Outcome assessment in inpatient pulmonary rehabilitation: clinical results and methodological aspects
van Stel, H.F.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Interpretation of change and longitudinal validity of the Quality of Life for Respiratory Illness Questionnaire (QoLRIQ) in inpatient pulmonary rehabilitation

Henk F. van Stel
A. Rianne Maillé
Vivian T. Colland
Walter Everaerd

1) Asthmacenter Heideheuvel, Hilversum
2) Julius Center for Health Sciences and Primary Care,
University Medical Center Utrecht, Utrecht
3) Department of Health Psychology, Utrecht University, Utrecht
4) Department of Clinical Psychology, University of Amsterdam, Amsterdam

Quality of Life Research 2003, 12 (in press)
The more I look at it,
the more I like it.
I do think it's good.

The fact is..
no matter how closely I study it,
no matter how I take it apart,
no matter how I break it down,
It remains consistent.
I wish you were here to see it.

King Crimson
Indiscipline, 1981
6.1 Abstract
The Quality of Life for Respiratory Illness Questionnaire (QoLRIQ) is an outcome measure for patients with asthma or chronic obstructive pulmonary disease (COPD). This study assessed the longitudinal validity, reliability of the change score and the interpretation of changes on the QoLRIQ in inpatient pulmonary rehabilitation, completed by 108 patients with moderate to severe asthma (39) or COPD (69). Domains and total score of the QoLRIQ changed significant (all p<0.0002) with standardised response means from 0.46 to 0.90. All QoLRIQ-change scores were significantly correlated with self-rated change in health and in disease symptoms and with change in self-assessed health status (r from 0.2 to 0.61). There were several significant correlations between QoLRIQ-change scores and change in experienced invalidity, emotional well-being, anxiety, depressive symptoms and Rand-36-domains (r from 0.2 to 0.68). The intraclass correlation coefficient of change was 0.90. The size of a minimal important difference (MID), computed from a retrospective global rating of change by the patients and with the standard error of measurement, was 0.5 points on a 7-point response scale. Computation of the MID from retrospective assessment of change may not be valid because this change was significantly correlated to post-treatment health status and significantly higher than serial assessment of change.

We conclude that the QoLRIQ is sensitive to change, longitudinally valid and reliable, with a MID of 0.5 points. These results enable the use of the QoLRIQ as an outcome measure in clinical trials with patients with moderate to severe asthma or COPD. The longitudinal measurement properties in less severe patients still need to be studied.

6.2 Introduction

The Quality of Life for Respiratory Illness Questionnaire (QoLRIQ) is a disease-specific health related quality of life (HRQL) questionnaire designed for both patients with asthma and patients with chronic obstructive pulmonary disease (COPD) [1]. The psychometric characteristics of the QoLRIQ are good: a high reliability, good validity, and sensitive to differences in disease severity (see below) [2]. The QoLRIQ was intended to be used as an outcome measure in clinical research. However, evaluation of the longitudinal measurement properties has not been performed yet.

There are different views on the optimal nomenclature of the (longitudinal) measurement properties of HRQL questionnaires. Guyatt et al [3] suggested that adequate responsiveness is a prerequisite for an outcome measure, next to validity and reliability. Husted et al [4] proposed the term ‘external responsiveness’ for the extent to which changes in a measure correspond with changes in external, related measures. The difference with longitudinal validity [3] is that the external measure has to represent a widely accepted indication of
change. Recently, Beaton et al. suggested a taxonomy for responsiveness based on context [5], with axes for who is being analyzed, which scores are contrasted and the type of change. Terwee et al. [6] argued that the concept of responsiveness as a separate property of evaluative instruments [3] is not necessary: all measures of ‘responsiveness’ used in the literature are measures of longitudinal validity or reliability, while some also contain information that can be used for interpretation. They proposed guidelines for assessment of longitudinal validity, reliability of change scores and interpretation of change [6]. In this paper, we describe the longitudinal measurement properties of the QoLRIQ, using the guidelines proposed by Terwee. First, we assessed the statistical significance and relative magnitude of changes [7] detected by the QoLRIQ by performing significance tests and computing standardised response means (SRM), a variant of the effect size (ES) [8]. Second, we studied the longitudinal validity of the QoLRIQ by computing correlation coefficients between change scores in QoLRIQ-domains and change scores from related outcome measures [3;4]. Third, we assessed the reliability of the change score by computing the intraclass coefficient of change [6;9]. Fourth, we assessed the size of a minimal important difference (MID) using a retrospective global rating of change question. Because several authors question the validity of retrospective assessment of change [9-11], we studied the validity of that method and determined the MID with alternative methods: computing the standard error of measurement (SEM) [12-14] and using the ES-benchmarks [15]. This study was performed in an inpatient pulmonary rehabilitation setting (IPR) including both patients with asthma and patients with COPD. We chose this setting because pulmonary rehabilitation is known to be an effective treatment for patients with both asthma and COPD [16;17]. Because IPR is a multidisciplinary treatment with multiple treatment goals, we expected clinically relevant change in several quality of life domains. Furthermore, improvement of quality of life is a major goal in pulmonary rehabilitation. A disease-specific HRQL questionnaire with established longitudinal measurement properties would serve as an important outcome measure in clinical trials about pulmonary rehabilitation.

6.3 Study design and subjects

Patients with asthma or COPD referred for inpatient pulmonary rehabilitation (IPR) at Asthmacenter Heideheuvel were recruited for this study. The IPR is a rolling programme, so patients were included consecutively from January 1996 to December 1997. Patients who did not complete the IPR or did not speak Dutch were excluded from this study. 108 patients (39 with asthma, 69 with COPD) were included. Diagnosis including assessment of disease severity was done by a pulmonologist according to criteria from the European
Respiratory Society [18] and the National Heart, Lung and Blood Institute [19]. The patients completed all questionnaires both pre- and post-treatment. Pre-treatment assessments were done in the first week of the observation period preceding the 3- to 6-month inpatient pulmonary rehabilitation programme. Post-treatment data were collected in the week prior to discharge. All patients gave written informed consent and the study protocol was approved by the institutional medical ethics committee.

