Dovepress

ORIGINAL RESEARCH

Is the St. George's Respiratory Questionnaire an Appropriate Measure of Symptom Severity and Activity Limitations for Clinical Trials in COPD? Analysis of Pooled Data from Five Randomized Clinical Trials

Angely Loubert[1]
Antoine Regnault [1]
Juliette Meunier [1]
Florian S Gutzwiller [2]
Stéphane A Regnier[2]

[1]Modus Outcomes, Lyon, France;
[2]Novartis Pharma AG, Basel, Switzerland

Purpose: The objective of this study was to examine the psychometric properties of the St. George's Respiratory Questionnaire (SGRQ) in patients with chronic obstructive pulmonary disease (COPD) using Rasch measurement theory (RMT) analyses.

Materials and Methods: RMT analysis was conducted on the baseline SGRQ data from five multi-national, Phase III randomized trials investigating a fixed-dose combination of a long-acting β2-agonist and a long-acting muscarinic antagonist in COPD patients. Analysis was performed for the SGRQ "Symptoms" and "Activity" domains. An exploratory analysis was also conducted using the different specific symptoms as defined in the reconceptualization of the SGRQ "Symptoms" domain. Differential item functioning (DIF) analysis was performed for geographical regions on the "Activity" domain, in order to explore cross-cultural validity of the SGRQ.

Results: Overall, the SGRQ "Activity" domain showed good measurement property, but two items ("Sitting or lying still making feel breathless" and "Playing sports or game making feel breathless") showed very high fit residuals. The SGRQ "Symptoms" domain demonstrated good targeting; however, two items showed disordered thresholds ("Coughed" and "Brought up phlegm"). In an exploratory RMT analysis, measures for "Cough and Sputum", "Breathing difficulties" or "Wheezing attacks" showed unsatisfactory measurement properties with poor reliability (person separation index = 0.35, 0.66 and 0.16, respectively) and targeting issues. The examination of cross-cultural performances of the SGRQ "Activity" items showed a great variability in the responses to these items in different global regions.

Conclusion: Our results indicated that SGRQ may not be an appropriate instrument to measure symptom severity or activity limitations in patients with COPD. Hence, there is a need to develop other relevant PRO instruments that can be used in conjunction with SGRQ to provide a holistic assessment of the health status of COPD patients in clinical research.

Keywords: patient-reported outcomes, health-related quality of life, health status, Rasch measurement theory, psychometric evaluation

Introduction

Patients with COPD may experience a range of symptoms (dyspnea, cough, mucus production, chest tightness, wheeze) of varying severities that dramatically impact their health-related quality of life (HRQoL), daily activities, physical functioning,

Correspondence: Angely Loubert
Modus Outcomes, 61 Cours De La
Liberté, Lyon 69003, France
Email angely.loubert@modusoutcomes.com

mental health (anxiety and depression) and sleep.[1] Measuring severity and frequency of symptoms and their impact, especially in terms of activity limitations, is therefore essential for the demonstration of the efficacy of new treatments in the context of clinical trials in COPD. The US Food and Drug Administration (FDA) draft guidance on drug development in COPD[2] acknowledges three patient-reported outcomes (PROs) measures of efficacy: symptoms, activity limitations, and HRQoL measures.

Numerous PRO instruments have been developed over the years in the field of COPD and selecting the most appropriate for inclusion in a clinical trial is not a straightforward question.[3] The St. George's Respiratory Questionnaire (SGRQ) is the most widely used in clinical research in COPD and was developed in 1992.[4,5] It is a multidimensional instrument that is composed of three domains: Symptoms, Activity and Impact. Thorough psychometric evaluation of the Symptoms and Activity domains using modern psychometrics methods, such as Rasch measurement theory (RMT), will allow a finer understanding on how they perform[6] and whether they could be good candidates to measures these two PRO domains of importance highlighted in the FDA draft guidance in COPD.[2]

The overall objective of the present analyses was to better understand the extent to which the items composing "Symptoms" and "Activity" domains of the SGRQ form scales that can efficiently capture and meaningfully characterize treatment efficacy in clinical trials, but also exploring the variability in measurement by these domains across cultures, which can occur in large multinational clinical trials. For this purpose, RMT analyses were performed using pooled data from five clinical trials investigating a fixed-dose combination of a long-acting β2-agonist (indacaterol) and a long-acting muscarinic antagonist (glycopyrrolate).

Materials and Methods
Study Sample

Data from five clinical trials investigating a fixed-dose combination of a long-acting β2-agonist (indacaterol) and a long-acting muscarinic antagonist (glycopyrrolate) in COPD patients were used: FLIGHT-1, FLIGHT-2,[7] LANTERN,[8] SHINE[9] and SPARK.[10] The selection of the clinical trial data to be used was pre-specified before any analysis was performed and no change in selection was done during the analysis. All trials were Phase III, multi-center, randomized, double-blind, parallel-group trials. Duration of follow-up

across the five trials ranged from 12 to 64 weeks. The comparators across these trials included indacaterol, glycopyrrolate, tiotropium, salmeterol/fluticasone combination or placebo. All of the trials collected the original version of the SGRQ. The SPARK trial included severe to very severe patients while all other trials included moderate to severe patients based on Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines.[11] The trials had various geographical scopes and covered North America, Europe, Asia and Africa.

