this is called *concurrent* or *convergent* validity. When the reliability and validity of health-related functioning measures have been established, these psychometric properties of new and more appropriate tests are generally accepted conditions for use of these measures in clinical settings and research. However, the appropriateness of the instrument designed to measure change in persons over time is not only determined by its reliability and validity. Measuring change in order to evaluate treatment efficacy requires the instrument to be sensitive to detecting change when patients improve in physical function after (for example) a coronary artery bypass surgery (CABS). Over the last 15 years, this property has become well known through the widely used concept of *responsiveness*. Responsiveness of health status measures has become one of the 'holy trinity' of necessary psychometric properties of health status instruments. To quantify responsiveness, several effect sizes are used as estimates of the amount of change detected with an instrument. In this respect, the most accurate approach is to ask the patient if the researcher is interested in understanding the patient's perception of the direction and amount of change in a domain of health-related functional status. This is common daily practice for clinicians. One of the aims of this thesis is to address some methodological issues relating to the assessment of change in health-related functional status and the meaning of the magnitude of assessed change in scores. Traditionally, the many generations of researchers who have evaluated the efficacy of medical interventions, base their decisions on the statistical significance of the within-group treatment-related change over time or any statistically significant difference in change from repeated measurements between experimental and control groups. In some cases, investigators eager for results are likely to detect a statistically significant (but very small) change in scores related to the intervention, simply due to large sample size. Consequently, even if change which is statistically significant, though trivial in magnitude, is detected, the $p < 0.05$ doctrine unwittingly pushes the question of how meaningful, important, relevant, or substantial the changer is into the background. Significance tests support the decision as to whether the change is due to chance fluctuation or can be functionally related to treatment. The observed statistical significance does not indicate the magnitude of change. In spite of this, some researchers implicitly suggest that smaller p-values represent larger, and thus more 'relevant', effects. [14]

Against this background, the objectives of this thesis can be formulated in terms of the following research questions:
How to determine the main psychometric properties of a new, disease-specific, health status measure?

How comparable are different operationalizations of effect sizes (ES) when outcome is interpreted as 'trivial' (ES < .20), 'small' (ES ≥ .20 < .50), 'moderate' (ES ≥ .50 < .80), or 'large' (ES ≥ .80) according to the well-known thresholds of Cohen [15]?
How concordant are the effect sizes, labelled by the researcher as 'trivial', 'small', 'moderate', or 'large' change in a domain of health-related function with the patient's perception of change in the same domain signified with the same qualitative terms?
How reliable and valid are multi-item scales of perceived change after treatment at follow-up as compared to longitudinal (before-after) assessments with scales comprised of identical items?

Besides this first chapter, this thesis consists of six other chapters. The main theme of this thesis deals with methodological problems in the assessment of treatment-related change in health-related functional status (HRFS). There is a large number of factors that potentially affect the interpretation of change in HRFS by the researcher and the perception of the direction and magnitude of change by patients who have undergone a particular medical intervention. Change over time in HRFS measures was assessed in patients with severe spinal spasticity and in patients whose symptoms were considered to belong to what is labelled 'heart failure'. Both groups underwent a treatment with known efficacy in order to detect treatment-related change. This thesis addresses the research questions stated above as follows:
in Chapter 2, the efficacy of a new treatment in health status is evaluated in a randomised clinical trial design. The analysis is representative of the 'classical' model of testing the null-hypothesis 'that differences are due to chance fluctuations'. Besides statistically significant p values, supplementary effect size indices are reported in order to indicate the relevance of change but no external criterion was used to decide what constitutes this relevance.

To assess change, an HRFS-instrument as a baseline measure must meet the criteria of *reliability* (in stable groups, it yields the same score each time) and *content or construct validity* (it reflects what it is supposed to measure), but when applied as a repeatedly assessed baseline measure, the additional most important and necessary property is the instrument's *responsiveness* (sensitivity to detect change over time). In Chapter 3, these psychometric properties (research question 1) are evaluated with the 'Minnesota Living with Heart Failure Questionnaire' (MLHF-Q) in a sample of patients who underwent treatment with known efficacy (DC electrical cardioversion).

In the evaluation of treatment efficacy, one of the most important properties of HRFS measures is its ability to detect change that is related to the treatment (and not

to regression to the mean, due to errors of measurement). This ability is well known as *responsiveness*, and is quantified by a variety of measures of effect magnitude. Cohen provided guidelines for interpreting the magnitude of his first effect size **d'** that was explicitly labelled as such, and expressed the size of treatment effect in units of the common population standard deviation estimated with the sample's pooled standard deviation. These guidelines are used for several indices called effect size, but the size of treatment effect is expressed in units of the sample standard deviation of either the baseline score or the change score. In Chapter 4, the risk of overestimation or underestimation of effect magnitude is evaluated for two comparable effect size indices (research question 2).

Change can be assessed prospectively with longitudinal assessments and retrospectively with global questions relating to perceived change. To validate the prospectively assessed change in HRFS, single global questions are used as an external criterion for interpreting those change scores valued as being 'important' by the patient (research question 3). In Chapter 5, a comparison is made between the intervals relating to the thresholds of Cohen of what constitutes small, moderate, or large longitudinal effects and the patients' judgement of what is perceived as small, moderate, or large improvement after treatment.

Patients' perception of the direction and magnitude of change in domains of health was assessed with single-item (global) scales, as well as with multiple items scales on perceived change derived from the original items from the Minnesota Living with Heart Failure Questionnaire. In Chapter 6, the concurrent or convergent validity is evaluated by comparison of the dimensions of HRFS in the repeatedly assessed baseline measure and the global, retrospective measure (research question 4). The 'known groups validity' is evaluated by comparison of both instruments between groups who improved or remained the same in angina pectoris.
The last chapter (Chapter 7) summarizes the results, conclusions, and implications for further research and development of the methodology for measuring change in health-related functional status.