6.3.1 IPR programme description
The main reasons for referral to the IPR programme were an unstable disease pattern and/or a high burden of disease, characterized by frequent hospitalization, a high medication usage and/or psychosocial problems. The inpatient programme aims at optimization of functioning in daily life. Because of the large variation in individual problems and the essential role of motivation in pulmonary rehabilitation [20], individualized treatment goals are formulated by the multidisciplinary treatment team in consultation with the patient. The key components of the programme are exercise training, optimization of the medication regimen, education, extensive psychosocial support and training of self-management skills, including self-pacing and adequate symptom perception. The duration of the IPR ranges from 3 up to 6 months, depending on the specific problems and treatment goals of a patient.

6.3.2 Outcome measures
HRQL was assessed with the QoLRIQ. This questionnaire consists of 55 items divided into seven domain subscales: breathing problems (9 items), physical problems (9), emotions (9), situations triggering or enhancing breathing problems (7), general activities (4), daily and domestic activities (10), and social activities, relationships and sexuality (7) [1]. The QoLRIQ uses a 7-point response scale ranging from "not at all" to "very severe" to assess the degree of trouble from symptoms or impediment in carrying out activities in the two previous weeks. A higher score represents a higher level of impairment. Test-retest reliability (intra-class correlation) with a one-month interval has been tested in stable primary care patients. Stable was defined as self-reported stability with no visits to the general practitioner for breathing problems. Test-retest reliability was 0.54 for the emotions domain and ranged from 0.71 to 0.85 for the other domains and the total score [2]. Construct validity was satisfying, with moderate to good correlations with the total score of the SIP (0.30 to 0.57), with domains from the MOS SF20 (-0.35 to -0.61), with subjective severity of attacks/dyspnea (0.62) and with the MRC-dyspnea score (0.44; all p<0.001) [2]. The QoLRIQ discriminated significantly between primary care patients, outpatients and pulmonary rehabilitation patients, both in asthma and in COPD [2].
Level of airways obstruction was assessed by the forced expiratory volume in one second (FEV₁) and by the forced expiratory volume in one second as percentage of the predicted value (FEV₁,%pred) (adjusted for age, gender and body weight) [21]. Self-reported dyspnea was assessed with the five-point MRC/ECCS dyspnea item (range 1 – 5) [22]. Overall health status was assessed with a single item for self-perceived health status (“How would you rate your health status at this moment”: very good, good, fair, moderate, poor) which was slightly modified from the Netherlands Health Survey Interview [23] (we changed the category “sometimes good, sometimes bad” into “moderate” due to misunderstandings of the former wording). At discharge a retrospective "global rating of change" question was added. Patients were asked to rate self-perceived change in disease symptoms on a 5–point response scale: “much improved – improved – the same – worse – much worse”. Well-being was assessed with two domains from the Medical Psychological Questionnaire for Lung Patients (MPQL) [24]: emotional well-being (range 13-39, higher=better) and experienced invalidity (range 11-33, lower=better). The Symptom Checklist 90 (SCL-90) [25] was used to assess anxiety and depressive symptoms (range 10-50 and 16-80, higher=more symptoms). During the last phase of the study, the Dutch version of the Rand-36, a generic quality of life questionnaire, was added (N=31) [26].

6.3.3 Statistical Analysis
The scores of each QoLRIQ-domain are standardized by dividing the sum of valid scores by the number of valid items. A domain should have at least \( \frac{1}{2}n + 1 \) valid items, otherwise the domain score is missing. A total score (“QoLRIQ-total”) is computed in a similar way from all valid items. Two domains were divided into subdomains: a) the “situations” domain because of the large difference in change between the 4 items about weather conditions (“triggers:weather”) and the 3 items about allergic triggers (“triggers:allergic”) and b) the “social activities, relations and sexuality” domain because a large number of patients skipped the 3 items about sexuality which caused missings for the whole domain (“social:activities” and “social:sexuality”). Descriptive statistics for the (sub)domains and total score include baseline mean score; baseline standard deviation; change score; percent of patients scoring at the floor or ceiling of the score range; and for reliability the internal consistency (Cronbach’s \( \alpha \)).

Normality of the distributions of pre- and post-treatment scores and change scores from all measures was assessed with the Shapiro-Wilk W test [27]. Differences in baseline scores between the asthma and COPD-groups were tested for significance with the Mann-Whitney U-test.
6.3.4 Significance of change and effect sizes

Differences in change between the groups of patients with asthma and patients with COPD were tested with an independent t-test on the change scores. Significance of change was assessed by the Wilcoxon matched pairs test and accepted at $p<0.05$. A variant of the effect size, the standardised response mean (SRM) was computed to assess the relative magnitude of observed changes [4;7]. The SRM is computed as the mean difference divided by the standard deviation of that difference [8]. SRM is interpreted using the benchmarks by Cohen: 0.2 represents a small change; 0.5 a moderate change and changes of 0.8 or higher are interpreted as a large change [28]. For change scores which were not normally distributed, a nonparametric SRM (SRMnp) was computed as the median change divided by the interquartile range (iqr) from that change [29]. Based on the main programme goals and clinical experience, we expected moderate to large changes in the QoLRIQ-total score and in domains representing daily functioning (general activities, daily and domestic functioning, social:activities) and emotional functioning (emotions); small to moderate changes in physical symptoms (breathing problems, physical problems); and no change in triggers:allergic.

To check whether the size of the change depended on the initial value, the correlation between the change score and the average of pre- and post-treatment QoLRIQ-total score was computed [27].

As the QoLRIQ was originally developed for patients with mild to moderate severe asthma or COPD, the item content may not be optimal for patients with severe asthma or COPD. Hyland et al [30] suggested that creating a purpose-specific version may improve the responsiveness of quality of life-questionnaires. They used the Breathing Problems Questionnaire in an outpatient pulmonary rehabilitation setting and limited it to the 10 items most sensitive to change (7 out of 33 items had an SRM>0.2). To assess whether this strategy would be useful for the QoLRIQ, we checked how many items showed at least a small change (SRM>0.2).