The analyses reported here were post hoc analyses performed on fully anonymized data from the five trials. All five trials were conducted in accordance with the declaration of Helsinki, and international conference harmonization good practice guidelines. All patients provided written informed consent for participation. Full information on the ethics statements can be found in the original publications on each trial.

There was no need for Ethics Committee approval for this work since this study was a post hoc analysis of randomized clinical trials that already received Ethics Committee approval.

St. George's Respiratory Questionnaire (SGRQ)

The SGRQ is a PRO instrument developed to assess the health status of patients with COPD and asthma.[4] It includes 50 items assessing three domains: Symptoms (severity and frequency of respiratory symptoms), Activity (effect of disease on common daily physical activities) and Impact (psycho-social effects of the disease). A total composite score can be calculated using all SGRQ items as well as three domain scores. A COPD-specific version, the SGRQ-C, includes a selection of 40 recoded items from the SGRQ.[12]

Our RMT analysis was underpinned by a conceptual examination of the content of the two SGRQ domains of interest as defined by the SGRQ developers. While the SGRQ "Activity" domain was deemed conceptually clear, the authors of this study deemed that "Symptoms" domain could be further broken down into three different types of symptoms: Cough and sputum; Breathing difficulty; and Wheezing attacks (Figure 1).

Rasch Measurement Theory (RMT)

RMT analyses use a mathematical model, the Rasch model, to evaluate the extent to which items from an
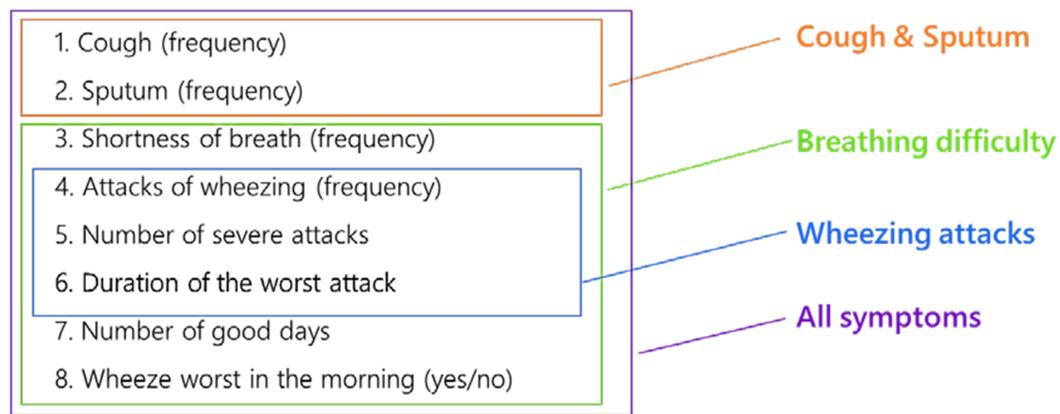
Figure 1 Conceptualization of the SGRQ "Symptoms" domain.

instrument can be summed to build a proper measurement of the underlying abstract concept.[13–15] RMT analyses explore the following properties:

- Targeting: With the Rasch model, the parameters for items and participants are estimated in the same continuum, allowing a direct comparison of the distributions of items and participants over this common continuum. Targeting addresses the matching of participants and items ensuring a sufficiently precise estimation of participant and item parameters. This is assessed by comparing the spread of person and item location estimates over the common continuum.

- Fit: Items must work together to define a clinically and statistically meaningful score. Otherwise, it is inappropriate to sum item responses to reach a total score and consider the total score an accurate measure of each target concept. When items do not work together in this way (ie, there is item misfit), the validity of an item set is questionable. Item fit is assessed based on ordering of item response options (ie, ordering of item thresholds)[16] and comparison of observed and expected responses using statistical indices and graphical examination of item characteristic curve (ICC).[17] Statistical indices include standardized fit residuals, which are recommended to lie in the range −2.50 to +2.50,[15] and chi-square tests.

- Reliability: The principle of reliability is that applying the patient-reported outcome measure on different occasions or by different observers produces consistent results.[18] It is assessed using the Person Separation

Index (PSI),[19] a reliability coefficient estimate. Reliability coefficients are commonly interpreted as follows: <0.70: unsatisfactory; 0.70–0.79: modest; 0.80–0.89: adequate; 0.90–1.00: good.[20]

- Differential item functioning (DIF): A key criterion to achieve a strong measurement is invariance implying that items mean the same within all patients, regardless of their characteristics (demographics, clinical, etc). In these analyses, we used DIF to examine cross-cultural invariance: the expected response to an item was compared for patients who have the same level of the measured concept but belong to the different global regions that are investigated in global clinical trials.

