Quality of Life Research (2006) 15: 1–14
DOI 10.1007/s11136-005-0126-x

Do clinical marker states improve responsiveness and construct validity of the standard gamble and feeling thermometer: A randomized multi-center trial in patients with chronic respiratory disease

Holger J. Schünemann[1,2,3], Roger Goldstein[4], M. Jeffery Mador[1], Douglas McKim[5], Elisabeth Stahl[6], Lauren E. Griffith[3], Ahmed M. Bayoumi[4], Peggy Austin[3] & Gordon H. Guyatt[3,7]

[1]*Division of Clinical Research Development and INFORMAtion Translation/INFROMA, Italian National Cancer Institute Regina Elena, Rome, Italy (E-mail: schuneh@mcmaster.ca);* [2]*Department of Medicine, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, New York, USA;* [3]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada;* [4]*Department of Medicine, University of Toronto, Toronto, Ontario, Canada;* [5]*University of Ottawa, Ontario, Canada;* [6]*AstraZeneca R & D, Lund, Sweden;* [7]*Department of Medicine, McMaster University, Hamilton, Ontario, Canada*

Accepted in revised form 22 June 2005

Abstract

Background: Optimizing the validity and responsiveness of utility measures will enhance their usefulness in randomized trials. We evaluated the impact of clinical marker state (CMS) rating prior to patients' rating their own health on two utility instruments (feeling thermometer (FT) and standard gamble (SG)) in patients with chronic respiratory disease (CRD). *Methods*: We randomized 182 patients with CRD to complete the FT (self-administered) and SG with CMS (FT+/SG+, n=91) or without marker states (FT−/SG−, n=91) before and after undergoing respiratory rehabilitation in a multi-center trial. *Results*: Use of CMS did not influence baseline utility scores. Improvement after therapy on the scale from 0 (dead) to 1.0 (full health) was 0.04 both in FT+ ($p=0.03$) and FT− ($p=0.02$; the difference between FT+ and FT− was 0.00, $p=0.83$). Improvement on the SG was 0.05 in both SG+ ($p=0.08$) and SG− ($p=0.04$; difference between SG+ and SG− 0.00, $p=0.95$). Correlations with other health related quality of life scores were highest for FT+. *Conclusion*: Administration of CMS did not improve responsiveness of the FT but may have improved construct validity. The SG showed limited construct validity and responsiveness that was not influenced by CMS use.

Key words: Clinical trial, Feeling thermometer, Preference measures, Quality of Life, Standard gamble, Utilities, Visual analogue scale

Introduction

Assessment of patients' utilities and health state preferences is receiving greater attention [1–5]. Utility and health state preference instruments yield data that are useful for both health related quality of life (HRQL) measurement and economic analysis [6–10]. Preference instruments are also of increasing interest for use in clinical trials.

Several different methods for measuring preferences are available [11–15]. Direct preference-based instruments generate a preference score for respondents' current health state, typically on a 0.0–1.0 scale where 0.0 indicates dead and 1.0

indicates full health. Many investigators regard the standard gamble (SG), a direct preference instrument, as the reference standard for preference measurement [13, 16, 17]. The SG, however, requires interviewer administration, may be conceptually challenging for patients, and, in comparison to other preference instruments, may be unresponsive to small but important changes in HRQL [18–20].

As a result of the limitations of the SG, investigators sometimes use other instruments to measure HRQL and obtain preference estimates. Alternatives include the feeling thermometer (FT), a visual analogue scale presented in the form of a thermometer [13]. When completing this instrument, patients choose the score on the thermometer that represents the value or preference they place on their health state. The FT is simpler than the SG and has shown good responsiveness and validity in several studies [21–26]. Although experience with the FT is predominantly based on interviewer administration, self-administration is possible and enhances the feasibility of the FT in the context of large clinical trials [25].

The use of direct preference instruments like the SG and FT requires that patients comprehend the task involved. Patients must understand how their own health state compares to other health states, including the extreme health states of ''full health'' and ''dead''. To optimize task understanding and to increase patients' thoughtfulness about their ratings, authors recommend the rating of clinical marker states (CMS) – patient scenarios or hypothetical health states – prior to rating their own health [27–30]. CMS may improve task understanding, because they remind patients to think about specific domains that affect their HRQL [30]. CMS may also stimulate patients to consider how their disease impacts on these HRQL domains and help clinicians interpret results of studies using preference instruments [31, 32]. Because chronic obstructive pulmonary disease (COPD) is a disease affecting primarily older patients, optimal administration of preference instruments in clinical studies may be particularly challenging in the context of COPD. CMS could, therefore, facilitate the completion of the FT and SG.

Patients typically rate three CMS representing mild, moderate and severe impairment of HRQL before rating their own health. Although the use of CMS is widespread, few studies have addressed their impact on responsiveness and construct validity of direct preference rating instruments. Indeed, we are aware of only two direct comparisons, our own prior studies [25, 26]. One study in patients with chronic respiratory disease (CRD) failed to demonstrate convincing improvements in measurement properties for either the SG or FT with CMS, but the study was small and possibly underpowered [26]. In another study in which we enrolled younger patients with gastroesophageal reflux disease, use of CMS improved the responsiveness and construct validity properties of the FT but not the SG [25]. However, patients in the latter study were younger than average patients with CRD and therapy with a proton pump inhibitor led to large improvements in HRQL, leaving uncertainty whether results are generalizable across patient populations and interventions.

Administration of CMS increases the amount of time required to complete a preference instrument and the complexity of the process, and thus discourages the use of the preference instruments. Efficiency of preference elicitation is a particularly important concern in clinical trials. Increased time and burden of administration in clinical trials would be desirable if response to an intervention, the primary endpoint in most clinical trials, could be detected more easily or with greater validity with the use of CMS, but not otherwise. Because we believe that preference instruments provide important information not otherwise available, we would like to encourage their use, and maximize their validity, responsiveness, and efficiency. Therefore, we conducted a randomized multicenter trial to compare the impact of CMS on the responsiveness (primary endpoint), and the validity, and time of administration (secondary endpoints) of the SG and FT in patients with CRD undergoing respiratory rehabilitation, an intervention of known effectiveness in improving HRQL [33].

