Assessment of change in clinical evaluation

Middel, L.J.
Assessment of change in clinical evaluation
Thesis University Groningen with summary in Dutch
ISBN 90 72156 94 3

Cover design: Mickey ©
Printed by Stichting Drukkerij C. Regenboog, Groningen

RIJKSUNIVERSITEIT GRONINGEN

ASSESSMENT OF CHANGE IN CLINICAL EVALUATION

Proefschrift

Ter verkrijging van het doctoraat in de Medische Wetenschappen
aan de Rijksuniversiteit Groningen op gezag van de
Rector Magnificus, dr. D.F.J. Bosscher,
in het openbaar te verdedigen op
woensdag 10 oktober 2001 om 16.00 uur

Door

Lambertus Johannes Middel
Geboren op 12 augustus 1946
te Groningen

.

Promotor:	Prof. Dr. W.J.A. van den Heuvel
Referent:	Dr. M.J.L. de Jongste
Promotiecommissie:	Prof. Dr. D. Post
	Prof. Dr. J.L. Peschar
	Prof. Dr. H.J.G.M. Crijns
Paranimfen:	Bert Hamersma
	Jan Just Middel

Damit das Mögliche entsteht,

Muss immer wieder das Unmögliche versucht werden

Hermann Hesse

Voor Aukje en Elke

..

Contents

..

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'.

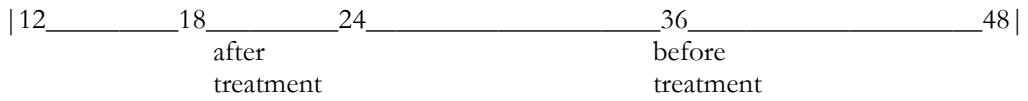
Sub 2. Many clinical studies have been conducted, that use different methods to estimate magnitude of change over time. These have indicated that there is no convincing evidence that either method offers any apparent advantages. [7,48] Any magnitude of change or responsiveness can be expressed by a 'd-index' 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.

2

Effect of intrathecal baclofen delivered by an implanted programmable pump on health related quality of life in patients with severe spasticity

Berrie Middel*, Hanna Kuipers-Upmeijer, Jelte Bouma*, Michiel Staal***, Dettie Oenema**; Theo Postma****, Sijmon Terpstra*****; Roy Stewart***

*Northern Centre for Health Care Research, Faculty of Medical Sciences, University of Groningen, **Departments of Neurology, ***Neurosurgery and *****Long-range Planning, University Hospital Groningen, **** Faculty of Economics, University of Groningen, the Netherlands.

Journal of Neurology, Neurosurgery and Psychiatry
Reprinted with permission of the publisher

ABSTRACT

Objectives
To compare clinical effectiveness and health related quality of life in patients with severe spasticity who received intrathecal baclofen or a placebo.
Methods
In a double-blind, randomised, multicentre trial 22 patients were followed up during 13 weeks and subsequently included in a 52 week observational longitudinal study. Patients were those with chronic, disabling spasticity who did not respond to maximum doses of oral baclofen, dantrolene, and tizanidine. After implantation of a programmable pump patients were randomly assigned to placebo or baclofen infusion for 13 weeks. After 13 weeks all patients received baclofen. Clinical efficacy was assessed by the Ashworth scale, spasm score, and self reported pain, and health related quality of life by the sickness impact profile (SIP) and the Hopkins symptom checklist (HSCL).
Results
At three months the scores of the placebo and baclofen group differed slightly for the spasm score (effect size = 0.20) and substantially for the Ashworth scale (effect size = 1.40) and pain score (effect size = 0.94); health related quality of life showed no significant differences. Three months after implantation the baclofen group showed a significant, substantial improvement on the SIP 'physical health', 'mental health', 'mobility' and 'sleep and rest' subscales and on the HSCL mental health scale; patients receiving placebo showed no change. After one year of baclofen treatment significant ($P < 0.05$) improvement was found on the SIP dimensions 'mobility' and 'body care and movement' with moderate effect sizes. Improvement on the SIP subscale 'physical health' ($P = < 0.05$; effect size 0.86), the SIP overall score (without 'ambulation'), and the 'physical health' and overall scale of the HSCL was also significant, with effect sizes > 0.80. Changes in health related behaviour were noted for 'sleep and rest' and 'recreation and pastimes' ($P = < 0.01$, $P = < 0.05$; effect size 0.95 and 0.63, respectively). Psychosocial behaviour showed no improvement.
Conclusions
Intrathecal baclofen delivered by an implanted, programmable pump resulted in improved self reported quality of life as assessed by the SIP, and HSCL physical health dimensions also suggest improvement.

Keywords: baclofen, health related quality of life, clinical outcomes

2.1 INTRODUCTION

Continuous intrathecal baclofen infusion via a subcutaneously implanted programmable pump has been used in the treatment of severe spasticity since 1984. Studies have evaluated neurological (Ashworth scale and spasm score), neurophysiological (EMG), urological (bladder function), and other clinically relevant outcomes, such as functional status activities of daily living.[1-12] Little attention has been paid, however, to health related quality of life, health status measures, and costs. This study addresses health related dimensions of quality of life as well as conventional outcome measures, including muscle tone (Ashworth scale) and frequency of spasms. Treatment outcomes were evaluated during one year after pump implantation to assess the long term effects of baclofen treatment, which is aimed at relieving symptoms and improving function. Because of the multiple causes of severe spasticity, no disease specific instruments were available and health related quality of life was assessed by generic measures covering a wide range of health status domains.

To our knowledge, this is the first time validated health status measures have been used in a randomised, controlled, clinical trial to evaluate the results of baclofen treatment. This paper presents the results of a first wave of 22 patients who were enrolled in a double blind, placebo controlled, clinical trial and randomly assigned to a placebo condition or effective drug (baclofen) treatment. Data collection of a second wave of patients, who received baclofen infusion immediately after implantation of the programmable pump, is in progress.

2.2 METHODS

2.2.1 Selection of Patients

Patients with severe spasticity caused by multiple sclerosis or spinal cord injury who had been referred by their general practitioner or specialist, were recruited from neurology, rehabilitation and neurosurgery departments of nine Dutch hospitals. Patients were included in the study when they met the following criteria: (1) aged 18 years or over, with chronic disabling spasticity of spinal origin inhibiting personal care, sitting, lying, and transfers, accompanied by pain and stiffness, or disturbed sleeping; (2) insufficient response to treatment with maximum doses of oral baclofen, dantrolene and tizanidine; (3) sufficient understanding of the consequences of the treatment. Patients were excluded when they were pregnant, had no neurological

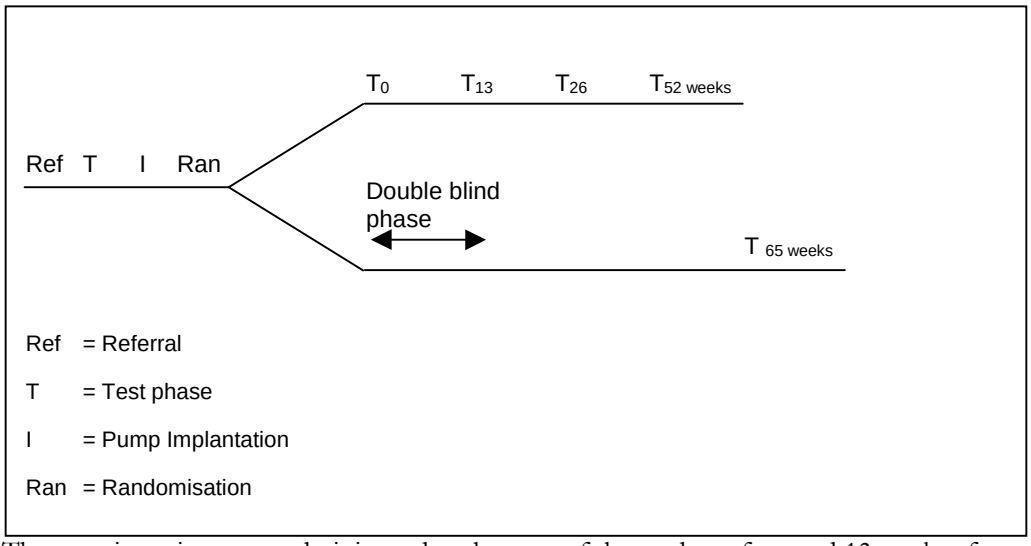
symptoms of supraspinal origin, or were allergic to baclofen.
After written consent was obtained, patients who fulfilled the inclusion criteria participated in a test phase to assess their responsiveness to baclofen. The maximum duration of the test phase was eight days. Every other day either baclofen or placebo was randomly administered by intrathecal bolus injections through a spinal catheter. Both doctor and patient were blinded during the test. Depending on the observed clinical effect consisting of improvement of at least 1 point on the Ashworth and spasm scales for eight hours, the test was repeated with an increased dose. All patients responded to one of the doses of baclofen (50, 75, 100, and 150 μg). At the start of the placebo controlled phase, patients were informed of the 50% chance of receiving a placebo for 13 weeks and of the possible risks and side effects of the treatment. Patients were aware that they could end their participation in the study and that this would not affect their care and treatment. All patients gave their consent in writing.
Using the Kolmogorov-Smirnov's test of normality, [13] we found that the normal distribution hypothesis had to be rejected for most of the variables used in the analysis. Therefore, we used the Wilcoxon matched pairs signed ranks test to estimate change scores between baseline and three months post-test. The difference in outcomes between the baclofen and placebo group at three months was analysed using the Wilcoxon-Mann-Whitney test for ordinal data. Effect sizes were calculated for the statistically significant results. According to Cohen, an effect size of 0.20 implies a small effect, 0.50 a medium effect, and > 0.80 a large effect. [14]
Due to lack of information on (clinical) indices from previous evaluations of patients with severe spasticity of spinal origin, we had no reliable figures to perform a power analysis and estimate the proper sample size.

2.2.2 Study Design and Treatment Assignment

A multicentre, randomised, double blind clinical trial was conducted to compare two groups of patients who were implanted with a programmable pump. During the first 13 weeks after implantation of a Synchro-Med programmable pump, the patients were randomly assigned to either baclofen (n = 12) or a placebo (n = 10). A balancing procedure was used to allocate the patients to the two conditions to achieve an equal distribution of patient characteristics with a potential effect on treatment outcomes over the two groups. [15] The balancing criteria were age, sex, and aetiology of spasticity.
Both patient and doctor were blinded during the first 13 weeks after implantation. In patients assigned to the baclofen condition the pump was telemetrically started after implantation. The initial pump velocity was based on the patient's response

during the test phase. If a patient's response had been satisfactory at 75 μg of baclofen, the initial day dosage was twice that dose (150 μg = 6.25 μg/h (150/24 = 6.25)). If response proved unsatisfactory, the velocity of the pump was increased by 10%. A maximum of two increases was made during the placebo-controlled phase. In patients assigned to the placebo condition, the same adjustment criteria were applied, but oral medication was maintained and at the end of the 13-week period the placebo was replaced by baclofen. Baclofen, placebo, and oral medication were supplied by the hospital pharmacist in a standard set of blank packages. The figure shows that the placebo-controlled phase was followed by a 52-week observational longitudinal follow up phase, which started as soon as the patient was put on continuous baclofen infusion. In patients receiving baclofen during the placebo controlled phase, the first phase coincided with the first 13 weeks of the second phase, that is, they were followed for a total of 52 weeks. In patients who were put on baclofen after 13 weeks of placebo, the two phases covered a period of 65 (13+52) weeks.

Figure 1.1 *Study design*



The questionnaires were administered at the start of the study, at four and 13 weeks after the start of the placebo-controlled phase, and at 26 and 52 weeks of the follow up phase.

The study received approval of the joint ethics committee of the Faculty of Medical Sciences, University of Groningen and University Hospital Groningen.

2.2.3 Measures

Ashworth scale and spasm score and self reported pain

The Ashworth scale and spasm score are clinical assessment scales for spasticity. To calculate the Ashworth score the grades for hip flexion/extension, hip abduction and adduction, knee flexion/extension and ankle dorsal flexion/extension on each side are summed and divided by eight. The modified Ashworth scale has 4 grades: grade 0 (no increase in tone), grade 1 (slight increase in tone, giving a "catch" when the affected part is moved in flexion or extension), grade 2 (more pronounced increase in tone, but affected part easily flexed), grade 3 (considerable increase in tone; passive movement difficult), and grade 4 (affected part rigid in flexion or extension). [16] The spasm score evaluates the frequency of spasms with scores: 0 (no spasm), 1 (mild spasms induced by stimulation), 2 (infrequent spasms occurring less than once per hour), 3 (spasms occurring more than once per hour), and 4 (spasms occurring more than 10 times per hour).

Pain was measured on a 10 point self-assessment scale with a sum score ranging from zero to 10, where 0 = having no pain and 10 = having unbearable pain.

The sickness impact profile

The sickness impact profile (SIP) is a behaviour based self report measure that is used to quantify sickness related dysfunction. [17] Patients are asked to complete a standardised questionnaire consisting of 136 items aggregated into 12 domains of daily functioning. It has a physical dimension consisting of three domains by aggregation of the item scores of the ambulation, mobility, and body care and movement scales, and a psychosocial dimension including four scales, that is, social interaction, alertness behaviour, emotional behaviour, and communication. The remaining, independent categories are not aggregated: sleep and rest, eating, work, home management, and recreation and pastimes. [18] Differential weights per item are aggregated for each category and for both dimensions, and standardised to a percentage of the maximum possible score ranging from 0 to 100 (0 = no functional limitation for the category and 100 = maximally possible limitation). As the patients in the study were unable to walk because of multiple sclerosis or spinal cord injury, the domains ambulation and home management were not considered. Only two patients had a part time paid job, so the category work was not included in the analysis.

The Hopkins symptom check list

The Hopkins symptom checklist (HSCL) was translated and validated in the Dutch situation by Luteijn et al. [19] It consists of 57 items with two subscales and an overall scale. The subscale physical health contains eight items with scores ranging from 0 to 24 (0 = no complaints at all) measuring the physical health experienced, for example, headache, low back pain, and dizziness. The subscale mental health measures psychoneurotic complaints and consists of 17 items with scores ranging from zero to 51 (0 = no complaints at all). Some examples of items of this scale are: 'I cannot get rid of nasty thoughts', 'I am feeling desperate about the future'. The overall scale covers all 57 items, including the 32 items of the subscales measuring psychoneurotic and somatic complaints, and ranges from zero to 171. According to Luteyn et al this scale is very sensitive to change in the evaluation of treatments.

2.2.4 Statistical Methods

The changes in clinical and health related quality of life at three months were analysed for the treatment and placebo group using the Wilcoxon matched pairs signed ranks test. The same test was used to analyse the results after one year of baclofen infusion. The differences at three months between the treatment and placebo group were analysed using the Wilcoxon-Mann-Whitney test for ordinal data. [20] Effect sizes (**d**) were calculated according to Cohen. [21] As the variance of the post-test measure is partly explained by the pretest scores, estimating the magnitude of the change between baseline and post-test in the treatment and control group required adjustment of the effect size **d'** for the correlation (**r**) between the scores of paired observations.

$$d = \frac{d'}{\sqrt{1-r}} \qquad d' = \text{effect size} = \frac{\overline{X}_{\text{baseline}} - \overline{X}_{\text{post - test}}}{SD\left(X_{\text{baseline}} \quad X_{\text{post - test}}\right)}$$

d' = effect size = mean change/pooled SD baseline and post-test score;
d = effect size adjusted for r;
r = correlation coefficient

2.3 RESULTS

Of 96 consecutive implantation candidates screened for inclusion in this study, 53 failed to meet the eligibility criteria because of suboptimal dosage of oral medication (n = 17), functional spasticity or effective oral medication (n = 13), no spasticity (n = 3), or because they fulfilled one of the exclusion criteria (n = 20). Five of the 43 eligible subjects refused to participate. Of the remaining 38 patients, 22 were randomly assigned to placebo or baclofen using a balancing procedure. After the first wave of 22 patients had been assigned to the double blind controlled conditions, all 16 patients of the second wave received baclofen immediately after pump implantation. The results of the evaluation of clinical efficacy and health related quality of life in all 38 patients are not yet available, but will be published in due course.

2.3.1 Demographics

Table 1 presents the overall characteristics. The mean (SD) age of the sample was 48.3 (12.7) years (range 19-70), 55% were women, and 59% and 41% had multiple sclerosis or spinal cord injury, respectively. Seventeen patients (77%) were married or divorced with an average number of 2 children. At the start of the study a relatively high proportion of patients with multiple sclerosis was enrolled. This was caused by a difference in consultancy function of the centres that first participated in the study and has led to a lower proportion of patients with spinal injury during the placebo controlled phase compared with the follow up phase.

Table 2.1 *Patient characteristics of study groups and balancing criteria*

	Baclofen	Placebo
Age[1] (mean)	45.8	46.3
Sex[1]		
male	5	5
female	7	5
Etiology[1]		
Multiple Sclerosis	7	6
Spinal cord	3	6
Children (mean n)	1.8	1.9

[1] Balancing criteria

2.3.2 Differences between the groups after three months

Our initial hypothesis was that the baclofen and placebo group would show differences in both clinical efficacy and physical and psychosocial functioning. To test the hypothesis we analysed the differences in mean scores on all the instruments during the first three months of the study (baseline to three months). At baseline, before implantation, no significant differences between the groups were found for the complete set of variables (Wilcoxon-Mann-Whitney test, $\alpha = 0.05$). For the three clinical efficacy measures, the null hypothesis - that is, equal mean scores at baseline and at three months post-test - could be rejected (table 2, columns 8 and 9). The estimated magnitude of the difference in the spasm score was small (effect size = 0.20); differences in the Ashworth scale (effect size = 1.40) and pain score (effect size = 0.94) were large.
However, the physical and psychosocial dimensions of the health related quality of life measures showed no significant differences between the placebo and treatment group at baseline and after three months.

2.3.3 Differences within the baclofen treatment and placebo group

Separate analysis of the two groups (table 2, columns 6 and 7) showed no significant changes for any of the outcomes in the placebo group after three months. However, the baclofen group showed significant changes in the following outcome measures: spasm score (P = 0.04); Ashworth scale (P = 0.04), the overall SIP score (P = 0.03); the physical dimension of the SIP (P = 0.02); the SIP mobility scale (P = 0.005); the SIP scale sleep and rest (P = 0.02); the SIP psychosocial behaviour scale (P = 0.04); the overall score of the HSCL (P = 0.002) and the mental health scale of the HSCL (P = 0.005). This trend is confirmed by the effect sizes which ranged from moderate to large (with values between 0.70 and 1.35) suggesting that baclofen infusion affected the domains of health related quality of life and clinical outcome in the predicted direction. Of the clinical efficacy data, the self reported pain score did not show a significant decrease during this period and the same applies for the HSCL physical health scale. Although the scores of the sickness impact profile dimensions eating, recreation and pastimes, body care and movement indicated an improvement after three months; the changes were not significant and are therefore not shown.

Table 2.2 *Three-month outcome of baclofen and placebo for severe spasticity (N = 22)*

	Baseline score		After 3 months			Wilcoxon Matched Pairs signed-ranks		Mann Whitney U Wilcoxon Rank Sum		
	Mean	(SD)	Mean	(SD)	es[1]	Z value	p[2] value	Z value	P value	es[3]
	1	2	3	4	5	6	7	8	9	10
Clinical efficacy:										
Spasm score:										
baclofen	2.23	0.54	1.65	1.11	0.74	-1.78	*	-2.32	*	0.20
placebo	1.83	0.66	1.81	0.76		-0.18	ns			
Ashworth scale:										
baclofen	2.51	0.70	1.51	1.20	1.12	-1.99	*	-2.49	**	1.40
placebo	3.07	0.41	2.87	0.57		-1.25	ns			
Self-reportedpain score:										
baclofen	4.20	2.98	2.75	3.22	0.72	-1.35	ns	-1.79	*	0.94
placebo	6.00	3.07	5.94	3.57		0.00	ns			
SIP[4]:										
Sleep and rest:										
baclofen	12.33	12.27	16.20	10.35	0.71	-1.99	*	-1.21	ns	
placebo	21.71	16.84	21.38	11.09		-0.10	ns			
Mobility										
baclofen	31.99	17.40	16.69	12.29	1.35	-2.50	**	-2.32	ns	
placebo	38.84	24.09	35.88	24.60		-0.42	ns			
Physical dimension:										
baclofen	39.98	9.78	35.10	5.41	1.07	-1.99	*	-1.31	ns	
placebo	42.80	12.73	39.53	11.49		-1.54	ns			

Table 2.2 *continued*

Psychosocial dimension:									
baclofen	16.03	13.69	12.26	9.87	0.74	-1.68	*	-0.73	ns
placebo	42.80	12.73	39.53	11.49		-0.06	ns		
SIP overall score:									
baclofen	31.72	9.80	27.79	5.32	1.00	-1.78	*	-0.82	ns
placebo	30.12	10.64	28.98	8.83		-0.18	ns		
HSCL[5]									
Physical health:									
baclofen	4.17	3.16	4.00	3.44		-1.01	ns	-0.076	ns
placebo	5.78	4.05	4.44	3.00		-0.82	ns		
Mental health:									
baclofen	7.83	4.97	5.00	4.28	1.28	-2.50	**	-0.086	ns
placebo	6.89	7.20	7.33	6.69		-0.10	ns		
HSCL overall score:									
baclofen	30.0	12.54	20.67	11.78	1.34	-2.79	***	-0.099	ns
placebo	31.0	21.62	28.22	18.43		-0.82	ns		

1 Effect size for paired observations
2 * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ns: not significant
3 Effect size for independent samples
4 Sickness Impact Profile
5 Hopkins Symptoms Check List

Table 2.3 *Clinical outcome measures, Health-related functional status, Perceived physical and mental health and depression at baseline, 1 year after baclofen infusion and changes in (sub)scale scores, and effect sizes (ES)*

Instrument subscale	Baseline score		1 year baclofen score				
	mean	sd	mean	sd	Z value	P value	Effect size
Ashworth scale:	2.87	0.54	0.44	0.51	-3.52	0.002	6.23
Self-reported pain score:	4.57	3.23	1.97	2.95	-2.35	0.009	1.07
Spasm score:	2.16	0.48	0.62	0.75	-3.42	0.003	3.05
Sickness Impact Profile:							
Categories:							
Sleep and rest	20.48	12.48	13.99	10.53	-2.20	0.01	0.95
Recreation and pastimes	42.47	22.47	30.53	22.35	-1.70	0.04	0.63
Ambulation[1]							
Mobility	35.10	19.64	25.16	19.50	-2.07	0.02	0.73
Body care and movement	50.62	19.30	41.44	18.72	-1.94	0.02	0.64
SIP Physical dimension:	41.48	8.07	33.44	12.73	-1.85	0.03	0.86
SIP Psychosocial dimension:	14.80	11.72	10.96	10.18	-1.54	ns[2]	
SIP Overall score:	31.28	7.93	25.13	9.61	-2.48	0.005	0.99
HopkinsSymptomCheck List:							
Physical health:	4.89	2.87	3.66	3.03	-2.19	0.01	0.86
Mental health:	7.17	5.26	5.44	4.57	-1.29	ns	
HSCL total score:	29.00	12.71	22.11	12.09	-2.22	0.01	0.87

1) The items of the SIP ambulation scale were not applicable for the patients in this study and were removed.

2.3.4 Results after one year of baclofen administration

Table 2.3 summarises the results of the evaluation of health outcome measures at baseline and one year after the start of intrathecal infusion of baclofen in the complete sample of patients with severe spasticity in the first wave of the study. Patients who were assigned to the 13 week placebo condition followed by baclofen treatment, were merged with the group who received baclofen from the start of the study. This observational longitudinal phase includes the entire initial sample of 22 patients, who were followed up during one year of intrathecal baclofen infusion. At one year patients showed substantial, significant improvement on clinical efficacy outcomes (self reported pain $P < 0.01$; effect size = 1.07, Ashworth scale and spasm score $P < 0.01$; effect size 6.23 and 3.05, respectively). Improvement was also found

for the physical dimensions mobility and body care and movement of the SIP, indicating a statistically significant ($P < 0.05$) change between baseline and post-test. The corresponding effect sizes suggest moderate changes in health related behaviour in these domains. Change was also significant and substantial for the physical health subscale of the SIP ($P < 0.05$; effect size 0.86). The SIP overall score (calculated without the ambulation items) and the physical health and overall scale of the HSCL showed a significant and substantial decrease (improvement) after 1 year, with large effect sizes > 0.80. Changes in health related behaviour was observed for the categories sleep and rest and recreation and pastimes ($P < 0.01$, $P < 0.05$; effect size 0.95 and 0.63 respectively). In striking contrast to the physical dimensions, the psychosocial dimensions of the SIP (social interaction, alertness behaviour, emotional behaviour and communication) and mental health of the HSCL did not show any significant improvement.

2.4 DISCUSSION

As expected, the mean scores of the clinical efficacy scales (muscle tone, spasm score, and self reported pain) before and after treatment (table 3) showed a clear change in the predicted direction after one year of intrathecal baclofen infusion.[22-24] These changes, which can be interpreted as an improvement in relevant clinical outcomes, are significant with large effect sizes.

