Relative responsiveness of the Chronic Respiratory Questionnaire, St. Georges Respiratory Questionnaire and four other health-related quality of life instruments for patients with chronic lung disease

Milo A. Puhan\textsuperscript{a,b}, Gordon H. Guyatt\textsuperscript{b,c}, Roger Goldstein\textsuperscript{d}, Jeff Mador\textsuperscript{e,h}, Douglas McKim\textsuperscript{f}, Elisabeth Stahl\textsuperscript{g,h}, Lauren Griffith\textsuperscript{b}, Holger J. Schünemann\textsuperscript{b,i,j,*}

\textsuperscript{a}Horten Centre, University of Zurich, Switzerland
\textsuperscript{b}Department of Clinical Epidemiology and Biostatistics, Hamilton, Ont., Canada
\textsuperscript{c}Department of Medicine, McMaster University, Hamilton, Ont., Canada
\textsuperscript{d}Department of Medicine, University of Toronto, Toronto, Ont., Canada
\textsuperscript{e}Veteran Affairs Medical Center, Buffalo, NY, USA
\textsuperscript{f}University of Ottawa, Ont., Canada
\textsuperscript{g}AstraZeneca R & D, Lund, Sweden
\textsuperscript{h}University of Aberdeen, Aberdeen, UK
\textsuperscript{i}Department of Medicine and Social and Preventive Medicine, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, NY, USA
\textsuperscript{j}Clinical Research Development and INFORMAtion Translation Unit, Italian National Cancer Institute Regina Elena, Rome, Italy

Received 6 July 2005; accepted 26 April 2006

KEYWORDS
Chronic Obstructive Pulmonary Disease; Quality of Life; Chronic Respiratory Questionnaire; St. Georges

Summary
Background: Selection of health-related quality of life (HRQL) instruments that are most responsive to changes in HRQL prevents investigators from drawing false-negative conclusions about the effectiveness of an intervention. The objective of this study was to compare the responsiveness of the Chronic Respiratory Questionnaire (CRQ), the St. Georges Respiratory Questionnaire (SGRQ) and four other HRQL instruments.

\textsuperscript{*}Corresponding author. Clinical Research Development and Information Translation Unit, Italian National Cancer Institute Regina Elena, Rome, Italy. Tel.: +3906 52665102.
E-mail address: schuneh@mcmaster.ca (H.J. Schünemann).
Introduction

Clinicians and investigators recognize the importance of measuring health-related quality of life (HRQL) as a patient-important outcome in clinical trials of patients with chronic obstructive pulmonary disease (COPD) and other chronic lung disease.\(^1\)\(^-\)\(^9\) Because there is no single gold standard for the measurement of HRQL, investigators need to select instruments that meet their research goal. If change in HRQL is the outcome of interest, as it is in the evaluation of treatments, responsiveness is a key measurement property. Responsiveness refers to an instrument’s ability to detect HRQL change, even if the magnitude of change is small.\(^{10}\) If instruments are not responsive, investigators and clinicians could draw false-negative conclusions from results of clinical research and monitoring of health programs, including respiratory rehabilitation.\(^{11}\)

A number of validated HRQL instruments for patients with chronic lung disease exist and investigators often encounter challenges in selecting the optimal instrument. In general, disease-specific instruments such as the Chronic Respiratory Questionnaire (CRQ)\(^{12}\) are more responsive than generic instruments such as the Sickness Impact Profile or SF-36. Indeed, earlier studies with COPD patients indicated superior responsiveness of disease-specific instruments.\(^{11,13,14}\) Thus, when responsiveness is key, investigators need to choose among disease-specific instruments such as the widely used CRQ and St. George Respiratory Questionnaire (SGRQ).\(^{15}\)

A previous study with COPD patients following respiratory rehabilitation found larger effect sizes for the CRQ,\(^{14}\) while another study assessing the effects of salmeterol and ipratropium bromide did not clearly favour the CRQ or SGRQ in terms of responsiveness.\(^{16}\) However, the latter study used a non-validated self-administered CRQ version. We recently validated a self-administered and standardized version of the original CRQ\(^{17-19}\) that facilitates the administration of the CRQ in a clinical trial or practice setting. Investigators have not yet compared this version of the CRQ with the SGRQ.