6.3.5 Longitudinal validity

Longitudinal validity was assessed by computing Spearman’s rank order correlations ($r_s$) between change in QoLRIQ-domains (and total score) and change in related outcomes: self-perceived change in disease symptoms, self-assessed health status, experienced invalidity, emotional well-being, anxiety, depressive symptoms and Rand-36 domains. Measures without change, such as FEV$_1$, were excluded from this analysis. A priori predictions regarding the magnitude of the correlations included high correlations (>0.5) [31] for change in QoLRIQ-total score with change in self-assessed health status and self-rated change in disease symptoms; moderate to high correlations (0.35 - 0.6) for change in emotions with change in mental health, well-being, anxiety and depressive symptoms;
moderate to high correlations for change in daily/domestic activities, general activities and physical problems with change in experienced invalidity and physical functioning; and a moderate to high correlation for social activities with social functioning. Significance of correlations was accepted at p<0.05.

6.3.6 Reliability of the change score
Reliability of the change score was computed as an intraclass correlation coefficient of change in the QoLRIQ-total score, using the formula provided by Streiner and Norman [9].

6.3.7 Interpretation of change
The size of clinically important differences can be assessed from the mean change in the categories of a retrospective ‘global rating of change’ question. This anchor-based method has been outlined by Jaeschke et al [32] and Juniper et al [33]. The minimal important difference (MID) can be estimated from the average change among patients who rate their health as somewhat improved or deteriorated minus the average change in patients who rate their health as unchanged [4]. In this study the self-rated change in disease symptoms was selected for computing the MID for the QoLRIQ domains and total score.

Several hypotheses regarding the validity of computing the size of MIDs from retrospective questions were tested. First, a one-sample t-test [27] was used to assess whether the mean change in QoLRIQ-total score in the group with rating ‘the same’ was significantly different from zero. 95% confidence intervals (95%CI) for mean changes and for MIDs were also computed [34].

Second, Norman et al [10] state that patients are unable to recall their initial health status, which causes global measures of change to be highly correlated with the present state and uncorrelated with the initial state. Spearman’s rank correlation coefficients between the retrospective rating of change in disease symptoms and pre- and post-treatment scores for QoLRIQ-total and self-assessed health status were computed to test this hypothesis.

Third, Fischer et al [11] showed that the retrospective assessment of change is significantly higher than the serial assessment of change (pre-treatment minus post-treatment scores). We repeated their analyses on our own data using the change in QoLRIQ-total score as serial change and the self-rated change in disease symptoms as retrospective assessment of change. The difference in sensitivity to change of both types of measurement [11] was assessed by comparing the SRM-serial (the mean change in QoLRIQ-total score divided by the standard deviation of that change) and the SRM-retrospective (the post-treatment self-rated change score divided by the standard deviation of that score). The significance of this difference, including a 95%CI, was also computed [35]. The concordance between serial and retrospective change was assessed by constructing a contingency table and a) assessing the crude % agreement from the on-diagonal agreement in the contingency table and b)
performing a McNemar test on this table to test if one measure was significantly higher than the other[11]. For the contingency table the serial change was divided into five categories: 
< -1.5: large negative change; -1.5 to -0.5: moderate negative change; -0.5 to +0.5: no change; +0.5 to +1.5: moderate positive change; > +1.5: large positive change. This categorization was based on the MID-thresholds for HRQL-questionnaires with a 7-point response scale [32;33], which have been suggested to be valid for any HRQL-questionnaire using a similar scale [33]. Additionally, the SRMnp of the Rand-36 retrospective question about health change in the past half year was compared to the SRMnp's of significantly improved Rand-36 domains.

6.3.8 Alternative computation of MID
We used three alternative methods to compute the MID. The first was to compute the standard error of measurement (SEM): the standard deviation of an instrument multiplied by the square root of one minus its reliability coefficient. The SEM was validated by Wyrwich et al. as a criterion for meaningful intra-individual changes in three chronic disease HRQL measures [12-14]. A one-SEM change corresponded with the MID of 0.5 point change on a 7-point response scale. The SEM was computed for each domain using the baseline standard deviation and the internal consistency (Cronbach’s α) found in this study. The second method was to use the benchmarks for effect sizes: what is the absolute change associated with an ES of 0.2 or 0.5 [6;15]? The third method was to compare serial change in the QoLRIQ to serial change (one-unit changes) in self-assessed health status, as an anchor-based alternative for categorizing change by retrospective change.

All statistical analyses were performed with Statistica for Windows version 5.1 (Statsoft Inc, Tulsa, OK, USA, 1998).

6.4 Results

6.4.1 Baseline characteristics
The study group consisted of severely limited patients. 68.5% of the patients had a diagnosis of severe asthma or COPD; the MRC-dyspnea score was high; 61% of the patients assessed their health status as moderate or poor (see table 6.1). The QoLRIQ-domainscores were not normally distributed, except for the QoLRIQ-totalscore and all change-scores. Two domains had noticeable floor or ceiling effects at baseline: almost 30% of the patients reported no impairment in the subdomain triggers:allergic, while almost 20% had the highest possible score for the subdomain social:activities. Baseline differences in quality of life between the asthma- and COPD-groups were only seen within domains representing
daily functioning: general activities (p=0.04) and daily/domestic activities (p<0.0001). Because of these differences the baseline QoLRIQ-total score tends to be somewhat worse in patients with COPD (p=0.06). Experienced invalidity was significantly worse (p=0.004) in patients with COPD, but emotional well-being, anxiety and depressive symptoms did not differ between patients with asthma and patients with COPD.