Summarising, in addition to reliability and validity, assessment of treatment-related change in HRFS is highly determined by the so-called 'third measurement property' of *responsiveness*. There is neither consensus on its 'theoretical' definition nor on its operationalization, i.e. the operations needed to quantify this property. The remainder of this chapter relates to terms and definitions of responsiveness, and consequently, to the corresponding methods of assessing it. Since no golden standard or reference range is available for indices of responsiveness, we address the patient's

perspective of the HRFS measures to get a better understanding of what a change in specific patient groups means.

1.3 TERMS AND DEFINITIONS

1.3.1 Responsiveness, a problematic construct

To give greater meaning to the interpretation of the amount of change in scores on health-related functional status instruments, the concept of *responsiveness* was introduced in publications. For clinical purposes, the usefulness of a HRFS-instrument depends on its ability to detect a change that is clinically meaningful. Clinically meaningful refers to a change that justifies alteration in management of the disease or to a change that indicates the efficacy of an innovative type of treatment in domains of health status. Responsive measures discriminate between trivial and substantial changes within clinical trial groups and consequently show the difference in change between those groups. Thus, the term *responsiveness* is used as an indicator of the instrument's sensitivity to change, as well as an indicator of the magnitude of treatment-related change over time. The term responsiveness is however, a confusing one for the beginner who encounters it in the literature, since papers addressing treatment-related change in health-related functional status may refer to a varying composite of aspects. As appears from a selection of scientific papers, the term *responsiveness* is used as an operational definition of: 1) an indicator of the sensitivity of an instrument to detect change over time [16-21] or even refer to the extent to which a measure is sensitive to *real* change [22]; 2) 'statistically significant change in an experimental group in which change should be present' [23]; 3) an indicator of the magnitude of treatment-related change [19-21,24-28]; and 4) a measure of clinically relevant change in health [29,30], although some investigators prefer the term 'clinically *significant* change' [31,32]. Qualitative terms such as 'clinically important' need at least a golden standard. As mentioned before, such a standard is not available for health-related functional status. The blinded observation of a clinician can be used as an external criterion for justifying the interpretation in terms of clinically relevant or important change in HRFS. Another external criterion or yardstick for the interpretation of changes in HRFS is the patient's perception of the importance of change after (for example) a specific treatment.

Husted *et al.* [33] distinguished internal responsiveness from external responsiveness by defining internal responsiveness as the ability of a measure to detect change over time, whereas external responsiveness was defined as the extent to which change in a measure relates to corresponding change in a reference measure. [12,34,35] Despite this

clarification of the concept of responsiveness by this recently published classification, the assessment of change in HRFS over time in clinical research is quantified using a variety of approaches. For the sake of clarity, we will therefore, in this thesis use the concepts in the following meaning:

- *responsiveness*: the psychometric property of a measurement instrument, namely its sensitivity to detect difference between two points in time (change over time) within groups;
- *meaningful or relevant difference*: the amount of change in scores or the magnitude of change within and between groups, according to statistical or other quantitative criteria (e.g. effect size indices);
- *clinically* relevant or *clinically* important change in scores on a health-related functional status measure (always linked to an external criterion of relevance).

The purpose of a study and its study design may require different psychometric properties of the outcome measure. Consequently, the measure must either have the property of being able to detect differences *between* subjects at a single point in time (discriminative instruments) i.e. the ability to differentiate between groups 'who have a better HRFS and those who have a worse HRFS '. [25,36,37] Other studies may require the instrument's ability to detect change over time *within* subjects (evaluative instruments). [38-40] Consequently, in randomised clinical trials (RCT), health-related functional status instruments should have both properties, namely: 1. the ability to reliably estimate change between baseline and post-test within an experimental and a control group, and 2. the ability to estimate the difference in change over time by comparing the average change assessed in treated and in non-treated subjects in order to determine treatment effect, since in general, subjects in the treatment group are expected to change (on the average) more than those in the control group do.

1.3.2 Responsiveness and the instrument's scope: generic verses. specific

An important criterion for choosing an instrument in order to detect change in health-related functional status is its generic or disease-specific scope, which will depend on the objectives of the specific study. Generic health status measures seek a broad perspective that is not specifically related to the restricted scope of the health-related functional status of the aspecific disease. Therefore, generic measures allow investigators to compare health status across different diseases. [41] Generic measures are health-related to the extent that disease, injury, treatment, or policy [42] influences them. Disease-specific measures focus on the disease being studied, allowing greater sensitivity to treatment-related change compared to generic measures. The responsiveness of a health status instrument is an important issue in

the decision to use disease-specific or generic measures of health-related functional state. For example, for those cases in which therapeutic effects are likely to be modest and undramatic,[18,43] a better sensitivity to change over time of an instrument is a necessary condition. It would seem that 'cardio specific' measures (for example) may be more appropriate to detect change in HRFS.[44] Although the question of whether instruments, that are tailored to the disease, are superior to measures of general function in terms of sensitivity to change, has not been settled definitely, a growing number of studies indicate that disease-specific measures seem to be more responsive than generic measures.[45-52] To evaluate a disease-specific instrument's concurrent responsiveness, the amount of change in scores of both instruments (often generic versus disease-specific measures) is assessed in relative terms under identical conditions. To standardise the comparison of alternative instruments, Relative Efficiency (RE) statistics are sometimes used. RE statistics are emphasised as a comparative measure of responsiveness. RE expresses the change score as the squared ratio of either t-scores from paired t-tests or the z-scores from the Mann-Whitney-Wilcoxon ranked pairs test that compares the assessed instrument to a standard. [16,48,50,53-59] Another method of standardising comparisons between generic and disease-specific measures is known as the Relative Validity (RV) coefficient. [29,60-63] This statistic is calculated for each pair of measures in the comparison and is defined as the ratio of their F-statistics (F-statistic for each measure is estimated by comparing change scores across groups that improved, stayed the same, or deteriorated). The RV coefficient indicates how much more or less valid each outcome measure is relative to the best outcome measure.

