Assessment of change in clinical evaluation

Middel, Lambertus Johannes

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 √ 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; ≥ . 20 < . 50 small effect; ≥ . 50 < . 80 moderate effect and ≥ . 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.