Table 6.1: Baseline characteristics

<table>
<thead>
<tr>
<th></th>
<th>Asthma</th>
<th>COPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>total N</td>
<td>39</td>
<td>69</td>
</tr>
<tr>
<td>diagnosis “severe” (n)</td>
<td>21</td>
<td>53</td>
</tr>
<tr>
<td>gender (n male / n female)</td>
<td>8 / 31</td>
<td>37 / 32</td>
</tr>
<tr>
<td>age (years) a</td>
<td>46.6 (16.6)</td>
<td>60.4 (11.0)</td>
</tr>
<tr>
<td>MRC-dyspnea score b,c</td>
<td>5 (0)</td>
<td>5 (0)</td>
</tr>
<tr>
<td>self-assessed health status b,d</td>
<td>4 (1)</td>
<td>4 (1)</td>
</tr>
<tr>
<td>FEV1 (liter) a</td>
<td>2.29 (0.82)</td>
<td>1.03 (0.50)</td>
</tr>
<tr>
<td>FEV1, % predicted d</td>
<td>78.7 (23.1)</td>
<td>36.6 (14.1)</td>
</tr>
</tbody>
</table>

a mean (sd); b median (interquartile range); c range: 1 (no dyspnea) to 5 (maximal dyspnea); d range: 1 (very good) to 5 (poor)

6.4.2 Significance of change and effect sizes

There were no significant differences between the patients with asthma and COPD in change in domains from the QoLRIQ, MPQL or SCL-90 (all p>0.2), in change in walking distance or in number of exacerbations during treatment (data not shown); SEM-values were very similar (see below). Therefore the change analysis was performed on all within-patient differences together. All QoLRIQ-domains showed statistically significant improvement, except for triggers:allergic (see table 6.2). Absolute differences ranged from 0.49 to 1.32 on a 7-point scale. The largest differences were seen in general activities, social:activities and daily/domestic activities. SRM’s of the QoLRIQ-domains ranged from 0.46 to 0.90, reflecting moderate to large effect sizes. 46 out of 55 items showed at least a small change (SRM>0.2).

The size of the change score did not depend on the initial value, as shown by a correlation of -0.12 (p=0.2) between the QoLRIQ-total change score and the average of pre- and posttreatment QoLRIQ-total scores.

All selected domains from other questionnaires improved significantly, especially emotional
well-being and self-assessed health status (see table 6.3), which showed highly significant changes and large effect sizes. Domain and change scores from the Rand-36 were not normally distributed. Five domains of the Rand-36 improved significantly: physical functioning, role-emotional, mental health, vitality and pain (see table 6.3). The groups with (n=31) and without Rand-36 were tested for differences in QoLRIQ baseline and change scores. There were no significant differences except for improvement in social:activities (p=0.003) which was higher in the group without Rand-36.

6.4.3 Longitudinal validity
The change scores from all domains of the QoLRIQ, except “triggers: allergic” which did not change, were significantly correlated with change in self-assessed health status ($r$ from 0.37 to 0.61, all $p<0.001$) and with self-rated change in disease symptoms ($r$ from 0.20 to 0.51, most $p<0.001$) (see table 6.4). All domains, except “triggering: weather”, were significantly correlated with change in experienced invalidity ($r$ from 0.25 to 0.42). There were some smaller correlations for emotional well-being with “physical problems” ($r=0.33$), “general activities” ($r=0.25$), “emotions” ($r=0.40$) and the QoLRIQ-total score ($r=0.27$) (all $p<0.05$). Change in “emotions” was significantly correlated to change in anxiety ($r=0.40$, $p<0.0001$) and to change in depressive symptoms ($r=0.47$, $p<0.0001$). Change in depressive symptoms was also correlated to other QoLRIQ-domains (see table 6.4). Change in all domains of the QoLRIQ except social relations was significantly correlated with the health change question from the Rand-36 (see table 6.5). There were several moderate to high and significant correlations between change in QoLRIQ-total score and change in the Rand-36-domains physical functioning, social functioning, role-physical, mental health and vitality (see table 6.5); there were no significant correlations with change in bodily pain, role-emotional and general health.
<table>
<thead>
<tr>
<th>Domain name</th>
<th>baseline score (sd)</th>
<th>N</th>
<th>internal consistency</th>
<th>% floor</th>
<th>% ceiling</th>
<th>change score</th>
<th>p-value of change</th>
<th>SRM</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breathing problems</td>
<td>3.53 (1.09)</td>
<td>108</td>
<td>0.81</td>
<td>0.00</td>
<td>0.00</td>
<td>0.71</td>
<td>&lt;0.0001</td>
<td>0.60</td>
<td>0.48</td>
</tr>
<tr>
<td>Physical problems</td>
<td>3.38 (1.02)</td>
<td>108</td>
<td>0.78</td>
<td>0.90</td>
<td>0.00</td>
<td>0.70</td>
<td>&lt;0.0001</td>
<td>0.74</td>
<td>0.48</td>
</tr>
<tr>
<td>Emotions</td>
<td>3.27 (1.24)</td>
<td>108</td>
<td>0.88</td>
<td>0.09</td>
<td>0.00</td>
<td>0.84</td>
<td>&lt;0.0001</td>
<td>0.72</td>
<td>0.43</td>
</tr>
<tr>
<td>General activities</td>
<td>4.26 (1.46)</td>
<td>108</td>
<td>0.80</td>
<td>2.80</td>
<td>3.70</td>
<td>1.32</td>
<td>&lt;0.0001</td>
<td>0.90</td>
<td>0.65</td>
</tr>
<tr>
<td>Triggering situations</td>
<td>3.45 (1.06)</td>
<td>106</td>
<td>0.78</td>
<td>1.90</td>
<td>0.00</td>
<td>0.49</td>
<td>&lt;0.0001</td>
<td>0.46</td>
<td>0.50</td>
</tr>
<tr>
<td>Triggers: weather</td>
<td>4.23 (1.22)</td>
<td>106</td>
<td>0.74</td>
<td>1.90</td>
<td>1.90</td>
<td>0.79</td>
<td>&lt;0.0001</td>
<td>0.58</td>
<td>0.62</td>
</tr>
<tr>
<td>Triggers: allergic</td>
<td>2.36 (1.42)</td>
<td>107</td>
<td>0.81</td>
<td>28.70</td>
<td>1.90</td>
<td>0.05</td>
<td>0.6</td>
<td>0.05</td>
<td>0.62</td>
</tr>
<tr>
<td>Daily/domestic activities</td>
<td>4.54 (1.40)</td>
<td>99</td>
<td>0.92</td>
<td>0.90</td>
<td>1.90</td>
<td>0.95</td>
<td>&lt;0.0001</td>
<td>0.75</td>
<td>0.40</td>
</tr>
<tr>
<td>Social activities</td>
<td>4.21 (1.64)</td>
<td>64</td>
<td>0.92</td>
<td>1.90</td>
<td>5.60</td>
<td>0.81</td>
<td>&lt;0.0001</td>
<td>0.54</td>
<td>0.46</td>
</tr>
<tr>
<td>Social: activities</td>
<td>4.70 (1.80)</td>
<td>101</td>
<td>0.90</td>
<td>2.80</td>
<td>18.50</td>
<td>0.96</td>
<td>&lt;0.0001</td>
<td>0.50</td>
<td>0.57</td>
</tr>
<tr>
<td>Social: sexuality</td>
<td>3.83 (1.83)</td>
<td>68</td>
<td>0.92</td>
<td>4.60</td>
<td>7.40</td>
<td>0.82</td>
<td>0.0002</td>
<td>0.46</td>
<td>0.52</td>
</tr>
<tr>
<td>QoLRIQ-total</td>
<td>3.77 (0.90)</td>
<td>108</td>
<td>0.94</td>
<td>0.00</td>
<td>0.00</td>
<td>0.82</td>
<td>&lt;0.0001</td>
<td>0.89</td>
<td>0.22</td>
</tr>
</tbody>
</table>