1.3.3 Effect size as responsiveness measure

Mean differences in outcomes of a test can be standardised to quantify an intervention's effect in units of standard deviation (SD). Consequently, standardising mean change over time with a standard deviation allows comparison of a particular intervention's different outcomes, independent of the measuring units. The resulting statistical measure is known as effect size index.

The effect size tells us something very different from the **p**-value, which indicates the obtained probability of a Type I error in a test of statistical significance. If a **p**-value is annotated as statistically significant, rejecting the null-hypothesis does not imply that the effect was important in any way nor does a non-significant **p**-value indicate a clinically trivial result. [64-67] Criticism of statistical hypothesis testing has a long history, [68] and even Jacob Cohen [14,69] played a prominent role in the anti-hypothesis-testing charge.[70] The adoption of a fixed level of significance may lead to the situation in which two researchers obtain identical treatment effects but obtain

different **p**-values (0.04 and 0.06) due to the effect of (slightly) different sample sizes leading to different decisions. Thus, **p**-values are confounded by the joint influence of sample size and the effect size [71] and make the rejection of the null-hypothesis not very informative. Another criticism of null hypothesis testing is 'that it is foolish to ask: 'Are the effects of A and B different?' "They are always different- in some decimal place- for any A and B". [72] Since then, quantitative investigators in medical and social sciences have proposed a variety of supplementary effect size indices, some of which we will clarify. Reporting effect sizes without appropriate statistical tests and associated p values is misleading and potentially dangerous if the number of observations that is required to detect a difference has not been estimated by means of a power analysis. Effect size statistics should be provided to supplement statistical testing (not as a substitute for it), and only when the outcome is sufficiently extreme from what would have been expected on the basis of chance ($p < \alpha$).
It should be noted that during the debate on 'significance testing', several vocal leaders in psychology and education research called for the universal reporting and interpretation of empirically produced effect sizes. [73,74]
There are myriad estimates of effect size out of which the researcher can make a choice [75] and the question arises as to which of the effect size measures 'that could be summoned up for a given problem should a researcher report?' [70,71] The most elegant solution for this problem would seem to be for authors to include the sufficient statistics so that every reader can compute whichever effect size index they believe is best suited to the situation. Table 1.1 gives an overview of responsiveness measures in repeated measurement study designs.

Table 1.1 *Formulas for responsiveness measures for change over time (Within-group standardised mean change)*

Paired t statistic	$\frac{\bar{X}1 - \bar{X}2}{SE^{*}}$
Effect size (1)	$\frac{\bar{X}1 - \bar{X}2}{SD_{pooled}^{**}}$
Effect size (2)	$\frac{\bar{X}1 - \bar{X}2}{SD_{baseline\ scores}}$
Effect size (3)	$\frac{(\bar{X}1 - \bar{X}2)_{treated\ subjects} - (\bar{X}1 - \bar{X}2)_{controls}}{SD_{pooled\ baseline}}$
Standardised Response Mean (1)	$\frac{\bar{X}1 - \bar{X}2}{SD_{change\ scores}}$
Standardised Response Mean (2)	$\frac{\bar{X}1 - \bar{X}2_{(improved\ subjects)}}{SD_{change\ scores\ (improved\ subjects)}}$
Standardised Effect size	$\frac{\bar{X}1 - \bar{X}2_{(improved\ subjects)}}{SD_{baseline\ (improved\ subjects)}}$
Responsiveness index (1)	$\frac{M.C.I.D^{***}}{SD_{change\ scores\ (stable\ subjects)}}$
Responsiveness index (2)	$\frac{\bar{X}1 - \bar{X}2}{SD_{baseline\ (stable\ subjects)}}$
Responsiveness index (3)	$\frac{\bar{X}1 - \bar{X}2}{SD_{change\ scores\ (stable\ subjects)}}$
Responsiveness coefficient	$\frac{\sigma^2(\bar{X}1 - \bar{X}2)}{\sigma^2(\bar{X}1 - \bar{X}2) + \sigma^2_{error}}$
Normalized ratio	$\frac{\bar{X}1 - \bar{X}2\ (improved\ subjects)}{SD_{baseline\ (stable\ subjects)}}$
Relative efficiency statistic	$(t\text{ - statistic }_{measure\ 1} / t\text{ - statistic }_{measure\ 2})^2$
Relative efficacy index ****	$(ES_P / ES_{P\ best})^2 \times 100$

* SE = standard error of the difference

** where pooled $SD = \sqrt{\frac{(SD_{baseline})^2 + (SD_{outcome})^2}{2}}$ for : $N_{baseline} = N_{outcome}$

*** Minimal Clinically Important Difference according to external criterion (i.e. the difference in change score between those who perceived no change and those who perceived little change) which is considered to be the minimal difference in change over time that patient's perceive as meaningful

**** Magnitude of change over time is estimated for each scale by dividing the mean change by the pooled variance of change, according to Cohen {154} denoted as ES_P. This relative efficacy statistic is computed by squaring the ration obtained by dividing each scale ES_P (numerator) by the scale having the largest ES_P (denominator). This statistic is then expressed as a percentage with respect to the best measure.

1.3.4 Effect size: a problematic statistic

Among researchers, who are not conversant with this method of estimating the amount of change over time, have made various critical comments about Cohen's work. [15] These include:

1. there is no consensus on the 'theoretical' meaning, or the conceptualisation of the effect size as an outcome variable;
2. there is no consensus on the mathematical way to determine the magnitude of the difference between scores gained on two different occasions: researchers classify the extent of responsiveness and magnitude with effect sizes using several standard deviations (SD), including the baseline SD, the SD of change (Cohen's effect size index **d**) and so on by using for each of them the thresholds of Cohen's effect size index **d',** which refers to the pooled samples' standard deviation.