RMT analyses were carried out using RUMM 2030 software (RUMM Laboratory, Perth, Australia) on the pooled baseline SGRQ data from the five trials (regardless of which treatment the individuals were assigned to). Separate analyses were performed for the SGRQ "Symptoms" and "Activity" domains: the item sets included in the Rasch model were iteratively modified based on the results, conceptual examination of the item content and previously published findings. The modification of the item sets included refinement of the item selection (eg, excluding items that would not fit the model) or recoding of the response options (eg, grouping response options). Given the large sample size and the sensitivity of the test used to sample size, non clinically meaningful difference could be statistically significant. Hence, statistical significance was consistently considered with caution. DIF analysis was performed for geographical regions on the "Activity"

domain, in order to explore cross-cultural validity of the SGRQ. In order to facilitate interpretation of the findings and to get more homogenous groups, the 42 countries included in the trials were grouped in 13 geographical regions according to the United Nations statistics division categorization:[21] Canada, Central and south America, China, Eastern Europe, India, Japan, North Africa and west Asia, Northern Europe, Oceania and south Africa, South and east Asia, Southern Europe, USA, Western Europe. Both uniform and non-uniform DIF were tested: DIF is said to be uniform if the difference in the expected response to an item between two groups is the same across the full range of the targeted concept being measured. A DIF is non-uniform if the difference between groups depends on the targeted concept being measured (eg, patients with low activity limitations from a given global region tend to endorse an item less than patients from other global regions while patients with high activity limitations of that same global region tend to endorse it more).

Results
Sample Description

The five trials included a total of 7,119 patients. Overall, the mean age at inclusion was between 63 and 65 years across all trials (Table 1). Most of the patients were male (73% versus 27% of female). This percentage was relatively similar among trials, except for LANTERN where 91% of patients were male, and FLIGHT-2 with only 58%. Most patients had moderate or severe COPD (Graded as 2 or 3 regarding GOLD system), with, respectively, 42% graded as moderate and 51% as severe. Only the SPARK trial aimed at including severe to very severe patients (79% and 21%, respectively). The RMT analysis of the SGRQ data was conducted using data from 7,116 patients (3 patients were excluded because all SGRQ items were missing for them).

RMT Analysis of the SGRQ "Activity" Domain

The analysis of the full SGRQ "Activity" item set showed good targeting with some gaps in the coverage of patients,

Table I Summary of the Demographics of the Samples from the Five Clinical Trials Used for the Rasch Measurement Theory Analysis of the SGRQ in Chronic Obstructive Pulmonary Disease

	Study					
	Shine (A2303) N=2135	**Spark (A2304) N=2207**	**Lantern (A2331) N=741**	**Flight-1 (A2336) N=1038**	**Flight-2 (A2337) N=998**	**Total N=7119**
Age – Mean (SD)	63.92 (8.83)	63.29 (7.99)	65.03 (7.87)	63.81 (8.31)	63.06 (8.39)	63.70 (8.35)
Male - n(%)	1610 (75.4%)	1650 (74.8%)	672 (90.7%)	705 (67.9%)	584 (58.5%)	5221 (73.3%)
COPD severity (GOLD) - n(%)						
Mild (GOLD 1)	2 (0.1%)	0 (0.0%)	3 (0.4%)	0 (0.0%)	0 (0.0%)	5 (0.1%)
Moderate (GOLD 2)	1358 (63.6%)	2 (0.1%)	387 (52.2%)	632 (60.9%)	605 (60.6%)	2984 (41.9%)
Severe (GOLD 3)	775 (36.3%)	1747 (79.2%)	344 (46.4%)	400 (38.5%)	384 (38.5%)	3650 (51.3%)
Very severe (GOLD 4)	0 (0.0%)	458 (20.8%)	0 (0.0%)	1 (0.1%)	0 (0.0%)	459 (6.4%)
Baseline FEV1 Post-Bronchodilator (% pred. FEV1) – Mean (SD)	51.94 (12.77)	35.16 (7.74)	47.74 (11.76)	54.61 (13.48)	54.57 (13.06)	47.03 (14.13)
Smoking history of subject - n(%)						
Ex-smoker	1288 (60.3%)	1378 (62.4%)	549 (74.1%)	526 (50.7%)	461 (46.2%)	4202 (59.0%)
Current smoker	847 (39.7%)	829 (37.6%)	192 (25.9%)	512 (49.3%)	537 (53.8%)	2917 (41.0%)
Baseline SGRQ Symptoms score – Mean (SD)	58.44 (20.97)	60.08 (21.29)	45.46 (19.79)	60.65 (21.93)	60.66 (21.04)	58.23 (21.56)
Baseline SGRQ Activity score – Mean (SD)	61.82 (20.54)	69.31 (19.27)	53.08 (17.48)	64.63 (18.54)	65.95 (18.85)	64.22 (19.91)
Baseline SGRQ Impact score – Mean (SD)	35.02 (20.92)	40.48 (20.52)	29.04 (17.55)	36.12 (18.92)	37.12 (19.93)	36.54 (20.32)

Abbreviations: COPD, chronic obstructive pulmonary disease; GOLD, global initiative for chronic obstructive lung disease; SD, standard deviation; SGRQ, St. George's Respiratory Questionnaire.

but adequate reliability, with a PSI of 0.80. However, most items showed misfit to the Rasch model (Table 2).