*percent at lowest degree of impairment (score = 1.0); percent with highest degree of impairment (score = 7.0); SRM= standardized response mean (change score / sd of change score); SEM=standard error of measurement (sd*N*(1-r_x), r_x=reliability coefficient, i.e. Cronbach’s α)
### Table 6.3: Change in health status and psychological functioning

<table>
<thead>
<tr>
<th>Domain</th>
<th>baseline score (iqr)</th>
<th>post-treatment score (iqr)</th>
<th>p-value of change</th>
<th>SRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-assessed health status</td>
<td>4 (1)</td>
<td>3 (1)</td>
<td>&lt;0.0001</td>
<td>0.91 *</td>
</tr>
<tr>
<td>Emotional well-being</td>
<td>19 (10)</td>
<td>29 (15)</td>
<td>&lt;0.0001</td>
<td>0.86 *</td>
</tr>
<tr>
<td>Experienced invalidity</td>
<td>30 (5)</td>
<td>28 (8)</td>
<td>0.002</td>
<td>0.33</td>
</tr>
<tr>
<td>Anxiety</td>
<td>17 (10)</td>
<td>16 (9)</td>
<td>0.013</td>
<td>0.20</td>
</tr>
<tr>
<td>Depressive symptoms</td>
<td>27 (14)</td>
<td>24 (14)</td>
<td>0.006</td>
<td>0.22</td>
</tr>
<tr>
<td>Rand-36: (n=31)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical functioning</td>
<td>25 (35)</td>
<td>25 (35)</td>
<td>0.018</td>
<td>0.20</td>
</tr>
<tr>
<td>Social functioning</td>
<td>44.4 (33.3)</td>
<td>44.4 (22.2)</td>
<td>0.15</td>
<td>0.25</td>
</tr>
<tr>
<td>Role-physical</td>
<td>0 (25)</td>
<td>0 (50)</td>
<td>0.12</td>
<td>0</td>
</tr>
<tr>
<td>Role-emotional</td>
<td>0 (66.7)</td>
<td>66.7 (100)</td>
<td>0.002</td>
<td>0.50</td>
</tr>
<tr>
<td>Mental health</td>
<td>52 (32)</td>
<td>64 (28)</td>
<td>0.04</td>
<td>0.50</td>
</tr>
<tr>
<td>Vitality</td>
<td>40 (25)</td>
<td>55 (15)</td>
<td>&lt;0.0001</td>
<td>0.75</td>
</tr>
<tr>
<td>Pain</td>
<td>55.1 (67.3)</td>
<td>67.3 (55.1)</td>
<td>0.01</td>
<td>0.35</td>
</tr>
<tr>
<td>General health</td>
<td>20 (16)</td>
<td>28 (20)</td>
<td>0.26</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Data are presented as medians (interquartile range) with a non-parametric SRM, except *: parametric SRM. Score ranges: self-assessed health status: 1 (very good) to 5 (poor); emotional well-being: 13-39, higher=better; experienced invalidity: 11-33, lower=better; anxiety: 10-50, higher=more symptoms; depressive symptoms: 16-80, higher=more symptoms; Rand 36 domains: range 0-100, higher=best.

#### 6.4.4 Reliability of the change score

Based on a pre-treatment variance of 0.81 and reliability (Cronbach’s α) of 0.94; a post-treatment variance of 0.96 and reliability of 0.96; and a correlation between pre- and post-treatment scores of 0.54, the intraclass correlation coefficient of change in the QoLRIQ-total score was 0.90.
Table 6.4: correlation of QoLRIQ-change scores with change in health status and psychological functioning

<table>
<thead>
<tr>
<th>change scores (spearman’s r)</th>
<th>self-assessed self-rated change health status in disease symptoms</th>
<th>emotional well-being</th>
<th>experienced invalidity</th>
<th>anxiety depressive symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breathing problems</td>
<td>.39***</td>
<td>.32**</td>
<td>.18</td>
<td>.31**</td>
</tr>
<tr>
<td>Physical problems</td>
<td>.45***</td>
<td>.34***</td>
<td>.34**</td>
<td>.37***</td>
</tr>
<tr>
<td>Emotions</td>
<td>.49***</td>
<td>.37***</td>
<td>.40***</td>
<td>.42***</td>
</tr>
<tr>
<td>General activities</td>
<td>.54***</td>
<td>.51***</td>
<td>.25</td>
<td>.25*</td>
</tr>
<tr>
<td>Triggers: weather</td>
<td>.37***</td>
<td>.20*</td>
<td>.12</td>
<td>.17</td>
</tr>
<tr>
<td>Daily/domestic activities</td>
<td>.48***</td>
<td>.48***</td>
<td>.15</td>
<td>.34**</td>
</tr>
<tr>
<td>Social activities</td>
<td>.42***</td>
<td>.42***</td>
<td>.17</td>
<td>.39**</td>
</tr>
<tr>
<td>QoLRIQ-total</td>
<td>.61***</td>
<td>.48***</td>
<td>.27*</td>
<td>.41***</td>
</tr>
</tbody>
</table>

*p<0.05  **= p<0.01  ***p<0.001.
The domain “triggers: allergic” was omitted because of a lack of change.