Sub 1. Regarding the use of the notion of effect size in HRFS research, several researchers have claimed that without an external criterion, the estimated amount of change measured by the effect size index can be denoted as *clinically* important change. [19,20,29,30,76] Other researchers assume that an effect size, estimated within a group of subjects, expresses the measure's ability to detect change over time (due to a therapeutic intervention) [16-21,29] without claiming that their effect size indicates that the instrument is sensitive or responsive to *clinically relevant* changes in patients' perceived health. When a HRFS instrument is used as an outcome measure, and the amount of change estimated with change scores (or quantified by an effect size) is defined as clinically relevant, the following question logically arises: 'What is meant by a clinically relevant change?' [77,78] Because patients and clinicians differ in the preferences or perceived relevance that they assign to particular aspects belonging to domains of health-related functional status, several authors have incorporated these perceptions or preferences into health status instruments [6,49,76,77,79-82] to give more significance to the term 'relevant'.

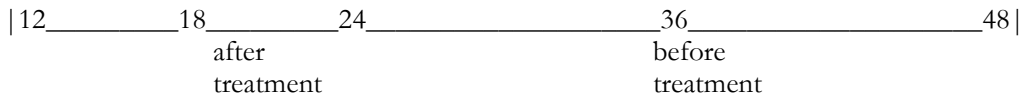
Sub 2. Many clinical studies have been conducted, that use different methods to estimate magnitude of change over time. These have indicated that there is no convincing evidence that either method offers any apparent advantages. [7,48] Any magnitude of change or responsiveness can be expressed by a 'd-index' estimate of magnitude of change; in other words, it measures the difference between two means in terms of their common standard deviation units. The literature shows that numerous quantitative indices belonging to the family of effect sizes (ES) [75] have been developed. However, there is no consensus on how to declare a difference in

terms of standard deviation units. The interpretation of the effect size is determined by the choice of the standard deviation used to standardise the mean change over time and, related to that, by the ready adoption of the interpretation guideline as set by Cohen. [15] Several effect size indices are used in quality of life research, which have in common that $\bar{X}_1 - \bar{X}_2$ is divided by a standard deviation. The researcher's decision as to which SD he will take is either a well-considered choice or one which is copied from well-reputed colleagues and has no further justification. However, in giving meaning to standardised mean change in terms of 'trivial', 'small', 'moderate', or 'large' effects using the thresholds that Cohen [15] provided us with some thirty years ago, it seems to have been forgotten that these cut-off points were calculated with the *pooled standard deviation*. Consequently, applying these thresholds for mean change scores standardised with the standard deviation of the change scores ($\bar{X}_{t-1} - \bar{X}_{t-2} / SD_{X1-X2}$) may lead to over- or underestimates of effects.
For his effect size (mean baseline scores minus mean follow-up scores, divided by the pooled standard deviation) Cohen came up with conventions for those values that constitute a 'trivial' (ES < .20), 'small' (ES ≥ 20 < .50), 'medium' (ES ≥ .50 < .80), and a 'large' effect (ES ≥ 80). However, for each of these effect size indices, these thresholds are used indiscriminately, which may have contributed to the confusion in this area. [33]

1.4 ESTIMATES OF CLINICALLY RELEVANT CHANGE

Ideally, to assess clinically relevant change, an external definition of what constitutes relevant is required. Clinicians, for instance, use reference values (reference range) for physiological 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 provide us with the opportunity to rate changes after treatment as being trivial, substantial, or clinically relevant in the expected direction. For example, for a reference range of normal values ranging from 12 to 24 units of measurement, an observation of 36 found before treatment (48 units is the maximum value this measure can acquire) would indicate the need for treatment. The seriousness of the deviation is 12 units from the upper limit of the reference range. When 18 units are measured after treatment, the amount of change in 18 units may be valued as clinically relevant, since this outcome is covered by the reference range (see Figure 1.1).

Figure 1.1 *Reference range*



In contrast, when health-related functional status 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 (for example) physical functioning in order to be able 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. As stated above, 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, often referred to as the external criterion. However, such an external criterion is not a 'golden standard' of health related functional status against which measurements could be calibrated. It is therefore not easy to place the magnitude of changes in health-related functional status in a context that is meaningful for researchers, clinicians, and other health professionals. Although the notion of clinically significant change, relevant change, or important change has been used to refer to clinicians' subjective judgements observed in patients undergoing treatment, it would appear that such a notion lacks an objective definition. A common method of interpretation of change magnitude in health status, and the one used in this thesis, is to compare the change in health status score with a global subjective judgement of the direction and amount of change by the patient. [30,39,83-94] This subjective judgement is obtained by asking the patient at the post-test stage about the extent to which deterioration or improvement has occurred since treatment. The measurement of retrospective perception of treatment-related change is known as a 'global question' or a transition question. Transition or global questions are used with verbal anchors varying from a dichotomous scale (e.g. improved vs. not improved [59,79]) to a 15-point scale ranging from -7 = 'a very great deal worse' to +7 = 'a very great deal better'. [91,92,95,96] 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 serial 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'. [91,92,97]
Numerous publications are devoted to the question of how the minimal clinically important change in scores with a repeated administered health status measure can be determined. [1,24,25,36,45,83,91,92,98-103] In the last decade, the concept of "minimal important change" has been quantified ambiguously. Some of the studies determine this minimal clinically important difference (MCID) from the perspective of clinicians. [104] Some of the studies relate serial change scores to global scales of perceived change after treatment to demonstrate that a change in score of 0.5 per item is the minimal clinically important change, if patients say 'I have improved (worsened) a little, or improved (worsened) somewhat'. Other studies advocate that any change of a patient's disease status should be considered 'minimally clinically relevant' if patients themselves think that they feel at least 'a little better'. [89,105] Consequently, the mean change in repeatedly measured scores will increase with the retrospective judgements of "I feel somewhat better", 'I feel a good deal better' and 'a very great deal better'. [32,84,91,92] Because of this, some studies use the mean difference between adjacent groups of those who experience no change and those who feel a little improved or a little worse as the best estimate of the minimal relevant change. [33,105] A weakness in this approach (although these verbal anchors can be used to estimate a relevant difference in an instrument's score over time) is that different distances between ordinal response categories will affect different estimates of the change score per item that constitutes minimal, moderate, or large change. [78] Varying distances on a global question or external criterion for what constitutes relevant change from the patient's perspective makes generalisability of outcome problematic.