We can conclude that intrathecal baclofen delivered by a subcutaneously implanted programmable pump resulted in a significant improvement in self reported health related quality of life regarding recreation and pastimes, rest and sleep, mobility, body care and movement as assessed with the sickness impact profile. The changes between the initial and final scores on the physical health dimension and the overall scores of the sickness impact profile and the Hopkins symptom checklist also point to improvement. No change was found for the SIP and HSCL psychosocial dimensions. Significant improvements are associated with effect sizes > 0.63. For non-significant changes the effect sizes ranged from 0.40 to 0.57.

Contrary to our expectations, three months after implantation the baclofen group and the placebo group did not differ significantly in the mean scores on the physical and psychosocial dimensions of health related quality of life instruments.

It was hypothesised that no significant changes would be found in the placebo group but a significant change was expected to have occurred in the baclofen group at three months. The group receiving baclofen immediately after implantation improved significantly on the clinical outcome measures, demonstrating the clinical efficacy of the treatment. This group showed significant changes in relevant physical and

psychosocial dimensions of self reported health status except for the HSCL physical health scale.

2.5 CONCLUSION

In interpreting the results of this study, one should bear in mind that the research design may have caused some underestimation of the results. The following considerations are important in this respect:
In the placebo phase of the study the pump could not be optimally programmed because the doctor who was responsible for the treatment was blinded. In patients assigned to the placebo condition this would have led to countless increases in the concentration of the contents or the velocity of the pump. Therefore, we decided to restrict the number of changes in velocity and/or concentration to two. This may have resulted in suboptimal doses for some of the patients in the baclofen group, which in turn may have affected treatment outcome. Thus the observed differences between the baclofen and placebo group may not be representative of optimal treatment results.
Three months is probably too short a period to find evidence of differences in dimensions of health- related quality of life between the treatment and baclofen group. Despite the significant and substantial observed change in clinical efficacy in the baclofen group, these patients continue to have other invalidating consequences of their underlying disease. This might explain the lack of significant differences in health related quality of life between patients receiving baclofen and placebo.
The necessity of blinding, even if the outcome seems too obvious, was shown by changes in spasticity and health related behaviour in one of the patients in the placebo group. For several weeks both patient and research team erroneously assumed that these changes were attributable to baclofen.
In cases where the optimum dosage was achieved after two corrections there is the probability of habituation causing a reduction in the effects after the first four weeks.
Aspects of physical health and daily functioning are probably associated with the degree of spasticity. Therefore, a reduction in severe spasticity is likely to induce a substantial change in the dimensions of physical health. The psychosocial dimensions, however, are probably more strongly associated with the unchangeable, underlying disease, which may explain the absence of a treatment effect in this respect even after one year.
The improvement on the psychosocial dimensions of quality of life during the first three months is probably associated with patients receiving increased attention from

medical professionals and their social network combined with (too) high expectations of the treatment. This effect is likely to disappear after one year of treatment.

Acknowledgment

This study was supported through a grant from the Dutch Sickfund Council. Appreciation is expressed to Ms. Mereke Gorsira for assistance in preparing the manuscript.

REFERENCES

1. Müller H, Zierski J, Dralle D et al. Intrathecal baclofen in spasticity. In: Müller H, Zierski J, Penn RD (eds) Local spinal therapy of spasticity. Berlin-Heidelberg: Springer Verlag, 1988.
2. Nance P, Schryvers O, Schmidt B et al. Intrathecal baclofen therapy for adults with spinal spasticity: therapeutic efficacy and effect on hospital admissions. Can J Neurol Sci 1995;**22**:22-9.
3. Albright L, Barron A, Fasick MP et al. Continuous intrathecal baclofen infusion for spasticity of cerebral origin. JAMA 1993;**270**:247-57.
4. Coffey RJ, Cahill D, Steers W et al. Intrathecal baclofen for intractable spasticity of spinal origin: results of a long-term multicenter study. J Neurosurg 1993;**78**:226-32.
5. Meythaler JM, Steers WD, Tuel SM et al. Continuous intrathecal baclofen in spinal spasticity: a prospective study. Am J Phys Med Rehabil 1992;**6**:321-27.
6. Steinbok P, Daneshvar H, Evans D, Kestle JRW. Cost analysis of continuous intrathecal baclofen versus selective functional posterior rhizotomy in the treatment of spastic quadriplegia associated with cerebral palsy. Pediatr Neurosurg 1995;**22**:255-65.
7. Saltuari L, Kronenberg M, Marosi MJ et al. Indication, efficiency and complications of intrathecal pump supported baclofen treatment in spinal spasticity. Acta Neurol 1992;**3**:187-94.
8. Ochs G, Delhaas EM. Long-term experience with intrathecal use of baclofen in severe spasticity. In: Lakke JPWF, Delhaas EM, Rutgers AWF (Eds.). Parental drug therapy in spasticity and Parkinson's disease. Carnforth: The Parthenon Publishing Group Limited, 1992.
9. Loubser PG, Narayan ChRK, Sandin KJ et al. Continuous infusion of intrathecal baclofen: long-term effects on spasticity in spinal cord injury. Paraplegia 1991;**29**:48-64.
10. Teddy P, Jamous A, Gardner B et al. Complications of intrathecal baclofen delivery. British J Neurosurg 1992;**6**:115-18.
11. Penn RD. Intrathecal baclofen for spasticity of spinal origin: seven years of experience. J Neurosurg 1992;**77**:236-40.
12. Patterson V, Watt M, Byrnes D et al. Management of severe spasticity with intrathecal baclofen delivered by a manually operated pump. J Neurol Neurosurg Psychiatry 1994;**57**:582-85.
13. Bennett C, Franklin N.L. Statistical analysis in chemistry and the chemical industry. New York: John Wiley, 1961.
14. Cohen J. A power primer. Psychological Bulletin 1992;**1**:155-59.
15. Zielhuis GA, Straatman H, Van 't Hof-Grootenboer AE et al. The choice of a balanced allocation method for a clinical trial in otitis media with effusion. Statistics in Medicine 1990;**9**:237-46.
16. Ashworth B. Preliminary trial of carisoprodal in multiple sclerosis. Practitioner 1964;**192**:540-2.
17. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;**19**:787-805.

18. De Bruin AF. The measurement of sickness impact: the construction of the SIP-68. Dissertation Maastricht: University of Limburg, 1994 (ISBN90-74421-02-4).
19. Luteyn F, Hamel LF, Bouman TK et al. Hopkins Symptom Checklist (Manual). Lisse: Swets & Zeitlinger, 1984.
20. Siegel S, Castellan N.J. Nonparametric statistics for the behavioural sciences.New York: McGraw-Hill, 1988.
21. Cohen J. Statistical power analyses for the behavioural sciences. New York: Academic Press, 1977.
22. McLean BN. Intrathecal baclofen in severe spasticity. Br J Hospital Medicine 1993;**49**:262-67.
23. Hugenholtz H, Nelson RF, Dehoux E, Bickerton R. Intrathecal baclofen for intractable spinal spasticity-a double blind cross-over comparison with placebo in 6 patients. Can J Neurol Sci 1992;**19**:188-95.
24. Azouvi P, Mane M, Thiebaut JB, Denys P, Remy-Neris O, Bussel B. Intrathecal baclofen administration for control of severe spinal spasticity: functional Improvement and long-term follow up. Arch Phys Med Rehabil 1996;**77**:35-9.

3
Psychometric properties of the Minnesota Living with Heart Failure -Questionnaire (MLHF-Q).

Berrie Middel, Msc,* Jelte Bouma, PhD,* Mike de Jongste, MD, PhD, Eric van Sonderen, PhD,* M.G.Niemeijer, MD, PhD,*** Harry Crijns, MD, PhD,** Wim van den Heuvel,PhD,***

* Northern Centre for Healthcare Research (NCH), School of Medicine, University of Groningen, The Netherlands

** Department of Cardiology and Thoracic Surgery, University Hospital Groningen, The Netherlands

*** Department of Cardiology Martini Hospital Groningen, The Netherlands

Clinical Rehabilitation 2001; 15: 380-391

ABSTRACT

Objective
The purpose of this study was to evaluate the psychometric properties of the Minnesota Living with Heart Failure Questionnaire (MLHF-Q) in patients with atrial fibrillation.
Design
This was a prospective study of the patients who underwent DC electrical cardioversion.
Setting
Clinics of Cardiology and Thoracic surgery of the University Hospital in Groningen, the Netherlands.
Main Outcome measures
The disease specific MLHF-Q and generic measures of quality of life were administered. The sensitivity to change over time was tested with effect sizes (ES). Internal consistency of MLHF-Q scales was estimated with Cronbach's alpha. To evaluate the construct validity multitrait-multimethod analysis was applied. The 'known group validity' was evaluated by the comparison of mean scores and effect sizes between two groups of the NYHA-classification (NYHA I versus II-III). Stability of MLHFQ-scales was estimated in a subgroup of patients, which remained stable. Perfect Congruence Analysis and factor analysis were applied to confirm the a priori determined structure.
Results
Cronbach's alpha was ≥ 0.80 of the MLHF-Q scales. Perfect Congruence Analysis (PCA) showed that the results resemble quite well the a priori assumed factor structure. Multitrait-multimethod analysis showed convergent validity coefficients ranging from .59 to .73 (physical impairment dimension); from.39 to .69 (emotional dimension). The magnitude of change can be interpreted as medium (ES = .50). The results of a "test-retest" analysis in a stable group can be valued as satisfactory for the MLHF-Q scales (Pearson's $r > .60$). The physical dimension and the overall score of MLHF-Q showed statistically significant differenc between NYHA I and II-III groups ($p < .001$) with large effect sizes (ES > 1.0).
Conclusions
The MLHF-Q has solid psychometric properties and the outcome of the current study indicate that the MLHF-Q is an effective and efficient instrument.

Keywords: Quality of life, outcome assessment, heart failure, validity

3.1 INTRODUCTION

In assessing health related quality of life and functional ability or health status, a distinction is made between disease-specific outcome measures developed to measure quality of life dimensions characteristic for patients having a particular disease, and generic instruments measuring more broadly defined dimensions of quality of life. Both types of instruments have their strengths and weaknesses.[1] An advantage of generic instruments is that they have a broad scope and can be used in many populations on a wide variety of diseases. A disadvantage is that general aspects of quality of life which are not significant for a specific disease will result in a less valid assessment of the concept of health-related quality of life in e.g. groups of (chronic) disease. Assessing only those aspects of quality of life, which are determined to be due to a particular disease, will result in a short instrument that will be more sensitive to detect change in disease-specific groups after (medical) interventions. A disadvantage of a disease-specific instrument is that study results are difficult to compare with those of other populations. In the current study health related quality of life was assessed with the specific Minnesota Living with Heart Failure Questionnaire (MLHF-Q), [2] the generic RAND-36 or SF-36 [3], the Multidimensional Fatigue Inventory (MFI-20) [4,5,6] and the Hospital Anxiety and Depression Scale (HADS). [7,8] The data were appropriate to conduct a validation study to estimate the sensitivity to change (responsiveness), the reliability and validity of the disease-specific MLHF-Q to obtain data for its future use in (Dutch) clinical evaluation studies.

The MLHF-Q consists of 21 items and addresses a wide range of health-related quality of life aspects. [9] In this article, the psychometric properties of the Dutch version of the MLHF-Q scales are evaluated and validated with conceptually similar dimensions of generic instruments: the RAND-36, the HADS and the MFI-20.

All instruments are self-report measures of quality of life on the dimensions of physical, mental or social well-being. The psychometric properties of the MLHF-Q have been evaluated already in its English version and the instrument has been used as outcome measure in clinical trials in the context of the American health care system. [2,10,11,12,13,14,15] In other countries the number of studies on the evaluation of the reliability and validity of the MLHF-Q is up till now not substantial .[16,17] The RAND-36 was chosen as the generic counterpart because it is a generally accepted and well-validated instrument, it is a short questionnaire with known psychometric properties, [18,19,20,21] it resembles closely the MLHF-Q dimensions, and is available in a Dutch version. [22,23]

The objectives of this study were:

- to compare the results from the MLHF-Q with the RAND-36, HADS and MFI-

20 in terms of reliability and sensitivity to detect change over time. It was hypothesised that the MLHF-Q would demonstrate a comparable magnitude of change over time;
- to compare the results of the questionnaire's clinical validity. It was hypothesised that the MLHF-Q would demonstrate that the more severely angina was rated by the NYHA-classification, the greater the deterioration in the patient's quality of life turned out to be. The change assessed in a group of patients who remained clinically unchanged or stable was hypothesised to be due to chance fluctuation;
- to find support for the factor structure originally found by Rector and Cohn [12] in our data;
- to provide empirical evidence that the MLHF-Q scale measures the underlying constructs of physical and emotional impairments it is reputed to represent.

The purpose of the present study was to use data of a treatment-outcome study to determine the performance of the MLHF-Q. The results of the clinical efficacy study will be published elsewhere.

3.2 METHODS

Consecutive patients scheduled for DC electrical cardio version were included in this study. Patients presented arterial fibrillation and arterial flutter and were treated at the department of Cardiology and Thoracic surgery of the University Hospital in Groningen.

Out of the 60 consecutive candidates for DC electrical cardio version screened for inclusion, five patients died within twelve months after the completion of the first questionnaire. One year after the first visit to the clinic, 44 patients out of 55 (80.0%) returned the questionnaire used for analysis of reliability and validity of the MLHF-Q.

All patients completed the questionnaires as a baseline assessment before the first treatment (DC electrical cardio version) in the department of Cardiology and Thoracic surgery of the University Hospital Groningen. The patients were invited to participate in the study by the cardiologist and after informed consent the patients completed the questionnaires, undisturbed, in a separate room. The cardiologist was blinded to the information of the questionnaires. The second and third assessment was at home, three and twelve months after the first electrical cardio version respectively. The questionnaires were returned in a pre-paid envelope to the Northern Centre for Healthcare Research of the University of Groningen.

3.2.1 Measurements

Both demographic characteristics of the patients and relevant medical background variables were administered with standard or usual questions and items in the medical examination procedure at the first visit to the outpatient clinic. To assess the impact of the treatment on daily physical, emotional and social functioning, four instruments were used. The RAND-36 is a generic instrument and consists of 36 items that contribute to eight scales that measure the following aspects of health: "physical functioning" (10 items), "social functioning" (2), "role limitations due to physical problems" (4), "role limitations due to emotional problems" (3), "mental health" (5), "energy/vitality" (4), "pain" (2), and "general health perception" (5). The one-item scale on change in perceived health was not used in the transformation of scores into a scale, because the MLHF-Q does not contain an item assessing change in perceived health. The RAND-36 item scores are summed and transformed to eight scales, each with scores between 0 and 100, where 0 represents the worst state of health and 100 the best state of health possible. [3,23] The Minnesota Living with Heart Failure Questionnaire is a disease-specific instrument and composed of 21 items and three scales that measure: the physical dimension (8 items), the emotional dimension (5 items) and the overall score on health-related quality of life (21 items). Eight separate items, which do not assess a single construct or dimension of health-related quality of life, measure social and economical impairments patients relate to their heart failure and are part of the overall score. The total score has a range between 0 and 105, the physical dimension (sub-scale) between 0 and 40, the emotional dimension (sub-scale) between 0 and 25 and the separate items on the socio-economic impairments between 0 and 40.
High scores on the MLHF-Q scales indicate a high negative impact of heart disease on the assessed aspects of quality of life.
The Multidimensional Fatigue Inventory (MFI-20) consists of twenty items and five sub-scales (General, Physical, Activity, Motivation, and Cognition). Each scale consists of four items and has a range from 4 to 20 and its total score ranges from 20 to 100. High scores indicate high fatigue. The subscales Anxiety and Depression of the Hospital Anxiety and Depression Scale (HADS) have a range between 0 and 21. A score of 7 or lower identifies 'non-cases', 8 to 10 'doubtful cases' and a score $\geq$ 11 'definite cases'.

3.2.2 Quantitative analysis

The features of the distribution of scores on the conceptually similar dimensions of the MLHF-Q, MFI-20, HADS, and RAND-36 were computed. Mean scores, standard deviations, and the percentage of patients with the maximal possible score

(ceiling) and the minimal possible score (floor) are represented.
In the examination of the construct validity of the MLHF-Q, scales of all instruments were used in the analysis. It was hypothesised that the scales, that are conceptually associated, would show strong correlations and scales, that are conceptually weaker associated, would demonstrate lower correlation coefficients.
In this study, the internal consistency of the MLHF-Q, RAND-36, HADS and MFI-20 scales was tested with Cronbach's α [24] to make comparisons between the instruments' mean alphas. An α-coefficient > 0.80 was considered as sufficient [32] irrespective of the number of items. Perfect Congruence Analysis and factor analysis were applied to confirm the a priori determined structure on which Rector and Cohn [12] have selected the items.
Test- retest stability of the MLHF-Q scales was assessed with correlation coefficients between baseline and 3 months after cardio version in a group in which the treatment was not successful (that showed no sinus rhythm three months after the first electrical cardio version), so their health status remained unchanged or stable. Although the test-retest procedure was not carried out by sending the questionnaire shortly after the first completion, we were interested in the variability of the MLHF-Q scores between two points in time within a group whose condition remained stable. However, high test-retest correlation coefficients as such do not give us information about the changes in time between baseline and 3-months outcome scores, and therefore we tested the hypothesis that the change over time in a stable group is due to chance fluctuations. The Wilcoxon Matched-Pairs Signed-Rank test was used due to the non-normal distribution of the outcome assessments.

To estimate the responsiveness, the ability of an instrument to detect the magnitude of change over time within one group, we used Cohen's effect size statistic d for paired observations.[25] As the variance of the post-test measure is partly explained by the pre-test scores, estimating the magnitude of the change between baseline and post-test in the treated group requires adjustment of the effect size d' for the correlation (r) between the baseline and post-test scores. [26,27]

$$d = \frac{d'}{\sqrt{1-r}} \qquad d' = \text{effect size} = \frac{\bar{X}_{baseline} - \bar{X}_{post\text{ - }test}}{SD\left(X_{baseline}\ \ X_{post\text{ - }test}\right)}$$

d' = effect size = mean change/pooled SD baseline and post-test score;
d = effect size adjusted for r;
r = correlation coefficient between repeated measurements.

An effect size of .20 has to be interpreted as a small effect, an effect size of .50 as a medium effect, and an effect size of > .80 as a large effect. [25,28] To evaluate the ability of the MLHF-Q to discriminate between subgroups of patients of which is known that they differ on an accepted classification of the seriousness of the disease, the 'known groups validity' of the MLHF-Q scales was tested. [29] The Man-Whitney U Wilcoxon rank sum test was used because of the non-normal distribution of the variables in the analysis. The grouping condition was NYHA classification I vs. II and III (due to the small number of observations class II and III were combined). [30] Cohen's effect size d' for unrelated samples, to estimate the magnitude of the difference in mean scores between these groups, was calculated by dividing the mean difference score by the pooled standard deviation for groups with unequal number of observations. [31]

$$d' = \frac{\overline{X}_{NYHA\ I} - \overline{X}_{NYHA\ II\text{-}III}}{SD\left(X_{NYHA\ I}\ X_{NYHA\ II\text{-}III}\right)}$$

3.3 RESULTS

In table 3.1 the descriptive statistics of the sample are shown. The mean (range) age of the patients in the study was 61.5 (range 28 - 87) years. The minority of patients was female (35%). The majority of patients had one or more heart diseases or other relevant diseases in addition to arterial fibrillation (AF). Only six persons had AF without any other disease. Almost half of the patients (46.7%) had two or more diseases next to AF. A relatively large group (41.7%) was treated for the first time for AF. The mean score on the NYHA classification (range 1-4) of 1.9 indicates a moderate severity of the underlying disease.

Table 3.1 *Patient characteristics at study enrolment (n=60)*

	No	(%)
Gender		
Men	39	(65.0)
Women	21	(35.0)
Mean age (y)	61.5	(SD 12.7)
Martial status:		
Married/living with partner	39	(60.0)
Widowed/unmarried/divorced	16	(24.6)
Missing value	5	(15.4)
Disease:		
Aortic Valve disease	12	(20.0)
Mitralic Valve disease	12	(20.0)
Hypertension	16	(26.7)
Congenital heart disease	7	(11.7)
Coronary Artery Disease	11	(18.3)
Cardiomyopathy	4	(6.7)
Hyperthyroidism	2	(3.3)
CARA	9	(15.0)
Miscellaneous	16	(26.7)
No disease	6	(10.0)
1 disease	26	(43.3)
2 diseases	21	(35.0)
3 -4 diseases	7	(11.7)
Mean NYHA-classification	1.9	(SD 0.6)

3.3.1 Distribution of scores, internal consistency and responsiveness

Mean baseline and post-test (1 year) scores, standard deviations and the percentages of patients with the maximum and minimum scores, are represented in table 3.2. A study of the distribution of scores of the MLHF-Q scales showed a skew ness in the direction of positive functioning or little or no impairment. The RAND-36 data showed the same tendency for four scales (social functioning, emotional role functioning, pain, and health perception). The RAND-36 scale 'physical role functioning' showed a tendency towards the opposite direction. Three conceptually related scales of the MFI ('physical', 'activity' and 'general' feelings of fatigue) were skewed in the direction of little impact on health-related quality of life while the cognition scale was skewed in the negative direction.

Table 3.2 *Means, standard deviations, minimum and maximum scale-scores, Crohnbach's alpha's, Pearson's correlations $t^1 – t^{12}$ and within groups effect size for paired observations (N = 44)*

	Pre-test					Post-test						
Dimension	Mean	SD	% Floor	% Ceiling	Reliability	Mean	SD	% Floor	% Ceiling	Reliability	Effect size. *	r
MHLF												
Physical dimension (0-40)	14.2	9.6	15.9	2.3	.88	10.4	10.3	15.9	2.3	.91	.65	.67
Emotional dimension (0-25)	5.9	5.7	20.5	2.3	.82	3.8	4.5	31.8	2.3	.81	.56	.51
Overall score (0-105)	28.5	19.6	11.4	2.3	.91	21.6	20.8	13.6	2.3	.94	.59	.67
Rand-36												
Physical functioning (0-100)	56.6	28.7	2.3	6.8	.93 (.92).**	66.1	27.1	11.4	2.3	.93	.63	.71
Social functioning (0-100)	60.4	25.3	13.6	4.7	.79 (.71)	72.3	25.2	31.8	2.3	.78	.72	.56
Role-physical (0-100)	27.3	39.3	15.9	56.8	.91 (.90)	51.7	45.1	38.6	34.1	.91	.77	.46
Role-emotional (0-100)	54.8	44.7	31.8	40.9	.90 (.86)	62.6	44.3	54.5	25.0	.90	.25	.52
Pain (0-100)	80.9	23.6	47.7	2.4	.90 (.93)	82.6	21.9	52.3	2.3	.89	.13	.67
Mental health (0-100)	64.6	21.7	6.8	4.7	.83 (.85)	72.5	17.8	2.3	2.3	.84	.54	.47
Energy/vitality (0-100)	48.2	24.2	2.3	4.9	.86 (.82)	58.6	22.4	2.3	2.3	.84	.65	.53
Health perception (0-100)	56.0	21.2	6.8	2.3	.79 (.82)	55.0	20.0	2.3	4.5	.76	.07	.51
HADS												
Anxiety	5.7	3.9	4.5	2.3	.83	4.2	3.3	15.9	2.3	.81	.61	.53
Depression	6.0	4.6	6.8	4.5	.84	5.3	4.5	11.4	2.3	.86	.26	.74
MFI-20												
General	12.6	5.2	6.8	11.4	.87	10.8	5.3	11.4	9.1	.89	.58	.66
Physical	12.1	4.5	7.0	11.4	.85	10.9	5.1	9.1	4.5	.90	.37	.49
Activity	12.3	5.1	4.5	9.1	.88	10.9	5.3	13.6	6.8	.90	.41	.54
Motivation	10.7	4.7	4.5	4.5	.76	10.2	4.5	15.9	2.3	.80	.20	.67
Cognitive	7.4	3.5	34.1	2.3	.82	8.0	3.9	27.3	2.3	.86	.24	.52
Overall fatigue	54.1	18.3	2.4	2.4	.94	49.5	20.2	2.3	2.3	.95	.40	.65

* Effect size d for paired observation [26]

** Reliabilities of a general Dutch municipality population [21,22]

The Cronbach's alpha's, the internal consistency coefficients, of the MLHF-Q, RAND-36, HADS and MFI-20 scales are also shown in table 3.2. The internal consistency of the MLHF -Q scales had a satisfactory level of reliability ($\alpha > .80$). [32] Only the RAND-36 scales "social functioning" and "general health perception" and the MFI-scale "cognition" were below this level (.79, .79 and .76 respectively). The reliability coefficients of the MLHF-Q scales remained satisfactory one year after enrolment.

The scales of the MLHF-Q at baseline assessment yielded internal consistency estimates (mean $\alpha = .85$; range = .82 to .88) equal to the RAND-36 (mean $\alpha = .86$; range = .79 to .93) and somewhat higher than those of the HADS (mean $\alpha = .83$; range = .83 to .84) and MFI-20 (mean $\alpha = .84$; range = .76 to .88).