Other outcomes have attained greater interest in clinical trials. For example, the comparison of clinical interventions across patient groups or evaluation of costs require generic HRQL instruments and preference-based measures such as the Feeling Thermometer (FT), Standard Gamble (SG), or Health Utilities Index 3 (HUI3).\(^{20}\) Preference-based instruments yield a single score between 0 and 1.0 and, in theory, allow comparison across different health programs and interventions.\(^{21-24}\) In addition, preference-based instruments are increasingly used as global measures of HRQL in trials so that responsiveness becomes important for these instruments as well. Indeed scores obtained with generic HRQL instruments have experienced transformation into utility scores because of the importance of a common metric that attempts to measure both HRQL and utilities. However, data addressing the responsiveness of preference-based instruments in relation to other HRQL tools are so far limited.\(^{25,26}\)

The aim of this study was to compare the responsiveness of the CRQ and other disease-specific, generic and preference-based instruments in patients with chronic lung disease following a respiratory rehabilitation program. There is strong evidence from meta-analyses that respiratory rehabilitation leads to patient-important improvements in HRQL among COPD patients.\(^{27}\) Therefore,
respiratory rehabilitation provides an ideal context for testing instrument responsiveness.

Methods

Patients and study design

For this analysis, we used data from a multicentre randomized trial that compared the measurement properties of the interviewer- (CRQ-IA) and self-administered CRQ (CRQ-SA) as well as those of individual and standardized dyspnea questions. In brief, we recruited patients with chronic lung diseases (mainly COPD) at four University respiratory rehabilitation programs in Toronto, Ottawa, Hamilton (all Canada) and Buffalo, NY (USA), and randomly assigned them to a group with interviewer- or self-administration of the CRQ. Patients followed an outpatient respiratory rehabilitation program of 8 weeks duration with an emphasis on physical exercise. Patients completed all outcome measures at the beginning of respiratory rehabilitation and approximately 12 weeks thereafter. We previously described that we randomized 281 patients to receive either the CRQ-IA or CRQ-SA, of whom we excluded two before the baseline assessment (inability to read) and of whom 102 did not complete the study for the following reasons: too much work or no time (n = 25), too sick to complete rehabilitation program (n = 21), SARS outbreak in Toronto (n = 17), not available for outcome assessment (n = 14), not satisfied with the rehabilitation program (n = 12), death of patients (n = 7), death of one site investigator (n = 3) or move to another place (n = 3). The administration of the HRQL instruments was identical for all instruments with the exception of the CRQ (administered as CRQ-IA and CRQ-SA) and the FT (administered with or without prior rating of 3 hypothetical health states) and is described below.

Disease-specific instruments

Chronic Respiratory Questionnaire

The original, interviewer administered CRQ has 20 items asking about the level of impairment in the domains of dyspnea, fatigue, emotional function and mastery. Patients express their degree of impairment from 1 (most severe impairment) to 7 (no impairment). In the present study, we randomized patients to complete either the original interviewer-administered or the self-administered version of the CRQ. All patients completed the original individualized and, in addition, the standardized dyspnea domain. We randomized the mode of administration and the order of individualized and standardized dyspnea questions to prevent order effects. The CRQ served as a model for many methodological studies and has shown good measurement properties including the self-administered format and the standardized dyspnea questions.

St. George’s Respiratory Questionnaire

The self-administered SGRQ consists of 76 items addressing the effect of respiratory disease on HRQL. A total score and three domain scores for symptoms, activity and impact summarize the responses and range from 0% (best) to 100% (worst). Previous studies showed good reliability, cross-sectional and longitudinal validity of the SGRQ.

Generic instruments

The Medical Outcomes Short Form 36 (SF-36): The SF-36 consists of 8 domains including physical functioning, role limitations-physical, bodily pain, general health, vitality, social functioning, role limitations-emotional, and mental health. The domains scores can be aggregated into two composite scores for physical and mental health. The SF-36 scores range from 0 to 100, with higher scores indicating better functioning and well-being. Both composite scores are standardized to have a mean score of 50 and standard deviation of 10 in the general US population. The SF-36 is valid and responsive to change in a variety of patient groups, including patients with COPD.

Preference-based instruments

Feeling Thermometer

The FT is a visual analogue scale presented in the form of a thermometer with 100 intervals, in which the best state is full health (equal to a score of 100) and the worst state is dead (a score of 0). We asked patients to reflect the health status in their score as it was during the last 7 days. We had administered the FT with or without clinical marker states. The latter occurred in random order in a 2 × 2 factorial design. The FT has demonstrated good measurement properties in patients with COPD.

Standard Gamble

The SG offers the patient two alternatives from which a choice must be made: Choice A is a hypothetical treatment with two possible outcomes: (1) returning to full health (probability p)
for $t$ years, at the end of which they die or (2) immediate death (probability $1 - P$). The alternative (choice B) is the certain outcome that he/she will stay in their own health state for $t$ years until death. The indifference probability, $P$, is the utility value for the health state in choice A in the interval from dead ($=0$) to full health ($=1$). The greater a respondent’s willingness to accept the risk of a worse outcome (e.g. dead) to avoid the health state in choice A, then the lower is the utility of the state in choice A to them.