Items 11 ("Sitting or lying still") and 17 ("Playing sport or games") showed very high fit residuals (much greater than 2.5). A visual inspection of the item characteristic curve (ICC) of these items confirmed that these items were not as sensitive to the change in activity limitation as expected (Supplementary Material 1). Based on these results, the Rasch model was applied to an item set excluding items 11 and 17. After excluding those items, we found good targeting albeit with some gaps in the coverage of patients, as displayed in Figure 2. Most of the items still showed misfit to the Rasch model, but none showed strong under-discrimination (Table 2). An adequate reliability with a PSI of 0.78 was observed.

The item location estimates obtained for each study independently were very stable (Figure 3). They illustrated a meaningful item hierarchy ranging from item 37 ("Can't or take a long time to take bath or shower") to item 44 ("Breathing makes intense activity (eg, run) difficult"). (Figure 4). The distribution of the patients over the continuum clearly shifted towards the left (ie, more severe impact) with GOLD stage, confirming that the activity

limitations captured by the SGRQ items were worsening with the general clinical severity of COPD (Supplementary Material 2).

All items were flagged as showing some form of DIF for geographic region, mostly non-uniform (Table 2). The examination of ICCs by geographical region showed that no global region was systematically different from the others but that the expected responses to the different SGRQ "Activity" items in each global region were distributed with a fair amount of variability around the "central" expected response from the Rasch model (Supplementary Material 3). This indicates that no specific SGRQ "Activity" item is problematic for the measurement of activity limitations across geographical regions but that there is still some heterogeneity in this measurement, which should be carefully considered.

## RMT Analysis of the SGRQ "Symptoms" Domain

The RMT analysis of the full SGRQ "Symptoms" item set showed good targeting with minor gaps in the coverage of patients and disordered thresholds for 6 of the 8 items: items 1, 2, 3, 4, 5 and 6. Seven of the 8 items showed

Table 2 SGRQ "Activity" Domain: Item Fit to the Rasch Model with and without Items 11 and 17 and Differential Item Functioning (DIF) by Geographical Region (Baseline Visits, Pooled Data from 5 Trials, N=7,116)
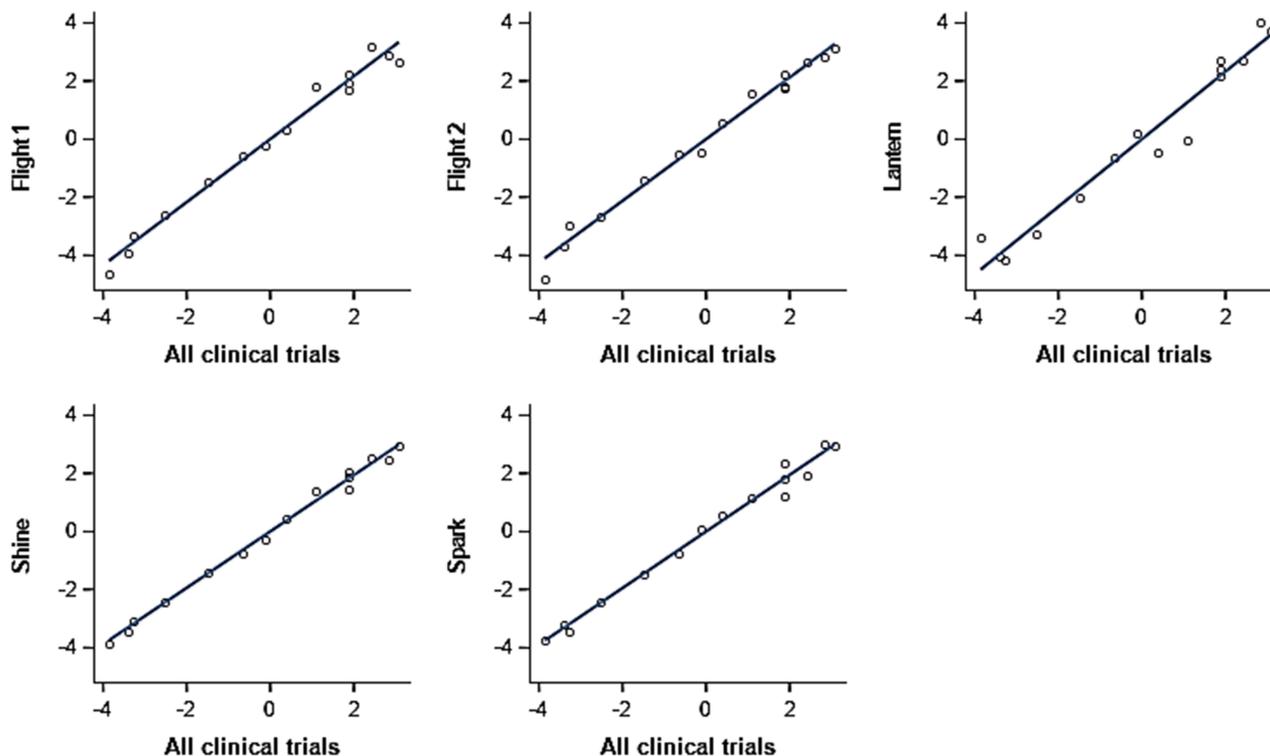
Item	Full SGRQ Activity Item Set		SGRQ Activity Item Set without Items 11 and 17		
	Fit Residual	χ^2	Fit Residual[1]	χ^2	DIF by Geographical Region
11. Sitting or lying still makes feel breathless	**8.74**	**337.7**			
12. Getting washed or dressed makes feel breathless	**−5.94**	**259.0**	−2.40	**196.9**	Non-uniform
13. Walking around the home makes feel breathless	−1.40	**144.1**	0.79	**109.5**	Non-uniform
14. Walking outside on the level makes feel breathless	−1.83	**167.6**	1.28	**158.1**	Uniform
15. Walking up a flight of stairs makes feel breathless	**−3.10**	**65.0**	−2.29	**52.1**	Non-uniform
16. Walking up hills makes feel breathless	1.54	**64.0**	**3.64**	**119.9**	Non-uniform
17. Playing sports or games makes feel breathless	**4.80**	**245.5**			
36. Long time to get washed/get dressed	**−5.45**	**250.9**	**−3.86**	**231.2**	Non-uniform
37. Cannot or take a long time to take bath or shower	−1.19	**96.3**	−0.33	**59.6**	Non-uniform
38. Walk slower than other people, or must stop for rest	**−4.04**	**186.9**	**−3.54**	**172.5**	Uniform
39. Housework takes a long time	**−6.86**	**266.4**	**−5.82**	**226.5**	Non-uniform
40. Must stop or slow down walking up one flight of stairs	**−4.25**	**218.8**	**−4.22**	**218.2**	Non-uniform
41. Must stop or slow down hurrying or walking fast	**−5.75**	**125.0**	**−5.99**	**106.9**	Non-uniform
42. Breathing makes mild activity (eg walk up hill) difficult	**−4.74**	**90.4**	**−4.22**	**80.6**	Non-uniform
43. Breathing makes moderate activity (eg jog) difficult	**−3.66**	**44.8**	**−3.62**	**37.1**	Non-uniform
44. Breathing makes intense activity (eg run or cycle) difficult	−0.48	**36.2**	0.52	**39.2**	Uniform