Methods

Patients and therapeutic intervention

We have followed Stalmeier and colleagues' suggestion regarding the presentation of information

Table 1. Marker states for mild, moderate and severe chronic airflow limitation

Mild	Moderate	Severe
– chronic lung problems mildly limit activities	– limited by chronic lung problems	– severely limited by chronic lung problems
– sometimes becomes more short of breath than used to when jogging, playing sports, or hurrying up a couple flights of stairs	– often becomes short of breath doing activities such as climbing a flight of stairs, or climbing a hill	– when speaking fast has to slow down because of shortness of breath
	– has to try to remember pacing during these activities and to consciously slow down, otherwise becomes very short of breath	– often becomes severely short of breath doing even ordinary activities such as bending, dressing or talking
		– gets very short of breath when angry or upset
		– must pace and slow down with every activity
– rarely feels more tired than used to be the case	– more tired than used to be, and feels low in energy some of the time	– feels very tired, and low in energy all of the time
	– in general, feels in control of breathing problems, but not always	– often feels out of control of breathing problems
– rarely feels frustrated and impatient about exercise limitation		– feels frustrated and impatient about breathing problems all of the time
– overall, generally happy and free from worry	– sometimes feels frustrated and impatient about breathing problems	– lung problems and the way they affect life have a large influence on emotions, and is almost always fretful, angry and irritable

in the methods section of studies dealing with preference-based instruments [34]. From September 2001 until June 2003, we recruited patients from four respiratory rehabilitation centers in Canada and the United States. Eligible patients included all inpatients and outpatients with chronic respiratory disease (CRD) enrolling in respiratory rehabilitation except for patients who were unable to complete the questionnaires due to language limitations. The institutions' ethic review boards approved the study and all patients provided written informed consent.

The rehabilitation programs were 8 weeks in duration, and included an exercise component. Recruitment terminated in Hamilton, Canada, in June of 2002 due to the death of the site investigator (Dr David Stubbing). The initial interviews took place during the patient's admission to the program and the follow-up interview took place at a clinic visit approximately 12 weeks thereafter. At each site, a single research assistant conducted or supervised interviews at baseline and follow-up according to an interviewer guide. Patients rated their health as it had been in the 2 weeks prior to the interviews. Follow-up appointments in Toronto proved unfeasible during the severe acute respiratory syndrome (SARS) outbreak because of temporary closure of the rehabilitation hospital. The hospital housing the Toronto rehabilitation facility served as one of the main care facilities during the SARS outbreak.

We randomized 280 patients with CRD to two different modes of administering the feeling thermometer and standard gamble. In a factorial design, we also randomized patients to the self-administered Chronic Respiratory Questionnaire (CRQ-SA) or the interviewer-administered CRQ (CRQ-IA). We will report the data of the latter, independent analysis in a separate manuscript [35].

Direct preference-based measures

Feeling thermometer with CMS (FT+)

The FT is a visual analogue scale depicted as a vertical thermometer in which the worst state is dead (a score of 0) and the best state is full health (equal to a score of 100) [13]. Patients completed a self-administered version with interval markings of the FT [25, 26]. Table 1 shows the three CMS representing mild, moderate and severe symptoms of COPD that we tested extensively in a pilot study [36].

Feeling thermometer without CMS (FT−)
In the alternative mode of administration, patients rated their own health on the self-administered FT without prior exposure to, or rating of, the three CMS.

Standard gamble
This instrument offers two options from which patients must make a choice: 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 the patient dies or (2) immediate death (probability $1 - p$). We varied t depending on the patient's age as follows: patients aged more than 80 years, $t =$ the rest of the patient's lifetime; age 76–80 years, $t = 10$ years; age 66–75 years, $t = 15$ years; age 56–65, $t = 25$ years; age 46–55 years, $t = 30$ years; and age 36–45 years, $t = 35$ years. Specifying duration of remaining life ensures that patients use the same time frame as others of their age, and reduces the random error that might result from patients inferring different time frames. Varying the time frame by age avoids an additional lack of realism if one chose a single time frame and either young patients have an unrealistically short duration of remaining life, or old patients have an unrealistically long duration. The alternative (choice B) is the sure outcome that the patient will stay in a health state (their own health state, or a marker state) for t years until death. We did not use the t years approach with the FT.

Interviewers used a chance board with the ping-pong approach varying the probability p in steps of 0.05 to obtain the value, p^*, where the patient considered choice A equal to choice B [13, 37]. This indifference probability, p^*, is the utility value for the health state or the patient's own health in the interval from dead ($=0$) to full health ($=1$). The greater a patient'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. We administered the SG in two formats. In one (SG+), patients rated the three CMS (mild, moderate and severe COPD) prior to rating their own health state. The CMS were identical to those of the FT. In the other approach (SG−), patients rated their own health state without prior exposure to the CMS.

Definition of full health and death
We defined full health and provided a definition similar to that of the Health Utilities Index 3 (HUI3) for both the FT and the SG [26, 38]. The descriptions included phrases such as "Able to walk around the neighborhood without difficulty, and without walking equipment", "Happy and interested in life", "Able to remember most things, think clearly and solve day to day problems" and "Free of pain and discomfort". We defined the worst health state as dead (equivalent to a score of 0) and we did not ask patients to rate the states full health or dead.

Validation instruments

Validation instruments included the CRQ. The CRQ is a disease specific instrument for patients with chronic airflow limitation measuring HRQL in the domains dyspnea, fatigue, emotional function and mastery on 7 point scales [39, 40]. We calculate domain scores as means scores on each domain. The St. George's Respiratory Questionnaire (SGRQ) is a widely used disease specific instrument for patients with respiratory disease [41]. The Short Form 36 version 2 (SF-36.v2) is a generic instrument including eight domains and it is one of the most widely used HRQL instruments [42]. The scoring for the original version of the SF-36 uses positive weights for scores from physical domains (for PCS) and negative weights for the mental health domains (for PCS) to compute component scores. In contrast the RAND-36 approach uses only positive weights. Investigators have noted that the negative weights can distort the results of the component scores [43, 44]. This distortion could be more evident in situations in which the health condition involves both substantial physical and mental health burdens. Because mental health may be impaired in patients with severe respiratory disease we computed summary scores using the RAND algorithm. RAND-36 weights exist for the original version of the SF-36 but not SF-36.v2, we developed our own weighting approach for the mental component score using the RAND36 strategy based on the program by Hays et al. [45]. The Health Utilities Index 3 (HUI3) is a multi-attribute utility instrument widely used as a generic HRQL instrument

[14]. All of these instruments are valid and responsive to change in HRQL.