**Health Utilities Index 3**

We used the self-administered Health Utilities Index 3 questionnaire with 15 items and asked patients to recall their health status over the week prior to the interview. The 8 attributes in the Health Utilities Index 3 multi-attribute utility measure are vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain. Each item has 5 or 6 response options. Based on the questionnaire responses we calculated a utility score from 0 to 1.0 where 0 represents dead and 1.0 full health.

**Statistical analysis**

To compare the responsiveness across instruments, we calculated standardized response means (SRMs) for each individual patient by dividing the change scores of each instrument (follow-up minus baseline score) by the standard deviations (SDs) of change scores of the whole study group. We compared SRMs among all instruments using paired $t$-tests and calculated 95% confidence intervals for SRMs using the approach proposed by Beaton et al. We performed all comparisons across instruments within one group including all patients. For the comparison of the SRMs, we combined both CRQ (CRQ-IA and CRQ-SA) and both FT groups to facilitate comparisons with the other HRQL instruments because the SRMs did not differ significantly between the CRQ-IA and CRQ-SA and the FT with or without clinical marker states. The larger sample size provided greater precision to estimate differences in relative responsiveness between the HRQL instruments.

The SRM, a widely used measure for responsiveness, differs from change scores in that it is corrected for the SD of the change score. We chose SRMs because it removes dependency on sample size as opposed to $t$-tests that depend on sample size. The SRM is superior to Cohen's effect size (change scores divided by SDs of baseline scores) because the signal (change score) relates to the noise term (SD of change score) directly. Using the SD of the change scores in the SRM rather than the SD of the baseline scores is an advantage over Cohen's effect size because baseline scores may be unrelated to change score variability. We performed the analyses using SAS 8.2 (SAS Institute Inc., Cary, NC, USA).

**Results**

Mean age of the 177 included patients was 69.0 years (SD 8.7), 59% were males, mean FEV$_1$ was 42.8 (19.2) % predicted and the proportion of patients with COPD was 93%. 7% of enrolled patients had other chronic lung diseases including idiopathic pulmonary fibrosis, chronic pulmonary aspergillosis, post-pulmonary resection and bronchiectasis.

Table 1 shows the SRMs of the disease-specific, generic and preference-based instruments. In general, SRMs were higher for the CRQ-SA compared to the CRQ-IA. The disease-specific CRQ and SGRQ showed larger SRMs compared to similar domains on the generic and preference-based instruments. However, the physical composite score of the SF-36 showed similar responsiveness (SRM = 0.37) was similar to the fatigue, emotions and mastery domains of the CRQ-IA and the activities and impact domain on the SGRQ. The CRQ dyspnea domains tended to have larger SRMs compared to the CRQ fatigue, emotions and mastery domains. The SGRQ total score showed a larger SRM than the SGRQ domains scores. While the signal of the SGRQ total score (mean change score of 5.4) was similar compared to the signal on the three domains (4.7–6.2), the standard deviation of change scores was lower (11.2 for total score versus 15.4 for the activities, 12.5 for the impacts and 17.6 for the symptoms domain).

Table 2 shows the comparisons of SRMs across instruments. The individualized dyspnea domain of the CRQ showed significantly higher responsiveness that all other instruments while the standardized dyspnea domain was significantly more responsive than the generic and preference-based instruments and the SGRQ domains. The mastery domain of the CRQ was more responsive compared to the generic and preference-based instruments, but did not differ from the SGRQ total and impacts score. The CRQ fatigue and emotions domain were significantly more responsive than the SF-36 mental composite score and the Health Utilities Index 3.

The SGRQ total score was significantly more responsive than the generic and preference-based instruments. However, only the impacts domain was significantly more responsive than the generic and preference-based instruments.
Among the generic and preference-based instruments, the SF-36 physical composite score was most responsive, but the differences with other instruments did not reach statistical significance. The FT showed a larger SRM compared to the SG and Health Utilities Index 3, but these differences were not statistically significant.

**Discussion**

This comparison of six HRQL and preference-based instruments showed that the self-administered CRQ, in particular the individualized and standardized dyspnea domains, are most responsive to changes in patients with chronic lung disease undergoing respiratory rehabilitation. The generic and preference-based instruments showed inferior relative responsiveness and did not differ significantly from each other.