1.4.1 Researcher's perspective versus patient perspective

In this thesis, the concordance between the patient's perception of the magnitude of change in domains of health-related functional status, the external criterion, and the magnitude of change estimated in terms of standardised mean change in scores over time is a major question (research question 3).
Change in scores on a health-related functioning scale is usually obtained by repeated baseline measurement. In order to discriminate between relevant and irrelevant change, so-called 'transition' or 'global questions' are used as the external criterion or standard: the patients are retrospectively asked how much they feel better or worse compared to the situation at baseline.
Consequently, we have two perspectives from which the direction and magnitude of change can be assessed, namely:

1. the *researcher's perspective* .Subtracting scores from repeated measurements using a health-related functional status instrument to determine change over time and interpreting the results in terms of statistical significance (**p**-value) or relevance (effect size);
2. the *patient's perspective*. If one is interested in understanding the patient's perception of change, direct transition questions are used to compare patient outcomes at one particular time (post-treatment) or over time. The patient gives a retrospective indication of his or her state of health before treatment, he/she compares it with the perceived present state of health after treatment and, by making a 'mental subtraction' of both states, signifies the extent of change (improved, unchanged, or deteriorated) on a global scale of transition.

These transition questions are put as retrospective questions after treatment and are aimed at determining the direction and magnitude of perceived change in general state of health or in domains of physical, emotional, and social health related functioning.

1.4.2 The patient's perspective: single global question

In some studies, HRFS items are used as a serial global rating to examine incremental perceived change between baseline and follow-up. [29,34-36,84,85,97,106-110]
Several studies discuss the accuracy, precision, reliability, and validity of *single* global ratings of health. [32,85,87,99,111-114] The main disadvantage of a single item of retrospectively perceived change in overall health is that the answer on a global rating scale indicated by "since the operation my state of health has worsened" does not cover domain-specific change in health. We can imagine that improvement in the domain of physical health is overshadowed by the perception of a worsening in emotional functioning. Therefore, domain-specific single transition questions are considered to be more valid indicators of perceived change in health status. [34,115,116]
Additionally, another disadvantage of a single question used to capture perceived change in specific domains of the patient's life (physical, emotional or social functioning) is that the internal consistency (reliability) cannot be estimated. Therefore, we have good reason to presume that multiple-item transitional scales tend to be more reliable than single-items [117]. Moreover, when the items of retrospective measures are conceptually identical with the repeatedly assessed items from the baseline measure, they will also have a better validity. [116]

1.4.3 Multi-item transition scales

As mentioned before, a common method of interpretation is to compare health status scores with a single, global transition judgement of the direction and amount of change made by the clinician or patient: this is often referred to as the external criterion. In clinical practice however, change after treatment is also assessed retrospectively by asking the patient to give an appraisal of the magnitude and the direction (improvement or deterioration) of change in health status or functioning. Given this practice, why not measure change directly (retrospectively) in evaluation studies of treatment efficacy?

In the interaction between clinician and patient, such a retrospective appraisal by the patient and physician on several clinically relevant components of health status has clinical relevance, as it determines the decisions made in the management of the disease. There is an ongoing debate about methods for estimating clinically relevant change. [34,112,118,119] In this debate, one of the assumptions is that changes inferred from repeated measurements approximate the change captured by the patient's retrospective perceptions of change over a period of time. [12,35] Other researchers have found that the retrospective recall of change in health status or symptoms is not as accurate as change found in pre-post designs because of the complexity of the question. When asked ' Have you got better or worse since your bypass operation?' patients firstly have to make a judgement of their 'present health state', then make a reconstruction of the situation before CABG, and then carry out a mental subtraction and come with an estimate of the direction and amount of change over time. This method has two weaknesses: the first is that when the time span is too large, people simply do not remember how they were before treatment or at the moment of their last visit at the clinic (the 'recall bias'). The second weakness is the correlation of the 'present state' with the retrospective estimate of change. [120] The retrospective assessment of treatment-related change may be invalid if patients feel prevented from living as they would like to by problems that are not related to the disease for which they are being treated. However, patients, who experience no limitation in their health-related functional status at follow-up, are likely to have been limited before treatment, and consequently they are likely to perceive improvement. Furthermore, if the time span is sufficiently large, we believe that retrospective recall is a very useful measurement if the measurement goal is to assess what the subject believes about the effect of treatment. Assessing change with single-transition judgements is a time-honoured approach, but there is a good reason to avoid single questions that are too global. With global transition items such as ' Have you got better or worse since your bypass operation?' the patient may refer only to a few symptoms which are manifest at that particular point in time; symptoms such as

'shortness of breath', 'pain in the chest', 'fatigue' etc.[115,116,121,122] Additionally, due to the relative coarseness of the single item compared with the multi-item scale, the single item is less well suited to detect minor differences in health perception which may still be clinically relevant. Multiple-item transition scales, on the other hand, enable patients to rate the extent to which they have changed on a number of disease-specific variables, thereby allowing for the possibility that not all aspects of functioning and health status will be given the same response. With the summed composite of transition items belonging to domains of HRFS, the constructed scale will yield more information reflecting meaningful change in the dimension than single items do. To the best of our knowledge, no studies have been detected in which a set of transition items is used to measure change in domains of health such as physical functioning, emotional functioning, and social functioning. The use of such small sets of multiple-item transition scales to measure change in domains of health provides an opportunity for an unequivocal representation of changes that are relevant for the patient. This method may also be considered in study designs where repeated measurement is not plausible, such as assessment of change after emergency referral to a hospital of patients who have had an acute heart attack.