At the follow-up visit, the same interviewer administered or supervised the administration of the instruments to each patient in the same order as at baseline. The CRQ, SGRQ, HUI3, FT and SF-36.v2 were self-administered under supervision; the SG was interviewer administered.

The FT and SG represent patients' preferences for the HRQL of their own, individually and subjectively defined health states while HUI3 overall utility scores represent community preferences for the patients' health states defined by the HUI3 health status classification system. The FT and SG differ in various ways, including the SG's presentation as a choice under uncertainty. All three, however, provide a global measure of patients' HRQL, and are therefore conceptually similar enough that, if all were valid, they would show substantial correlations with one another. The SF-36 is a generic instrument that measures all important domains of HRQL. Given that the extensive process of refinement of the SF-36 focused on retaining the most important items, we would anticipate moderate correlations between the preference measures and the physical and mental function scores of the SF-36.

The CRQ and the SGRQ are disease-specific measures that focus on the most important problems in physical and emotional function of patients with COPD. A large body of literature suggests that patients with severe COPD experience major physical and emotional function problems, and that they find such problems extremely distressing [46]. The more severe the COPD, the greater the impact of the specific problems associated with COPD on patients' HRQL. Our patients' decision to invest large amounts of time and energy in respiratory rehabilitation provides testimony to the impact of their COPD-associated impairments on their HRQL. These considerations led us to anticipate that the construct validity of the preference measures would be manifest in moderate correlations with the domains of the CRQ and SGRQ that reflect important physical and emotional dysfunction, and change in that physical and emotional dysfunction.

Study design

We randomized patients to either the FT+ and SG+, or the FT− and SG− format of direct preference measurement. Patients rated their own health as it had been in the 2 weeks prior to the interview both at baseline and follow-up. In the FT+/SG+ group, patients first rated the three CMS on the FT before rating their own health on the FT, followed by the three CMS and their own health on the SG. In the FT−/SG− group patients rated their own health on the FT followed by the SG. This order avoided exposure to any other HRQL instruments before patients rated the FT and therefore allowed us to evaluate the FT as a freestanding self-administered instrument. In addition, we randomized the order of the CMS in two possible sequences (either moderate, severe and mild or moderate, mild and severe, respectively) to account for possible effects of administration order. We administered all instruments in the same order both at the baseline and the follow-up visit.

An experienced research coordinator from the methods center trained all site interviewers in a day-long session. We performed randomization at the methods center in blocks of eight stratified by center. Ethic review boards at all study sites approved the study protocol and all patients signed an informed consent form prior to enrollment in the study.

To avoid bias in the interpretation of the results, the authors (with the exception of the statistician, LG) remained blinded to patients' group assignment (CMS vs. no CMS) during the period in which we formulated our interpretation of the results. We broke the code only after agreeing on the main conclusions for this manuscript.

Statistical analysis

For the comparisons between patients included in this analysis and patients who did not complete the study we used unpaired *t*-tests for continuous variables and chi-square test for categorical outcome variables. For better comparison with the SG we performed a linear transformation of the FT scores by dividing the scores by 100. Thus, in

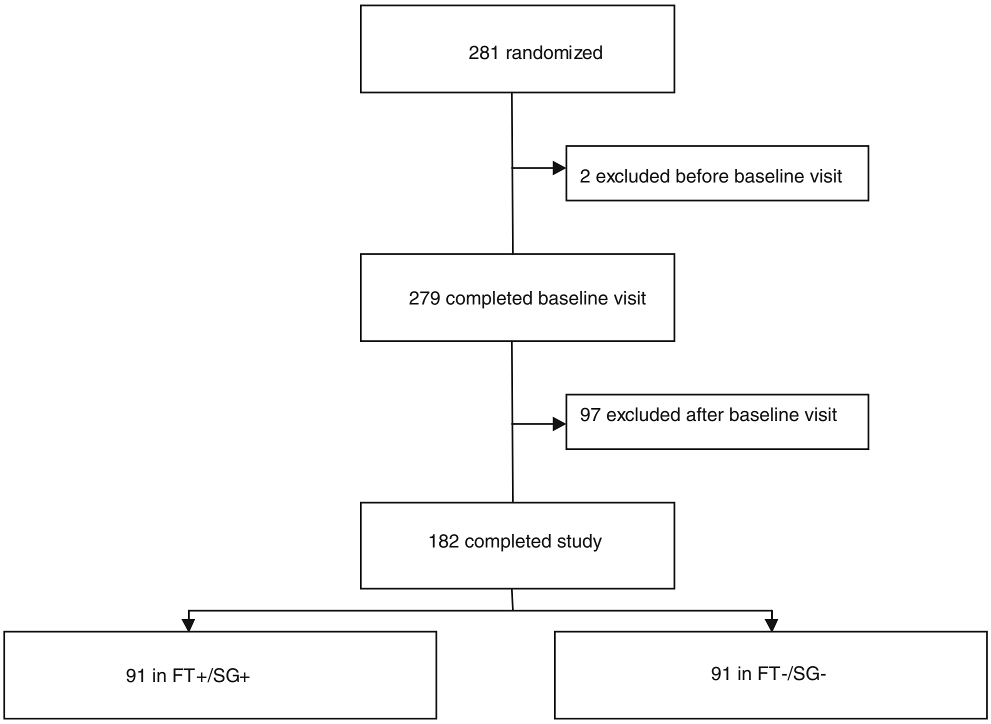
281 randomized
2 excluded before baseline visit
279 completed baseline visit
97 excluded after baseline visit
182 completed study
91 in FT+/SG+
91 in FT-/SG-

Figure 1. Flow diagram describing patient enrollment and completion of the study. This flow diagram describes patient enrollment and flow through the study. The reasons for exclusion were: 2 patients were withdrawn because of inability to read (prior to the baseline visit); 25 patients reported that the completion of the instruments was too much work, no time or uncomfortable with the questions, 21 patients did not complete the rehabilitation program because they were too sick, we could not follow-up 17 patients because of severe acute respiratory syndrome (SARS), 12 were not satisfied with the rehabilitation program (possibly related to SARS), 6 could not be reached for follow-up (possibly related to SARS), 7 patients died, 3 patients stopped because of the death of one of the investigators, 3 patients moved, and 3 did not complete the feeling thermometer or standard gamble at both visits.

this analysis the reported scores on the FT ranged from 0 to 1. We calculated mean scores and standard deviation for the scores of the FT and SG CMS for the group of patients randomized to the administration of the FT and SG prior to rating their own health. Except for the SGRQ, higher scores on any of the HRQL instruments represent better HRQL. For the calculation of the correlation coefficients we multiplied the SGRQ scores with 1 to facilitate comparison.