Notes: In bold, fit residuals outside the range of −2.5/2.5; χ^2 values with p-values<0.001. Adapted with permission from St George's University of London. Available from: http://www.healthstatus.sgul.ac.uk/SGRQ_download/SGRQ%20Manual%20June%202009.pdf. Accessed September 02, 2020.[27]
Abbreviations: SGRQ, St. George's Respiratory Questionnaire; DIF, differential item functioning.
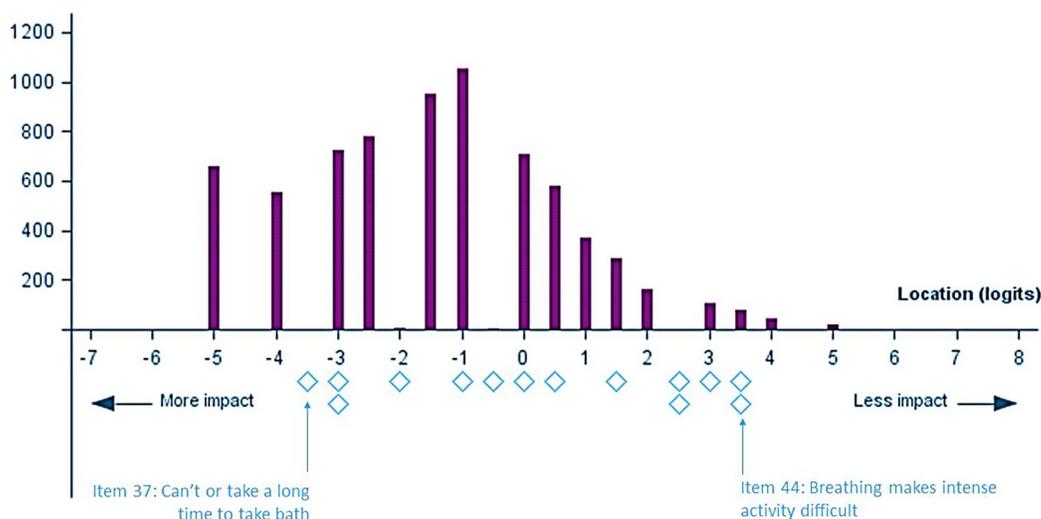
Figure 2 Person-item distribution of the SGRQ "Activity" domain without items 11 and 17 (Baseline visits, pooled data from 5 trials, N=7,116). The top part of the figure (purple) shows the distribution of impact on activity level in the sample, and the lower part (blue) shows the distribution of impact on activity level in the SGRQ "Activity" domain item thresholds. The blue diamonds corresponds to the "thresholds" between two adjacent item response categories (presented in Figure 4).