Table 6.5: correlation of QoLRIQ change scores with Rand-36 change scores

<table>
<thead>
<tr>
<th>change scores (spearman’s r)</th>
<th>health change</th>
<th>physical function</th>
<th>social function</th>
<th>role-physical</th>
<th>role-mental</th>
<th>mental health</th>
<th>vitality</th>
<th>bodily pain</th>
<th>general health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breathing problems</td>
<td>.49**</td>
<td>.28</td>
<td>.35</td>
<td>.47**</td>
<td>.25</td>
<td>.31</td>
<td>.31</td>
<td>.08</td>
<td>.29</td>
</tr>
<tr>
<td>Physical problems</td>
<td>.35*</td>
<td>.25</td>
<td>.05</td>
<td>.19</td>
<td>.23</td>
<td>.19</td>
<td>.18</td>
<td>.09</td>
<td>.14</td>
</tr>
<tr>
<td>Emotions</td>
<td>.41*</td>
<td>.22</td>
<td>.47**</td>
<td>.61***</td>
<td>.12</td>
<td>.54**</td>
<td>.32</td>
<td>.09</td>
<td>.19</td>
</tr>
<tr>
<td>General activities</td>
<td>.45*</td>
<td>.66***</td>
<td>.46*</td>
<td>.57**</td>
<td>.29</td>
<td>.56**</td>
<td>.46**</td>
<td>.23</td>
<td>.13</td>
</tr>
<tr>
<td>Triggers: weather</td>
<td>.41*</td>
<td>.29</td>
<td>.29</td>
<td>.19</td>
<td>.13</td>
<td>.41*</td>
<td>.62***</td>
<td>.06</td>
<td>.14</td>
</tr>
<tr>
<td>Daily/domestic activities</td>
<td>.54**</td>
<td>.51**</td>
<td>.43*</td>
<td>.46*</td>
<td>.11</td>
<td>.42*</td>
<td>.38</td>
<td>.30</td>
<td>.33</td>
</tr>
<tr>
<td>Social activities</td>
<td>.38</td>
<td>.68**</td>
<td>.45</td>
<td>.34</td>
<td>.21</td>
<td>.46</td>
<td>.36</td>
<td>.18</td>
<td>.22</td>
</tr>
<tr>
<td>QoLRIQ-total</td>
<td>.54**</td>
<td>.52**</td>
<td>.56**</td>
<td>.57***</td>
<td>.25</td>
<td>.55**</td>
<td>.52**</td>
<td>.16</td>
<td>.22</td>
</tr>
</tbody>
</table>

* = p<0.05;  ** = p<0.01;  ***p<0.001.
The domain “triggers: allergic” was omitted because of a lack of change.
6.4.5 Interpretation of change
The mean change in QoLRIQ-total in improved and deteriorated patients was 0.93 and -0.07 respectively; the mean change in patients who rated themselves as unchanged was 0.42 points (see table 6.6). This results in a MID for improvement of 0.51 points (95% CI: 0.04 to 0.98) and a MID for deterioration of 0.49 (95% CI: -0.11 to 1.09). MIDs for the domains ranged from 0.06 to 0.97 points.

<table>
<thead>
<tr>
<th>global rating of change: disease symptoms*</th>
<th>worse</th>
<th>the same</th>
<th>improved</th>
<th>much improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean change (SD) in QoLRIQ-total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MID</td>
<td>-0.49</td>
<td></td>
<td>+0.51</td>
<td></td>
</tr>
<tr>
<td>range of change</td>
<td>-1.07 to 0.81</td>
<td>-0.72 to 2.23</td>
<td>-0.53 to 3.46</td>
<td>-0.22 to 3.88</td>
</tr>
<tr>
<td>95% CI of mean</td>
<td>-0.54 to 0.40</td>
<td>0.04 to 0.80</td>
<td>0.65 to 1.21</td>
<td>0.95 to 1.45</td>
</tr>
<tr>
<td>(n=10)</td>
<td></td>
<td>(n=18)</td>
<td>(n=37)</td>
<td>(n=40)</td>
</tr>
</tbody>
</table>

* Category “much worse” with only 2 patients was omitted; MID = minimal important difference

Table 6.6: Categorizing QoLRIQ-change by global rating of change

Testing the hypotheses about the validity of retrospective assessment of change showed that the mean change in the ‘unchanged’ group had a large 95% confidence interval and was significantly different from zero (p<0.05). The retrospective rating of change in disease symptoms was significantly correlated to the post-treatment QoLRIQ-total score ($r_s=0.55$, $p<0.0001$) and to post-treatment self-assessed health status ($r_s=0.51$, $p<0.0001$) but not to the pre-treatment QoLRIQ-total score ($r_s=0.15$, $p=0.13$) and pre-treatment self-assessed health status ($r_s=0.19$, $p=0.06$). The comparison of effect sizes for retrospective assessment and serial assessment of change showed that the SRM-retrospective was 1.95 and SRM-serial 0.89. The difference between these SRM’s was significant ($p<0.0001$, 95% CI of difference: 0.86 to 1.24). The crude percentage agreement in the contingency table was 39.3%. The McNemar test of the probability that one measure gives higher scores than the other was significant ($p<0.05$). The retrospective question from the Rand-36 about health change had a SRMnp of 1.0 while the Rand-36 domains (except vitality) had SRMnp’s of 0.2 to 0.5.