Responsiveness

We focused on the ability of the FT+, FT−, SG+, and SG− to detect improvement with respiratory rehabilitation therapy by performing paired *t*-tests comparing baseline and follow-up values on the utility measures with and without the randomly allocated CMS. To evaluate relative responsiveness of the instruments with and without CMS, we compared the differences between scores from the baseline and follow-up visit for the utility measures with and without use of the CMS using an unpaired *t*-test.

Construct validity

Using data from the baseline visit we evaluated the cross-sectional construct validity of the instruments. We calculated Pearson's correlation coefficients for the pairs of scores for all pair-wise combinations on the FT, SG, HUI3, CRQ, SGRQ and the SF-36.v2. We assumed that higher correlations of the FT and SG with the validation instruments would indicate greater construct validity. To assess longitudinal construct validity of the instruments we calculated Pearson's correlation coefficients for the differences between scores from the baseline and follow-up visit for FT+, FT−, SG+ and SG− with the change on the

Table 2. Demographic information for patients randomized to feeling thermometer and standard gamble with and without marker states

Patient characteristic	*With* marker states (n = 91)	*Without* marker states (n = 91)
Gender (% female)	37 (40.7)	36 (39.6)
Age[#] (SD)	68.2 (8.1)	67.5 (8.0)
Diagnosis		
COPD[§]	83 (92.2)	85 (93.4)
Other[*]	7 (7.8)	6 (6.6)
FEV_1 % predicted (n with data)	44.2 (20.8) (n = 75)	42.8 (16.0) (n = 76)
Time since diagnosis in years[#] (SD); median (IQR)	11.6 (13.8) 6.5 (3, 15)	10.0 (10.5) 7 (3, 13)
Smoking history (%)		
Never	7 (7.7)	2 (2.2)
Current smoker	2 (2.2)	1 (1.1)
Quit	82 (90.1)	88 (96.7)
Living alone (%)	27 (29.7)	28 (30.8)
Employed (%)	9 (10.0)	9 (10.0)
HUI-3	0.62 (0.27)	0.56 (0.27)
CRQ - Dyspnea	4.29 (1.34)	4.13 (1.16)
CRQ - Fatigue	4.06 (1.27)	3.86 (1.29)
CRQ - Emotional function	4.89 (1.18)	4.69 (1.21)
CRQ - Mastery	4.91 (1.30)	4.74 (1.26)
St George's - Activities	71.20 (21.46)	75.91 (13.61)
St. George's - Impacts	36.81 (18.25)	36.67 (15.41)
St. George's - Symptoms	57.26 (22.85)	63.46 (17.79)
SF36-Physical Component Score	34.32 (8.34)	32.33 (7.56)
SF36- Mental Component Score	50.06 (10.65)	49.94 (12.66)

[#]Mean (standard deviation).

[§]Chronic Obstructive Pulmonary Disease.

*Other diagnoses include: idiopathic pulmonary fibrosis, chronic pulmonary aspergillosis, post-pulmonary resection and bronchiectasis.

HRQL instruments. We considered correlations of less than 0.2 as very weak, from 0.2 to 0.35 as weak, from greater than 0.35 to 0.5 as moderate and of more than 0.5 as strong. We compared Pearson's correlation coefficients using the *z*-test.

Sample size

The primary outcome of interest in this analysis was the relative responsiveness of the FT + /SG + compared to FT−/SG−. In our previous study, the mean scores in the FT− group (on the scale from 0 to 1) increased by 0.07 (SD 0.20) after respiratory rehabilitation [47]. The corresponding change for the administration of the FT in the FT + group was 0.14 (SD 0.21). As a result, the difference in the magnitude of change between the two groups was 0.07 (SD 0.20). We previously also found that the minimal important difference (MID) on the FT is approximately 0.05–0.08 [47]. We determined that by enrolling 200 patients (100 in each group) we would be able to detect a difference of 0.08 (equal to approximately 0.4 SD of the change score) in the change between the two groups with a 2-tailed test, an alpha of 0.05 and a beta of 0.2. Due to the 12-week follow-up we ended recruitment when the projected numbers indicated a sample size of 100 per group. However, shortly after we concluded recruitment of new patients the primary recruitment center lost a large number of patients due to the SARS outbreak. Therefore, we did not reach the projected sample size of 100 per group. The actual power of the test to detect a difference of 0.08 in the mean preference scores between the FT + and FT− groups with the reduced number of participants (91 per group) was in fact 85% due to the smaller than expected variance.

Multiple testing

The primary objective was the comparison of the FT + and FT− as well as the SG + and SG− groups, respectively. Because of the intended use in clinical trials for which responsiveness is of primary concern, we used the standard α of 0.05 as

Table 3. Responsiveness of the feeling thermometer (FT) and the standard gamble (SG) with and without marker states

Group	N	Baseline Mean ± standard deviation	Follow-up Mean ± standard deviation	Differences: follow-up - baseline (95% confidence interval)	*p* value for difference between the baseline and follow-up
FT *with* marker states	91	0.61 ± 0.18	0.65 ± 0.14	0.04 (0.004, 0.071)	0.03
FT *without* marker states	91	0.60 ± 0.17	0.64 ±0.12	0.04 (0.007, 0.079)	0.02
p-value for difference between groups	0.74	0.90	0.83		
SG *with* marker states	91	0.68 ± 0.28	0.73 ± 0.22	0.05 (−0.007, 0.010)	0.08
SG without marker states	91	0.65 ± 0.21	0.70 ± 0.18	0.05 (0.002, 0.010)	0.04
p-value for difference between groups	0.46	0.43	0.95		

cut-off for statistical significance for this objective. For the evaluation of construct validity we used several HRQL instruments, some with several domains. As a result we performed many statistical tests to compare correlation coefficients, thereby increasing the probability of statistically significant findings by chance alone in the absence of adjustment of the *p*-values. However, in the present report we made no adjustment for multiple comparisons because we used validation instruments with correlated scores and because of the lack of consensus on how one should adjust for such analysis. We acknowledge that the multiple testing weakens the inferences with respect to differences in correlation coefficients between the two study arms of the study.