REFERENCES

1. Croft P. Measuring up to shoulder pain. Ann Rheum Dis 1998; 57:65-66.
2. Testa MA, Nackley JF. Methods for quality-of-life Studies. Annu.Rev.Public Health 1994; 15:535-559.
3. Hunt SM. The problem of quality of life. Qual.Life Res. 1997; 205-212.
4. Gill TM, Feinstein AR. A critical appraisal of the quality of quality of life measurements. JAMA 1994; 619-626.
5. Spilker B. Quality of Life and Pharmacoeconomics in Clinical Trials. 2nd.Ed. Philadelphia: Lippincott-Raven, 1996.
6. Browne JP, McGee HM, O'Boyle CA. Conceptual approaches to the assessment of quality of life. Psychology and Health 1997; 12:737-751.
7. Bonomi AE, Patrick DL, Bushnell DM, Martin M. Quality of life measurement. Will we ever be satisfied? J Clin Epidemiol 2000; 53:19-23.
8. Anderson KL, Burckhardt CS. Conceptualization and measurement of quality of life as an outcome variable for health care intervention and research. Journal of Advanced Nursing 1999; 29:298-306.
9. Fitzpatrick R. A pragmatic defence of health status measures. Health Care Analysis 1996; 4:265-272.
10. Kempen GIJM, Steverink N, Ormel J, Deeg DJH. The assessment of ADL among frail elderly in an interview survey: Self-report versus performance-based tests and determinants of discrepancies. Journal of Gerontology:Psychological Sciences 1996; 51B:254-260.
11. Van Heuvelen MJG. Physical activity, physical fitness and disability in older persons.(Dissertation). Groningen: Rijksuniversiteit Groningen, 1999.
12. Emery CF, Blumenthal JA. Perceived change among participants in an exercise program for older adults. The Gerontologist 1990; 30:516-521.
13. Rector TS, Kubo SH, Cohn JN. Patients' self-assessment of their congestive heart failure.Part 2: Content, reliability and validity of a new measure, The Minnesota Living with Heart Failure Questionnaire. Heart Failure 1987; 3:198-209.
14. Cohen J. The earth is round ($p < .05$). American Psychologist 1994; 49:997-1003.
15. Cohen J. Statistical power analysis for the behavioural sciences. revised edition. New York: Academic Press, 1977.
16. 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; 51:137-145.
17. 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.
18. 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; 18:139-144.

19. Wiebe S, Rose K, Derry P, McLachlan R. Outcome assessment in epilepsy: comparative responsiveness of quality of life and psychosocial instruments. Epilepsia 1997; 38:430-438.
20. Russel MGVM, Pastoor CJ, Brandon S, Rijken J, Engels LGJB, Van der Heijde DMFM, 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; 58:282-288.
21. 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; 44:417-421.
22. 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; 37:56-67.
23. Wasserfallen JB, Gold K, Schulman KA, Baraniuk JN. Development and validation of a rhinoconjunctivitis and asthma symptom score for use as an outcome measure in clinical trials. J.Allergy Clin.Immunol. 1997; 100:16-22.
24. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal Chron.Dis. 1987; 40:171-178.
25. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control.Clin.Trials. 1991; 12:142S-158S.
26. 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; 32:1127-1133.
27. Middel B, Kuipers-Upmeijer H, Bouma J, Staal MJ, Oenema D, Postma Th, et al. Effect of intrathecal baclofen delivered by an implanted programmable pump on health related quality of life in patients with severe spasticity. J Neurol Neurosurg Psychiatry 1997; 63:204-209.
28. de Beurs E, van Balkom AJLM, Lange A, Koele P, van Dyck R. Treatment of Panic Disorder With Agoraphobia: Comparison of Fluvoxamine, Placebo, and Psychological Panic Management Combined With Exposure and of Exposure in Vivo Alone. American Journal of Psychiatry 1995; 152:683-691.
29. 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; 51:617-631.
30. 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; 57:82-87.
31. Bain BA, Dollaghan CA. Clinical Forum: Treatment efficacy. The notion of clinically significant change. Language, Speech, and Hearing in Schools 1991; 22:264-270.
32. 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; 16:139-144.
33. Husted JA, Cook RJ, Farewell VTGDD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000; 53:459-468.