The ability to detect change over time within one group with paired observations was estimated with the effect size proposed by Cohen. [25] An effect size of .20 has to be interpreted as a small effect, an effect size of .50 as a medium effect and as an effect size of $\geq$.80 as large effect. Large effect sizes were not found. The MLHF-Q scales showed medium effect sizes. The RAND-36 scales 'role limitations due to emotional problems' and 'pain' demonstrated small effect sizes and 'general health perception' showed no ability to detect change between baseline and one-year outcome assessment. The HADS-anxiety scale and the MFI-20 'general fatigue' scale showed medium effect sizes. The physical and emotional dimensions of the RAND-36, HADS, and MFI-20 demonstrate comparable indicators of change over time within this particular group.

3.3.2 Item analysis

The MLHF-Q contains three dimensions or scales: a physical dimension, an emotional dimension, and a global quality of life dimension. A comparison was made with the results of the factor analysis of Rector and Cohn. [12] Their data provided us with an a priori assumed four-factor structure that was forced in order to evaluate the congruence of our data with the original structure. Therefore, a computer program for Simultaneous Component Analysis (SCA) for variables measured in two or more populations was applied. [33] The four a priori assumed factors based on the structure in the data of Rector and Cohn explained 58% of the total variance as a result of the SCA-Perfect Congruence Analysis (PECON). [34] A principal component analysis with rotation according to the varimax criterion was performed without the constraints of the structure elaborated by Rector and Cohn. In this analysis the four factors explained 61 % of the total variance. This difference of 3% indicates an acceptable discrepancy, but still indicates an insufficient recognition in our data. A fourth socio-economic dimension of impairments, that patients relate to their heart failure, was

suggested by Rector et al., [2] but in the current study the items did not load on a socio-economic component. As is demonstrated in the matrix (table 3.3), 6 out of the 21 items had very high loadings (> .70) and 13 items had high loadings (> .50 - < .70) on their respective factors. Only one item had a high loading on two factors (impairment because of ankle oedema). On face value we may conclude that the results closely resemble the findings of Rector and Cohn. [12] A closer inspection of the four factor solution, however, shows some deviations from the original factor structure: two items of the physical dimension identified by Rector and Cohn ("making your sleeping well at night difficult" and "your relating to or doing things with your friends or family difficult") have a high loading on factor three and four representing the impairments on a heterogeneous set of health-related aspects of heart failure. Factor 2 demonstrates high loadings of the items on the physical dimension. Although all the items of the emotional dimension had high loadings on factor 1, the following items showed also high loadings: "going away from home" (physical dimension), "ankle oedema", "hospitalisation", and "medical costs"(socio-economic impairments) on this factor.

Table 3.3 *Principal-Components factor Analysis with Varimax rotation of the MLHF-Q*

	Factor I	Factor II	Factor III	Factor IV
Making you:				
stay in a hospital	r .69	.23	.10	-.13
feel you are a burden to your family	e .62	.12	.36	-.34
feel depressed	e .70	.06	.33	.26
Worry	e .69	.16	-.05	.14
feel a loss of self control in your life	e .64	.37	.17	.09
going away places away from home difficult	p .67	.35	.11	.21
making it difficult for you to concentrate or remember things	e .65	.01	.01	.33
costing you money for medical care	r .47	.18	.25	-.29
causing swelling your ankles, legs, etc.	r .52	.52	-.01	-.02
walking about or climbing stairs difficult	.26	p .83	.10	.02
working around the house or yard	.31	p .79	.26	.19
sit or lie down to rest during the day	.40	p .73	.11	-.01
tired, fatigued or low on energy	.27	p .54	.31	.25
short of breath	-.06	p .60	.31	.26
sexual activities difficult	.02	.13	r .82	.16
eating less of the foods you like	.06	- .07	r .76	.16
recreational pastimes, sports/hobbies difficult	.13	.49	r .68	.17
your relating to or doing things with with your friends or family difficult	.28	.26	p .47	.24
side effects from medications	.22	.32	r .56	-.06
working to earn a living difficult	.11	.10	.21	r .73
sleeping well at night difficult	.10	.22	.39	p .60

e = emotional dimension
p = physical dimension
r = single items used in the construction of the overall score

3.3.3 Construct validity

In this study we attempted to provide evidence that the Minnesota Living with Heart Failure Questionnaire scales are measuring the underlying constructs of physical and emotional impairments it is reputed to represent.
The multitrait-multimethod approach outlined by Campbell and Fiske [35] was used to assess the convergent and discriminant validity of the MLHF-Q measures of physical and emotional impairment.
Convergent validity (i.e. evidence that we are measuring what we purport to measure) is provided by data that show that different measures of conceptually related dimensions of health-related-quality of life are highly correlated. [36,37] In addition, we expect that each of the measures of physical and emotional dimensions of quality of

life measures a different construct (i.e. that the 'physical fatigue' scale does not measure depression (discriminant validity).
In Table 3.4, multitrait-multimethod matrices were constructed for each of the assessed dimensions of quality of life (physical and emotional). Evidence of convergent validity is drawn from examination of the coefficients in the heterotrait-heteromethod triangles, enclosed by solid lines in Table 3.4. We also expect some association between the scales measuring dimensions of, for example, physical quality of life, if the same questionnaire (method) was used and items were not presented in a randomised order (correlated measurement error). These heterotrait-monomethod coefficients are depicted in bold. In the area enclosed by broken lines, the coefficients between variables that have no trait in common are shown.
The correlations between the three generic methods and the MLHF-Q scale assessing physical impact on quality of life are, as expected, high and have, compared to the heterotrait-monomethod coefficients, the same magnitude. The correlations between the three generic methods and the MLHF-Q scale assessing emotional impact, while statistically significant, are moderate (except the correlation between the HADS-depression scale with the RAND-'role emotional' scale).
To demonstrate divergent validity the multitrait-monomethod correlation coefficients must be higher than correlation coefficients for variables that have neither trait nor method in common. The values that represent relations between the components of physical and emotionally impaired quality of life, which are represented in the area enclosed by broken lines, are of interest. Most of the scales that are supposed to measure different constructs are weakly correlated, regardless the method used. In accordance with our expectation, some correlations were of moderate magnitude simply due to shared method variance (printed in bold).
This analysis provides reassurance that with the MLHF-Q we are measuring physical impairment and that there is convergence among methods. The emotional impairment component, however, is moderately associated with the other methods.

Table 3.4 *Multitrait-Multimethod Matrix for the emotional and mental dimensions of health related quality of life (N=60)*

Constructs	1	2	3	4	5	6	7	8	9	10
1. MLHF physical dimension										
2. Rand-36 physical functioning	-.65									
3. Rand-36 role-physical	-.60	**.64**								
4. Rand-36 energy/vitality	-.73	**.66**	**. 61**							
5. MFI-20 physical	.63	-.72	-.59	-.70						
6. MLHF-emotional dimension	**.44**	-.28	-.35	-.34	.37					
7. Rand-36 mental health	.32	**.35**	**.31**	**.53**	-.35	-.46				
8. Rand-36 role-emotional	.33	**.27**	**.49**	**.43**	-.27	-.47	**.50**			
9. HADS anxiety	.30	-.24	-.30	-.25	.42	.43	-.60	-.59		
10. HADS depression	.44	-.46	-.37	-.46	.35	.44	-.59	-.39	.55	

all correlations $p < .01$; corresponding dimensions are printed bold
The areas surrounded by solid lines are the hetero-trait-heteromethod triangles.
The area surrounded by broken lines comprises the coefficients for variables that have no trait in common.
(the hetereo-trait monomethod coefficients are depicted bold in both areas)

3.3.4 Test-retest

If a quality of life instrument like the MLHF-Q is developed to be used as an evaluative instrument in clinical trials, one of the conditions, which should be fulfilled is that it has the ability to demonstrate stability over time in subjects whose health status does not change (test-retest reliability). [34] Table 3.5 shows the test-retest correlation coefficients after a period of three months of stability in health status without serious cardiac events. The results can be valued as satisfactory for all MLHF-Q scales. However, although we can interpret the test-retest correlation coefficients as satisfactory, these estimates of linear relationships do not provide information about the existence of significant change in a selected group of stable patients. To test the statistical significance of the change between baseline and three-month outcome the Wilcoxon Matched-Pairs Signed-Rank test was used, because of the non-normal distribution of the MLHF-Q scales. None of the MLHF-Q scales demonstrated significant change.

Table 3.5 *Means, standard deviations (sd) test-retest correlations (r) and difference in scores between baseline and 3-months outcome in a group of patients that did not show improvement in sinus rhythm.*

	t1		T2				
	mean	sd	Mean	Sd	r	Z-score	P
MHLF-Q							
Physical dimension	14.39	9.59	15.91	12.91	.70	-0.23	.82
Emotional dimension	6.05	5.77	4.63	4.69	.63	-1.47	.14
Overall score	29.79	18.65	26.00	20.34	.73	-1.28	.20

3.3.5 Known group validity

In order to evaluate the ability of the MLHF-Q dimensions to discriminate between so-called 'known groups', which should show differences based on the cardiologists (blinded) classification of the severity of the disease, the study sample was divided into two subgroups: NYHA classification I vs. II-III. The results of the analysis of the ability of the MLHF-Q scales to discriminate between 'known groups' are presented in table 3.6. The physical dimension and the overall score of the MLHF-Q discriminated sharply between the NYHA II-III and I groups ($p < .001$) with large effect sizes. The MLHF-emotional dimension discriminated also clearly between these groups ($p = .01$) but with a moderate effect size.

Table 3.6 *Discriminative ability of the Minnesota Living with Heart Failure Questionnaire between NYHA-classification Groups*

	NYHA class I (n=15)		NYHA class II and III (n=38)				
					rank sum[1]		
	mean 1	sd 2	Mean 3	sd 4	z-value 5	p-value 6	es[2] 7
MHLF-Q							
Physical dimension	5.5	7.1	16.7	9.0	-3.8	.0001	1.31
Emotional dimension	3.5	6.4	6.5	5.3	-2.5	.01	0.53
Overall score	11.7	16.5	33.9	18.1	-3.6	.0003	1.25

[1] Mann-Whitney U, one-sided

[2] Estimation of the effect size used Cohen's d for independent samples when $n_1 \# n_2$, which is defined as the difference in mean scores divided by the pooled standard deviation:
est. $\sigma = \sqrt{\ } (N_1 - 1)s_1^2 + (N_2 - 1)s_2^2 / (N_1 - 1) + (N_2 - 1)$

3.4 DISCUSSION

To what extent does the Dutch version of the MLHF-Q measure the desired underlying concept or reflect what it is supposed to measure? In this study, the MLHF-Q construct validity was determined by higher and significant correlation coefficients between the MLHF-Q scales and corresponding dimensions of the MFI-20, HADS, and the RAND-36 and by lower correlations with non-corresponding dimensions of health related quality of life of these instruments. The MLHF-Q 'physical' dimension showed higher correlations with the RAND-36 scales 'social functioning', 'energy-vitality', 'health perception', and 'pain' indicating that these domains of quality of life, which are not tagged by the MLHF-Q, are more likely to be associated with physical limitations in this study group.

One of the great advantages in clinical trials is that the MLHF-Q is short; but its disadvantage is that it does not cover other relevant domains of quality of life impairment, such as impairment of social functioning or vitality. In the detection of change over time (pre- and post-test) the MLHF-Q performs equally well compared with the RAND-36 estimating the same standardised mean change-score expressed in effect sizes that are interpreted as medium effect for both instruments on physical and emotional functioning. We hypothesised a greater responsiveness, because the MLHF-Q should have greater precision due to the disease specific operationalized items of the domains' physical and emotional impairment. An alternative explanation for not detecting greater changes may be related to the selected group of patients: firstly, the questionnaire is developed to assess health-related quality of life associated with heart failure, which is not existent in every subject within this group; secondly, disappearance of the arterial fibrillation (AF) probably hasn't a strong impact on health related quality of life because of the fact that in ninety percent of the subjects the underlying diseases in addition to AF still exists. However, it is to be expected that in 'before - after' intervention studies, the MLHF-Q will show the ability to detect the appropriate magnitude of change over time. This expectation is based on the result of our study, namely that the MLHF-Q showed to be sensitive to detect change within and between groups, even if the differences are small.

In the ability to discriminate between 'known groups' the magnitude of the difference on the physical dimension of the MLHF-Q was large (Effect size >1) and statistically significant ($p < .001$). The emotional impact on quality of life showed a statistically significant difference (accompanied with a moderate effect size) between NYHA-I and II-III classified subjects. The substantial difference between both estimates of the magnitude of the difference between NYHA-I and II-III may be determined by

the dominant physical component of the NYHA classification. The correlations with conceptually related emotional dimensions (scales) were, while significant, of moderate magnitude. The cultural differences between the American and Dutch society, in combination with semantic differences in the translation of the items, are probably the explanatory factors. In the Dutch translation, 'making your going places away from home difficult' and 'making your stay in a hospital' are probably more associated with the emotional impact of disturbing the relationship with significant others, than with physical inhibition. The results of the current study indicate that the application of the MLHF-Q will enable Dutch researchers to assess health-related quality of life in clinical trials in which clinically relevant change will occur.

Acknowledgement:
We would like to thank Dr. Thomas S. Rector for giving us permission to use the Minnesota Living with Heart Failure Questionnaire for scientific purposes.

Clinical message:
Patients, considered for cardio version of arterial fibrillation, were studied to validate the Dutch version of the Minnesota Living with Heart Failure Questionnaire (MLHF-Q), by investigating responsiveness, reliability, validity, and effect size. Outcomes showed that the MLHF-Q is an effective and efficient instrument to assess clinically important change in health-related quality of life.

Reference List

1 Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Medical Care 1989; 27S:S217-S232.

2 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.

3 International Resource Center for Health Care Assessment (IRC). How to score the MOS 36-item Short Form Health Survey (SF36): SF-36™ scoring rules. 1991. Boston, New England Medical Center Hospitals.

4 Smets.E.M.A., Garssen B, Bonke B. Het Meten van vermoeidheid met de Multidimensionele Vermoeidheids Index (MVI-20)/Multidimensional Fatigue Inventory. Een Handleiding. 1995. Amsterdam, Afdeling Medische Psychologie AMC,Universiteit van Amsterdam.

5 Smets.E.M.A., Garssen B, Bonke B, de Haes JCJM. The Multidimensional Fatigue Inventory (MFI): Psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995; 39:315-325.

6 Smets.E.M.A., Garssen B, Cull A, de Haes JCJM. The application of the Multidimensional Fatigue Inventory (MFI-20) in cancer patients receiving radiotherapy. British Journal of Cancer 1996; 73(2):241-245.

7 Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale . Acta Psychiatr Scand 1983; 67:361-370.

8 Spinhoven PH, Ormel J, Sloekers PPA, Kempen GIJM. A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. Psychological Medicine 1997; 27(2):363-370.

9 Guyatt GH. Measurement of health-related quality of life in heart failure. JACC 1993; 22(4 (Supplement A)):185A-191A.

10 Kubo SH, Gollub S, Bourge R, Rahko P, Cobb F, Jessup M et al. Beneficial effects of Pimobendan on exercise tolerance and quality of life in patients with heart failure. Circulation 1992; 85(3):942-849.

11 Rector TS, Kubo SH, Cohn JN. Validity of the Minnesota Living with Heart Failure Questionnaire as a measure of therapeutic response to Enalapril or placebo. American Journal of Cardiology 1993; 71(May,1):1106-1107.

12 Rector TS, Cohn JN. Assessment of patient outcome with the Minnesota Living with heart Failure questionnaire: Reliability and validity during a randomized, double blind, placebo-controlled trial of pimobendan. American Heart Journal 1992; October,124(4):1017-1025.

13 Colucci WS, Packer M, Bristow MR, Gilbert EM, Cohn JN, Fowler MB et al. Carvedilol inhibits clinical progression in patients with mild symptoms of heart failure. US Carvedilol Heart Failure Study Group. Circulation 1996; 94(11):2800-2806.
14 Bulpitt CJ. Quality of life with ACE inhibitors in chronic heart failure. J Cardiovasc Pharmacol 1996; 27 Suppl 2:S31-S35.
15 Noe LL, Vreeland MG, Pezzella SM, Trotter JP. A pharmacoeconomic assessment of Torsemide and Furosemide in the treatment of patients with congestive heart failure. Clinical Therapeutics 1999; 21(5):854-866.
16 Cohen-Solal A, Caviezel B, Laperche T, Gourgon R. Analyse critique des échelles de qualité de vie en cardiologie; applications à l'ínsuffisance cardiaque. Arch Mal Coeur 1994; 87(IV):71-77.
17 Metra M, Nodari S, Garbellini M, Boldi E, Rosselli F, Milan E et al. [The effects of mid- and long-term administration (3-4 years) of carvedilol in patients with idiopathic dilated cardiomyopathy]. Cardiologia 1997; 42(5):503-512.
18 Jenkinson C, Layte R, Lawrence K. Development and testing of the Medical Outcomes Study 36-Item Short Form Health Survey summary scale scores in the United Kingdom. Results from a large-scale survey and a clinical trial. Med Care 1997; 35(4):410-416.
19 McHorney CA, Ware JEJ, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993; 31(3):247-263.
20 Jenkinson C, Coulter A, Wright L. Short form 36 (SF36) health survey questionnaire: normative data for adults of working age [see comments]. BMJ 1993; 306(6890):1437-1440.
21 Van der Zee K, Sanderman R, Heyink JW, De Haes H. Psychometric qualities of the RAND 36-item Health Survey 1.0: a multidimensional measure of general health status. Int J Behav Med 1996;(3):104-122.
22 Van der Zee K, Sanderman R, Heyink JW. [The psychometric properties of the SF-36 in a Dutch population] De psychometrische kwaliteiten van de MOS Short Form health Survey (SF-36) in een Nederlandse populatie. Tijdschrift voor Sociale Gezondheidszorg 1993; 71:183-191.
23 Van der Zee K, Sanderman R. [The assessment of general health status with the RAND-36]Het meten van de algemene gezondheidstoestand met de RAND-36, een handleiding. 1993. Noordelijk Centrum voor Gezondheidsvraagstukken.
24 Cronbach LJ. Coefficient Alpha and the internal structure of tests. Psychometrica 1951; 16:297-334.
25 Cohen J. Statistical power analysis for the behavioural sciences. revised edition ed. New York: Academic Press, 1977.

26 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.
27 Vulink NCC, Overgaauw DM, Jessurun GA, Ten Vaarwerk IAM, Kropman TJB, Van der Schans CP et al. The effects of spinal cord stimulation on quality of life in patients with therapeutically chronic refractory angina pectoris. Neuromodulation 1999; 2:33-40.
28 Cohen J. A Power Primer. Psychological Bulletin 1992; 112(1):155-159.
29 Crocker L, Algina J. Introduction to classical and modern test theory. New York: Holt,Rinehart and Winston, 1986.
30 New York Heart Association. Diseases of the heart and blood vessels: Nomenclature and criteria for diagnosis. 1964. Boston, Little Brown. Ref Type: Report
31 Lipsey MW. Design sensitivity. Statistical power for experimental research. SAGE Publications,London., 1990.
32 Nunnally JC. Psychometric Theory. 2nd ed. New York: Mc Graw Hill, 1978.
33 Kiers HAL. SCA, a program for simultaneous components analysis of variables measured in two or more populations. (Manual). 1990. ProGAMMA, University of Groningen.
34 Ten Berge JMF. Rotation to perfect congruence and cross-validation of component weights across populations. Multivariate Behavioural Research 1986; 21:262-266.
35 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 1959; 56:81-105.
36 Teresi JA, Golden RR, Gurland BJ, Wilder DE, Benett RG. Construct validity of indicator-scales developed from the comprehensive assessment and referral evaluation interview schedule. Journal of Gerontology 1984; 39(2):147-157.
37 Varni JW, Seid M, Rode CA. The PedsQL™: measurement model for the pediatric quality of life inventory. Medical Care 1999; 37(2):126-139.
38 Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Disease 1985; 38(1):27-36.

4

How to interpret the magnitude of change in health-related quality of life? A study on the use of Cohen's thresholds for effect size estimates.

Berrie Middel, MSc., Eric van Sonderen, Ph.D.

Submitted

ABSTRACT

This paper aims to identify problems in the evaluation of the magnitude of treatment-related change, or responsiveness of health status or health-related quality of life instruments, which are induced by standardizing change over time with the standard deviation of the difference score. This effect size is widely used and is represented as the Standardised Response Mean (SRM), and interpretation is problematic when it is used to estimate the magnitude of change over time with Cohen's rule of thumb for effect size (ES) which is based on standardisation with the pooled standard deviation. In the case of standardizing mean change with the SD of that change, application of the well-known cut-off points for pooled standard deviation units ('trivial' (ES < .20), 'small' (ES ≥ .20 <. 50), 'moderate' (ES ≥ .50 < .80), or large (ES ≥. .80) may lead to over- or underestimation of the magnitude of change over time due to the correlation between assessments.

.

Keywords: Responsiveness, Health Status, Sensitivity to change, Methodology, Effect size, Standardised Response Mean

4.1 INTRODUCTION

In the practice of health-related quality of life research, most researchers remain primarily interested in the statistical significance of the change in health-related functional status or quality of life in pre post designs. In combination with, e.g., the T-test approach, substantial effects can be detected [1-3] with an estimate of effect size. If a p-value is annotated as statistically significant, rejecting the null hypothesis does not imply an effect of important magnitude; likewise, a non-significant **p**-value does not indicate a trivial result, [4-7] although some researchers implicitly deem more important those results with smaller p-values.

In the last decade, however, a growing number of longitudinal intervention studies are focussed on questions like "If the change between baseline and outcome is statistically significant, what can we say about the magnitude (or amount) of change over time that has been detected? Can we interpret this difference in terms of an important difference or as a relevant (substantial) change?" To answer these questions, the responsiveness, i.e. the ability of quality of life outcome measures to detect change over time, has become crucial in the past decade. However, the responsiveness estimation is neglected in many clinical studies in which it could give information on the importance of change due to treatment effects supplementary to the statistical significance of change over time (e.g. before and after intervention) [8,9] Reporting effect sizes without appropriate statistical tests and associated p-values is misleading and potentially dangerous when the number of observations that is required to detect a difference has not been estimated with a power analysis. Effect size statistic should be provided to supplement (not as a substitute for) statistical testing, and only then, when the outcome is sufficiently extreme from what would have been expected on the basis of chance ($p < \alpha$).

Noteworthy in this respect is that in the field of psychological research, editorial policy indicates that "until there is a real impediment to doing so, authors should routinely present an effect size estimate along with the outcome of a significance test". [10,11]

Several quantitative indices have been developed [10-20] that belong to this family of effect sizes or standardized differences, each calculated with a different denominator in the

($\bar{X}1 - \bar{X}2$ /SD) formula, namely the SD of stable subjects, the SD of the baseline assessment, the SD of the observed difference score and the pooled standard deviation (SDp). Obviously, there is no consensus on how to declare a difference in terms of standard deviation units. Only in a small number of publications is this lack of consensus on the most appropriate effect size indicator signalled. [21-25] Despite the fact that different opinions exist on the method to estimate magnitude of difference

between groups or the magnitude of change within groups, researchers use the straitjacket of thresholds Cohen provided us with some 30 years ago. [26] However, these thresholds are taken for granted by many researchers for every version of effect size index. With regard to the correct use and interpretation of effect size indices as estimates of treatment related magnitude of change, we must revisit some basic assumptions:

1. the ES is developed and elaborated by Cohen to estimate power or the necessary sample size to detect relevant change with the basic principle of independent, equal size samples with common within-population standard deviation σ ;
2. in the case that this ES is used in paired samples or in a repeated measurement-design it must be adjusted for correct use of power tables and sample size tables;

4.1.1. Independent samples

Cohen represented the effect size (ES) on some dependent or outcome measure used in an experiment in terms of the difference (using the symbol d' to denote this ES) between the treatment and control group expressed in units of common within-population standard deviation (in samples this standard deviation is estimated with the pooled standard deviation) as follows:

[A]

$$ES = d' \quad \frac{\left(\bar{X}_{treatment} - \bar{X}_{control}\right)}{\sigma}$$

With this estimate of effect size, after analysing a wide sampling of behavioural research, Cohen developed his rules of thumb and reported that effect of .8σ being on the large end of the range, .5 σ was the medium, and .2 σ was at the small end of the range. [27]

4.1.2. Dependent samples or paired observations

The difference or change in matched observations within subjects is standardized by the common within-population σ, according to Cohen's [1977,p.13], but due to the removal of the variation in many extraneous characteristics of the subjects, the index must be adjusted (see appendix), dividing d' by $\sqrt{(1-r)}$. Cohen used the symbol d to denote this adjusted ES. [1]

1 As we will demonstrate, the effect size d is equivalent to the Stadardised Response Mean (SRM), i.e. mean change or difference divided by the standard deviation of that change of difference (see appendix)

[B]

$$d = \frac{d'}{\sqrt{(1-r)}}$$

d' = effect size for independent samples
d = adjusted effect size
r =correlation between baseline and outcome

This √(1-r) – correction of the denominator of formula A is necessary for a proper use of power and sample size tables since these assume 2(n-1) degrees of freedom where, in the case of paired observations, only n-1 are actually available. [26] This consequence for power and sample size estimation is something different from the use of the effect size d in evaluating efficacy of a new treatment in terms of amount of change in health status, which was not the aim of Cohen's work. Therefore, we did not abstract data from effect size estimates of health-related quality of life scales when they were used for the sole purpose of power analysis to draw conclusions from the results of the statistical analysis, or to answer the question whether the investigators had sufficient sample size to allow the detection of a relevant difference. [10,28,29]

Effect Size as an evaluative indicator of magnitude of difference inhealth-related functional status: Independent samples versus repeated measures

When effect sizes are calculated as the standardized difference in mean score to evaluate the efficacy of a new treatment with the use of Cohen's thresholds, for example between a treatment group and a control group, formula [A] should be used. The effect size can be calculated by pooling the estimates (pooled standard deviation) derived from sample data. In contrast to this independent sample case, effect sizes are also used in evaluation studies (pre- post study designs) as estimates of the responsiveness of (for example) a new outcome measure. Effects are often used to give meaning to change over time in terms of 'trivial' (ES < .20), 'small' (ES ≥ .20 < .50) ,'moderate' (ES ≥ .50 < .80) or 'large' (ES ≥ .80) change. Cohen [26] introduced this 'matched pairs' effect size (see appendix equation A2), which was later renamed the standardised response mean (SRM) by Liang et al. [28] to avoid confusion concerning other effect size indices. However, several researchers seem to have adopted the idea that **every** standardized difference is subject to Cohen's definitions of trivial, small, moderate and large effect. Such a belief could lead to

misinterpretations in studies focussing on treatment-related outcome in paired samples since these cut-off points of the magnitude of the difference were not established as a rule of thumb with the effect size d (dependent samples) but with the index d' (independent samples). Thus we argue that Cohen's thresholds are based on the assumption of common within-standard deviation (with matched pairs sample data we use the raw within-group pooled SD), resulting in an effect size we annotate as ES_P. Consequently, in matched pairs studies these thresholds cannot be used interchangeably for the SRM due to the role of the correlation between repeated measures or paired samples. In this article the attention is focussed on the standardized change in mean score between two points in time **within** a single group, estimated with the within-group effect size. In relation to the use of Cohen's rule of thumb for effect size interpretation, we evaluate the consequences of the calibration of the SRM with the ES_P and the role of the correlation between pre and post test scores.