Results

We initially enrolled 281 patients in this study. Figure 1 describes the flow of participants through each stage of the randomized trial and the reasons for exclusion from the study.

A total of 182 patients (n = 91 in the FT + /SG + group and n = 91 in the FT−/SG− group) completed the study. Important reasons for exclusion were the death of the investigator responsible for the Hamilton rehabilitation program and the outbreak of SARS precluding follow-up of many patients in the Toronto center. However, there was no difference in gender, age, length and type of respiratory diagnosis, smoking history and employment status in the patients who are

Table 4. Cross-sectional construct validity of the FT and SG groups at baseline (Pearson correlation coefficient)

Type of instrument	Instrument	FT *with* marker states (n = 91)	FT *without* marker states (n = 91)	SG *with* marker states (n = 91)	SG *without* marker states (n = 91)
Utility measures	SG with marker states	0.42	X	X	X
	SG without marker states	X	0.36	X	X
	HUI-3	0.49	0.46	0.28	0.29
Disease-specific questionnaires	CRQ - Dyspnea	0.58	0.46	0.22	0.27
	CRQ - Fatigue	0.62	0.47	0.23	0.30
	CRQ - Emotional function	0.56	0.46	0.29	0.42
	CRQ - Mastery	0.54	0.47	0.16	0.35
	St.Georges's - Symptoms	0.39	0.21	0.18	0.22
	St.George's - Activities	0.55	0.32	0.29	0.12
	St. George's - Impacts	0.61	0.46	0.33	0.21
	SF-36 Physical Component Score	0.59	0.44	0.28	0.22
	SF-36 Mental Component Score	0.54	0.39	0.24	0.39

$p < 0.05$ for $r > 0.21$, there was no statistically significant difference in *r* between the FT+ and FT− or SG+ and SG− groups, respectively.

Table 5. Longitudinal construct validity of the FT and SG. Correlations between changes in the FT, SG and other instruments (Pearson correlation coefficient)

Type of instrument	Instrument	FT *with* marker states (n = 91)	FT *without* marker states (n = 91)	SG *with* marker states (n = 91)	SG *without* marker states (n = 91)
Utility measures	SG with marker states	0.31	X	X	X
	SG without marker states	X	0.28	X	X
	HUI-3	0.44	0.32	0.05	0.21
Disease-specific questionnaires	CRQ - Dyspnea	0.44	0.28	0.18	0.07
	CRQ - Fatigue	0.48	0.35	0.18	0.08
	CRQ - Emotional function	0.33	0.26	0.11	0.03
	CRQ - Mastery	0.30	0.21	0.02	−0.02
	St.Georges's - Symptoms	0.11	0.07	−0.01	0.06
	St.George's - Activities	0.15	0.04	0.22*	−0.15*
	St. George's - Impacts	0.37	0.27	0.12	−0.04
	SF-36 Physical Component Score	0.23	0.19	0.08	0.04
	SF-36 Mental Component Score	0.43*	0.14*	0.13	−0.06

$p < 0.05$ for $r > 0.21$.
*$p <$ 0.05 for differences between correlation coefficients of FT+ and FT− or SG+ and SG–, respectively.

included in this analysis compared with those not included in this analysis (the lowest *p*-value for differences between these groups was 0.16 for gender). Table 2 shows the characteristics of included patients. The two groups were similar in baseline characteristics and baseline HRQL scores. The scores on the HRQL questionnaires indicated significant physical impairment. The mental component scores on the SF-36 indicated no impairment with the original scoring. However, when we used the modified RAND scoring approach, the results indicated some impairment on the mental component score (43.3, SD 11.1, in the FT+ vs. 42.9, SD 11.4, in the FT− groups). All preference based instruments detected important impairment of HRQL indicated by the low scores the participants exhibited. The mean scores and SD for the CMS at follow-up in the FT+/SG+ were 0.81 (0.12) for the mild, 0.60 (0.13) for the moderate and 0.36 (0.16) for the severe CMS on the FT. The corresponding ratings of the CMS on the SG were 0.84 (0.12) for the mild, 0.70 (0.20) for the moderate and 0.52 (0.25) for the severe CMS.

Responsiveness

Table 3 presents the responsiveness of the FT and SG with and without CMS. There were no significant differences in baseline scores on the FT and SG. The improvements in scores after therapy were small in that the mean change approached the MID on the FT. The mean change was 0.04 ($p = 0.03$ vs. baseline) in the FT+ group and 0.04 ($p = 0.02$ vs. baseline) in the FT− group (difference between FT+ and FT− 0.00, $p = 0.83$). The corresponding improvements were 0.05 in the SG+ group ($p = 0.08$ vs. baseline) and 0.05 in the SG− group ($p = 0.04$ vs. baseline) (difference between SG+ and SG− 0.00; $p = 0.95$). Although the mean change scores were slightly greater for the SG than for the FT, only the FT groups and SG−, but not SG+ showed statistically significant change scores. The latter is a result of the greater variance around the change scores of the SG.

Construct validity

Table 4 shows the cross-sectional construct validity of the FT and SG at baseline. The correlations were moderate to strong for the FT regardless of whether or not the patients completed the CMS. Individual correlations were not statistically significantly different between the FT+ and FT− groups, but all correlation coefficients were higher in FT+. Eight out of 11 correlations were strong and 3 correlations were moderate in the FT+ group, while there were no strong correlations, 9 moderate and 2 weak correlations in the FT− group. Because we observed similar correlations between the FT and the other instruments at follow-up, we present only the results for the baseline visit.

The two right-most columns of Table 4 examine the cross-sectional construct validity of the SG at baseline. Overall the correlations were substantially weaker compared to the FT, and there was no clear difference between the SG+ and SG− group (3 out of 10 correlations were higher for the SG+ group). As with the FT, the correlations between the SG and other instruments at follow-up were similar to those at baseline.

Table 5 shows the correlations of the change scores from baseline to follow-up for the FT and SG with the other measures. Correlations between the FT and other instruments were generally moderately strong and were stronger in the FT+ group compared to the FT− group. Compared to the FT− group all correlations were higher in the FT+ group. Differences between the FT+ and FT− groups in the strength of correlations with the validation instruments reached statistical significance for the SF-36.v2 mental component score.

The correlations of change in the SG with change in other instruments were weak. Compared to omission of CMS, the administration of CMS with the SG resulted in similar correlation coefficients for most questionnaires although the difference was statistically significant for the SGRQ-Activities domain. Out of 10 correlations 8 were stronger in the SG+ group.