For computation of the standard error of the mean (SEM), the internal consistency is needed. The internal consistency ranged from 0.78 for ‘physical problems’ and ‘triggering situations’ to 0.94 for the QoLRIQ-total score (see table 6.2). The SEM for the domains ranged from 0.4 points for ‘daily/domestic activities’ to 0.65 for ‘general activities’, with a mean of 0.49 points (see table 6.2). The SEM for the QoLRIQ-total score had a value of
0.22 points. SEM-values did not differ between the groups with asthma and COPD. Computation of the size of a MID from the ES-benchmarks and the pre-treatment SD of the QoLRIQ-total score gave MID-values of 0.18 points for a small ES (0.2) and 0.45 points for a moderate ES (0.5). For the QoLRIQ-domains, the value for a MID ranged from 0.2 to 0.33 points for a small ES and from 0.51 to 0.82 points for a moderate ES. Categorizing change in the QoLRIQ-total score by one-unit changes in self-assessed health status gave an MID for improvement of 0.37 and MID for deterioration of 0.64 (see table 6.7). The group of patients with no change in self-assessed health status showed a mean change in QoLRIQ-total score of 0.34, which is significantly different from zero (p=0.008).

<table>
<thead>
<tr>
<th>change in self-assessed health status*</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean change (SD) in QoLRIQ-total</td>
<td>-0.3 (0.54)</td>
<td>0.34 (0.64)</td>
<td>0.71 (0.67)</td>
<td>1.37 (0.96)</td>
<td>1.89 (0.77)</td>
</tr>
<tr>
<td>(n=6)</td>
<td>(n=29)</td>
<td>(n=36)</td>
<td>(n=25)</td>
<td>(n=9)</td>
<td></td>
</tr>
</tbody>
</table>

* Categories +4 and -2, each with only one patient, were omitted

### 6.5 Discussion

We have assessed the longitudinal measurement properties of the Quality of Life for Respiratory Illness Questionnaire [1], a disease-specific questionnaire for HRQL, using the guidelines proposed by Terwee et al [6]. The study group consisted of both patients with asthma and patients with COPD who completed an inpatient pulmonary rehabilitation program. Diagnosis, MRC-dyspnea score and self-assessed health status showed that most of these patients were severely limited. All domains of the QoLRIQ showed statistically significant changes, with moderate to large effect sizes, except for the subdomain ‘triggers: allergic’ which had a large baseline floor effect. The QoLRIQ seems to be very sensitive to change: most items showed at least a small change, which makes a restricted version[30] unnecessary. As expected, domains representing daily functioning (‘general activities’, ‘daily/domestic activities’ and ‘social: activities’) showed the largest absolute improvement. This improvement resembles the high impairment patients’ report in this area and the treatment focus on improving daily functioning. The observed changes were highly correlated with related measures of function such as change in self-assessed health status and self-rated change in disease symptoms. There were lower correlations with change in experienced invalidity. As expected, the “emotions” domain correlated good with change in anxiety and depressive symptoms as assessed with the SCL-90 and with change in
emotional well-being assessed with the MPQL. QoLRIQ-change scores were also significant correlated to change in several Rand-36 domains, especially with the health change question. The reliability of the change score, assessed with the intraclass correlation coefficient of change in the QoLRIQ-total score, was high.

We selected the SRM as the most appropriate effect size statistic in this study, because it accounts for the variation in change by using the standard deviation of the observed change [8;36]. There are variants of the effect size which use the standard deviation of stable subjects. We had serious doubt that the patients who rated their disease symptoms as "the same" in this study were truly unchanged: their mean QoLRIQ-total change score was significantly different from zero. Therefore the responsiveness ratio by Guyatt [37] (which uses the standard deviation of difference scores in stable subjects) or the calibrated responsiveness ratio [29] (which uses the difference in change between self-rated clinically improved and stable subjects, divided by the standard deviation of stable subjects) could not be used.

6.5.1 Interpretation of change
The minimal important difference or MID was assessed with both anchor-based and distribution-based methods [7;38]: from retrospective assessment of change in disease symptoms, from change in self-assessed health status, by computation of the standard error of the mean (SEM) and from the ES-benchmarks. The ‘retrospective’ method gave a MID for the QoLRIQ-total score of 0.5 point in both positive and negative direction, although the range and 95% CI of the mean change in each category of self-rated change were wide and the MIDs for the domains ranged from 0 to 1 point. Categorizing by change in self-assessed health status gave positive and negative MIDs for the QoLRIQ-total score of 0.4 and 0.6 respectively. The mean SEM-values for the QoLRIQ-domains was similar to the retrospective MID. The SEM for the QoLRIQ-total score was much smaller, which is caused by the larger number of items used to compute the QoLRIQ-total score, resulting in a much smaller standard deviation and a high reliability. Using a moderate effect size gave MIDs quite similar to the SEM-based MIDs. So, 0.5 seems the best point estimate for the MID, within a range from 0.4 to 0.6.

6.5.2 Retrospective MID
Our results regarding the size of the retrospectively computed minimal important differences (MIDs) were to some extent similar to the findings of Jaeschke et al [32] and Juniper et al [33]. They used a global rating of change question with a 15-point Likert scale to assess the MID for the Chronic Respiratory Questionnaire (CRQ), Chronic Heart Questionnaire (CHQ) and Asthma Quality of Life Questionnaire (AQLQ). For all three questionnaires a change of 0.5 points may be considered as the MID [32;33]. As an
example, in the Juniper study on the AQLQ patients who rated themselves as the same changed 0.11 points, while patients who improved or deteriorated a little/somewhat changed 0.41 resp. -0.62 points [33]. Similar values were found for the CRQ and CHQ [32]. We also found an anchor-based MID of 0.5 for the QoLRIQ-total score, but a) the mean change in patients with rating ‘the same’ was significantly different from zero and identical to the mean change in ‘improved’ patients in the Juniper study, and b) the MIDs for the domains varied between 0 and 1 unit. There are several possible explanations why our findings differ. First, we used a 5-point response scale for the global rating of change question. The results of this scale may differ from the 15-point scale used in the studies by Jaeschke and Juniper, although their scale was abbreviated to a 7-point scale. Second, our retrospective question about change in disease symptoms is not specific, which may explain the large variation in MIDs for the domains: overall change in disease symptoms is not likely to be similar to change in a specific domain. Third, differences in study group, time frame and intervention may explain some variation. Our study group consisted mostly of patients with severe asthma or severe COPD who completed 3 to 6 months of inpatient treatment. The Juniper study used patients with symptomatic asthma, doing three assessments in an 8-week period, without intervention. However, our results also differ from the study by Jaeschke, who combined patients from pulmonary rehabilitation and trials on salbutamol and digoxin.