To investigate how serious discrepancies can appear in effect size interpretation we first elaborate a theoretical example and used a sample of studies to evaluate the seriousness of these differences in practice. To evaluate the seriousness of the discrepancies between SRM and ES_P, the correlation of the subject's repeated measurements was needed. Empirical data were collected for the purpose of secondary analysis to draw conclusions in terms of the relative size of the SRM to the ES_P in relation to the size of the correlation. Applying Cohen's thresholds, which are based on the pooled estimate of effect, to interpret the SRM on the one hand may lead to similar results or subtle and trivial differences, but on the other hand also to meaningful shifts in classification of the amount of estimated change. In this article we analysed 148 SRMs interpreted using Cohen's rule of thumb and compared these SRMs with Cohen's ES_P from which these thresholds were derived. Furthermore, we calculated for the range of the correlation coefficient 0.01 to 0.99 the SRM adjusted for Cohen's cut-off points 0.20, 0.50 and 0.80 of the pooled effect size.

4.2 MATERIALS AND METHODS

To study the consequences of the impact of the association or correlation between repeated measures, we restrict the analysis to two effect size indices suitable for the evaluation and interpretation of magnitude of change over time (or responsiveness) within one group, namely the SRM and the ES_P. In this study we use the pooled SD

(SDP) as the standardizing unit (denominator) of mean change score over time (nominator) to calculate the effect size (ES_P)[2].

[C]

$$ES_p = \frac{\overline{X}_{change}}{SD_{(pooled)}}$$

The ES_P introduced by Cohen was made comparable to the SRM where the $SD_{(X\text{-}change)}$ is used as the denominator in which, as we will demonstrate below, the correlation between baseline and outcome scores is involved.

The SRM is the ratio between the mean change score and the variability (the standard deviation) of that change score within the same group.

[D]

$$SRM = \frac{\overline{X}_{change}}{SD_{(Xchange)}}$$

The relationship between ES_P (d') and SRM (d) and the correlationbetween baseline and outcome scores

One of our purposes was to get an indication of how the SRM varies in accordance with the size of the correlation between pre and post test scores when the correct pooled effect size estimate is used. An example may illustrate the role of r, the correlation of a person's health status measurements over time: In a study in which the outcome of a medical intervention was evaluated with a health-related quality of life measure, and in the case of improvement, a lower mean score after intervention was hypothesized. The investigator finds at baseline a mean score of 11.12 with a standard deviation of 4.43 and a mean score of 9.16 (SD: 4.88) at follow up. The estimate of the common within-standard deviation, which is the square root of $(SD_{baseline})^2 + (SD_{outcome})^2 / 2$), thus 4.66, and the pooled effect size (ES_P) is then 0.420 (11.12 – 9.16 /4.66). Before we compare the ES_P and SRM in relation to the correlation between repeated measurements, we must solve the problem of the

[2] $\text{effect size} = \frac{\left(\overline{X}_{baseline} - \overline{X}_{outcome}\right)}{\text{pooled SD}}$ where $\text{pooled SD} = \sqrt{\frac{(SD_{baseline})^2 + (SD_{outcome})^2}{2}}$ for $N_{baseline} = N_{outcome}$

equation of both formulas C and D. According to Cohen, the difference between means for **dependent** samples is standardised by a value "which is √2 (1-r) as large as would be the case were they independent". Cohen 1977, p.49
From equation A4 in the appendix, (d'/√2) / √(1-r) is equivalent to the SRM and alternatively SRM * √ 2 * √(1-r) is equivalent to d' and both indices will vary with the size of r . In table 4.1 we have elaborated the hypothetical example in which the effect size ES_P (d') = 0.42, is transformed into the SRM for a series of values of r. Both effect sizes are equal in the case that r = 0.50): ES_P = (0.42/√2) / √(1-0.50) = SRM, and the SRM for r .50 is then (0.42/1.41) / 0.71 = 0.42. In table 7.1 it is shown that the SRM gets larger for larger values of r. For example, an effect size of 0.42 indicating ' small effect' corresponds with a 'medium effect' (SRM = 0.50) if the correlation between the repeated measurements is approximately .64. This small effect estimated with the ES_P corresponds with a 'large effect' (SRM ≥ .80) if this correlation is approximately .86.

Table 4.1 *The conversion of an effect size calculated with the pooled SD (ES_P) of 0.42 into a SRM with correlation coefficients ranging from .00 - .90*

corr.	.00	0.10	.20	.30	.40	.50	.60	.65	.70	.80	.86	.90
$\frac{(.42/\sqrt{2})}{\sqrt{(1-r)}}$	.297	.313	.332	.355	.384	.420	.470	.502	.543	.664	.794	.940

4.2.2. The sample of studies

In examining the role of the correlation in the estimation of a within-group effect size index, we searched Medline and Psyclit for the years 1984-1999 and Current Contents for 1996 and 2000. We searched for studies with key words 'quality of life', 'health status', and 'questionnaire' and articles were scanned for the terms 'responsiveness', or 'sensitivity to change'.
The primary selection consisted of 151 publications and showed the existence of differing opinions about the appropriateness of the effect size as originally proposed by Cohen, which has led to the introduction of new methods of estimating the magnitude of change assessed over time. Due to the variation in the definition of the mean change scores in the nominator and in the definition of the standard deviation in the denominator of the effect size index formula, not all the studies were appropriate for this study. Therefore, 29 studies suitable for our analysis were selected using the following inclusion criteria:

1. the research should encompass repeated (self reported) health outcome assessments evaluating change within one group (paired observations);

2. a SRM must be represented accompanied with Cohen's thresholds;
3. health outcome must be assessed by (self reported) questionnaires with a disease specific or general mode.

Some other causes of the large reduction to 29 studies appropriate for the purpose of this article are:

- standardised Response Means were used without referring to Cohen's thresholds;
- missing information inhibited us from calculating the effect sizes needed;
- the topics of responsiveness and sensitivity to change were discussed purely from methodological or statistical perspective.

4.3 RESULTS

The original sample comprised 142 or scales belonging to scattered dimensions of health-related quality of life or health status measures. The current selection of 29 papers was determined by the condition that, a SRM had to be represented with referral to Cohen's thresholds for interpretation of change magnitude. [1,21,22,28,30-35,35-53] These 29 publications comprised 411 Standardised Response Mean indices which sizes were interpreted with referring to Cohen's thresholds. From this sample of SRM indices, 148 were published together with sufficient information to estimate the ES_P calculated with an estimated correlation coefficient (r) (see Appendix equation A5). The correlation coefficient is needed to compare the SRM that was shown with reference to Cohen's thresholds, with the ES_P as the correct yardstick for SRM interpretation. Additionally, 263 Standardised Response Means were detected with the investigator's reference to Cohen's rule of thumb (which is derived from pooled estimates of standard deviation) but unfortunately, not sufficient information was given to estimate the correlation between assessments.

4.3.1. The classification of treatment effect with the SRM with Cohen's thresholds for ES_P

In the interpretation of these two effect size indices, the thresholds proposed by Cohen [26-54] as operational definitions of magnitude of change cannot be used interchangeably as is generally assumed. Table 4.2 summarizes the results of the SRM's substituted into ES_P 's (using equation A4 in the appendix). It shows clearly that mistakes can be made in the classification of the magnitude of detected change in health-related quality of life if Cohen's rule of thumb of the ES_P is assumed in the interpretation of the SRM. In 148 SRM's that were adjusted, the magnitude of the

correlation coefficient caused no change in classification in 77%. Approximately, 34 of the 148 estimated effect sizes (23 percent) did not fall in the category indicating the same magnitude of change. Underestimation of effect size according to Cohen's thresholds occurred in 5 SRM's (3.4%), whereas 29 SRM's (19.6%) were overestimates of effect size.

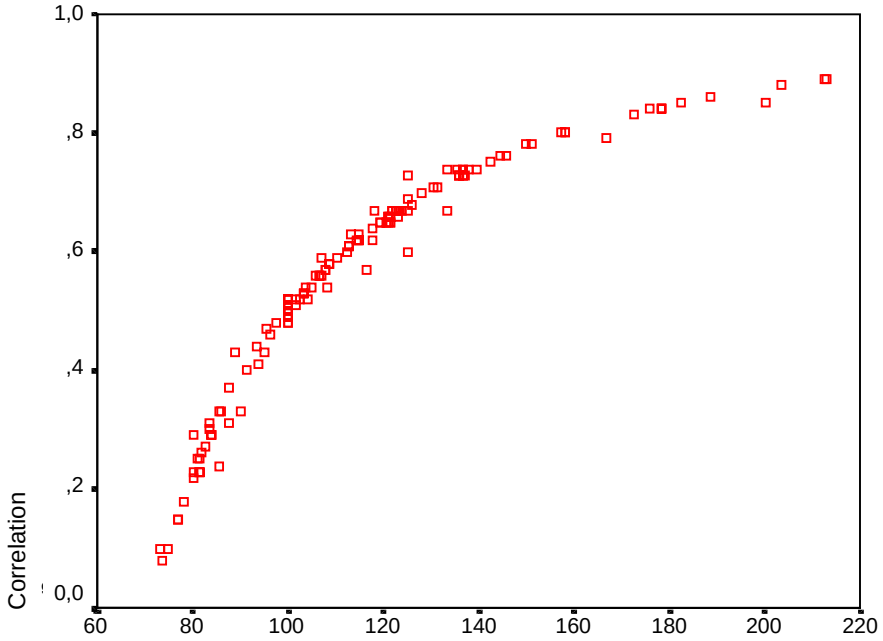
Table 4.2 *Similarities and differences between the Standardised Response Mean (SRM) and pooled effect size (ES_P) interpreted using Cohen's thresholds (N=411: 148 + 263)*

Calculated by equation A4[1]						Equation A4[2] not applicable
ES_{pooled}	ES < .20 Trivial effect	ES ≥ .20< .50 Small effect	ES ≥50< .80 Medium effect	ES ≥ .80 Large effect	Total	
SRM						
< .20	43	2			45	33
≥.20 <. 50	6	35	2		43	92
≥50 <. 80		11	13	1	25	69
≥ .80			12	23	35	69
total	49	48	27	24	148	263

[1] See appendix.
[2] SRM used with interpretation according to Cohen's thresholds for ES_{pooled}

To get a better understanding of the role of the calculated correlation between baseline and follow-up score in the relationship between these two effect size indices, we have, for these 148 estimates, expressed the SRM as the percentage of the value of the ES_P . In figure 4.1 it is shown that the SRM covers the ES_P 100% at the x-axis with the calculated r = .50 at the y-axis for each of the instrument scales of which the pre-post test correlation r was recalculated. The depicted curve shows, irrespective of the values of the effect sizes estimated in our sample of health-related quality of life scales, that the relative distance from the SRM to the ES_P varies with the size of the baseline-follow-up correlation.

Figure 4.1 *The relationship between the correlation and the relative ratio of the SRM and ES_P (N=148)*

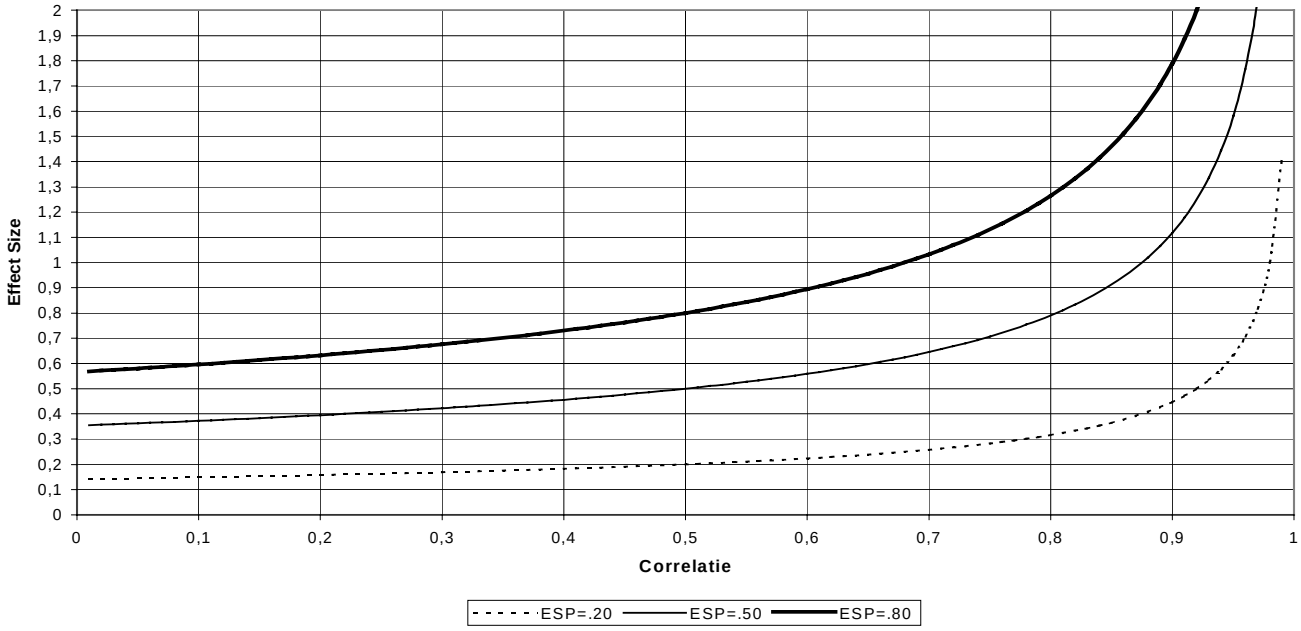


(SRM / ES_P * 100)

To avoid invalid interpretations in the evaluation of responsiveness with SRM index we have, for every value of the correlation between baseline and follow-up score, calculated the corresponding ES_P 's for Cohen's thresholds of .20 = small, .50 = medium, and .80 = large. Indices that lie within the interval that corresponds with these thresholds are not depicted. To classify the magnitude of change estimated with the SRM more precisely, this effect size index is adjusted for every value of the correlation coefficient (r) between baseline and follow-up assessments and brought into line with Cohen's thresholds for effect size. Figure 4.2 shows that SRM' s of 0.20, 0.50 and 0.80, don't deviate after calibration with Cohen's ES_P taken as the original standard, when r =.50. A SRM of 0.20 must be tagged as trivial effect as long as the correlation coefficient ranges from r=.01 to r=.49. With large corresponding correlation coefficients a small SRM of 0.20 must be tagged as moderate (.20/√2 / √1- .92 = .50) or large (.20/√2 / √1-.97 = .80) The class midpoint 0.35 of the 'small

effect' range of effect (not depicted) has to be classified as moderate or large effect with correlation coefficients of 0.76 (.35/√2 / √1-.76 = .50) and 0.91 (.35/√2 / √1-.91 = .80) respectively.
SRM's of 0.80 has to be tagged as 'moderate' effect if the correlation ranges from r = 0.01 to 0.49. The SRM ≥ 0.80 cannot drop below the cut-off points of small and trivial due to the correlation magnitude between baseline and outcome measurements. 'Moderate' effect (SRM = 0.50) must be tagged as 'small' if the correlation between repeated measures is below 0.49 and has to be classified as 'large' in case of r = .81. The class midpoint 0.65 (not depicted) of the 'moderate effect' range of effect must be valued as 'small' with a r = 0.14 (.65/√2/√1-.14 = .49).
In contrast with the fixed threshold values .20, .50 and .80 in figure 4.2, in the analysis of 148 effect size estimates from which the correlation of a person's health status measurements over time was calculated, we found SRM values ranging from 0.04 to 2.42. Correlation coefficients ranged from .08 to 0.89 and 70% of the 148 coefficients were larger than 0.50. Overestimates of effect size (see table 4.2) are not depicted in figure 2, but are easily estimated. For example A SRM of 0.85 interpreted by the researcher as large effect, changed into a moderate effect according to Cohen's thresholds, due to a correlation of 0.12 between repeated measurements

Figure 4.2 *Cohen's threshold's for effect size SRM corrected for the size of the correllation coefficent between repeated measurements*



(

4.4 DISCUSSION

The values used in effect size classification for difference between means as small, medium, and large was arbitrary but seemed reasonable, Cohen stated some 30 years ago. In the debate over which standardizing unit of the difference one should take in a within- group situation, we propose that estimating the magnitude of change by using either the SD of the change score or the pooled SD is preferable to the use of the SD at baseline as proposed by Kazis et al., [12] although the SRM must be adjusted to make correct use of Cohen's thresholds when magnitude of change over time is estimated in evaluation research. These thresholds of Cohen are now being cited without distinguishing between the units by which the assessed change over time is standardized. This is surprising since there is unequivocally no doubt that his rule of thumb was derived from the pooled SD as the estimate of the common within variance. Moreover, routine action in calculating effect sizes may have led to a reduced awareness of factors originally considered only in the calculation of power and sample size. For instance, the calculation of power of the detected change or difference without using the information of r can lead to the wrong inferences. Cohen, p. 50

In evaluation research on treatment-related quality of life, researchers seem to overlook the fact that, in assessing change over time within one subject, the experimental technique of 'self-matching' reduces the proportion of the total variance due to extraneous variables not related to the treatment or intervention per se. [55]

We may conclude that the rule of thumb proposed by Cohen can induce differences in the interpretation of the size of estimated effects. At present it does not appear to us that a single set of rules that are unequivocal or normative at some level is available. We have begun to explore alternative methods in effect size estimation and have assessed the interrelation between two effect sizes as estimates of magnitude of change over time within groups. As we have demonstrated, errors can easily be made and different interpretations of the magnitude of detected change may occur. In analysing the data from our sample of published studies on change over time in health-related quality of life, we saw meaningful shifts in magnitude of detected change in relation to the size of the correlation between pre- and post-test scores. In this article we have attempted to draw the attention to the problem of over- or underestimation of effect sizes when the Standardized Response Mean is used. Studies in which the mean change over time is standardized with the SDbaseline according to Kazis et al. [12] should report the ES_P to show that the results were not dependent on the choice of deominator in the d-index formula.

Due to their increasing appearance, it is important that all aspects of estimating the

magnitude of change be inspected. One of these aspects is the consequence of the hidden role of the correlation coefficient between repeated measurements, which increases the risk of incorrect conclusions. This initial effort may provide a moderate step toward the development of a precise and useful index in quality of life assessment in clinical trials.

Acknowledgements

Appreciation is expressed to Drs. Roy Stewart for providing valuable assistance with several aspects of the analysis, and a critical review of the manuscript. Prof.dr. Wim van den Heuvel and dr. Mike de Jongste provide helpful reviews of the manuscript.

APPENDIX

Given Cohen's formula 1 for the Effect Size index for means from matched samples:

$$(A1) \quad d_z' = \frac{m_z}{\sigma_z} = SRM$$

where:

$$\sigma z = \sigma(X_{baseline} - X_{outcome}) = \sqrt{(\sigma_{xbaseline})^2 + (\sigma_{xoutcome})^2 - 2r\sigma_{xbaseline}\sigma_{xoutcome}}$$

and assumed equal variance, i.e.:

$$\sigma x^2_{baseline} = \sigma x^2_{outcome} = \sigma x^2$$

(A2) gives:

$$\sigma_z = \sigma(x_{baseline} - x_{outcome}) = \sqrt{2\sigma^2 - 2r\sigma^2} = \sigma\sqrt{2(1-r)}$$

and for the Effect Size index for means of independent samples the standardizing unit is:

$$(A3) \quad d4' = \frac{m_{xbaseline} - m_{xoutcome}}{\sigma_P} = ES_P$$

where:

$$SD_P = \sqrt{\frac{(\sigma x_{baseline})^2 + (\sigma x_{outcome})^2}{2}} \qquad \text{for : } N_{baseline} = N_{outcome}$$

Now from equation A2 and A3 we use the difference between the standardizing unit for difference in means for matched samples (SRM) being $\sigma\sqrt{2(1-r)}/\sigma = \sqrt{2(1-r)}$ as large as would be in the case of independent samples (ES_P) [26, p.48-52]. Now we can subsitute SRM into ES_P by:

(A4)

$$SRM = d'_z = \frac{mz}{\sigma} \qquad ES_P = d_4' \frac{m_{baseline} - m_{outcome}}{\sigma}$$

$$d = d'_z \times \sqrt{2} \qquad d = \frac{d_4'}{\sqrt{(1-r)}}$$

$$d = SRM \times \sqrt{2} \qquad d = \frac{ES_P}{\sqrt{(1-r)}}$$

$$SRM \times \sqrt{2} = \frac{ES_P}{\sqrt{(1-r)}} \left\{ \text{with } r = 0 : SRM \times \sqrt{2} = ES_P \right\}$$

$$SRM \times \sqrt{2} \times \sqrt{(1-r)} = ES_P \quad \text{and} \quad \left(ES_P/\sqrt{2}\right)/\sqrt{(1-r)} = SRM$$

we note that r is estimated in cases in which the standard deviation at baseline, outcome, as well as the standard deviation of the difference or change score were published:

(A5)

$$\sigma_{change} = \sqrt{(\sigma_{baseline})^2 + (\sigma_{outcome})^2 - 2r\sigma x_{baseline}\sigma x_{outcome}}$$

$$\sigma_{change}{}^2 = (\sigma_{baseline})^2 + (\sigma_{outcome})^2 - 2r\sigma x_{baseline}\sigma x_{outcome}$$

$$\sigma_{change}{}^2 - (\sigma_{baseline})^2 - (\sigma_{outcome})^2 = -2r\sigma x_{baseline}\sigma x_{outcome}$$

$$2r = \frac{(\sigma_{baseline})^2 + (\sigma_{outcome})^2 - (\sigma_{change})^2}{(\sigma x_{baseline})(\sigma x_{outcome})} \approx r = 1/2 \frac{(SD_1)^2 + (SD_2)^2 - (SDchange)^2}{(SD_1)(SD_2)}$$

REFERENCES

1. Leon AC, Shear K, Portera L, Klerman GL. Effect Size as a Measure of Symptom-Specific Drug Change in Clinical Trials. Psychopharmacology Bulletin 1993;29(2):163-7.
2. Pulver AE, Bartko JJ, McGrath JA. The Power of Analysis: Statistical Perspectives. Part 1. Psychiatry Research 1988;23:295-9.
3. Brewer JK. Effect Size: The most troublesome of the hypothesis testing considerations. CEDR Quarterly 1978;11(4):7-10.
4. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. American Psychologist 1989;44:1276-84.
5. Rosenthal R. Progress in clinical psychology: Is there any? Clinical Psychology: Science and Practice 1995;2:133-50.
6. Rosenthal R, Rubin DB. The Counternull value of an effect size: a new statistic. Psychological Science 1994;5(6):329-34.
7. Bartko JJ, Pulver AE, Carpenter WT. The Power of Analysis: Statistical Perspectives. Part 2. Psychiatry Research 1988;23:301-9.
8. Borenstein M. A Note on the use of confidence intervals in psychiatric research. Psychopharmacology Bulletin 1994;30(2):235-8.
9. Cooper HM. On the significance of effects and the effects of significance. Journal of Personality and Social Psychology 1981;41(5):1013-8.
10. Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. J.Clinical Epidemiology 1991;44(4/5):417-21.
11. Thompson B. Editorial policies regarding statistical significance tests: Futher comments. Educ.Res. 1997;26(5):29-32.
12. Kazis LE, Anderson JJ, Meenan RF. Effect Sizes for Interpreting Changes in Health Status. Medical Care 1989;27(3,Supplement):S178-S189
13. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal Chron.Dis. 1987;40(2):171-8.
14. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control.Clin.Trials. 1991;12(4 Suppl):142S-58S.
15. Pfennings LEMA, van der Ploeg HM, Cohen L, Polman CH. A comparison of responsiveness indices in multiple sclerosis patients. Qual.Life Res. 1999;8:481-9.
16. Wright JG and Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50(3):239-46.
17. Hillers ThK, Guyatt GH, Oldridge N, Crowe J, Willan A, Griffith L, Feeny D. Quality of life after myocardial infarction. Journal of Clinical Epidemiology 1994;47(11):1287-96.