Figure 3 Comparison of SGRQ Activity item location estimates obtained from each trial separately vs obtained from pooled five trials.

misfit to the Rasch model (Table 3). A modest reliability, with a PSI of 0.72 was observed.

As the response categories were not functioning as expected, a recategorization of the response scales of the six items with disordered thresholds was explored: for items 1 to 4, the two response options reflecting chest infection were grouped with "not at all" because they were conditional of having had chest infections, which was a different concept; for item 5, a category "1, 2 or 3 attacks" was created by merging the three corresponding original categories; for item 6, a category "more than 1 day, but less than 1 week" was created by merging the
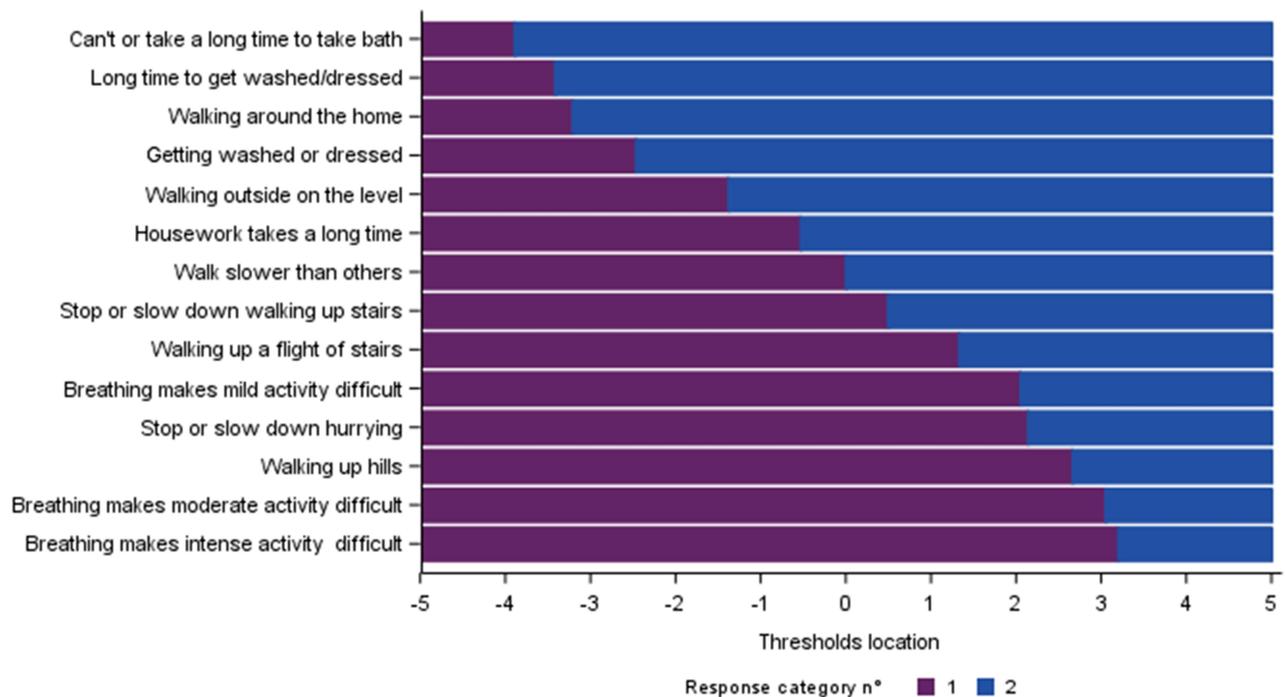
Figure 4 RMT analysis of the SGRQ "Activity" domain without item 11: item hierarchy (Baseline visits from 5 trials, N=7,116).

categories "1 or 2 days" and "3 days or more". The resulting items were analyzed using the Rasch model, leading to good targeting with some gaps in the coverage of patients (Figures 5 and 6). Two items showed disordered thresholds: Items 1 "Coughed" and 2 "Brought up phlegm (sputum)". Six items showed misfit to the Rasch model (Table 3). Items 6 ("Duration of worst attack") was the most problematic as it had a high standardized residual statistic, indicating that it was not as sensitive to change in overall symptom severity as expected. A modest reliability with a PSI of 0.70 was observed. The distributions of the patients over the continuum were not markedly different for GOLD stages 2, 3 and 4 (Supplementary Material 2), which suggests that the symptom score does not discriminate well between clinical severity of COPD.