We think that the long duration of our IPR-program introduces recollection error, which threatens the validity of the retrospective assessments [39]. Furthermore, the treatment probably causes response shift [40] by actively intervening in mechanisms for coping with disease and thereby modifying the internal standards, values or conceptualisation of HRQL of a patient.

6.5.3 Validity of retrospective assessment of change

We studied the validity of retrospective assessment of change because some authors state that patients are unable to recall their initial health status [9;10]. Our results confirmed this statement: the retrospective change question was indeed significantly correlated to the post-treatment scores for QoLRIQ-total and self-assessed health status but not to the pre-treatment scores.

Fischer et al [11] showed that the retrospective assessment of change is significantly higher than the serial assessment of change (pre-treatment minus post-treatment scores). Their observations were in detail confirmed in our study. There was a large difference in sensitivity to change: the retrospective SRMs were double the size of the serial SRMs. The contingency table showed a very similar crude agreement, while the McNemar test on the contingency table was also significant. This means that retrospective assessment of change results in significantly higher change scores than serial assessment, which was also found by
Norman et al [10]. We agree with Fischer and coworkers that retrospective assessment of change is different from (and complementary to) serial assessment of change in quality of life. However, these findings implicate that the global rating of change method for determining MIDs is not valid, at least not in this study. Despite the lack of validity, we found a MID of 0.5 points (although only for the total score), which is similar to the MID for the CRQ and AQLQ.

6.5.4 Alternative computation of MID
Recently Wyrwich et al. validated the standard error of measurement (SEM) as a criterion for meaningful intra-individual changes in three chronic disease quality of life measures [12-14]. The SEM has two useful properties: it is sample-independent and it is expressed in the original metric of the instrument[12]. They concluded that a one-SEM change corresponded well to the anchor-based MID standards for domains from the CRQ, CHQ and AQLQ: 0.5 points on a 7-point response scale. The SEM-values for the QoLRIQ-domains, with a mean of 0.49 points, are very similar to the SEM-value proposed by Wyrwich. This is the fourth study that supports the value of the SEM-criterion of 0.5 points for quality of life-domains with a 7-point response scale. However, a clinical validation of the 0.5 MID value for the QoLRIQ is still necessary. Although the retrospective anchor-based method did give the same value, the apparent lack of validity of retrospective assessment makes its contribution doubtful.

Because the SEM-values for the QoLRIQ-domains lie closely around 0.5 points, it seems sensible to use this value also for the total score. Its lower SEM-value may however imply that the QoLRIQ-total score is sensitive to very small changes in quality of life. This also shows from the small MID for the QoLRIQ-total score when computed from the small ES-benchmark. We do however not think that a small ES or the accompanying absolute change of about 0.20 points resembles a clinically relevant change after an intensive treatment like pulmonary rehabilitation.

6.5.5 Other HRQL-measures in asthma and COPD
The longitudinal measurement properties of other disease-specific and generic QoL-questionnaires used for patients with asthma or COPD have been assessed mostly by the ability to detect statistically significant changes and by the longitudinal validity [30;41-48]. The size of MIDs has been assessed only in some instruments. For the CRQ and AQLQ-Juniper the global rating of change method as described above was used. This resulted in thresholds of 0.5 (small change), 1.0 (moderate change) and 1.5 points (large change) on a 7-point response scale. The MID of 0.5 points for for 7-point scales was confirmed in a study by Redelmeier et al on the CRQ which used patients' interpersonal judgment of change [49] and in studies of Wyrwich et al on the size of the SEM-criterion[12-14].
Multiple regression analysis on physician- and nurse-derived hypothetical clinically relevant differences in wheeze, cough, 6-minute walking distance, dyspnea and depression was used to assess the MID of the St George’s Respiratory Questionnaire [50]. A change of 4 units on a 100-unit scale indicated a clinically significant difference, which has been confirmed by patient and physician estimates of treatment efficacy [51]. The size of MIDs of other respiratory questionnaires have not been reported. Because we did not compare our questionnaire to other respiratory HRQL-questionnaires, we can not conclude if the QoLRIQ performs better or worse than other questionnaires. The good cross-sectional [1;2] and longitudinal measurement properties, the ease of scoring and the current development of a short version for monitoring, make the QoLRIQ a good candidate for use in both clinical practice and research settings.

6.5.6 Limitations
There are some limitations to this study. We did not include patients with less severe asthma or COPD (general practice, outpatient clinic, outpatient pulmonary rehabilitation), so the question remains if the QoLRIQ has good longitudinal measurement properties in less severely ill patients. The low level of impairment measured in primary care patients [2] may impede the detection of improvement, although Terwee et al suggested that the MID is not a fixed property of an instrument and that less intensive treatments may have a smaller MID [52].

We did not control for multiple comparisons in the part on longitudinal validity. Because of the small number of patients completing the Rand-36, we left the level of significance at 0.05, because correlations can not be expected to be significant at a more stringent level with only 31 patients. However, our conclusion that the longitudinal validity of the QoLRIQ is good, does not change when applying a far more stringent level of 0.001, because a considerable amount of correlations are still significant at that level.

6.6 Conclusion

In conclusion, we have shown that the QoLRIQ is sensitive to change, has a good longitudinal validity and reliability, and has a MID of 0.5 points (with a range from 0.4 to 0.6) in pulmonary rehabilitation for patients with moderate to severe asthma or COPD. The method of computing MID’s with global rating of change questions was not valid in this study because the retrospective assessment of change was significantly correlated to post-treatment health status and significantly higher than serial assessment of change. The SEM-value of 0.5 points as a threshold for meaningful change (MID) in domains of questionnaires with a 7-point response scale was confirmed in this study. These results enable the use of
the QoLRIQ as an outcome measure in clinical trials with patients with moderate to severe asthma or COPD. The longitudinal measurement properties in less severe patients still need to be studied.
6.7 Reference List