18. 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(5):683-91.
19. Taylor S, Woody S, McLean PD, Koch WJ. Sensitivity of outcome measures for treatments of generalized social phobia. Assessment 1997;4(2):181-91.
20. Wiebe S, Rose K, Derry P, McLachlan R. Outcome assessment in epilepsy: comparative responsiveness of quality of life and psychosocial instruments. Epilepsia 1997;38(4):430-8.
21. Beurskens AJHM, de Vet HCW, Koke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71-6.
22. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the Rehabilitation Activities Profile and the Barthel Index. Journal of Cliical Epidemiology 1996;49(1):39-44.
23. Lachs MS. The more things change... Journal of Clinical Epidemiology 1993;46(10):1091-2.
24. 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.
25. Kraemer HC. Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. Psychoneuroendocrinology 1992;17(6):527-36.
26. Cohen J. Statistical power analysis for the behavioural sciences. revised edition ed. New York: Academic Press; 1977.
27. Lipsey MW. Design sensitivity. Statistical power for experimental research. SAGE Publications,London.; 1990.
28. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Medical Care 1990;28(7):632-42.
29. Diehr P, Psaty BM, Patrick DL. Effect size and power for clinical trials that measure years of healthy life. Stat.Med. 1997;16(11):1211-23.
30. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Medical Care 1992;30(10):917-25.
31. 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;(5):223-34.
32. Jacobs HM, Touw-Otten FWMM, de Melker RA. The Evaluation of Changes in Functional Health Status in Patients with Abdominal Complaints. Journal of Clinical Epidemiology 1996;49(2):163-71.
33. Doeglas D, Krol B, Guillemin F, Suurmeijer Th, Sanderman R, Smedstad LM, van den Heuvel WJA. The Asessment 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(10):1834-43.

34. Bouchet C, Guillemin F, Briancon S. [Comparison of 3 quality of life instruments in the longitudinal study of rheumatoid arthritis] Comparaison de trois instruments de qualité de vie pour l'etude longitudinale de la polyarthrite rheumatoide. Rev.Epidemiol.Sante.Publique. 1995;43(3):250-8.
35. Koes BW. Efficacy of manual therapy and physiotherapy for back and neck complaints. (dissertation). Maastricht: University of Limburg; 1992.
36. Vliet-Vlieland ThPM, Zwinderman AH, Breedveld FC, Hazes JMW. Measurement of morning stiffness in rheumatoid arthritis clinical trials. J Clin Epidemiol 1997;50(7):757-63.
37. Husted JA, Cook RJ, Farewell VTGDD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459-68.
38. Middel B, Kuipers-Upmeijer H, Bouma J, Staal MJ, Oenema D, Postma Th, Terpstra S, Stewart R. 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;63:204-9.
39. 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.
40. Wells G, Boers M, Shea B, Tugwell P, Westhovens R, Saurez-Almazor M, Buchbinder R. 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(1):217-21.
41. O'Carrol RE, Cossar JA, Couston MC, Hayes PC. Sensitivity to change following liver transplantation. A comparison of three instruments that measure quality of life. Journal of Health Psychology 2000;5(1):69-74.
42. Macduff C, Russell E. The problem of measuring change in individual health-related quality of life by postal questionnaire: use of the patient-generated index in a disabled population. Qual.Life Res. 1998;7:761-9.
43. Brunner HI, Feldman BM, Bombardier C, Silverman ED. Sensitivity of the systemic lupus erythematosus disease activity index, british isles lupus assessment group index,and systemic lupus activity measure in the evaluation of clinical change in childhood-onset systemic lupus erythematosus. Arthritis and Rheumatism 1999;42(7):1354-60.
44. Sneeuw KCA, Aaronson NK, Sprangers MAG, Detmar SB, Wever LDV, Schornagel JH. Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. J Clin Epidemiol 1998;51(7):617-31.
45. Vulink NCC, Overgaauw DM, Jessurun GA, Ten Vaarwerk IAM, Kropman TJB, Van der Schans CP, Middel B, Staal MJ, De Jongste MJL. The effects of spinal cord stimulation on quality of life in patients with therapeutically chronic refractory angina pectoris. Neuromodulation 1999;2:33-40.
46. Gliklich RE, Hilinsky JM. Longitudinal sensitivity of generic and specific health measures in chronic sinusitis. Quality of Life Research 1995;4:27-32.
47. Sneeuw KCA, Aaronson NK, Osoba D, Muller MJ, Hsu M-A, Yung AWK, Brada M, Newlands ES. The Use of Significant Others as Proxy Raters of the Quality of Life of Patients with Brain Cancer. Medical Care 1997;35(5):490-506.

48. Vaile JH, Mathers M, Ramos-Remus C, Russel AS. Generic health instruments do not comprehensively capture patient perceived improvements in patients with carpal tunnel syndrome. The Journal of Rheumatology 1999;26(5):1163-6.
49. Bruin AFd, Diederiks JPM, De Witte LP, Stevens FCJ, Philipsen H. Assessing the Responsiveness of a Functional Status Measure: The Sickness Impact Profile Versus the SIP68. J Clin Epidemiol 1997;50(5):529-40.
50. De Witte LP. After the rehabilitation centre; a study into the course of functioning after discharge from rehabilitation. (dissertation). Amsterdam/Lisse: Zwets en Zeitlinger; 1992.
51. Janssen M. Personal Networks of chronic patients. (dissertation). Maastricht: University of Limburg; 1997.
52. Hidding A, Van der Linden Sj, Boers M, Gielen X, Kester A, De Witte LP, Dijkmans B, Moonenburgh J. Is group physical therapy superior to individual therapy in ankylosing spondylitis, A randomized controlled trial. Arthritis Care and Research 1993;6(3):117-25.
53. Courtens AM. Kenmerken van zorg en kwaliteit van leven bij patienten met kanker (Characteristics of Care and Quality of Life in cancer patients) (dissertation). Maastricht: University of Limburg; 1993.
54. Cohen J. A Power Primer. Psychological Bulletin 1992;112(1):155-9.
55. Winer BJ. Statistical principles in experimental design. second ed. Tokyo: McGraw-Hill Kogakusha; 1962.

5

How to validate clinically important change in health-related functional status. Is the magnitude of the effect size consistently related to magnitude of change as indicated by a global question rating?

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

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

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

Submitted

SUMMARY

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

5.1 INTRODUCTION

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

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

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

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

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

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

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

5.1.1. The researcher's perspective

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

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

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

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

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

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

5.1.2. The patient's perspective

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

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

5.2 PATIENTS AND METHODS

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

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

5.2.1. Patient selection

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

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

5.2.2. Data Collection and measures

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

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

5.3 DETERMINATION OF CONCORDANCE BETWEEN TWO INTERVALS OF MAGNITUDE OF CHANGE

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

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

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

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

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

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

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

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

5.3.1. Effect size concordant, according to the external criterion

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

5.3.2 Effect size discordant, overestimation according to the external criterion

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

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

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

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

5.4 RESULTS

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

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

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

Table 5.1 *Sociodemographic Characteristics*

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

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

5.5 ANALYSIS

5.5.1. Item-analysis

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

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

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

5.5.2. Scale analysis

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

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

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

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

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

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

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

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

5.6 DISCUSSION

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

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

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

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

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

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

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

REFERENCES

1. Stockler MR, Osoba D, Goodwin P, Corey P, Tannock IF. Responsiveness to change in health-related quality of life in a randomized clinical trial: A comparison of the Prostate Cancer Specific Quality Of Life Instrument (PROSQOLI) with analogous scales from the EORTC QLQ-C30 and a Trial Specific Module. J Clin Epidemiol 1998;51(2):137-45.
2. Murawski MM, Miederhoff PA. On the generalizability of statistical expressions of health related quality of life instrument responsiveness: a data synthesis. Quality of Life Research 1998;7:11-22.
3. Sneeuw KCA, Aaronson NK, Sprangers MAG, Detmar SB, Wever LDV, Schornagel JH. Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. J Clin Epidemiol 1998;51(7):617-31.
4. Taylor R, Kirby B, Burdon D, Caves R. The assessment of recovery in patients after myocardial infarction using three generic quality-of-life measures. J Cardiopulmonary Rehabil 1998;18:139-44.
5. Wiebe S, Rose K, Derry P, McLachlan R. Outcome assessment in epilepsy: comparative responsiveness of quality of life and psychosocial instruments. Epilepsia 1997;38(4):430-8.
6. Russel MGVM, Pastoor CJ, Brandon S, Rijken J, Engels LGJB, Van der Heijde DMFM, and et al. Validation of the dutch translation of the Inflammatory Bowel Disease Questionnaire (IBDQ): A health related quality of life questionnaire in inflammatory bowel disease. Digestion 1997;58282-8.
7. Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. J.Clinical Epidemiology 1991;44(4/5):417-21.
8. Parkerson GR, Willke RJ, Hays RD. An International Comparison of the reliability and responsiveness of the Duke Health Profile for measuring health-related quality of life of patients treated with Alprostadil for erectile dysfunction. Medical Care 1999;37(1):56-67.
9. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal Chron.Dis. 1987;40(2):171-8.
10. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control.Clin.Trials. 1991;12(4 Suppl):142S-58S.
11. Katz JN, Gelberman RH, Wright EA, Lew RA , Liang MH. Responsiveness of Self-Reported and Objective Measures of Disease Severity in Carpal Tunnel Syndrome. Medical Care 1994;32(11):1127-33.
12. Van der Windt DAWM, Van der Heijden GJMG, De Winter AF, Koes BW, Deville W, Bouter LM. The responsiveness of the Shoulder Disability Questionnaire. Ann Rheum Dis 1998;57:82-7.
13. Beurskens AJHM, de Vet HCW, Koke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71-6.

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

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

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

6

Why don't we ask patients with coronary artery disease directly how much they have changed after treatment? A comparison of retrospective multi-item change scales with serial change in domains of health related functional status.

Berrie Middel, Msc*, Eric van Sonderen, PhD*, Mathieu de Greef, PhD*, Harry, J.G.M. Crijns, MD,PHD+, Mike J.L. de Jongste, MD, PhD+, Roy Stewart, Msc*, Wim J.A. van den Heuvel, PhD*.

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

\+ Department of Cardiology, Thoraxcenter, University Hospital of groningen, The Netherlands

Submitted

ABSTRACT

Purpose
In the literature on the measurement of treatment-related change some authors advocate the use of change scores as the best approach; others prefer the approach to measure change by asking patients how much they have changed after treatment. The aim of this study was to compare these two methods in measurement of changes in identical domains of health –related functional status (HRFS).

Methods.
Analogous to the widely used single global or transition questions which assess the patient's retrospectively perceived change, we modified every item belonging to domains of health of the Minnesota Living with Heart Failure-Questionnaire (MLHF-Q), into transition questions. Subjects were 217 patients with heart failure, who underwent a treatment with known efficacy, were selected.

Results.
The reliability of the scale 'physical functioning' as the composite of change scores, and of the concordant scale composed of identical transition items were estimated and yielded a Cronbach's alpha of 0.86 and 0.92, respectively, whereas the change in 'emotional functioning' and the concordant transition scale yielded identical alphas of 0.76. Factor analysis yielded similar and clearly interpretable dimensions for both methods. The canonical correlation (R_c) between the composite of the change score items and the composite of the concordant transition items that belong to the domain of physical functioning was R_c=.63 ($p < .001$), whereas the combination of the emotional functioning-items yielded a R_c=.48 ($p< .001$) with a percentage of linearly explained variance between the two dimensions (40% and 23%, respectively).

Conclusions.
The patient's retrospective assessment of change after intervention appears to provide reliable and valid information compared to prospective change scores when items are used beloning to validated measures health-status.

6.1 INTRODUCTION

Clinicians and other health professionals have difficulty to find Health-Related Functional Status (HRFS) measures clinical important change that is expected to occur after treatment evaluated in cardiological trials. To evaluate clinically relevant change reliable and valid, the patient's perception or opinion of what constitutes relevant change in health status domains must be captured in a multi-dimensional mode. However, treatment-related change over time is usually based on change or difference scores that are calculated simply by subtracting baseline scores from post-treatment scores. The interpretation of the amount of change over time as clinically relevant depends on a subjective judgement of either clinician or patients as a substitute for a golden standard. Therefore, a widely accepted solution to discriminate between relevant and irrelevant change is the use of so-called 'transition' or 'global questions' as external criterion or standard, by asking the patients retrospectively how much they feel better or worse compared to the situation at baseline.

These transition questions are used in several modes: 1) single, retrospective questions at follow-up about the direction and magnitude of perceived change in general or 2) after treatment in domains of health, e.g. physical, emotional and social functioning, 3) about satisfaction with the change after treatment, 4) perceived degree of change in difficulty to accomplish a specified task or 5) as a serially assessed retrospective global rating to examine incremental perceived change between baseline and follow-up [1-12]. Another reason to use transition questions is to compare the results of assessing change in health-related quality of life with both repeated measurement and with the patient's retrospective perception of change after treatment.

In several studies, the accuracy, precision, reliability and validity of *single* global or transition questions of health have been discussed [6,13-19]. The main disadvantage of a single item of retrospectively perceived change in overall health is that the answer "since the operation my health state has worsened" does not cover domain-specific change in health. We can imagine that the improvement in the domain of physical health is overshadowed by the perception of a worsening in emotional functioning and in such situations a global single transition question is not a valid indicator of change in health status. 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 or reliability cannot be estimated. Therefore, we have good reasons to presume that multiple-item transitional scales tend to be more reliable than single-items [20] and, when these retrospective measures are conceptually identical with the longitudinally assessed change in health domains,

they would also have a better validity [21]. On the one hand, several studies have provided evidence that change in health-status domains estimated with direct transition questions were more accurate when compared with estimates derived from change scores [6,11,16,17]. On the other hand, the direction of prospective or serial change does not always correspond with the content of the single transition question(s) and therefore may miss some contrary changes [11,21,22]. To solve this, it has been suggested that patients should be asked a series of transition judgements about functional limitations related to their disease or to areas of health status such as physical, emotional and social function [10,17]. However, we have not been able to find studies in which retrospective change in domains of health was measured with multi-transition items intended to contribute to the assessment of change in an underlying dimension of health status. Two studies we have found used a set of multiple transition items but without a relationship with an underlying dimension of health [11,23]. No study was found in which an attempt was made to investigate the question whether both methods measure the same dimensions of health status by means of the summed composite of item scores (scales). Also no study was found in which the concordance between the perceived change and change scores, calculated in a before-after design, was evaluated simultaneously in a one-to one relationship with items and with scales. Before we can reasonably claim that multiple item transition scales measure change in health-related functional status (HRFS), we have to demonstrate that this new operational procedure yields substantially the same results as measuring change with the standard repeated measurement procedure. To this end, his paper examines the reliability, the convergent validity and 'known groups' validity of the patient's view of change in domains of health as outcome measures. The following questions are addressed:

1. Do we measure change in the same domains of HRFS if we use multi-item transition scales assessing (retrospective) perceived change compared with (prospective) change assessed with repeated measures?
2. To what extent does concordance exist between health status scales composed of prospective change scores and those composed of retrospective transition scores?
3. How do both instruments compare in their ability to discriminate between groups known to differ on a measure external to the questionnaire (i.e. disappearance of angina pectoris)?

6.2 METHODS

To ensure that change in health status occurred we selected a group of patients undergoing a treatment with known efficacy and selected a disease-specific instrument with known sensitivity to detect change over time [24-26]. The study sample was composed of two groups: 1. patients who, after a diagnostic Coronary Angio Graphy (CAG), were scheduled for an invasive treatment Percutaneous Transluminal Coronary Angioplasty (PTCA) or Coronary Artery Bypass Grafting (CABG) and 2. patients who needed no operative intervention after CAG and were treated with pharmacotherapy.
The aim of the current study was to compare potentially similar strategies of measuring treatment-related change in two well-defined important domains of HRFS: physical functioning, and emotional functioning. Consequently, items that are not purported to measure either the domain of physical or emotional functioning are not used in the comparison. To assess serial change in the domain of physical functioning we used the 8-item scale from a disease-sepcific HRFS instrument (namely the Minnesota Living with Heart Failure Questionnaire (MLHF-Q) [27,28] and 3 items from the MOS-20. [29] To assess serial change in the domain of emotional functioning the 5-item scale of the MLHF-Q was used. To assess perceived change in these domains of health with a direct method we modified the selected items of the MLHF-Q physical functioning scale and MLHF-Q emotional functioning scale and the MOS-20 items into direct questions (transition questions). These multi-item transition scales were administered at post-treatment.

6.2.1. Patient selection

Consecutive patients who, following a Coronary Angiography (CAG), were scheduled for Percuteneous Transluminal Coronary Angioplasty (PTCA) or Coronary Artery Bypass Grafting(CABG) or who needed no operative intervention but were treated with a (modified) pharmaco-therapy, were recruited from January to December 1998 from the Groningen University Hospital, the Martini Hospital, Groningen, and the Weezenlanden Hospital in Zwolle in the Netherlands. Patients with other incapacitating diseases, with cognitive impairments, aged 75 or older, or who did not speak Dutch were excluded. Ethical approval was obtained from the ethics committee at each participating hospital.

After inclusion patients received a mailed questionnaire accompanied by a written informed consent form. The questionnaire was serially administered at baseline and 6 weeks after the decision for non-invasive intervention or 6 weeks after the day a PTCA /CABG-intervention was executed. After the questionnaires were received, they were routinely checked on completeness at baseline as well as at follow-up. If questions or pages had not been filled in, either a copy was sent with a kind request to complete the questions or, in the cases of it being one question, patients were interviewed by telephone. Because the completeness of the questionnaire was monitored by a computer-programme both at baseline and follow up, we effectively reduced the non-response on questions, and consequently, on scales.
To ascertain the assessment of substantial treatment-related change we approached patients treated with interventions with known efficacy, i.e. invasive treatments PTCA or CABG and non-invasive pharmacotherapy. We presumed that at baseline i.e. prior to CAG, both patients and cardiologists had no information about decision concerning either intervention and would not affect the assessment of subjective health and should reduce the risk of floor and ceiling effects. However, this control for potential bias resulted in logistic problems and, six months after the start of the study, we were forced to select patients waiting for outpatient treatment (PTCA) or waiting for hospital admission (CABG) somewhat later after the decision was taken.

6.2.2. Data Collection and measures

The Minnesota Living with Heart Failure Questionnaire (MLHF-Q) is a disease-specific instrument which is composed of 21 items and three scales that measure the following: the physical functioning dimension (8 items), the emotional functioning dimension (5 items) and the overall score on health-related quality of life (21 items). Eight separate items do not assess an underlying dimension of health-related quality of life and therefore were not used for the current paper. These eight items measure several meaningful social and economic impairments that patients relate to their heart failure, although these 'socio-economic' items are used as a part of the overall score [27,30-33]. However, one item from the MLHF-Q had no correlation with the physical functioning scale, as predefined by Rector et al [28] both in a previous Dutch sample [26] and in the current study. Therefore, the item "'did your heart failure prevent you from living as you wanted by making your relating to or doing things with your friends or family difficult?" was skipped for scale construction and not used in further analysis. Finally, both the items from the MLHF-Q and the MOS-20 (10 items) were used in the analysis of the concordance between two methods of measuring change in the domain of physical functioning.

The response options range from "no" (score 0); very little (score 1) to very much (score 5). The total score of, the physical dimension (sub-scale) ranges from 0 to 40, the emotional dimension (sub-scale) ranges from 0 to 25. To investigate whether differences between the MLHF-Q and a generic measure of physical functioning would exist we have extended the questionnaire with 3 items from the MOS-20 [29] but with the response options analogue to the questionnaire's format. These three items had the following format: "Did your heart failure prevent you from living as you wanted **during the last month** by making it difficult for you 1) to bend, stoop or lift light objects?, 2) lift heavy objects, like moving a table? And 3) run at a fast pace? "

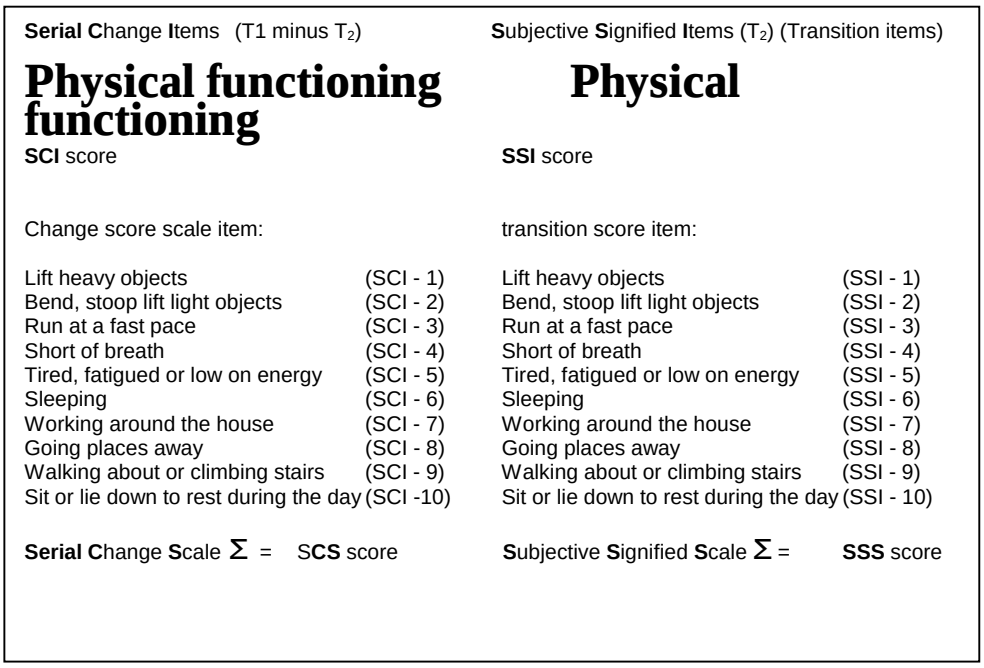
Two methods of assessing change in health-related quality of life (HRQL) with multi-item scales were applied with the study data: The first method, repeated baseline measurement of HRQL-domains with items from the MLHF-Q and MOS-20. In addition, the repeatedly measured battery of questions was transformed in a retrospective 'transition question' mode. Consequently, the patient's perception or subjective significance of change was captured at follow-up of each of the MLHF-Q and MOS-20 items in terms of the extent of feeling improved, deteriorated or not changed. Hence, the degree to which they perceived that change had occurred, on that particular item was rated on a 7-point Likert scale. Following Osoba et al. [15], we have chosen to use the term "subjective significance" because it indicates whose judgement was used to determine the direction and magnitude of change. These transition items are designed to elicit information regarding perceived change over time in specific aspects belonging to domains of HRFS.

Figure 6.1 shows, with the physical functioning items, these two strategies of assessing change in HRFS:

1. The original MLHF-Q items were phrased as follows: "Did your heart failure prevent you from living as you wanted **during the last month** by making your sleeping well at night difficult?" Serial change scores on items (SCI-scores) were calculated by subtracting the follow-up score from the baseline score to get positive numerical change data indicating improvement and negative numerical change data indicating deterioration. A change score of zero was considered to indicate neither improvement nor deterioration. With the summed composite of the SCI scores we constructed the serial change scale (SCS).
2. The Subjective Signified Items (SSI) questionnaire was used to classify patients according to whether they had improved or deteriorated on each item of the questionnaire belonging to the dimension of physical and emotional functioning. The questions were phrased as follows: "Since the last time I filled out the questionnaire (or: since my operation), my problems with walking about or climbing stairs related to my heart failure have become.

3. For every SSI item each patient was asked to circle the answer that best described the perception of the direction and magnitude of change at follow-up on a 7-point Likert scale: 1) a great deal worse; 2) moderately worse; 3) a little worse; 4) no change; 5) a little better; 6) moderately better and 7) a great deal better. With the summed composite of the SSI scores we constructed the Subjective Signified Scale (SSS). Figure 6.1 shows a general representation of prospective and retrospective methods with the physical functioning scale items.

Figure 6.1 *General representation of the prospective and retrospective method of assessing change in physical functioning.*



Functional impairment due to angina was assessed using a set of questions corresponding to the Canadian Cardiovascular Society (CCS), and to the New York Heart Association (NYHA) respectively.

6.2.3. Statistical analysis

To investigate whether the prescribed underlying domains of health of the baseline scale-items by Rector et al. [28] were measured with the same results with the change score items and the transition items, LISREL analysis was used to test the equality of factor structures (principal component analysis). Cronbach's α was used to examine the internal consistency of the transition scales that emerged from these analyses. Canonical correlation analysis was applied as a general procedure for investigating the relationships between two sets of variables. With canonical correlation analysis we transformed the prospective change item scores (PCI scores) from the set of, for example, physical functioning items into a linear combination, which is called the canonical variable. The linear combination, or canonical variable, was also constructed from the concordant set of subjective signified items (SSI scores). These linear combinations were composed so that the correlation between both composed canonical variables was maximised. This correlation is called the canonical correlation (R_c). In other words, this investigated the research question to what extent the set of PCI scores be predicted or 'explained' by the pendant SSI scores. Transformation into z-scores was used to convert different raw scores of the items and scales in both batteries to share the same measurement unit with a mean of 0 and a standard deviation of 1 in order to make comparisons between sub-groups. Data analysis was performed using SPSS for Windows [34], LISREL [35] and SAS [36].