Time of administration

The time for completion was 9.2 min (95% CI: 8.5–9.9) in the FT+, 4.4 min (95% CI: 4.1–4.7) in the FT−, 13.1 min (95% CI: 12.0–14.2) in the SG+ and 6.0 min in the SG− (95% CI: 5.3–6.6) groups at baseline. The differences between groups were all statistically significant. On average the time for completion was 1.0 min (FT−) to 1.8 min (SG+) shorter at follow-up ($p < 0.05$ for all groups compared with baseline).

Discussion

This RCT evaluated the influence of administering hypothetical health or clinical marker states on responsiveness and construct validity of two instruments that investigators use to obtain preference scores, the FT and SG. The FT detected small improvements in HRQL, but there was no increase in responsiveness with administration of CMS. Results suggested higher construct validity in the FT+ group compared to the FT− group (Tables 4 and 5). The results also suggested that the SG detected small improvements in HRQL, but these results were not statistically significant for the SG+. Overall, the differences in responsiveness and construct validity between the SG+ and SG− were small and in general not statistically significant.

The strength of this study includes its randomized multi-center design. The study's parallel group design eliminated the possibility of one format influencing response to the other. We used an intervention of known effectiveness in regard to improving health-related quality of life in patients with CRD [33]. Another important strength of our study was the use of multiple validation instruments and choice of a self-administered version of the FT.

The study has some weaknesses. We limited ourselves to addressing the impact of CMS on the responsiveness and construct validity of the FT and SG. Even if CMS did not improve the construct validity or responsiveness of the FT and SG, they may still be useful in enhancing their score interpretability for those evaluating effectiveness of interventions [31, 32], an issue that we did not address in this study. Another way in which investigators may use marker states is, when HRQL impairment is minimal, to 'chain' SG responses. Here, patients first rate their health state in relation to the mild impairment CMS state. Subsequently, they rate the mild CMS against death and full health.

While we developed the CMS carefully, it is possible that different, equally thorough ways of developing and defining CMS might lead to different results. In addition, due to the outbreak of SARS we did not reach the projected sample size (91 instead of 100 per group). However, the power of the tests we employed to evaluate responsiveness did not suffer because the variance around the mean was lower than we anticipated. Our power to detect small but statistically significant differences in correlation coefficients was limited. The possibility exists that we observed statistically significant results in the secondary endpoints that are due to multiple testing. Therefore, readers should interpret the statistically significant differences in the correlation coefficients with caution.

We confirmed previous studies showing that utilities obtained with visual analogue scales, such as the FT, are lower than utilities obtained with the SG [13]. Patients place a lower value on impaired health states when using the FT than when using the SG. However, all three preference based instruments detected important impairment of HRQL in these patients with CRD at baseline.

The FT showed moderate to strong correlations with change scores on the other questionnaires (Table 5). The correlations were all stronger with use of the CMS, but the differences reached statistical significance only for the SF-36.v2 mental component score (Table 5), and interpretation of significant correlations is difficult in the presence of multiple comparisons. On the other hand, failure to show more and stronger statistically significant differences between FT+ and FT− may have been a function of limited sample size and consequently limited power. Moreover, chance is an unlikely explanation for the finding that all of the correlation coefficients were higher in the FT+ group than in the FT− group (Tables 4 and 5).

The SG showed weaker cross-sectional correlations than the FT (Table 4) and very poor longitudinal construct validity in terms of the constructs measured with the HRQL validation instruments; correlations were near 0 (Table 5). Correlations appeared uninfluenced by use of CMS. Overall, 11 out of 20 correlations were higher in the SG+ group (3 out of 10 correlations in Table 4 and 8 out of 10 in Table 5).

One could argue that correlations of the SG with results of disease-specific instruments are not relevant to the construct validity of the SG. This would be the case if one did not anticipate that CRD-related impairments would have a strong influence on patient preferences. There is strong evidence from other studies that severe CRD has a profound effect on patients' HRQL. Moreover, the patients in our study were willing to spend considerable time and energy on respiratory rehabilitation, another testimony to the impact of the condition on their lives. These considerations suggest that very low correlations between SG and the validation tools in this study reflect limitations with SG validity.

Our previous smaller trial in which we randomized 84 older patients with COPD to an interviewer administered FT and SG with or without administration of CMS are generally consistent with the results of this study [26]. As in the current study, in our previous analysis of patients undergoing respiratory rehabilitation, results suggested superior construct validity of the FT was superior when patients rated CMS prior to rating their own health. Although we observed a trend toward superior responsiveness in the FT+ group in that study, the difference compared with the FT− group was not statistically significant. The prior study also failed to show clear differences in construct validity between the SG+ and SG−. The only trend that did exist was toward inferior cross-sectional and longitudinal construct validity of the SG+ compared with the SG− [26].

In another study enrolling patients with moderate to severe GERD, we found evidence for improved responsiveness and longitudinal construct validity of FT+, but not for SG+, in patients randomized to rating CMS prior to rating their own health [25]. Furthermore, as in the current study, correlations with disease-specific instruments were moderate to high, providing strong support for FT validity. Nevertheless, the evidence for construct validity does not establish that the FT has interval scale properties. Improvements in responsiveness with marker state administration may be population or intervention specific in that we have found strong evidence for the phenomenon in GERD, but not CRD patients. One possible explanation is that the GERD patients' mean age was more than 15 years younger than the CRD patients; younger patients may gain more from CMS in terms of contextualizing their subsequent responses than do older patients.

The traditional use of the SG includes administration of CMS. Our three randomized studies provide consistent evidence that use of CMS does not improve SG responsiveness and they may compromise its validity. The results of the three studies suggest that investigators can minimize respondent burden by using the SG in the least burdened fashion, without prior rating of CMS. Our approach to this investigation was to start from a neutral position with regard to the CMS. That is, we made no assumptions about their impact on efficiency or measurement properties, and proceeded accordingly. We found the obvious in terms of efficiency – CMS adds appreciable time and respondent burden – but failed to observe compelling evidence of benefit or harm in terms of measurement properties. An alternative philosophical approach to the investigation would have been to assume that the CMS has beneficial effects on measurement properties, and thus to recommend its use even in the face of increased respondent burden unless results showed definitively excluded benefit, or demonstrated definitive compromise of measurement properties with the CMS. Observers who find this alternative approach compelling might conclude that our results fail to definitively demonstrate equivalence or detrimental effects of the CMS on measurement properties and that, as a result, investigators should continue its use despite the associated administrative burden.