34. Ziebland S. Measuring changes in health status. In: Jenkinson C, editor. Measuring health and medical outcomes. London: UCL Press, 1999:
35. Ziebland S, Fitzpatrick R, Jenkinson C, Mowat A, Mowat A. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the Health Assessment Questionnaire (HAQ) and modified HAQ. Annals of the Rheumatic Diseases 1992; 1202-1205.
36. 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; 48:1369-1378.
37. Vliet-Vlieland ThPM, Zwinderman AH, Breedveld FC, Hazes JMW. Measurement of morning stiffness in rheumatoid arthritis clinical trials. J Clin Epidemiol 1997; 50:757-763.
38. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J.Clin.Epidemiology 1992; 45:1341-1345.
39. Norman G. Issues in the use of change scores in randomized trials. J.Clin.Epidemiology 1989; 42:1097-1105.
40. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Annals of Internal Medicine 1993; 118:622-629.
41. Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, et al. Functional status and well-being of patients with chronic conditions: results from the medical outcome study. JAMA 1989; 262:907-913.
42. Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Medical Care 1989; 27S:S217-S232
43. Middel B, Bouma J, Crijns HJGM, De Jongste MJL, Van Sonderen FLP, Niemeijer MG, et al. The psychometric properties of the Minnesota Living with Heart Failure Questionnaire (MLHF-Q). Clinical Rehabilitation 2000; accepted for publication:
44. Hillers ThK, Guyatt GH, Oldridge N, Crowe J, Willan A, Griffith L, et al. Quality of life after myocardial infarction. Journal of Clinical Epidemiology 1994; 47:1287-1296.
45. Juniper EF. Measuring health-related quality of life in rhinitis. J.Allergy Clin.Immunol. 1997; 99:S742-9.
46. Hawker G, Melfi C, Paul J, Green R, Bombardier C. Comparison of a generic (SF-36) and a disease-specific (WOMAC) instrument in the measurement of outcomes after knee replacement surgery. J.Rheumatol. 1995; 22:1193-1196.
47. Gliklich RE, Hilinsky JM. Longitudinal sensitivity of generic and specific health measures in chronic sinusitis. Quality of Life Research 1995; 4:27-32.
48. Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997; 50:(3)239-246.
49. Bessette L, Sangha O, Kuntz KM, Keller RB, Lew RA, Fossel AH, et al. Comparative responsiveness of generic versus disease-specific and weighted versus unweighted health status measures in carpal tunnel syndrome. Medical Care 1998; 36:491-502.
50. 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; 51:827-835.
51. 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; 26:1163-1166.

52. Wells G, Boers M, Shea B, Tugwell P, Westhovens R, Saurez-Almazor M, et al. Sensitivity to change of generic quality of life instruments in patients with rheumatoid arthritis: preliminary findings in the generic health OMERACT Study. The Journal of Rheumatology 1999; 26:217-221.
53. Doeglas D, Krol B, Guillemin F, Suurmeijer Th, Sanderman R, Smedstad LM, et al. The Assessment of Functional Status in Rheumatoid Arthritis: A Cross Cultural, Longitudinal Comparison of the Health Assessment Questionnaire and the Groningen Activity Restriction Scale. The Journal of Rheumatology 1995; 22:1834-1843.
54. Gordon JE, Powell C, Rockwood K. Goal attainment scaling as a measure of clinically important change in nursing-home patients. Age and Ageing 1999; 28:275-281.
55. Hurny C, Bernhard J, Coates A, Peterson HF, Castiglione-Gertsch M, Gelber RD, et al. Responsiveness of a Single-Item Indicator Versus a Multi-Item Scale; Assessment of Emotional Well-Being in an International Adjuvant Breast Cancer Trial. Medical Care 1996; 34:234-248.
56. 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; 28:542-547.
57. MacKnight C, Rockwood K. A Hierarchical Assessment of Balance and Mobility. Age and Ageing 1995; 24:126-130.
58. Rockwood K, Joyce B, Stolee P. Use of goal attainment scaling in measuring clinically important change in cognitive rehabilitation patients. J Clin Epidemiol 1997; 50:581-588.
59. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the Rehabilitation Activities Profile and the Barthel Index. Journal of Cliical Epidemiology 1996; 49:39-44.
60. Roberts R, Hemingway H, Marmot M. Psychometric and clinical validity of the SF-36 General Health Survey in the Whitehall II study. British J of Health Psychology 1997; 285-300.
61. Vickrey BG, Hays RD, Genovese BJ, Myers LW, Ellison GW. Comparson of a generic to disease-targeted health-related quality of life measures for multiple sclerosis. J Clin Epidemiol 1997; 50:557-569.
62. Ware JE, Kemp JP, Buchner DA, Singer AE, Norman G. The responsiveness of disease-specific and generic health measures to changes in the severity of asthma among adults. Qual.Life Res. 1997; 7:235-244.
63. Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek AE. Comparison of methods for the scoring and statistical analysis of SF-36 health profiles and summary measures: summary of results from the Medical Outcome Study. Med Care 1995; 33(Suppl. 4):AS264-AS279.
64. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. American Psychologist 1989; 44:1276-1284.
65. Rosenthal R. Progress in clinical psychology: Is there any? Clinical Psychology: Science and Practice 1995; 2:133-150.
66. Rosenthal R, Rubin DB. The counternull value of an effect size: a new statistic. Psychological Science 1994; 5:329-334.

67. Bartko JJ, Pulver AE, Carpenter WT. The Power of Analysis: Statistical Perspectives. Part 2. Psychiatry Research 1988; 23:301-309.
68. Rozeboom WW. The fallacy of the null-hypothesis significance test. Psychological Bulletin 1960; 57:416-428.
69. Cohen J. Things I have learned (so far). American Psychogist 1992; 45:1304-1312.
70. Levin JR, Robinson DH. Further reflections on hypothesis testing and editorial policy for primary research journals. Educational Psychology Review 1999; 11:143-155.
71. Thompson B. If statistical significance tests are broken/misused, what practices should supplement or replace them? Theory & Psychology 1999; 9:165-181.
72. Tukey JW. The philosophy of multiple comparisons. Statistical Science 1991; 6:100-116.
73. Thompson B. Editorial policies regarding statistical significance tests: Futher comments. Educ.Res. 1997; 26:29-32.
74. Murphy KR. Editorial. Journal of Applied Psychology 1997; 82:3-5.
75. Kirk RE. Practical significance: A concept whose time has come. Educational and Psychological Measurement 1996; 56:746-759.
76. Naylor CD, Llewellyn-Thomas HA. Can there be a more patient-centered approach to determining clinically important effect sizes for randomized treatment trials? J.Clin.Epidemiology 1994; 47:787-795.
77. Lachs MS. The more things change... Journal of Clinical Epidemiology 1993; 46:1091-1092.
78. Wright JG. The minimal important difference: Who's to say what is important? J Clin Epidemiol 1996; 49:1221-1222.
79. 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; 14:446-451.
80. Mitchell PH. The significance of treatment effects: significance to whom? Medical Care 1995; 33:AS280-AS285
81. 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:229-234.
82. 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; 46:1113-1118.
83. Deyo RA, Patrick DL. The significance of treatment effects: The clinical perspective. Medical Care 1995; 33:AS286-AS291
84. 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; 50:79-93.
85. Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients; The floor phenomenon. Medical Care 1990; 28:1142-1152.
86. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an anology to diagnostic test performance. J.Chronic Disease 1986; 39:897-906.
87. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. Transition questions to assess outcome in rheumatoid arthritis. British Journal of Rheumatology 1993; 32:807-811.