6.3 RESULTS

Sample

The source population consisted of patients who were referred by their general practitioner for a CAG in three hospitals in the northern part of the Netherlands. Of the 398 candidates screened for inclusion in this study, 139 (34.9%) did not return the first questionnaire. A questionnaire was received from the remaining 259 patients. We could not test the probability of systematic differences between non-responders and the study sample because no information was accessible without written informed consent from the patients who did not return the first questionnaire.

42 patients (16.2%) dropped out before the follow-up assessment. The reasons for not responding at follow-up were because the patient died (n= 7), had no heart failure (n=9), refused further participation (n=9), was too ill at follow-up (n=3), had moved (n=3) or did not react at all (n=11). To ensure that the patients who dropped out at follow-up did not deviate systematically from the study group, the characteristics of these patients at the time they returned the first questionnaire were

compared with the baseline characteristics of those who completed the questionnaire at follow-up. Except for educational level (the study sample had a statistically significant higher education), the demographic characteristics of the two groups were similar. This comparison also showed no statistically significant differences in mean scores on baseline health-status scales.
Analyses were based on 217 subjects (83.8%) who filled in the questionnaires at baseline and at post- test.

Demographics
The mean age of the patients was 60.6 (SD 9.43) with a range of 25 to 75. Sixty-one (28%) were female and 156 (72%) were male. Men were more likely to have a partner, to live together with someone, to have a higher education, and to be in employment. Five percent, 44%, 21% and 27%, respectively, had a self-reported NYHA-class I to IV at baseline. At follow-up, sixty-four (29%) had undergone a CABG, 71 (33%) a PTCA, and 82 (38%) were being treated with pharmacotherapy. Additional sample characteristics are presented in Chapter 5, table 5.1

Internal consistency
The homogeneity or unidimensionality of prospective change items and their corresponding transition items was estimated with the internal consistency coefficient (Cronbach's α). The multi-item transition scale 'physical functioning' yielded a somewhat higher internal consistency estimate (Cronbach's $\alpha = 0.92$) than the same scale composed of items' change-score (Cronbach's $\alpha = 0.86$), whereas these coefficients were identical for both versions of the scale 'emotional functioning' (Cronbach's $\alpha = 0.76$)

Convergent validity of prospective and retrospective measures of change
Two different operationalizations to capture the same concept of self-rated health were used. First, the classical repeated measurement change scores, derived from items from the MLHF-Q and the MOS-20. Second, a new set of operations to measure change in HRFS- domain by the means of transition items. We will consider measuring global self rated health with multi-item transition scales and hypothesized that, compaired to the serial change method, the underlying dimensions of physical and emotional functioning could be measured with the same results. One way to answer this question is to look at the concurrent or convergent validity of retrospective transition measures.
To investigate the convergent validity of the two domains assessed with repeated measurement as well with transition questions, we performed a factor comparison

by means of a principal component analysis with baseline items, item's difference scores and pendant transition items.
Table 6.2 presents the results of the principal component analysis (PCA) with items from three item-sets: 1. Items tapping the dimensions 'physical functioning' and 'emotional functioning' assessed at baseline, 2) the change score as the difference between baseline and follow-up for both dimensions (Serial Change Items) and 3) directly assessed change with the same items but modified to measure the perceived direction and magnitude of change at follow-up (Subjective Signified Items). Table 6.2 shows that in each battery, the items that cover the underlying construct of physical functioning have the expected factor loadings, [28,26] which were satisfactory >.50, except for the item with reference to the impact of heart failure on sleeping well at night. In each set of variables, the items that cover the domain of emotional functioning were, without exception, unambiguously related to the expected factor. The factor loadings of the prospective change scores on items were systematically lower than those estimated with the concordant baseline items and transition items, and the factor loadings of both baseline items and concordant transition items were of the same magnitude. Scales composed of multi-item change scores are usually more unreliable than either the baseline or follow-up measures on which they are based. The unreliability of the change score 'typically reduces its correlation with anything, including retrospective assessments' [37,38]. Notwithstanding these differences, the analysis of the three sets of items yielded similar factor loadings (Lambda in LISREL-notation) not only by this elaboration. The factor structure as prescribed by Rector et al. was confirmed by the baseline measure and the similarity between the factor structures of each mode of measurement series was shown by the overlap of the 95% confidence intervals.

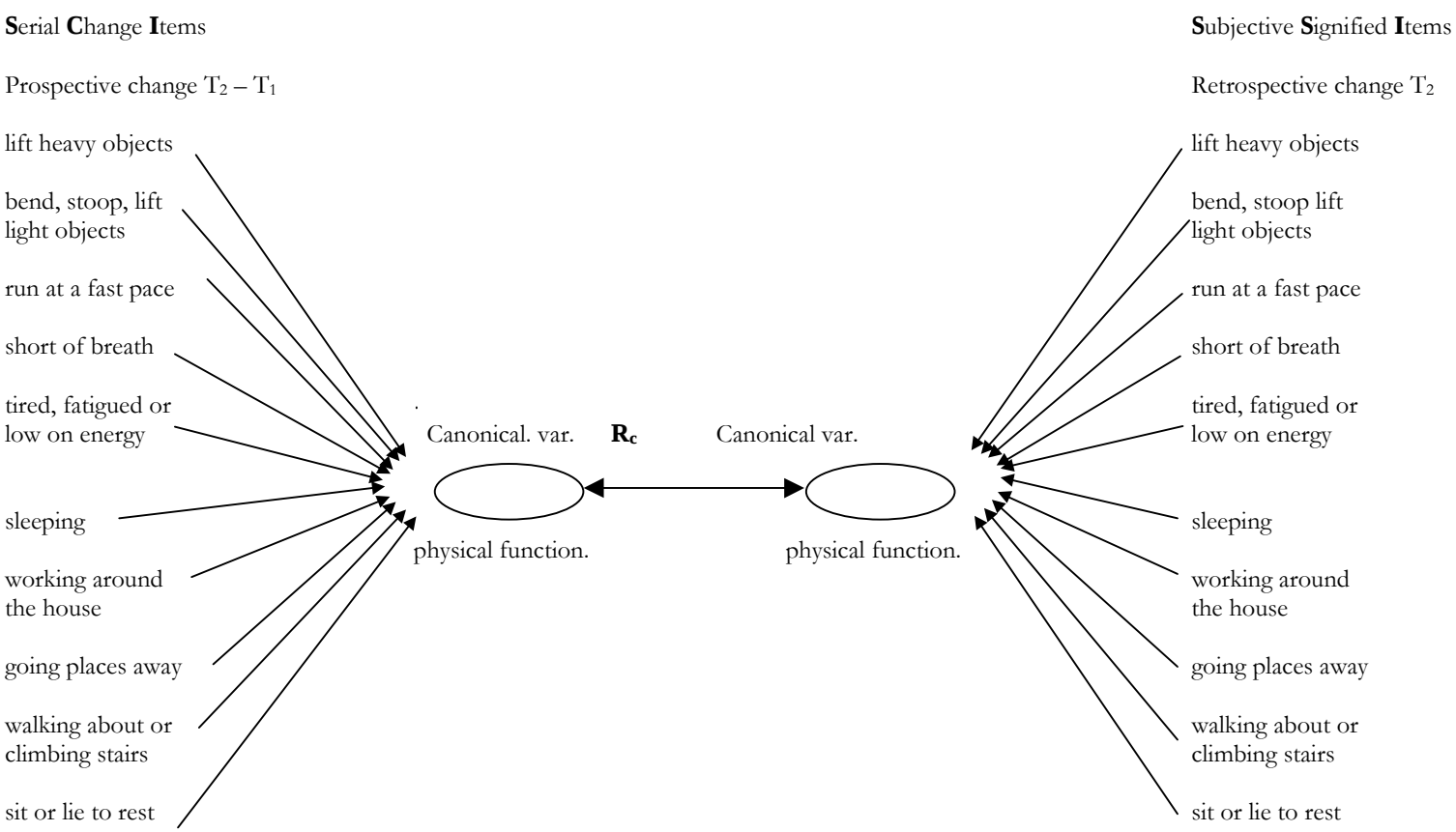
Table 6.2. *The loadings* of the scale components (item scale correlations) of baseline items, change scores of repeatedly assessed baseline items (Serial Change Items) and transition items (Subjective Signified Items) with 95% Confidence Intervals (C.I.)* (* Lambda in LISREL notation)

	Baseline Items			Serial Change Items (SCI)			Subjective Signified Items (SSI)		
	Loading	95%	C. I.	loading	95%	C.I.	loading	95%	C.I.
Physical functioning									
Sit or lie down to rest during the day	0.76	0,86	0,62	0,67	0,69	0,41	0,75	0,85	0,61
Walking about or climbing stairs	0.82	0,94	0,70	0,74	0,80	0,56	0,83	0,95	0,71
Working around the house or yard	0.86	0,98	0,74	0,71	0,76	0,52	0,79	0,89	0,69
Going places away from home	0.69	0,78	0,54	0.55	0,64	0,36	0,62	0,74	0,50
Sleeping well at night	0.54	0,63	0,35	0,39	0,47	0,19	0,49	0,56	0,28
Tired, fatigued or low on energy	0.74	0,83	0,59	0,71	0,76	0,52	0,68	0,80	0,56
Short of breath	0.80	0,92	0,68	0,70	0,82	0,58	0,81	0,93	0,69
Bend, stoop, or lift light objects	0.79	0,91	0,67	0,73	0,85	0,61	0,81	0,93	0,69
Lift heavy objects, like moving a table	0.79	0,91	0,67	0,75	0,84	0,60	0,78	0,90	0,66
Run at a fast pace	0.69	0,80	0,56	0,66	0,74	0,46	0,69	0,81	0,57
Emotional functioning MLHF-Q									
Feel you are a burden to the family	0,66	0,74	0,42	0,51	0,67	0,39	0,61	0,73	0,37
Feel a loss of self-control	0,85	0,97	0,73	0,88	1,00	0,76	0,84	0,98	0,70
Worry	0,72	0,84	0,60	0,69	0,83	0,55	0,64	0,78	0,50
Making it difficult for you to concentrate	0,60	0,72	0,41	0,59	0,70	0,35	0,57	0,65	0,37
Feel depressed	0,71	0,83	0,59	0,61	0,75	0,47	0,64	0,78	0,50

Canonical correlation between composites of change and transition items

Another way to answer the question 'do both operationalizations capture the same change in domains of health?' we also investigated with canonical correlation coefficients the magnitude of association between two concordant sets of items. Given that the transition items were forced into line with the content of the items used to assess prospective change we analysed both batteries of items with canonical correlation analysis (see Figure 6.2). Canonical coefficients only reflect the extent to which each item (given the other items) contributes to the composite of the items in the set of variables to which the item belongs. Therefore, we prefer to present the results with the canonical loadings as follows [39,40] : The correlation between the canonical variable and the items that belong to the underlying dimension of health status which were calculated with prospective change scores of scale items and corresponding transition items.

Figure 6.2: *Hypothetical model of canonical correlation analysis with serial and retrospective change-items of the physical functioning dimension.*



The canonical correlation (R_c) between the composite of the change score items and the composite of the concordant transition items that belong to the domain of physical functioning was R_c=.63 ($p < .001$), whereas the association between the canonical variables as the combination of the emotional functioning-items yielded a R_c=.48 ($p< .001$). These canonical correlations in terms of the percentage of linearly explained variance is fairly large between the two dimensions (40% and 23%, respectively).

Table 6. 3 *Correlations between the Serial Change-Items and the Subjective Signified Items and their corresponding canonical variable.*

	Serial Change Items (SCI)	Subjective Signified Items (SSI)
Physical functioning MLHF-Q		
	.55	.48
Sit or lie down to rest during the day		
Walking about or climbing stairs	.79	.90
Working around the house or yard	.68	.73
Going places away from home	.35	.34
Sleeping well at night	.36	.35
Tired, fatigued or low on energy	.55	.69
Short of breath	.41	.40
Bending, stooping, or lifting light objects	.55	.45
Lifting heavy objects, like moving a table	.77	.56
Running at a fast pace	.51	.44
Emotional functioning		
Feel you are a burden to the family	.48	.44
Feel a loss of self control	.64	.65
Worry	.61	.79
Making it difficult for you to concentrate or remember things	.60	.40
Feel depressed	.94	.94

Table 6.3 shows the correlations between the canonical variables with the original items: the canonical loadings. The prospective change items as well as the concordant transition items showed canonical loadings which were satisfactory according to the criterion > .30 of Levine [39] and, regardless the method of measuring change over time, we could unambiguously identify the items assessing the underlying dimensional configuration demonstrated with factor analysis. This result was remarkable because in the canonical correlation analysis the criterion of the linear combination is aimed at maximising the explained variance between two sets of variables, whereas in Principal Components Analysis (PCA), the linear combination criterion is aimed at maximising the explained variance within a set of variables. In spite of these divergent criteria underlying the method of data reduction, in the set with the change score items (PCI) as well as in the concordant set with transition items (SSI), identical items covered change in the expected domain of health. This result justifies, additional to the results from lisrel-analysis, the comparability of the summed composite of prospective change scales (PCS) with the summed composite of the retrospectively perceived change items (SSS).

Known groups validity

Another question was 'do both operationalizations of the measures have an equal ability to discriminate between subgroups known to differ on a clinically relevant variable?'. Therefore, improvement of angina pectoris (AP) was used as an external criterion to distinguish patients who improved from patient whose AP class stayed the same. Hence. the sample was divided into two groups who should differ based on the improvement of angina pectoris according to the NYHA classification [41]: patients who improved and patients who showed no shift in their NYHA classification. To test the hypothesis that both instruments of prospective change scales and retrospective transition scales have an equal ability to discriminate between these subgroups with known change in AP the Mann-Whitney U test was employed. For this analysis we have, apart from the overall scale of 10 all items belonging to the physical functioning dimension, used the MOS-20 items as an additional measure of physical functioning.

The results of the evaluation of the ability of the prospective and retrospective scales to discriminate between 'known groups' are reported in table 6.4 All p-values were beyond < .01, and effect sizes indicated small or moderate differences on both scales (small effect: ES < 0.20; moderate effect: ES > 0.20 < 0.50 and large effect: ES > 0.80). The difference in change in disease-specific physical functioning derived from repeated measurement detected a small difference between the groups whereas the generic scale estimated a moderate effect. Retrospective scales composed of the same

items showed effect size in reverse order. Change in the domain of the emotional functioning showed small difference between improved and stable AP groups for both methods.
This finding makes it reasonable to suppose that both methods are similar in several respects and that it is difficult to prove that one of them has superior qualities.

Table 6.4 *Discriminative ability of Serial Change Scales and corresponding retrospective transition scales between groups differing in change on the NYHA-classification (improved patients vs. patients who remained the same).*

	Improved Mean (SD)	N	stable mean (SD)	N	z-value	p-value	effect size[1]
Prospective							
Physical functioning:							
MLHF-Q scale	0.22 (1.01)	124	- 0.17 (0.97)	86	- 2.69	< 0.01	0.40
MOS-20 scale	0.27 (1.08)	119	- 0.24 (0.87)	86	- 3.42	< 0.01	0.53
Overall scale	0.25 (1.02)	124	- 0.20 (0.95)	86	- 2.93	< 0.01	0.46
Emotional functioning:							
MLHF-Q scale	0.21 (0.84)	123	- 0.19 (1.05)	87	- 2.42	< 0.01	0.41
Retrospective							
Physical functioning:							
MLHF-Q scale	0.39 (1.01)	123	- 0.27 (0.91)	86	- 4.74	< 0.01	0.71
MOS-20 scale	0.24 (1.11)	124	- 0.16 (0.88)	86	- 3.39	< 0.01	0.42
Overall scale	0.36 (1.03)	123	- 0.25 (0.91)	86	- 4.50	< 0.01	0.64
Emotional functioning:							
MLHF-Q scale	0.22 (1.03)	124	- 0.17 (0.96)	86	-.3.37	< 0.01	0.39

1 Effect size for independent samples: $(\bar{X}_1 - \bar{X}_1 / SD_{pooled})$ $SD = \sqrt{\frac{(S^2_1 + S^2_2)}{(N_1\text{-}1 + N_2\text{-}1)}}$

6.4 DISCUSSION

In patients suffering acute and or traumatic events, the measurement of health outcome due to clinical intervention is often hampered by the absence of a baseline measure. Reliable and valid measures of retrospective change would provide a solution to this problem. This study provides a method to assess perceived change with multi-item transition scales and we suggest that, in contrast with the single-item approach, it can provide a more valid determination of the change score that is signified as relevant by patients.

In most of the studies, transition questions are used as a single item to assess retrospective perceived change. One of the problems with the global assessment of change with one item is that the reliability cannot be established since Cronbach's alpha cannot be computed for a single item. Following Norman et al. [1], we could not locate studies that have examined the reliability of transition scales measuring a hypothesised underlying dimensional configuration. In our study, in which we applied multi-item measures of prospective and retrospective change in domains of HRFS, the scales had a satisfactory level of reliability [42]. The lower reliability of the emotional functioning scales may be due to a smaller number of items as the primary way to make scales more reliable is to make them longer [42].

Both the methods we applied in this study have several strengths and criticisms. Working with repeated measures and directly derived change scores has the major problem that they are ridden with a regression effect and prone to measurement error [43-45]. Change scores derived from repeated measurement may also be flawed by floor and ceiling effects [6,13], carryover effects of learning if the retest intervals are too short, specific events occurring between the first and second assessment, the 'natural' course of the disease, acquiescence and social desirability, and so on. [46]. Another threat to the validity of change scores is the assumption of researchers that subjects have an internalised perception of their level of functioning with regard to, for example, the domain of physical health status and that this internalised standard will not change from baseline to follow-up. This confounding is associated with response-shift bias [47]. There are also several threats to the reliability and validity of our findings based on the use of multi-transition items. First, retrospective perception of change may not be accurate due to recall bias. Patients may equate their present state with change in health status: if a respondent is doing poorly after treatment he might be inclined to think that on the whole things are getting worse even if his health state improved or did not change [1,48,49]. However, in the study of Fitzpatrick et al. [17], the transition questions were shown not to be determined by the patients' mood at follow-up or their present state. Additionally, patients who have suffered a large decline in health may overestimate their perception of baseline health

if they long for the time when their health was better [13,49]. Second, respondents may recalibrate their baseline situation due to the clinical intervention or may feel inclined to give socially desirable answers or they may change the "anchors" for their ratings over time, the so-called called 'response shift' [13,18]. Inaccurate recall seems to be determined by the time interval since exposure or intervention and by the degree of detail required [50], but the significance, the vividness and meaningfulness of events also contribute to recall. Some of the findings of Aseltine et al. [49] suggest that their measures assessing more concrete aspects of patient's condition provided greater correspondence between prospective and retrospective assessment than the more abstract measures of general health. Despite the limitations of transition questions, there is a growing realization that patients can be more directly involved in judging for themselves whether treatments have improved their health status or that relative to the observed health status of other patients by directly asked transition questions [5,6,17,51-53]. Moreover, transition questions were shown to be more sensitive to changes over time in health-related quality of life than were change scores [10,11,16]. The result of the analysis of equal factor loadings indicated that the multi-item transition indices (scales) measure phenomena similar to those measured by the serially assessed dimensions to which they were paired. This could arouse criticism because some items may not have contributed optimally to the assessment of underlying constructs. This may have been caused by the translation of items into transition questions, which may have changed their content, leading to the association of an item with the domain to which it did not belong. For example, the modification of the impact of heart failure on sleeping as an item belonging to the dimension of physical functioning may have become associated with sleeping problems caused by worries and anxiety. In several studies, the agreement between retrospective assessments and serial assessments were poor if single items were used. Therefore, the results of this study argue for multi-item batteries of transition items measuring (disease) specific and relevant domains of HRFS, since one-item transition questions do not cover the sum of aspects of health that belong to the underlying construct or dimension.
Further studies should address the psychometric aspects of transition scales used repeatedly in longitudinal studies, such as test-retest reliability and so on.
In conclusion, retrospectively assessed perception of change after intervention, appears to provide reliable and valid information compared to prospective change scores derived from repeated baseline measurement of health-status dimensions.

Reference List

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients; The floor phenomenon. Medical Care 1990; 28:1142-1152.
7. 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.
8. 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.
9. 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.
10. 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.
11. Ziebland S. Measuring changes in health status. In: Jenkinson C, editor. Measuring health and medical outcomes. London: UCL Press, 1999:
12. MacKenzie RC, Charlson ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. Arch Intern Med 1986; 146:1325-1329.
13. 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.
14. 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.
15. 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.
16. 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.

17. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. Transition questions to assess outcome in rheumatoid arthritis. British Journal of Rheumatology 1993; 32:807-811.
18. Manusco CA, Charlson ME. Does recollection error threaten the validity of cross-sectional studies of effectiveness? Medical Care 1995; 33:AS77-AS88
19. 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.
20. Cunny KA, Perri M. Single-item vs. multiple-item measures of health-related quality of life. Psychol Rep 1991; 69:127-130.
21. 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.
22. 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.
23. Emery CF, Blumenthal JA. Perceived change among participants in an exercise program for older adults. The Gerontologist 1990; 30:516-521.
24. Guyatt GH. Measurement of health-related quality of life in heart failure. JACC 1993; 22:185A-191A.
25. Guyatt GH. Measurement of health-related quality of life in heart failure. Special Issue: Heart disease: The psychological challenge. The Irish Journal of Psychology 1994; 15:148-163.
26. 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; 15: 380-391
27. Rector TS, Tschumperlin LK, Kubo SH, Bank AJ, Francis GS, McDonald KM, et al. Use of the Living With Heart Failure questionnaire to ascertain patients' perspectives on improvement in quality of life versus risk of drug-induced death. J.Card.Fail. 1995; 1:201-206.
28. Rector TS, Cohn JN. Assessment of patient outcome with the Minnesota Living with heart Failure questionnaire: Reliability and validity during a randomized, double blind, placebo-controlled trial of pimobendan. American Heart Journal 1992; October,124:1017-1025.
29. Stewart AL, Hays RD, Ware JE. The MOS Short-form General Health Survey: Reliability and validity in a patient population. Medical Care 1988; 26:724-735.
30. Noe LL, Vreeland MG, Pezzella SM, Trotter JP. A pharmacoeconomic assessment of Torsemide and Furosemide in the treatment of patients with congestive heart failure. Clinical Therapeutics 1999; 21:854-866.
31. Kubo SH, Gollub S, Bourge R, Rahko P, Cobb F, Jessup M, et al. Beneficial effects of Pimobendan on exercise tolerance and quality of life in patients with heart failure. Circulation 1992; 85:942-849.
32. Rector TS, Kubo SH, Cohn JN. Validity of the Minnesota Living with Heart Failure Questionnaire as a measure of therapeutic response to Enalapril or placebo. American Journal of Cardiology 1993; 71:1106-1107.

33. Rector TS. Effect of ACE inhibitors on the quality of life of patients with heart failure. Coron.Artery.Dis. 1995; 6:310-314.
34. Statistical Package for the Social Science. SPSS® for Windows,V7.5.3.Chicago:SPSS,inc. 1997;
35. LISREL 7® A guide to the program and applications.Chicago:Jöreskog and Sörbom SPSS inc. 1989;
36. SAS Institute,Inc.,SAS/STAT® User's Guide,Version 6,Fourth Edition,Volume 1,NC:SAS Institute Inc. 1990;
37. Cronbach LJ, Furby L. How we should measure "change"-or should we? Psychological Bulletin 1970; 74:68-80.
38. Bausell RB, Berman BM. Assessing patients' views of clinical changes [letter]. JAMA 2000; 283:1824
39. Levine MS. Canonical analysis and factor comparison. 1977; Beverly Hills: Sage Publications. 0-8039-0655-2.
40. Thompson B. Canonical correlation analysis: Uses and interpretation. 1984; Beverly Hills: Sage Publications. 0-8039-2392-9.
41. Criteria Committee of the New York Heart Association. Nomenclature and criteria for diagnosis of diseases of the heart and blood vessels: 1973; Boston: Little Brown.
42. Nunnally JC. Psychometric Theory. 2nd ed. New York: Mc Graw Hill, 1978.
43. Nunnally JC. The study of change in evaluation research: principles concerning measurement, experimental design, and analysis. In: Struening EL, Brewer MB, editors. Handbook of evaluation research. SAGE, 1983:231-269.
44. Gottman JM, Rushe RH. The Analysis of Change: Issues, Fallacies, and New Ideas. Journal of Consulting and Clinical Psychology 1993; 61:907-910.
45. Hsu LM. Regression Toward the Mean Associated With Measurement Error and the Identification of Improvement and Deterioration in Psychotherapy. Journal of Consulting and Clinical Psychology 1995; 63:141-144.
46. Cook TD, Campbell DT. Quasi-experimentation. Design & analysis issues for field settings. Chicago: Rand McNally College Publishing Company, 1979.
47. Schwarz CE, Sprangers MAG. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc.Sci.Med. 1999; 48:1531-1548.
48. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. second edition ed. Oxford: Oxford University Press, 1995.
49. Aseltine RH, Carlson KJ, Fowler FJ, Barry MJ. Comparing prospective and retrospective measures of treatment outcomes. Medical Care 1995; 33(uppl.):AS67-AS76
50. Coughlin SS. Recall bias in epidemiologic studies. J.Cinical Epidemiology 1990; 43:87-91.
51. 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; 21:417-439.
52. 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.

53. Redemeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An Illustration in Rheumatology. Arch Intern Med 1993; 153:1337-1342.

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.

8

Summary

In this thesis we used several methods that provide different perspectives on estimation of the magnitude of change in health status and studied the criteria for establishing the relevance or importance of that change.

Chapter 1 introduces the question of to what extent statistically significant changes can be interpreted as relevant, or to what extent substantial changes can be related to clinical interventions.
Trivial change assessed with a health status instrument is more likely to be statistically significant in large samples than in small samples. To give meaning to differences in scores, a widely used method to quantify the amount of change over time has become known as the concept of 'responsiveness'. There is no consensus about the operationalisation of this concept. In studies validating new health status measures are validated, the concept of responsiveness is often defined as the psychometric property of the instrument's sensitivity to measure change over time. In order to demonstrate that the new instrument is more able to detect change compared to a concurrent instrument (for example, generic vs. disease-specific measures), the concept of 'relative efficiency' was introduced in health status measurement. However, when the efficacy of medical interventions is evaluated in terms of change in health status, the term 'responsiveness' is also used to indicate the magnitude of treatment-related change. The relevance of the change between baseline and post-test is estimated with effect size indices (ES). Large effects (ES > .80) are often determined in terms of being *clinically* relevant or important. One might pose the question whether this is a valid interpretation of an effect size. In order to interpret change over time in terms of being clinically important, a golden standard or external criterion for what constitutes importance is necessary. Therefore, in order to interpret change scores as clinically relevant, the external criterion used in this thesis is the patient's perception of the relevance of change in corresponding domains of health status. This external criterion is often called a global question or a transition question, and its score represents a retrospective appraisal of change in terms of the extent of improvement or deterioration or unchanged after treatment. In this thesis health-related functional status was assessed longitudinally with the questionnaire items and, after intervention, the perceived change was assessed with the same items in retrospective mode, or with transition questions. Little is known about the relationship between the longitudinal difference in scale scores, such as the researcher's indicator of the extent and direction of change, and the retrospective transition scale scores assessing the patient's appraisal of the extent and direction of change in the same domains. Furthermore, even less is known about the psychometric properties of measures assessing perception of change with transition

questions. The aim of this thesis is to evaluate the concordance between scores derived from repeated measurement and corresponding transition scores, the overestimation and underestimation of effect size indices in dependent samples, and the psychometric properties of transition scales.

In *Chapter 2*, the changes in health-related functional status are described in a randomised, controlled, clinical trial among patients who responded insufficiently to treatment with maximum doses of pharmaceuticals aimed at reducing severe spasticity. By means of continuous intrathecal baclofen infusion via a subcutaneously implanted programmable pump, a substantial reduction in spasticity and an improvement in health-related functional status was expected. In order to evaluate change over time between baseline and outcome within groups, and the differences in change between experimental and control group, non-parametric statistical tests and effect size indices were used.
Intrathecal baclofen delivered by an implanted programmable pump resulted in statistically significant and relevant improvement in both measures of clinical efficacy (Ashworth scale, spasm score and pain) and self-reported health status (Sickness Impact Profile and the Hopkins Symptoms Check List).

In *Chapter 3*, the psychometric properties of he Minnesota Living with Heart Failure Questionnaire (MLHF-Q) are evaluated in a longitudinal study. The Dutch version of the MLHF-Q was chosen to measure changes in health status among patients who were treated with DC electrical cardioversion for atrial flutter. The one-year follow-up included a self-administered disease-specific MLHF-Q and generic measures of health-related functional status. Internal consistency of MLHF-Q scales was estimated with Cronbach's alpha; the construct validity was evaluated with a multitrait-multimethod analysis, using scales from the RAND-36, the Hospital Anxiety and Depression Scale (HADS) and the Multidimensional Fatigue Inventory (MFI-20). The 'known group validity' was evaluated by comparing mean scores and effect sizes between two groups having low and higher severity of angina pectoris, as assessed by the cardiologist (NYHA classification I versus II-III). Stability of MLHF-Q scales was estimated in a subgroup of patients that remained stable three months after baseline assessment. Perfect Congruence Analysis showed that the results quite closely resemble the previously assumed factor structure from the original American instrument. Cronbach's alpha of the MLHF-Q scales were satisfactory and the overall score of the MLHF-Q discriminated statistically significant between the NYHA I and II-III groups.
Multitrait-multimethod analysis showed a good concordance between the MLHF-Q

physical functioning scale and physical functioning scales from concurrent instruments. The emotional functioning scale showed a weaker concordance with concurrent scales. The MLHF-Q was sensitive to detecting change over time in a group of patients undergoing treatment to improve sinus rhythm. The results of a 'test-retest' analysis in a stable group can be valued as satisfactory for the MLHF-Q scales (Pearsons r > .60). The MLHF-Q has solid psychometric properties and the outcome of the current study indicates that the MLHF-Q is a reliable, valid and efficient health status
instrument.

In *Chapter 4,* we evaluated Cohen's thresholds to interpret an effect size index, namely in terms of 'trivial effect', 'small effect', 'medium effect' and 'large effect'. Cohen developed these thresholds with an effect size index based on standardisation of differences between mean scores, using the pooled standard deviation (SD_p). The mean difference (*d*) between independent groups, expressed in units of pooled standard deviation, was denoted as the effect size *d'*. The effect size *d'* has to be adjusted in cases of paired observations, as Cohen's tables for power and sample size estimation for independent samples assume 2(n-1) degrees of freedom, in contrast with the case of paired observations, where only n-1 are available. Consequently, Cohen adjusted differences in mean scores in dependent samples or matched observations by dividing *d'* by $\sqrt{1-r}$ (r = correlation coefficient T_1 and T_2). He used the symbol *d* to denote this adjusted effect size index. This effect size is identical to an effect size known as the Standardised Response Mean (SRM = the mean change score divided by its standard deviation). This SRM belongs to a large family of standardised mean differences used as effect size indices. In this chapter we demonstrate that when Cohen's rule of thumb for effect size interpretation based on *d'* are applied to SRM, there is a twenty percent risk of overestimation or underestimation. We used publications which calculated SRM indices are calculated for our analysis when the correlation coefficient from T_1 and T2 could be calculated in order to estimate *d'* (148 of 411 SRM's). Effect sizes estimated by using other standard deviations in the denominator of the ratio could not be converted into *d'*. Consequently, applying Cohen's thresholds for effect size interpretation to every standardised mean difference score may lead to over- or underestimation of the magnitude of change over time.

Chapter 5 decribes the concordance between the researcher's interpretation of the direction and extent of change (using an effect size) and the patient's appraisal of the direction and extent of change in the same domains of health status. Patients from

this study sample (N=217) underwent a treatment with known efficacy. After treatment, twenty patients indicated deterioration. Due to the small number of patients who deteriorated, the analyses are restricted to those who improved or remained stable after treatment: Percutaneous Transluminal Coronary Angioplasty (PTCA), Coronary Artery Bypass Grafting (CABG) or pharmacotherapy. In this study, the Dutch version of the Minnesota Living with Heart Failure Questionnaire was used, supplemented with three MOS-20 items from the physical function scale, to assess change in health related functional status. Since we obtained from each item both a change score and a score on the same corresponding transition question, the concordance between both scores was analysed for 23 items. Every questionnaire item was linked to a global question addressing the same health status aspect, and for 23 items the change scores were standardized and broken down according to the item-related global question rating. Thus, irrespective of an item's domain, we calculated 4,798 response combinations out of a total of 4,991 (217 x 23), representing missing data of less than 4%. To evaluate the concordance between the magnitude of change in domains of health-related quality of life and an external criterion, the standardized change scores of scales and a single global question intended to correspond with the repeated measures of physical and emotional functioning were used in the analysis. Global questions were phrased as follows: "Since the last time I filled out this questionnaire (or: Since my operation), my physical problems are …..1) a great deal worse; 2) moderately worse; 3) a little worse; 4) unchanged; 5) a little better; 6) moderately better and 7) a great deal better". In this chapter, an attempt has been made to develop a simple criterion to determine the extent of concordance or discordance between the size of effect according to Cohen's thresholds (researchers appraisal of the amount of change) and the external criterion (the patient's appraisal) with a global question. The ratings of the global question appeared to be in keeping with Cohen's thresholds ($< .20$ trivial; $\geq .20 < .50$ small effect; $\geq .50 < .80$ moderate effect and $\geq .80$ large effect at item level (irrespective of domains) as well as at scale level.

In *Chapter 6,* we evaluated the relationships between the longitudinally assessed change in scores from the physical and emotional functioning scales and the scales of perceived change (transition scales) in the same domains. In this study, the Minnesota Living with Heart Failure Questionnaire was used with 3 supplementary items from the MOS-20, to assess change over time in health-related functional status. Perceived change in physical and emotional functioning were assessed by modified items added to the questionnaire's scale items and were assessed at T_2 after treatment (Percutaneous Transluminal Coronary Angioplasty (PTCA), Coronary Artery Bypass

Grafting (CABG) or pharmacotherapy). This modification was composed of the retrospective mode of the items (denoted as transition items), by rating the extent and direction of perceived change on that particular item (for example, 'climbing stairs'). Internal consistency estimated with Cronbach's alpha yielded satisfactory coefficients for both longitudinally assessed scales' change scores as well as for scores of transition scales in the same domains. Factor analysis of baseline items, change scores of these items and their corresponding transition items demonstrated identical factor loadings. To avoid unreliable eyeball interpretation, 95% confidence intervals were calculated for the corresponding items in these data sets. The results indicate that no differences between factor loadings exist, although the factor loadings for the item change scores were, as expected, systematically lower, as compared to the loadings of baseline and transition items. The canonical correlation between the composite of the change score items and the composite of the concordant transition items belonging to the domain of physical and emotional functioning were fairly large and explained 40% and 23 % of the variance, respectively.

Chapter 7 summarises the main findings from the preceding chapter and critically evaluates them based on insights gained afterwards. The main conclusions are that wholehearted adoption of Cohen's thresholds for interpretation of the Standardised Response Mean (SRM) may lead to over- and underestimation of effect size.
Chapter 4 presents a simple method to adjust these thresholds for SRM estimates. A secondary analysis on the data from Chapters 2 and 3 showed that - after applying the adjustment rule from Chapter 4 on calculated SRM' s - almost twice the number of over- and underestimation of effect size, according to the thresholds belonging to Cohen's *d*'. The small number of observations in both samples may have determined this deviation from the results of Chapter 4. Another main result regards the concordant relation between the amount of longitudinally assessed change in health-related functioning and the patient's perception of the extent of change among patients with heart failure. These results from Chapter 5 were confirmed by another study [1] among cancer patients. Effect sizes estimated with transition scales and their corresponding longitudinal change scales showed no differences between groups who differed in treatment effect (improved vs. stable) nor in groups who differed in the severity of treatment (invasive vs. pharmacotherapy). Administration of the questionnaire's items modified into transition items showed no differences comparing longitudinally assessed items regarding internal consistency of scales, effect size estimates with scales, and principal components factor structure. The

1 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.

results regarding the assessment of perceived change in health status may have been confounded by 'recall bias' (patients cannot remember exactly the extent of limitation in, for example, climbing stairs before the intervention) and 'present state bias' (the extent of limitation at the moment of assessment after the intervention). These two confounding variables are part of this study and the outcome will be published after publication of this thesis.

8

Samenvatting

Dit proefschrift is het resultaat van een aantal artikelen waarin, met behulp van verschillende ragenlijstmethoden, verandering in gezondheidstoestand is gemeten. Het doel is na te gaan wat de betekenis van zo'n verandering in gezondheidstoestand is. Hoofdstuk 1 is een inleiding op de wijze waarop onderzoekers vat proberen te krijgen op de vraag: in hoeverre zijn statistisch significante verschillen ook relevante verschillen, d.w.z. ook substantiële verschillen. Immers, in tegenstelling tot kleine steekproeven, hebben triviale verschillen in grote steekproeven een grotere kans om statistisch significant zijn. Een veel gebruikte methode om de relevantie van verschillen te kwantificeren en te interpreteren, wordt in de literatuur aangeduid met het begrip 'responsiveness'. Het concept responsiveness wordt niet eenduidig gehanteerd. In studies waarin nieuwe instrumenten worden gevalideerd wordt vaak met responsiveness 'de gevoeligheid van een meetinstrument om verandering te meten' bedoeld als een psychometrische eigenschap van het instrument. Om aan te tonen dat het ene meetinstrument meer gevoelig is verandering te meten dan een soortgenoot (bijvoorbeeld een ziekte-specifiek versus generiek) is het begrip 'relatieve responsiveness' geïntroduceerd. In studies waarin (medische) interventies op werkzaamheid worden geëvalueerd, wordt het begrip responsiveness ook gebruikt maar dan als indicator voor ook de grootte van het aan de behandeling gerelateerde effect. Bij dit type onderzoek is dus de relevantie van het gevonden verschil tussen voor- en nameting bij klinische interventies bepaald door de hoogte van de effect size. Een groot effect (effect size > .80) wordt dan als een klinisch relevant of klinisch belangrijk verschil wordt geïnterpreteerd door onderzoekers. De vraag is of dit terecht is. Om een verschil als klinisch belangrijk te kunnen interpreteren is een extern criterium nodig voor die belangrijkheid of relevantie. Om verschillen of veranderingen in gezondheidstoestand die gemeten zijn met verschilscores verkregen door longitudinale meting met vragenlijsten te kunnen interpreteren als klinisch relevant of belangrijk, is in dit proefschrift het oordeel van de patiënt als extern criterium gebruikt. Een dergelijk criterium wordt doorgaans gevormd door een score op retrospectieve vraag aan de patiënt in welke mate hij of zij vindt onveranderd, verbeterd of verslechterd te zijn na de behandeling.

In dit proefschrift is, met alle items van een vragenlijst verandering in de gezondheidstoestand longitudinaal gemeten en is, na de interventie met dezelfde items in een retrospectieve vorm, de gepercipieerde verandering gemeten met transitie-items. Er is nog veel onbekend over de relatie tussen de longitudinale verschilscore die de onderzoeker gebruikt om te kunnen oordelen over de richting en grootte van de verandering, en de retrospectieve transitie scores op schaal-items

waarmee het oordeel van de patiënt over de richting en omvang van de verandering wordt gemeten. Tevens is er zeer weinig bekend over de psychometrische eigenschappen van meetinstrumenten die in retrospectieve vorm de gepercipieerde verandering meten. Het tweede doel van dit proefschrift is om na te gaan in welke mate beide oordelen met elkaar overeenstemmen, welke over- en onderschattingen van effectgrootte er gemaakt kunnen worden bij afhankelijke steekproeven en welke psychometrische eigenschappen het meetinstrument in retrospectieve vorm heeft.

In hoofdstuk 2 is worden de veranderingen bestudeerd in een gerandomiseerde studie bij een groep patiënten waarbij oraal toegediende medicatie in maximale doses geen resultaat meer gaf in de verwachtte vermindering van ernstige vormen van spasticiteit. Door de medicatie intrathecaal toe te dienen, door middel van een subcutaan geïmplanteerde programmeerbare pomp, werd een substantiële vermindering van spasticiteit en verbetering in gezondheidstoestand verwacht. Om de verandering tussen baseline en post-test als ook de verschillen in verandering tussen de groep intrathecaal toegediende placebo en de groep met werkzame stof (Baclofen) te kunnen beoordelen, zijn non-parametrische toetsen en effect sizes toegepast. In deze studie zijn effect sizes alleen gepresenteerd indien de verschillen niet aan toevalsfluctuaties toe te schrijven zijn. De toedieningswijze van medicatie voor ernstige spasticiteit resulteerde na een jaar in statistisch significante en relevante verschillen in zowel klinische parameters als in scores op de lichamelijke dimensies van gezondheidstoestand gemeten met de Sickness Impact Profile en de Hopkins Symptoms Checklist. Deze studie illustreert het klassieke gebruik van statistische toetsen en effect sizes om daarmee uitspraken te doen over de relevantie van het gevonden effect.

In hoofdstuk 3 worden de psychometrische eigenschappen geëvalueerd van het meetinstrument dat in deze studie gekozen is om verandering in gezondheidstoestand te meten, de Nederlandse versie van de Minnesota Living with Heart Failure Questionnaire (MLHF-Q). De steekproef voor dit onderzoek betrof een groep patiënten die behandeld werd voor boezemfibrileren. De interne consistentie werd geschat met behulp van Cronbach's alpha; de construct validiteit werd beoordeeld met behulp van multitrait-multimethod analyse met gebruikmaking van schalen uit de Rand-36, uit de Hospital Anxiety and Depression Scale (HADS) en uit de Multidimensional Fatigue Inventory (MFI-20). Known-Group validiteit werd geëvalueerd door de verschillen te analyseren tussen de groep met de minst ernstige vorm van angina pectoris met de groep met de meer ernstige vorm aan de hand van het oordeel van een cardioloog (NYHA classificatie). De test-hertest betrouwbaarheid is geschat met een sub-groep in de steekproef waarvan de

gezondheidstoestand gedurende drie maanden na de baseline meting geen verandering had ondergaan. Om na te gaan of de dimensies van de factoranalyse uit deze steekproef, uit een Nederlandse populatie patiënten met een vorm van hartfalen, overeenkomen met die uit de oorspronkelijke factor analyse in de Amerikaanse studie, is gebruik gemaakt van Perfect Congruence Analysis (PCA). De MLHF-Q schalen hadden een bevredigende Cronbach's alpha en discrimineerden goed tussen de groepen met ernstige en minder ernstige angina pectoris. Het resultaat van de multitrait-multimethod analyse laat een goede overeenstemming zien tussen de 'fysiek functioneren' schaal van de MLHF-Q en de fysieke dimensies van de andere (concurrente) instrumenten. Voor de schaal 'emotioneel functioneren' is deze overeenstemming zwakker. De factor ladingen in de oorspronkelijke schaal constructie met een Amerikaanse steekproef kwamen goed overeen met de ladingen in de Nederlandse steekproef. De MLHF-Q bleek gevoelig om verandering te meten in een groep patiënten die een behandeling ondergaat die primair gericht is op verbetering van sinusritme. De test-hertest van de MLHF-Q resulteerde in bevredigende correlaties groter dan .60. De MLHF-Q is dus een betrouwbaar en valide instrument om veranderingen in gezondheidstoestand te meten bij patiënten met hartfalen.

In hoofdstuk 4 wordt de toepassing van de grenzen van Cohen om een effect size te interpreteren in termen van 'triviaal', 'klein', 'medium' of 'groot' nader geanalyseerd. Deze grenzen zijn door Cohen ontwikkeld op basis van een index gebaseerd op het verschil in gemiddelden tussen twee onafhankelijke steekproeven gedeeld door de gepoolde standaarddeviatie (SDp) Dit in eenheden van de gepoolde standaarddeviatie uitgedrukte verschil d (difference) is door Cohen geannoteerd als d'. Omdat Cohen in zijn standaardwerk over poweranalyse en effect size de tabellen waarmee de omvang van de steekproef te bepalen zijn heeft gebaseerd op twee steekproeven, corrigeert hij d' voor de situatie waarin er sprake is van bijvoorbeeld een herhaalde meting binnen 1 steekproef. Hiertoe corrigeert hij d' met $\sqrt{1-r}$ waarbij r de correlatie is tussen de meting op T1 en T2. Deze effect size is gelijk aan het gemiddelde verschil tussen T1 en T2 gedeeld door de standaarddeviatie van dat verschil. Deze effect size wordt veel gebruikt en staat bekend als de Standardised Response Mean (SRM). Daarnaast is er een aantal effect sizes ontwikkeld waar het gemiddelde verschil in eenheden van andere standaarddeviaties wordt uitgedrukt die van deze SDp en SRM afwijken. In dit hoofdstuk wordt aangetoond dat, als men de grenzen van Cohen, behorend bij de d', toepast op de SRM, er een risico is van over- en onderschatting van de geschatte effectgrootte in ongeveer één op de vijf in de gebruikte steekproef van effect sizes. Aangezien de SRM alleen te herleiden is tot d' als de correlatiecoëfficient tussen T1 en T2 berekend kon worden op grond van de

in de betreffende publicatie gerepresenteerd gegevens (148 van de 411 SRM's) heeft de analyse zich tot deze moeten beperken. Effect sizes indices geschat door het gemiddelde verschil te standaardiseren met de standaarddeviatie (SD) van o.a. de baseline scores (in subgroepen) of de SD van de veranderingsscores (in subgroepen) zijn wiskundig niet te herleiden tot de effect size d'. Derhalve is het kritiekloos toepassen van de grenzen van Cohen niet vrij van het risico van over- en onderschatting van de grootte van het effect.

In hoofdstuk 5 is de overeenkomst tussen het oordeel van de onderzoeker over de mate van verbetering (met behulp van effect size indices) en het oordeel van de patiënt over de mate van verbetering beschreven. De patiënten in deze steekproef (N=217) ondergingen een behandeling waarvan bekend is dat deze de gezondheidstoestand verbeterd. Twintig patiënten gaven na de behandeling aan verslechterd te zijn. De analyses zijn door dit kleine aantal beperkt tot patiënten die van mening waren te zijn verbeterd of onveranderd te zijn gebleven en na de interventie die zij ondergingen: Percutaneous Transluminal Coronary Angioplasty (PTCA), Coronary Artery Bypass Grafting (CABG) of farmacotherapie). In deze studie is de Nederlandse versie van de Minnesota Living with Heart Failure Questionnaire (MLHF-Q), aangevuld met enkele MOS-20 items, gebruikt om verandering in gezondheidstoestand te meten. Aangezien van elk item in de vragenlijst zowel een verschilscore als een bij dit item behorend retrospectief oordeel gemeten is, is in eerste instantie een vergelijking tussen onderzoekersoordeel (verschilscore) en patiëntoordeel (retrospectief oordeel) op item niveau uitgevoerd voor 23 items. Geen rekening houdend met de dimensies waartoe items behoren zijn 4798 response-combinaties berekend van een totaal van 4991 (217 x 23) als gevolg van 4% missing data.

Voor de bepaling van de concordantie tussen de gestandaardiseerde verschillen (effect grootte) van schalen of domeinen, is als extern criterium voor de relevantie van de verbetering in de schaalscores gebruik gemaakt van zogenoemde 'global questions' naar gepercipieerde verandering in deze domeinen. Een voorbeel van zo'n global question is 'Sinds de bypass operatie is mijn beperking in het trappenlopen: 1) sterk verbeterd, 2) nogal verbeterd, 3) weinig verbeterd, 4) onveranderd, 5) een beetje slechter, 6) nogal slechter en 7) sterk verslechterd'.
In dit hoofdstuk is geprobeerd een eenvoudig criterium te ontwikkelen om te bepalen wanneer het oordeel van de onderzoeker over de grootte van de verandering (effect size) in overeenstemming is met het externe criterium, of daarvan afwijkt in termen van een over- of onderschatting. De vuistregel van Cohen (< .20 triviaal; ≥ .20 < .50 klein; ≥ .50 < .80 medium; en ≥ .80 groot effect) waarmee effect grootte

doorgaans door onderzoekers worden geïnterpreteerd, lijken synchroon te lopen met het de oordelen van de patiënt in termen van gepercipieerde verbetering (respectievelijk: 'geen verandering'; 'een beetje verbeterd'; 'nogal verbeterd' en 'veel verbeterd'.

In hoofdstuk 6 is de relatie onderzocht tussen: 1) de gemeten verandering in de schalen 'lichamelijk' en 'emotioneel functioneren' met longitudinale meting en 2) de gepercipieerde verandering in deze domeinen. In deze studie is eveneens de Nederlandse versie van de Minnesota Living with Heart Failure Questionnaire (MLHF-Q), aangevuld met enkele MOS-20 items, gebruikt om verandering in gezondheidstoestand te meten. De gepercipieerde verandering in lichamelijk en emotioneel functioneren zijn gemeten met gemodificeerde schaal-items die opgenomen zijn in de vragenlijst die na de behandeling (Percutaneous Transluminal Coronary Angioplasty (PTCA), Coronary Artery Bypass Grafting (CABG) of farmacotherapie) is ingevuld. Deze modificatie bestaat uit de retrospectieve vorm van het betreffende item (transitie item genoemd) waarmee gevraagd werd naar de richting en de mate van verandering na de behandeling. Analoog aan de interne consistentie van de schaal van veranderingscores werd de interne consistentie van de transitie schalen geschat. De interne consistentie uitgedrukt in Cronbach's alpha van de met herhaalde meting verkregen veranderscores en die van de corresponderende transitieschaal 'fysiek functioneren' waren bevredigend. Factoranalyses van de baseline items, de verschilscore van deze items en de hiermee corresponderende transitie items resulteerden in dezelfde factorstructuur. Om onbetrouwbare 'eyeball' interpretatie te voorkomen zijn 95% betrouwbaarheidsintervallen rond de ladingen van de items binnen elke dataset berekend. De resultaten laten geen verschillen zien ondanks de verwachtte systematisch lagere ladingen van de item verschilscores vergeleken met die van de baseline items en transitie items. De canonische correlatie tussen de lineaire combinatie van de verschilscores per item en de lineaire combinatie van de transitie-items resulteerde voor beide schalen in 40% (verandering in fysiek functioneren) respectievelijk 23% (verandering in emotioneel functioneren) verklaarde variantie. De conclusie is dat beide meetmethoden overlap vertonen maar ook verschillende veranderingen registreren. In situaties waarin geen baselinemeting mogelijk is, zou met de transitiemethode volstaan kunnen worden.