Possible explanations for the lack of CMS' usefulness for the SG include fatigue in patients who performed multiple ratings on the SG [13, 48]. In our study, rating the SG with CMS required approximately 4–5 min more than rating the FT with CMS which patients on average completed in 6–9 min. Furthermore, the increased complexity of the SG might lead to difficulties with marker state rating. To the extent that patients find rating a hypothetical health state challenging in itself, the SG may add to this possible confusion. Analyses of data from the first of the three trials we conducted indicate that the reliability of CMS is lower for the SG than for the FT [49]. In that study, the smaller randomized trial of 84 patients we described above, 64.2% of the CMSs ratings were in the correct order of severity during two measurements on the FT but only 11.3% on the SG.

In addition, detecting small changes in FT scores as a result of rehabilitation provides further support for the responsiveness the FT had demonstrated in other studies [21–24, 47]. In the previous study in patients with COPD undergoing a similar intervention, the changes measured with the FT were larger than in the current study. Changes measured with the validation instruments were also proportionally larger in that study. For example, CRQ domain change scores ranged from 0.36 to 0.83 on the 7-point scale in the current study, but they ranged from 0.91 to 1.42 in the previous study. The MID that corresponds to patients' experience of an important change in HRQL corresponds to a change of 0.5 on the 7-point scale of the CRQ [50, 51]. However, interpretation of the MID should always acknowledge that if the mean change in HRQL scores is below the MID, a proportion of patients may still have experienced a change that is above the MID [52–54]. Our results based on the mean change scores of the CRQ and the SGRQ indicate that, in this study, a large proportion of patients did indeed experience important changes to their HRQL [40]. The correlations between the change scores on the FT and the validation instruments suggest that the FT is indeed measuring changes in HRQL.

In conclusion, we demonstrated in this and two other studies that CMS possibly improve the construct validity, but not the responsiveness of the FT. Thus, decisions about use of CMS with the FT will depend on issues of feasibility, efficiency, and the relative priority given to responsiveness and validity. CMS did not, however, improve responsiveness or construct validity of the SG. Fully elucidating the impact of CMS will require further investigations in other populations and could include other preference instruments. Until this evidence is available, investigators could consider omitting CMS when administering the SG in a clinical trial setting when responsiveness and validity are the key measurement properties required.

Acknowledgements

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 and she reviewed drafts of the original manuscript making minor wording and presentation suggestions.

References

1. Ward MM, Javitz HS, Smith WM, Bakst A. Direct medical cost of chronic obstructive pulmonary disease in the U.S.A. Respir Med 2000; 94(11): 1123–1129.

2. Sculpher M. Using economic evaluations to reduce the burden of asthma and chronic obstructive pulmonary disease. Pharmacoeconomics 2001; 19(Suppl 2): 21–25.
3. Regueiro CR, Hamel MB, Davis RB, Desbiens N, Connors AF Jr., Phillips RS. A comparison of generalist and pulmonologist care for patients hospitalized with severe chronic obstructive pulmonary disease: Resource intensity, hospital costs, and survival. SUPPORT Investigators. Study to understand prognoses and preferences for outcomes and risks of treatment [comment]. Am J Med 1998; 105(5): 366–372.
4. Molloy DW, Guyatt GH, Russo R, et al. Systematic implementation of an advance directive program in nursing homes: A randomized controlled trial. JAMA 2000; 283(11): 1437–1444.
5. Plant PK, Owen JL, Parrott S, Elliott MW. Cost-effectiveness of ward based non-invasive ventilation for acute exacerbations of chronic obstructive pulmonary disease: Economic analysis of randomised controlled trial. Br Med J 2003; 326(7396): 956.
6. Wennberg J. Outcomes research, cost containment, and the fear of health care rationing. N Engl J Med 1990; 323: 1202–1204.
7. Patrick D, Erickson P. Health status and health policy: Quality of life in health care evaluation and resource allocation. New York: Oxford University Press, 1993.
8. Torrance GW, Feeny D. Utilities and quality-adjusted life years. Int J Technol Assess Health Care 1989; 5(4): 559–575.
9. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med 1993; 118(8): 622–629.
10. Feeny DH, Torrance GW. Incorporating utility-based quality-of-life assessment measures in clinical trials. Two examples. Med Care 1989; 27(3 Suppl): S190–S204.
11. Kaplan RM, Bush JW, Berry CC. Health status: Types of validity and the index of well-being. Health Serv Res 1976; 11(4): 478–507.
12. EuroQolGroup. EuroQol – a new facility for the measurement of health-related quality of life. Health Policy 1990; 16: 199–208.
13. Bennett KJ, Torrance G. Measuring health state preferences and utilities: Rating scale, time trade-off, and standard gamble techniques. In: Spilker B (ed.), Quality of Life and Pharmacoeconomics in Clinical Trials, 2nd ed. Philadelphia: Lippincott-Raven, 1996: 259.
14. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the Health Utilities Index mark 3 system. Med Care 2002; 40(2): 113–128.
15. Brazier JE, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol 1998; 51: 1115–1128.
16. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002; 21(2): 271–292.
17. Green C, Brazier J, Deverill M. Valuing health-related quality of life. A review of health state valuation techniques. Pharmacoeconomics 2000; 17(2): 151–165.
18. Llewellyn-Thomas HA, Thiel EC, McGreal MJ. Cancer patients' evaluations of their current health states: The influences of expectations, comparisons, actual health status, and mood. Med Decis Making 1992; 12(2): 115–122.
19. Llewellyn-Thomas HA. Health state descriptions. Purposes, issues, a proposal. Med Care 1996; 34(12 Suppl): DS109–DS118.
20. Juniper EF, Thompson AK, Roberts JN. Can the standard gamble and rating scale be used to measure quality of life in rhinoconjunctivitis? Comparison with the RQLQ and SF-36. Allergy 2002; 57(3): 201–206.
21. Llewellyn-Thomas HA, Williams JI, Levy L, Naylor CD. Using a trade-off technique to assess patients' treatment preferences for benign prostatic hyperplasia. Med Decis Making 1996; 16(3): 262–282.
22. Juniper E, Thompson A, Roberts J. Can the standard gamble and rating scale be used to measure quality of life in rhinoconjunctivitis? Comparison with the RQLQ and SF-36. Allergy 2002; 57: 201–206.
23. Fries JF, Ramey DR. "Arthritis specific" global health analog scales assess "generic" health related quality-of-life in patients with rheumatoid arthritis. J Rheumatol 1997; 24(9): 1697–1702.
24. Bakker CRM, Vansantenhoeufft M, Bolwijn P, Vandoorslaer E, Bennett K, Vanderlinden S. Patient utilities in fibromyalgia and the association with other outcome measures. J Rheumatol 1995; 22: 1536–1543.
25. Schünemann HJ, Armstrong D, Fallone C, Barkun A, Degli'Innocenti A, Heels-Ansdell D, Wiklund I, Tanser L, Chiba N, Austin P, Van Zanten S, El-Dika S, Guyatt GH. A randomized multi-center trial to evaluate simple utility elicitation techniques in patients with gastro-esophageal reflux disease. Medical Care 2004; 42(11): 1132–1142.
26. Schünemann HJ, Griffith L, Stubbing D, Goldstein R, Guyatt GH. A clinical trial to evaluate the measurement properties of 2 direct preference instruments administered with and without hypothetical marker states. Med Decis Making 2003; 23(2): 140–149.
27. Bennett KJ, Torrance GW, Moran LA, Smith F, Goldsmith CH. Health state utilities in knee replacement surgery: The development and evaluation of McKnee. J Rheumatol 1997; 24(9): 1796–1805.
28. Bennett K, Torrance G, Tugwell P. Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. Control Clin Trials 1991; 12(4 Suppl): 118S–128S.
29. Froberg DG, Kane RL. Methodology for measuring health-state preferences–III: Population and context effects. J Clin Epidemiol 1989; 42(6): 585–592.
30. Mohide EA, Torrance GW, Streiner DL, Pringle DM, Gilbert R. Measuring the wellbeing of family caregivers using the time trade-off technique. J Clin Epidemiol 1988; 41(5): 475–482.
31. Guyatt G. Making Sense of Quality-of-Life Data. Med Care 2000; 38: 175–179.
32. Feeny D, Juniper E, Ferrie P, Griffith L, Guyatt G. Why not just ask the kids? Health-related quality of life in children with asthma. In: Drotar D (ed.), Measuring Health-Related Quality of Life in Children and Adolescents: Implications for Research, Practice, and Policy. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers, 1998: 171–185.