An exploratory RMT analysis was also conducted using the different specific symptoms as defined in the reconceptualization of the SGRQ "Symptoms" domain (Figure 1). The resulting measures for "Cough and Sputum", "Breathing difficulties" or "Wheezing attacks" showed unsatisfactory measurement properties with poor reliability (PSI= 0.35, 0.66 and 0.16, respectively) and targeting issues (a large number of patients having the maximum "Cough and Sputum" measure or the minimum "Wheezing attacks" score) (data not

shown). These results were probably due to the small number of SGRQ items available for each type of symptom.

Table 3 SGRQ "Symptoms" Domain: Item Fit to the Rasch Model Before and After Recoding Response Options (Baseline Visits, Pooled Data from 5 Trials, N=7,116)

Item	SGRQ Symptoms Item Set		SGRQ Symptoms Item Set With Recoded Response Options	
	Fit Residual	Chi^2	Fit Residual	Chi^2
1. Coughed	**−4.00**	**181.5**	**−6.44**	**160.5**
2. Brought up phlegm (sputum)	**2.71**	**100.1**	−0.69	**75.3**
3. Shortness of breath	**−5.68**	**205.4**	**−6.40**	**155.7**
4. Attacks of wheezing	−2.38	**212.6**	**−5.41**	**185.0**
5. Number of severe attacks	**−2.55**	**236.1**	−0.68	**249.9**
6. Duration of worst attack	**8.13**	**294.6**	**9.01**	**238.1**
7. Number of good days	**4.01**	**52.5**	**3.44**	16.4
8. Worse in the morning	**11.30**	22.3	**12.47**	27.8

Notes: In bold, fit residuals outside the range of −2.5/2.5; χ^2 values with p-values<0.001. Adapted with permission from St George's University of London. Available from: http://www.healthstatus.sgul.ac.uk/SGRQ_download/SGRQ%20Manual%20June%202009.pdf. Accessed September 02, 2020.[27]
Abbreviation: SGRQ, St. George's Respiratory Questionnaire.
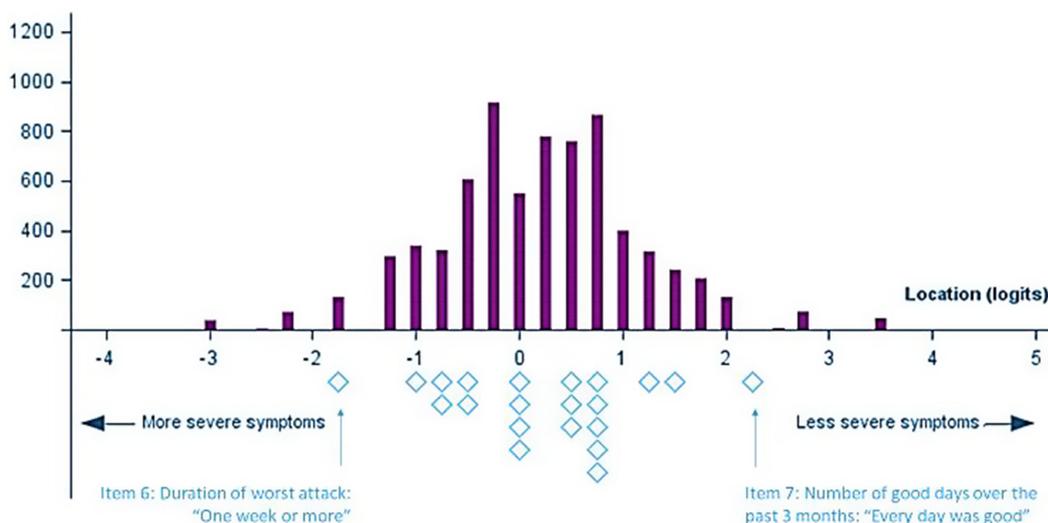
Figure 5 Person-threshold distribution of the SGRQ "Symptoms" domain after item recoding (Baseline visits from 5 trials, N=7,116). The top part of the figure (purple) shows the distribution of symptoms level in the sample, and the lower part (blue) shows the distribution of symptoms level in the SGRQ 'Symptoms' domain item thresholds. The blue diamonds corresponds to the "thresholds" between two adjacent item response categories (presented in Figure 6).
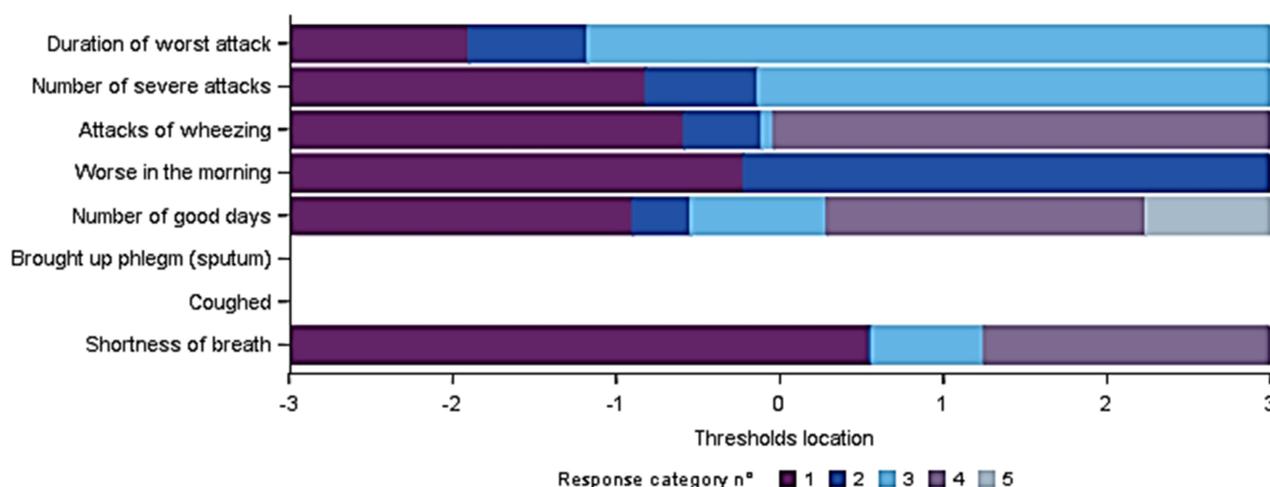
Figure 6 Thresholds map of the SGRQ "Symptoms" domain after item recoding (Baseline visits from 5 trials, N=7,116).