Strengths of this study include the standardized administration or supervision of instruments by trained interviewers present in all centers. In addition we included a battery of the most widely used disease-specific, generic and preference-based instruments available for patients with chronic lung disease and randomized the order of administration for the individualized and standardized CRQ dyspnea domains to eliminate order effects.

### Table 1

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Standardized response mean (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease-specific instruments</strong></td>
<td></td>
</tr>
<tr>
<td>CRQ-IA ( (n = 86) )</td>
<td>Dyspnea (Individualized) 0.66 (0.45, 0.87)</td>
</tr>
<tr>
<td></td>
<td>Dyspnea (Standardized) 0.50 (0.29, 0.71)</td>
</tr>
<tr>
<td></td>
<td>Fatigue 0.25 (0.04, 0.46)</td>
</tr>
<tr>
<td></td>
<td>Emotional function 0.24 (0.03, 0.45)</td>
</tr>
<tr>
<td></td>
<td>Mastery 0.38 (0.17, 0.59)</td>
</tr>
<tr>
<td>CRQ-SA ( (n = 91) )</td>
<td>Dyspnea (Individualized) 0.84 (0.63, 1.05)</td>
</tr>
<tr>
<td></td>
<td>Dyspnea (Standardized) 0.69 (0.48, 0.90)</td>
</tr>
<tr>
<td></td>
<td>Fatigue 0.60 (0.39, 0.81)</td>
</tr>
<tr>
<td></td>
<td>Emotional function 0.56 (0.35, 0.77)</td>
</tr>
<tr>
<td></td>
<td>Mastery 0.70 (0.49, 0.91)</td>
</tr>
<tr>
<td>SGRQ ( (n = 177) )</td>
<td>Total Score 0.51 (0.36, 0.66)</td>
</tr>
<tr>
<td></td>
<td>Activities 0.33 (0.18, 0.48)</td>
</tr>
<tr>
<td></td>
<td>Impacts 0.46 (0.31, 0.61)</td>
</tr>
<tr>
<td></td>
<td>Symptoms 0.34 (0.19, 0.49)</td>
</tr>
<tr>
<td><strong>Generic instrument</strong></td>
<td></td>
</tr>
<tr>
<td>SF-36 ( (n = 177) )</td>
<td>Physical composite score 0.37 (0.22, 0.52)</td>
</tr>
<tr>
<td></td>
<td>Mental composite score 0.17 (0.02, 0.32)</td>
</tr>
<tr>
<td></td>
<td>Physical functioning domain 0.30 (0.15, 0.45)</td>
</tr>
<tr>
<td></td>
<td>Role physical domain 0.21 (0.06, 0.36)</td>
</tr>
<tr>
<td></td>
<td>Bodily pain domain 0.22 (0.07, 0.37)</td>
</tr>
<tr>
<td></td>
<td>General health domain 0.20 (0.06, 0.35)</td>
</tr>
<tr>
<td></td>
<td>Vitality domain 0.37 (0.22, 0.52)</td>
</tr>
<tr>
<td></td>
<td>Social functioning domain 0.21 (0.06, 0.36)</td>
</tr>
<tr>
<td></td>
<td>Role emotional domain 0.07 (-0.08, 0.22)</td>
</tr>
<tr>
<td></td>
<td>Mental health domain 0.15 (0.002, 0.30)</td>
</tr>
<tr>
<td><strong>Preference-based instruments</strong></td>
<td></td>
</tr>
<tr>
<td>Feeling Thermometer ( (n = 177) )</td>
<td>0.28 (0.13, 0.43)</td>
</tr>
<tr>
<td>Standard Gamble ( (n = 177) )</td>
<td>0.21 (0.06, 0.36)</td>
</tr>
<tr>
<td>Health Utilities Index 3 ( (n = 177) )</td>
<td>0.20 (0.05, 0.35)</td>
</tr>
</tbody>
</table>
A recent systematic review comparing the responsiveness of specific and generic HRQL instruments in RCTs showed superior responsiveness of specific instruments across a wide range of patients and interventions. To appraise studies that included comparisons of responsiveness between specific, generic and preference-based instruments in COPD patients (on which our study focused), we performed a PubMed search. We used the following terms “QUALITY OF LIFE [MESH]”, “instrument* OR scale* OR tool* or questionnaire*”, “COMPARATIVE STUDY [MESH]” and “Lung Diseases, Obstructive [MESH] OR COPD [MESH]”. Of 119 potentially relevant articles, 14 reported on comparisons of relative responsiveness. The majority of studies reported superior responsiveness of disease-specific instruments (CRQ, SGRQ, Respiratory Quality of Life Questionnaire and Pulmonary Functional Status Instrument) compared to generic instruments (SF-36, Sickness Impact Profile, Nottingham Health Profile, Quality of Well Being), while two studies did not find differences. One study found superior responsiveness of the Nottingham Health Profile compared to the SGRQ. Two studies comparing the CRQ and SGRQ concluded that the CRQ is more responsive, while two other studies did not favour one over the other instrument in terms of responsiveness.