33. Lacasse Y, Brosseau L, Milne S, et al. Pulmonary rehabilitation for chronic obstructive pulmonary disease. Cochrane Database of Systematic Reviews 2002(3): CD003793.
34. Stalmeier PF, Goldstein MK, Holmes AM, et al. What should be reported in a methods section on utility assessment?. Med Decis Making 2001; 21(3): 200–207.
35. Schünemann HJ, Goldstein R, Mador J, McKim D, Stahl E, Griffith L, Puhan M, Grant BJB, Austin P, Collins R, Guyatt GH. A randomized controlled trial to evaluate the self-administered standardized CRQ. Eur Respir J 2005; 25(1): 31–40.
36. Schünemann HJ, Stahl E, Austin P, Aki E, Armstrong D, Stubbing, D, Guyatt GH. A comparison of narrative and table formats for presenting hypothetical health states to patients with gastrointestinal or pulmonary disease: Results from two randomized studies. Med Decis Making 2004; 24: 53–60.
37. Furlong W, Feeny D, Torrance G, Barr R. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. Ann Med 2001; 33: 375–384.
38. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions. Health Utilities Index. Pharmacoeconomics 1995; 7(6): 503–520.
39. Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. A measure of quality of life for clinical trials in chronic lung disease. Thorax 1987; 42(10): 773–778.
40. Schünemann H, Goldstein R, Mador J, et al. A randomized controlled trial to evaluate the self-administered standardized CRQ. Eur Respir J 2005; 25(1): 31–40.
41. Jones PW, Quirk FH, Baveystock CM. The St George's Respiratory Questionnaire. Respir Med 1991; 85(Suppl B): 25–31; discussion 33–37.
42. Ware JE Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30(6): 473–483.
43. Birbeck G, Kim S, Hays R, Vickery B. Quality of life measures in epilepsy. How well can they detect change over time?. Neurology 2000; 54(9): 1822–1827.
44. Simon G, Revicki D, Grothaus L, Vonkorff M. SF-36 Summary Scores. Are Physical and Mental Health Truly Distinct?. Med Care 1998; 36(4): 567–572.
45. Hays R, Sherbourne C, Spritzer K, Dixon W. A micro-computer program (sf36.exe) that generates SAS Code for scoring the SF-36 Health Survey. Proceedings of the 22nd Annual SAS Users Group International Conference 1996; pp. 1128–1132.
46. Guyatt GH, Townsend M, Berman LB, Pugsley SO. Quality of life in patients with chronic airflow limitation. Br J Dis Chest 1987; 81(1): 45–54.
47. Schünemann H, Griffith L, Jaeschke R, Stubbing D, Goldstein R, Guyatt GH. Evaluation of the minimal important difference for the feeling thermometer and St. Georges Respiratory questionnaire in patients with chronic airflow limitation. J Clin Epidemiol 2003; 56(12): 1170–1176.
48. Llewellyn-Thomas HA, McGreal MJ, Thiel EC. Cancer patients' decision making and trial-entry preferences: The effects of "framing" information about short-term toxicity and long-term survival. Med Decis Making 1995; 15(1): 4–12.
49. Puhan MA, Guyatt G, Montori V, Devereaux PJ, Bhandari M, Griffith L, Goldstein R, Schünemann HJ. The standard gamble demonstrated lower reliability than the feeling thermometer. J Clin Epidemiol 2005; 58(5): 458–465.
50. Jaeschke R, Guyatt GH, Keller J, Singer J. Interpreting changes in quality-of-life score in N of 1 randomized trials. Control Clin Trials 1991; 12(4 Suppl): 226S–233S.
51. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10(4): 407–415.
52. Norman GR, Sridhar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. Med Care 2001; 39(10): 1039–1047.
53. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002; 55(9): 900–908.
54. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002; 77(4): 371–383.

Address for correspondence: Dr Schünemann, Division of Clinical Research Development and Information Translation, Italian National Cancer Institute Regina Elena, Via Elio Chianesi 53, 00144 Rome, Italy
Phone: +39 06 5266 5102
E-mail: schuneh@mcmaster.ca.