88. Fitzpatrick R, Albrecht G. The plausibility of quality-of-life measures in different domains of health care. In: Nordenfelt L, editor. Concepts and measurements of quality of life in health care. Kluwer Academic Publishers, 1994:201-227.
89. Fortin PR, Stucki G, Katz JN. Measuring relevant change: an emerging challenge in rheumatologic clinical trials. Arthritis Rheum. 1995; 38:1027-1030.
90. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J.Clinical Epidemiology 1989; 42:403-408.
91. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimally clinically important difference. Controlled Clinical Trials 1989; 10:407-415.
92. 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; 47:81-87.
93. Lydick E, Epstein RS. Interpretation of quality of life changes. Quality of Life Research 1993; 2:221-226.
94. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of Change Scores in Ordinal Clinical Scales and Health Status Measures: The Whole May Not Equal the Sum of the Parts. Journal of Clinical Epidemiology 1996; 49:711-717.
95. 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; 37:469-478.
96. 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; 52:861-873.
97. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997; 50:869-879.
98. Guyatt GH, Townsend M., Pugsley SO, Keller JL, Short HD, Taylor DW, et al. Bronchodilators in chronic airflow limitation: Effects on airway function, exercise capacity and quality of life. American Rev Respir Disease 1987; 1069-1074.
99. Baker DW, Hays RD, Brook RH. Understanding changes in health status;Is the floor phenomenon merely the last step of the staircase? Medical Care 1997; 35:1-15.
100. Wells GA, Tugwell P, Kraag GR, Baker PRA, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: The patient's perspective. The Journal of Rheumatology 1993; 20:557-560.
101. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for Clinically Important Changes in Outcomes: Development, Scoring and Evaluation of Rheumatoid Arthritis Patient and Trial Profiles. The Journal of Rheumatology 1993; 20:561-565.
102. Mahajan P, Pearlman D, Okamoto L. The effect of fluticasone on functional status and sleep in children with asthma and on the quality of life of their parents. J Allergy Clin Immunol 1998; 102:19-23.
103. Juniper EF. Quality of life questionnaires: Does statistically significant = clinically important? J Allergy Clin Immunol 1998; 102:16-17.
104. Burback D, Molnar FJ, St-John P, Man-Son HM. Key methodological features of randomized controlled trials of Alzheimer's disease therapy. Minimal clinically

important difference, sample size and trial duration. Dement.Geriatr.Cogn.Disord. 1999; 10:534-540.
105. Eberle E, Ottillinger B. Clinically relevant change and clinically relevant diference in knee osteoarthritis. Osteoarthritis and Cartilage 1999; 7:502-503.
106. Guyatt GH, Eagle DJ, Sackett B, Willan A, Griffith L, McIlroy W, et al. Measuring quality of life in the frail elderly. J.Clin.Epidemiol. 1993; 46:1433-1444.
107. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the Minimal Important Difference in Symptoms: A Comparison of Two Techniques. Journal of Clinical Epidemiology 1996; 49:1215-1219.
108. Garratt AM, Ruta DA, Abdalla MI, Russell T. Responsiveness of the SF-36 and a condition-specific measure of health for patients with varicose veins. Quality of Life Research 1996; 223-234.
109. Deyo RA, Inui TS. Toward Clinical Applications of Health Status Measures: Sensitivity of Scales to Clinically Important Changes. Health Services Research 1984; 19:275-289.
110. MacKenzie RC, Charlson ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. Arch Intern Med 1986; 146:1325-1329.
111. MacKenzie RC, Charlson ME, DiGioia D, Kelley K. Can the Sickness Impact Profile measure change? An example of scale assessment. J Chron Dis 1986; 39:429-438.
112. Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. JAMA 1999; 282:1157-1163.
113. Manusco CA, Charlson ME. Does recollection error threaten the validity of cross-sectional studies of effectiveness? Medical Care 1995; 33:AS77-AS88
114. Doll HA, Black NA, Flood AB, McPherson K. Criterion validation of the Nottingham Health Profile: Patient views of surgery for benign prostatic hypertrophy. Soc.Sci.Med. 1993; 37:115-122.
115. 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; 70:608-610.
116. 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; 51:11-18.
117. Cunny KA, Perri M. Single-item vs. multiple-item measures of health-related quality of life. Psychol Rep 1991; 69:127-130.
118. Mahler DA, Weinberg DH, Wells CK, Feinstein AR. The measurement of Dyspnea. Contents,Interobserver agreement, and physiologic correlates of two new clinical indexes. Chest 1984; 85:751-758.
119. Osoba D. Interpreting the meaningfulness of change in health-related quality of life scores: lessons from studies in adults. Int.J.Cancer 1999; 12:132-137.
120. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Second edition. Oxford: Oxford University Press, 1995.
121. Read JL, Quin RJ, Hoefer MA. Measuring overall health: an evaluation of three important approaches. J Chron Dis 1987; 40:7S-19S.
122.Leavey R, Wilkin D. A comparison of two health survey measures of health status. Soc.Sci.Med. 1988; 27:269-275.