In hoofdstuk 7 worden de belangrijkste resultaten van de voorgaande hoofdstukken samengevat en, op basis van de in een later stadium verkregen inzichten, kritisch geëvalueerd. De belangrijkst conclusies zijn dat het kritiekloos toepassen van de vuistregel van Cohen voor effect size op de Standardised Response Mean tot over en onderschattingen leidt. In hoofdstuk 4 is daarvoor een eenvoudige correctie

methode ontwikkeld. In een secundaire analyse van de gegevens uit de hoofdstukken 2 en 3 werden, in afwijking van het resultaat uit hoofdstuk 4, bijna twee maal zoveel over- of onderschattingen gevonden ten opzichte van Cohen's grenzen behorend bij effect size d'. De veel kleinere steekproef kan hier een rol hebben gespeeld. Een andere belangrijke conclusie is dat grootte van verschilscores verkregen uit herhaalde meting, in grote mate concordant zijn met de door de patiënt gepercipieerde grootte van de verandering. De in hoofdstuk 5 gevonden overeenstemming tussen het onderzoekers-oordeel over de effect grootte en de door de patiënt gepercipieerde verandering werd door een andere studie [1] bevestigd. De effect size in domeinen van de gepercipieerde verandering en die van de longitudinaal gemeten verandering blijken niet van elkaar te verschillen, zijn even groot voor groepen waar een gotere verandering verwacht wordt en waar deze niet verwacht wordt. Door de items uit een vragenlijst na de operatie af te nemen in de vorm van transitie-items en deze op schaalniveau te vergelijken met de veranderingscores, is aangetoond dat schalen die gepercipieerde veranderingen meten vergelijkbare effect sizes, factorladingen en interne consistentie hebben. De invloed van mogelijke confounders in de meting van gepercipieerde verandering in domeinen van gezondheid, is object van analyse en een publicatie na dit proefschrift. De beïnvloeding van het retrospectieve oordeel over het effect van behandeling door 'recall bias' (men weet niet meer in welke mate men voor de behandeling beperkt was in bijvoorbeeld trappenlopen) en de 'present state bias' (de mate van (niet) beperkt zijn op het moment van herhaalde meting na de behandeling) is onderwerp van lopend onderzoek.

---

1. 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.

Dankwoord

In 1969 begon mijn eerste leermeester Jan Jessen een onderzoek naar Medische Consumptie onder een representatieve steekproef uit de Nederlandse bevolking. In 1970 kwam ik als leerjongen en duvelstoejager voor het veldwerk, het uitbetalen van student-interviewers, de dataverzameling, de codering en verwerking op ponskaarten. Mijn tweede leermeester Ivan Gadourek nam de leiding van dit onderzoek over na het overlijden van Jessen. Beide leermeesters hebben mij, voordat ik mijn avondopleiding aan het Groninger Avond Atheneum had afgerond, gesocialiseerd op de werkvloer van het medisch-sociologisch onderzoek. De mogelijkheden die ik in deze periode onder leiding van Jan en Ivan heb gekregen en de persoonlijke betrokkenheid die ik mocht ervaren zijn tot op de dag van vandaag voor mijn professionele ontwikkeling van de grootste betekenis. Ik ben Herman Walta, student-assistent van het allereerste uur nog altijd dankbaar dat hij op zijn rustige manier heeft weten te voorkomen dat ik op die eerste schreden op het smalle pad der sociologie reeds zou zijn verdwaald door de veelheid van nieuwe kennis en vooral door het voor mij toen onbegrijpelijke jargon.

De groep studenten die, in 1973-74 in de fase van analyse en rapportage, de belangrijkste peilers onder het Medische Consumptie project vormden bestond uit Hans Knol, Hans Ormel, Johan Groothoff, Sineke Ten Horn, en Jelte Bouma die allen hun weg vervolgden met het afronden van een dissertatie. Ik vind het leuk de laatste van het "Medische Consumptie" team te zijn die de serie van proefschriften compleet mag maken.

Het schrijven van een proefschrift is een avontuur waarvan het reisdoel bekend is, de koers is uitgezet, maar het anticiperen op de gevaren die het slagen van een dergelijke expeditie haalbaar maakt, lukt altijd maar ten dele. Zonder professionele koersbewaking en zonder emotionele ondersteuning van mensen om je heen zou het een barre tocht geworden zijn. In de eerste plaats wil ik Wim van den Heuvel op deze plaats dank zeggen voor de samenwerking vanaf 1985 bij de oprichting van het NCG. In het jaar daarop gaf al ik te kennen te willen promoveren, maar de opbouw van het NCG in oprichting had prioriteit. "Als de NCG organisatie compleet en operationeel is, dan kun je aan een proefschrift denken", was het standpunt van het dagelijks bestuur destijds. Ik had toen geen enkele moeite met deze beslissing die jij destijds met Wouter van Rossum in ons eerste functioneringsgesprek naar voren bracht. Het was ook in de begintijd van zo'n klein onderzoekinstituut 'in oprichting' dat ik mij wel eens liet ontvallen "een directeur van niks te zijn". Zonder de ondersteuning van Liesbeth Massaut, in deze beginjaren uitgebreid met de komst van Jacqueline Nannenberg, was het met het NCG niks geworden. Ondanks de geweldige inzet van dit kleine team, bleven er elk jaar steeds te weinig vrijheidsgraden

over voor het schrijven van een proefschrift. Vijf jaar na dit eerste functioneringsgesprek kon het predikaat 'in oprichting' weg worden gelaten op het briefpapier en vijf jaar later werd de managementstructuur van het volwassen geworden onderzoekinstituut gewijzigd waardoor er ruimte kwam om over dat plan uit '86 na te denken. En dank zij jou kon ik een jaar later beginnen. Eind 1997 was het protocol gereed en konden we de dataverzameling eind 1998 afsluiten en in het begin van 1999 met de analyses beginnen. In de hectische periode die daarop volgde ben je voor mij het belangrijkste ijkpunt geweest voor de hypothesen en de analyse technieken waarvoor ik meende te moeten kiezen. Op deze plaats wil ik zeggen het als een voorrecht te beschouwen met je te hebben mogen samenwerken en je bedanken voor het vertrouwen dat je in mij hebt gehouden in de afgelopen vier jaar.
Beste Mike. Toen we in '97 met jou het idee bespraken om deze studie uit te voeren en dat we daarvoor bij de afdeling Cardiologie en Thoraxchirurgie een patiëntengroep zochten die op een wachtlijst stond voor een behandeling met aangetoond effect, ben jij degene geweest die zijn nek durfde uitsteken en collega cardiologen heeft weten te overtuigen. Ik ben je zeer dankbaar voor kritische, maar altijd relevante, commentaren op veel onderdelen van het manuscript die voor een medicus niet tot de dagelijkse kost behoren.
Een stille kracht die altijd een oplossing wist creëren, en met lang doorvragen de vraagsteling helder wist te krijgen was van doorslaggevende betekenis als ik weer eens vastgelopen was. Beste Roy, je hebt me nooit in de kou laten staan en met jouw inventiviteit is menig analyse tot stand gekomen.
In de beginfase van de eerste analyses op item-niveau, met nog niet zo goed doordachte veronderstellingen en hypothesen, was Willem Lok met zijn credo ' Berry, alles kan' voor mij een grote steun.
Mathieu de Greef en Jitse van Dijk, zijn collega's waar je, als je het even niet meer ziet zitten, altijd weer weggaat met een goed gevoel.
Eric van Sonderen en Jelte Bouma poogden het schip op koers te houden in werkbesprekingen met discussies die er soms niet om logen, maar waar ook altijd plaats was voor relativering door humor met een hoog 'Bouma-gehalte'.
Prof. dr. D.Post, prof. dr. J.L. Peschar en prof. dr. H.J.G.M. Crijns wil ik bedanken voor de snelle beoordeling van dit manuscript.
Zonder het 'monitoring' programma dat Danny Barnhoorn voor dit onderzoek schreef, zouden we nooit zo accuraat en efficiënt het veldwerk hebben kunnen uitvoeren. Telkenmale als de pc werd aangezet wisten we wie de tweede vragenlijst moest hebben, wie benaderd moest worden omdat de vragenlijst niet volledig was ingevuld. Aan hem hebben we het vooral te danken dat we zo weinig missende waarden in ons bestand hebben.
Wat is het een voorrecht samen te kunnen werken met studenten die geïnteresseerd

zijn in empirisch onderzoek, de handen uit de mouwen steken bij het veldwerk, hun intellectuele talent inzetten om met de data scripties te schrijven en hun analyse technieken toe te passen. Op deze plaats bedank ik in tijdsvolgorde waarop ze het 'erf van Middel betraden: Marc Boerma, Egbert Hofhuis, Rutger Jansen, Marian Klene en Ineke Koopmans en last but not least, Heike Goudriaan.
Ik ben grote dank verschuldigd aan de respondenten die mij het vertrouwen hebben gegeven hun naam en adres te mogen gebruiken om ze twee keer lastig te mogen vallen met een moeilijke vragenlijst.
Ziekenhuis de Weezenlanden te Zwolle en het Martini Ziekenhuis te Groningen hebben aan de instroom van respondenten in belangrijke mate bijgedragen door patiënten toestemming te vragen aan het onderzoek deel te nemen.
Wat een geweldige opluchting was het dat er collega's zijn die een gevoel voor (en plezier aan) vormgeving van de door mij niet erg professioneel opgemaakte tekst. Wilma Warmelink de spil in de organisatie van de sectie Zorgwetenschap en Nettie de Rade van de afdeling Bewegingswetenschap wil ik bedanken voor het fatsoeneren van zoveel slordigheden.
Elke onderzoeker heeft een scientific community nodig om zijn onderzoek uit te kunnen uitvoeren. Ik bewaar goede herinneringen aan de discussie in de beginfase en aan het eind van dit promotietraject met collega's die deelnemen aan het onderzoeksprogramma 'Public Health and Health Sevices Research'. Ik zie er dan ook naar uit het onderzoek van de sectie Zorgwetenschap in dit programma te kunnen ontwikkelen en uitbouwen.
In de afgelopen 30 jaar hebben Heit en Mem met grote betrokkenheid en steun mijn wel en wee aan deze universiteit gevolgd. Wat zou Heit het een gebeurtenis hebben gevonden hierbij te zijn. Ik mis hem zeer.
Lieve Kim. Dank je wel voor al je nauwgezette werk bij het bewerken en corrigeren van de tekst.
Lieve Esther, lieve Mickey. Dank je wel voor jouw manier van vormgeven van dit boekje.
Lieve Jan-Just. Wie had nog kunnen denken dat jij met mij deze dag zo zou kunnen vervolmaken. Ik ben trots op je.
Lieve Aukje. In de jaren dat wij met ons gezin een hechte band vormden heb jij het mij mogelijk gemaakt te studeren en ook in de moeilijke tijden in ons leven heb je mij gesterkt in het idee dat het nooit te laat is om af te studeren.
Lieve Elke. Het is door mijn 'late roeping' dit werkstuk te willen voltooien voor jou de laatste twee jaar niet gemakkelijk geweest je leven met mij te moeten delen. Zonder jou had ik het echt niet kunnen voltooien.

Northern Centre for Healthcare Research (NCH) and previous dissertations

The Northern Centre for Healthcare Research (NCH) was founded in 1990 as a research institute of the University of Groningen (RUG), The Netherlands. Researchers from both the Medical and Social Faculty, with various professional backgrounds, are members of the NCH. These include medical sociologists, medical doctors, psychologists and human movement scientists. Research of the NCH is aimed at optimising quality of life of patients and quality of healthcare, and focuses on (a) determinants of health and illness, (b) consequences of illness, (c) the effects of medical treatment and decision making, and (d) the evaluation of health services and various types of interventions. At the time that this thesis is published, the NCH comprises five research programmes.

Until 1998, the NCH covered two research programmes, i.e. Determinants of Health and Medical Decision Making and Evaluation of Healthcare. The first programme was reformulated in 1996 and was continued as Disorder, Disability and Quality of Life (DDQ). Hence, previous dissertations in this area are listed as part of the present DDQ-programme. The second programme was subdivided in 1998 into two new programmes, i.e. Public Health and Public Health Services Research and Rational Drug Use.
Dissertations published earlier within the second programme are listed retrospectively under these new headings. In 1998, two new programmes, The Outcome and Evaluation of Interventions in Patients with Motor Problems and Process and Effects of Movement Programs, were formulated and officially integrated in the NCH in January 1999. The accomplished dissertations since the start of the programmes in 1998 are included in the list. In 2000 the Department of General Practice joined the NCH and together with the Rational Drug Use group initiated a new research programme, i.e. Implementation of Evidence Based Medicine in the Medical Practice.

More information regarding the institute and its research can be obtained from our internet site: http://www.med.rug.nl/nch

Public Health and Public Health Services Research

Lucht F van der (1992) *Sociale ongelijkheid en gezondheid bij kinderen.*
PROMOTOR: prof dr WJA van den Heuvel. REFERENT: dr JW Groothoff

De Man FH (1992) *Gezien de spoedeisendheid van het geval. Beoordeling van kwaliteit en effectiviteit van spoedeisende medische hulpverlening; modellen, methoden en uitvoering.* PROMOTORES: prof drs B. Binnendijk, prof dr ThP van Hoorn,prof dr H.P.M. Jägers.
REFERENT: dr HJ ten Duis

Engelsman C & Geertsma A (1994) *De kwaliteit van verwijzingen.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr FM Haaijer-Ruskamp, prof dr B Meyboom-de Jong

Puttiger PHJ (1994) *De medische keuring bij gebruik van persluchtmaskers.*
PROMOTORES: prof dr D Post,prof dr WJA Goedhard. CO-PROMOTOR: dr JW Groothoff

Dekker GF (1995) *Rugklachten-management-programma bij de Nederlandse Aardolie Maatschappij B.V.: ontwerp, uitvoering en evaluatie.*
PROMOTORES: prof dr D Post, prof WH Eisma. CO-PROMOTOR: dr JW Groothoff

Mulder HC (1996) *Het medisch kunnen: technieken, keuze en zeggenschap in de moderne geneeskunde.*PROMOTOR: prof dr WJA van den Heuvel

Mink van der Molen AB (1997) *Carpale letsels: onderzoek naar de verzuimaspecten ten gevolge van carpale letsels in Nederland 1990-1993.* PROMOTORES: prof dr PH Robinson, prof WH Eisma. CO-PROMOTOR: dr JW Groothoff. REFERENT: dr GJP Visser

Tuinstra J (1998) *Health in adolescence: an empirical study of social inequality in health, health risk behaviour and decision making styles.* PROMOTORES: prof dr D Post, prof dr WJA van den Heuvel. CO-PROMOTOR: dr JW Groothoff

Dijkstra A (1998) *Care dependency: an assessment instrument for use in long-term care facilities.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr ThWN Dassen

Wijk P van der (1999) *Economics: Charon of Medicine?* PROMOTORES: prof dr WJA van den Heuvel,
prof dr L Koopmans, prof dr FFH Rutten. REFERENT: dr J Bouma

Hospers JJ (1999) *Allergy and airway hyperresponsiveness: risk factors for mortality.*
PROMOTORES: prof dr D Post, prof dr DS Postma, prof dr ST Weiss

Goossen WTF (2000) *Towards strategic use of nursing information in the Netherlands.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr ThWN Dassen, prof dr ir A Hasman

Pal TM (2001) *Humidifiers disease in synthetic fiber plants: an occupational health study.*
PROMOTORES: prof dr JGR de Monchy, prof dr D Post, prof dr JW Groothoff

Beltman H (2001) *Buigen of barsten? Hoofdstukken uit de geschiedenis van de zorg aan mensen met een verstandelijke handicap in Nederland 1945 – 2000.* PROMOTORES:: prof dr D Post, prof dr AthG van Gennep

Dalen I van (2001) Second opinions on orthopaedic surgery.
PROMOTORES: prof dr JR van Horn, prof dr PP Groenewegen, prof.dr. JW Groothoff

Implementation of Evidence Based Medicine in the Medical Practice

Zijlstra IF (1991) *De regionaal klinisch farmacoloog.* PROMOTORES: prof dr H Wesseling, prof dr FWJ Gribnau, prof dr C van Weel. REFERENTEN: dr FM Haaijer-Ruskamp, dr H Wollersheim

Jong-van den Berg LTW de (1992) *Drug utilization studies in pregnancy: what can they contribute to safety assessment?* PROMOTORES: prof dr MNG Dukes, prof dr H Wesseling.
REFERENT: dr FM Haaijer-Ruskamp

Denig P (1994) *Drug choice in medical practice: rationals, routines, and remedies.*
PROMOTORES: prof dr FM Haaijer-Ruskamp, prof dr H Wesseling

Boerkamp E (1995) *Assessing professional services quality: an application in health care.*
PROMOTORES: prof dr JC Reuijl, prof dr FM Haaijer-Ruskamp

Trigt AM van (1995) *Making news about medicines.*
PROMOTORES: prof dr TFJ Tromp, prof dr FM Haaijer-Ruskamp

Dijkers FW (1997) *Repeat prescriptions: a study in general practice in the Netherlands.*
PROMOTORES: prof dr B Meyboom-de Jong, prof dr FM Haaijer-Ruskamp, prof dr AF Casparie

Bosch JL (1997) *Outcome assessment of the percutaneous treatment of lilac artery occlusive disease.*
PROMOTORES: prof dr MGM Hunink, prof dr WPThM Mall, prof dr L Koopmans

Vries SO de (1998) *Management strategies for intermittent claudication.*
PROMOTOR: prof dr MGM Hunink. REFERENT: dr JB Wong

Veehof LJG (1999) *Polypharmacy in the elderly.*
PROMOTORES: prof dr B Meyboom-de Jong, prof dr FM Haaijer-Ruskamp

Veninga CCM (2000) *Improving prescribing in general practice.*
PROMOTOR: prof dr FM Haaijer-Ruskamp.REFERENT: dr P Denig

The Outcome and Evaluation of Interventions in Patients with Motor Problems (from 1998 onwards)

Sluis CK van der (1998) *Outcomes of major trauma.* PROMOTORES: prof dr HJ ten Duis, prof WH Eisma

Geertzen JHB (1998) *Reflex sympathetic dystrophy: a study in the perspective of rehabilitation medicine.*PROMOTORES: prof WH Eisma, prof dr HJ ten Duis. CO-PROMOTOR: dr JW Groothoff.
REFERENT: dr PU Dijkstra

Halbertsma JPK (1999) *Short hamstrings & stretching: a study of muscle elasticity.*
PROMOTORES: prof WH Eisma, prof dr LNH Göeken. CO-PROMOTOR: dr JW Groothoff. REFERENT: dr ir AL Hof

Rommers GM (2000) *The elderly amputee: rehabilitation and functional outcome.*
PROMOTOR: prof WH Eisma. CO-PROMOTOR: dr JW Groothoff

Process and Effects of Movement Programs (from 1998 onwards)

Berkhuysen MA (1999) *Toward tailor-made cardiac rehabilitation: getting at the heart of the matters.*
PROMOTORES: prof dr AP Buunk, prof dr P Rispens. REFERENT: dr R Sanderman

Heuvelen MJG van (1999) *Physical activity, physical fitness and disability in older persons.*
PROMOTOR: prof dr P Rispens. CO-PROMOTORES: dr WH Brouwer, dr GIJM Kempen.
REFERENT: dr MHG de Greef

Lettinga AT (2000) *Diversity in neurological physiotherapy: a comparative analysis of clinical and scientific practices.* PROMOTORES: prof dr P Rispens, prof dr PJM Helders, prof dr A Mol

Stevens M (2001) *Groningen Active Living Model (GALM): development and initial validation.*
PROMOTOR: prof dr P Rispens. REFERENTEN: dr KAPM Lemmink, dr MHG de Greef

Disorder, Disability and Quality of Life

Sanderman R (1988) *Life events, mediating variables and psychological distress: a longitudinal study.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr PE Boeke, prof dr PMG Emmelkamp.
REFERENT: dr J Ormel

Kempen GIJM (1990) *Thuiszorg voor ouderen: een onderzoek naar de individuele determinanten van het gebruik van wijkverpleging en/of gezinsverzorging op verzorgend en huishoudelijk gebied.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr W Molenaar. REFERENT: dr ThPBM Suurmeijer

Sonderen FLP van (1991) *Het meten van sociale steun.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr FN Stokman. REFERENT: dr J Ormel

Heyink JW (1992) *Levertransplantatie: psycho-sociale aspecten.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr MJH Slooff. REFERENT: dr Tj Tijmstra

Gerritsen JC (1993) *Onafhankelijkheid van ouderen: mogelijkheden en voorwaarden.*
PROMOTOR: prof dr WJA van den Heuvel

Reitsma B (1994) *The end of the line? Evaluation of a multidisciplinary team approach to chronic pain.* PROMOTORES: prof dr EC Klip, prof dr JWF Beks, prof dr JP Hennis

Ranchor AV (1994) *Social class, psychosocial factors and disease: from description towards explanation.* PROMOTORES: prof dr WJA van den Heuvel, prof dr AP Buunk.
REFERENTEN: dr R Sanderman, dr J Bouma

Oosterhuis A (1994) *De gedragstherapeutische behandeling van slaapklachten.*
PROMOTOR: prof dr EC Klip

Linschoten CP van (1994) *Gezondheidsbeleving van ouderen: een longitudinale studie.*
PROMOTOR: prof dr WJA van den Heuvel. CO-PROMOTOR: dr J Ormel

Linden-van den Heuvell GFEC van (1994) *Voorbereiding op medische ingrepen.*
PROMOTOR: prof dr EC Klip

Uitenbroek DG (1995) *Exercise behavior*. PROMOTOR: prof dr WJA van den Heuvel

Steverink N (1995) *Zo lang mogelijk zelfstandig: naar een verklaring van verschillen ten aanzien van opname in een verzorgingstehuis onder fysiek kwetsbare ouderen.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr TAB Snijders, prof dr J Ormel

Ruiter JH de (1995) *Sociale ondersteuning en kwaliteit van leven bij patiënten met kanker.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr H Schraffordt Koops.
REFERENTEN: dr FLP van Sonderen, dr R Sanderman

Krol B (1996) *Quality of life in rheumatoid arthritis patients: the relation between personality, social support and depression.* PROMOTOR: prof dr WJA van den Heuvel. REFERENTEN: dr R Sanderman, dr ThPBM Suurmeijer

Kooiker SE (1996) *Illness in everyday life: a health diary study of common symptoms and their consequences.* PROMOTORES: prof dr WJA van den Heuvel, prof dr J van der Zee

Zwanikken CP (1997) *Multiple sclerose: epidemiologie en kwaliteit van leven.*
PROMOTOR: prof dr J Minderhoud. CO-PROMOTORES: dr JW Groothoff, dr ThPBM Suurmeijer

Scaf-Klomp W (1997) *Screening for breast cancer: attendance and psychological consequences.*
PROMOTOR: prof dr WJA van den Heuvel. REFERENT: dr R Sanderman

Nieboer AP (1997) *Life-events and well-being: a prospective study on changes in well-being of elderly people due to a serious illness event or death of the spouse.*
PROMOTORES: prof dr SM Lindenberg, prof dr J Ormel

Eijk LM van (1997) *Activity and well-being in the elderly.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr SM Lindenberg. REFERENT: dr GIJM Kempen

Alberts JF (1998*) The professionalized patient: sociocultural determinants of health services utilization.*
PROMOTOR: prof dr WJA van den Heuvel. REFERENT: dr R Sanderman

Jong GM de (1999) Stress, stress management and issues regarding implementation. PROMOTORES: prof dr PMG Emmelkamp, prof dr JL Peschar. REFERENT: dr R Sanderman

Tiesinga LJ (1999) *Fatigue and Exertion Fatigue: from description through validation to application of the Dutch Fatigue Scale (DUFS) and the Dutch Exertion Fatigue Scale (DEFS).*
PROMOTORES: prof dr WJA van den Heuvel, prof dr ThWN Dassen. CO-PROMOTOR: dr RJG Halfens

Nijboer C (2000) *Caregiving to patients with colorectal cancer: a longitudinal study on caregiving by partners.* PROMOTORES: prof dr GAM van den Bos, prof dr R Sanderman. CO-PROMOTOR: dr AHM Triemstra

Doeglas DM (2000) *Functional ability, social support and quality of life: a longitudinal study in patients with early rheumatoid arthritis.*
PROMOTORES: prof dr WJA van den Heuvel, prof dr R Sanderman. CO-PROMOTOR: dr ThPBM Suurmeijer

Hoekstra-Weebers JEHM (2000) *Parental adaptation to pediatric cancer.*
PROMOTORES: prof dr EC Klip, prof dr WA Kamps. REFERENT: dr JPC Jaspers