However, only one study formally compared responsiveness within patients across instruments while most studies assessed responsiveness for each instrument separately. For example, the studies assessed whether the difference between baseline and follow-up scores was statistically significant for each instrument using one-sample t-tests and declared similar responsiveness when instruments detected significant differences. Other studies used Cohen’s effect sizes or SRMs and declared similar responsiveness within patients across instruments. Consensus does not exist about the most appropriate method to assess and compare responsiveness. Unlike most earlier studies, we did not rely on t-tests or Cohen’s effect sizes to assess and compare responsiveness. Significance testing using t-tests depends partly on the observed change scores and variability in change scores. However, this test also depends on sample size, which should not influence judgments about responsiveness. Cohen’s effect size is calculated by dividing the change score by the standard deviation of the baseline score. In contrast to variability in change, there is no...
necessary or logical relationship between between-person variability at baseline and instrument responsiveness. We used the SRM to compare responsiveness within patients across several instruments, because the SRM deals with the aforementioned limitations in that the denominator captures variability in change (the noise component of the signal-to-noise ratio) and removes the dependency on sample size. However, the multiple comparison we used in this analysis do raise the possibility of a type I error. One solution to this problem is reliance on lower P-values and readers may focus on the P-values that are less than 0.01 in our analysis. Other methods of correction for multiple testing have limitations because the comparison we conducted were not independent.

The superior responsiveness of the SGRQ total score compared to its domain scores deserves attention. The increased number of items in the total score resulted in decreased variability in change, and thus an increased signal (change score) to noise (SD of change scores) ratio. Thus by calculating total scores, the SRM becomes larger compared to the domain scores. There are, however, limitations to using a total score. If effects of an intervention vary across domains of HRQL, the total score obscures potentially important information from HRQL domains and may result in misleading interpretations. In the present study, the effect of respiratory rehabilitation was similar in all domains and the SGRQ total score may adequately represent the overall treatment effect. In the present article, we did not calculate CRQ total scores because use of CRQ total scores requires additional research.

Investigators should not base the selection of the instrument on responsiveness alone, but also on longitudinal validity. In order to understand if an instrument measures changes in HRQL that one intends to measure, one needs to consider correlations of change scores with those of other instruments. We have previously described the longitudinal construct validity for the CRQ-IA and CRQ-SA and SGRQ. The results of the latter and other studies suggest that the CRQ and SGRQ measure similar constructs.

In recent years, investigators suggested that HQRL instruments can be used in clinical practice for evaluation purposes but also to facilitate the discussion of HQRL issues between patients and clinicians. Clinicians should also choose instruments based on evidence about the measurement properties. But there are practical aspects that are important for a successful introduction of HQRL instruments into clinical practice. First, the duration and mode of administration (interviewer versus self-administration) should require as few resources as possible. Second, calculation of domain or summary scores should be straightforward and not involve complicated algorithms. Third, there should be evidence about the minimal important difference to guide interpretation of domain or summary scores. Considering these practical aspects, the CRQ-SA is likely to be the most prudent choice as a HQRL instrument for clinical practice.

This study confirms that the disease-specific HRQL instruments (CRQ and SGRQ), in particular the CRQ dyspnea domain, are substantially more responsive than generic measures when used for the evaluation of respiratory rehabilitation. It is possible that the relative responsiveness of HRQL instruments may depend on the population studied and contexts other than respiratory rehabilitation. Studies using fully validated instruments evaluating the extent to which responsiveness depends on patients or interventions would enhance our understanding of the measurement of change in HRQL.

Acknowledgments

The work was in part supported by an unrestricted educational grant from AstraZeneca, Lund, Sweden. The sponsor was not involved in the study design, conduction of the trial and analysis of data. One of the authors is an employee of AstraZeneca and participated in the study through discussion and she reviewed drafts of the original manuscript making wording and presentation suggestions. Pfizer Limited, UK, provided additional funding. The CRQ-IA and CRQ-SA are copyright by McMaster University, Hamilton, Canada; Principal Authors Dr. Gordon Guyatt and Dr. Holger Schünemann. Use of the CRQ requires permission by McMaster University and the authors.

References