Discussion

The objective of this work was to determine whether the SGRQ could be an appropriate instrument to support the definition of clinical trial endpoints of symptom severity and activity limitations in COPD. The SGRQ "Activity" domain showed good measurement property overall. The main issue was the inclusion of two items in the scoring ("Sitting or lying still making feel breathless" and "Playing sports or game making feel breathless"). This finding was consistent with previous findings reported in the literature,[12,22] that led to the creation of the SGRQ-C.[12]

The examination of cross-cultural performances of the SGRQ "Activity" items showed a great variability in the responses to these items in different global regions.

The SGRQ 'Symptoms' domain showed good targeting, but some major issues related to its response scales, composition and reliability. An exploratory recoding of the response scale of the SGRQ "Symptoms" items led to some improvement in its measurement performance. Yet, it did not fully address the existing issues. A key question in this context is whether COPD symptoms map on a single continuum. Given the heterogeneity of COPD

symptoms, the SGRQ "Symptoms" score may reflect different symptomatic manifestations in a single composite index. While this kind of index can be useful, it does not fit in the measurement paradigm of the Rasch model. With this in mind, we explored the possibility of "unpacking" the SGRQ Symptoms items to create independent measures of three COPD symptoms: "Cough and sputum", "Breathing difficulty" and "Wheezing attacks". However, these tentative measures did not show satisfactory performance, likely due to the limited number of items available in the SGRQ to characterize each symptom.

The findings of our analyses are consistent with previously published literature[12,22] and highlight some interpretation risks if the SGRQ "Symptoms" and "Activity" scores were used in clinical trials to measure symptom severity and activity limitations in patients with COPD, respectively. First, the SGRQ "Activity" domain covers a wide range of severity, by including a wide variety of activities, which was reflected by the breadth of the continuum resulting from the RMT analyses of this domain. While it is a strength for the versatility of use of the SGRQ, it may be an issue in the context of a clinical trial where the objective is to demonstrate the benefit of a treatment in a targeted population. In a clinical trial sample where patients are rigorously selected, only a few items are well targeted to the very level of severity of these patients while certain items may be targeted for one severity range (eg, GOLD stage 1), it is another few items that will be relevant to another severity (eg, GOLD stage 2). Therefore, small improvements in the activity levels that may still be meaningful for patients may not be captured by the SGRQ "Activity" overall domain score because only one or two items are likely to change for them. Second, even though the cross-cultural examination of the SGRQ "Activity" domain did not identify a global region that was systematically different from the others, it revealed a substantial variation in how the patients respond to the SGRQ items depending on where they live. Hence, the generalizability of the SGRQ measurement properties across countries may be questioned, and, more importantly, for global clinical trials, the cross-cultural variability in the measurement of the targeted concept could hinder the ability of the trial to detect a true treatment benefit.[23] Finally, as it stands, the SGRQ "Symptoms" domain may generate results that are difficult to interpret, as it is composed of a heterogenous set of symptoms: if a change is observed in the SGRQ symptom score, it would not be possible to distinguish whether it is more an improvement in cough or sputum production, breathing difficulties, or wheezing attacks.

PRO endpoints, specifically about symptoms and activity limitations, are critical for the demonstration of the benefit of new treatments in COPD clinical trials. Our results indicated that the SGRQ may not be an appropriate instrument for this purpose. Other PRO instruments should therefore be identified for this purpose. An appropriate instrument to measure the impact of COPD on activity level in a clinical trial setting should focus on targeted activities, defined according to the targeted population severity. For symptoms, using specific measures of the various COPD symptomatic manifestations would allow changes in each type of symptoms to be efficiently captured and meaningfully interpreted. Instruments such as the comprehensive cough and sputum assessment questionnaire (CASA-Q),[24] or the Evaluating Respiratory Symptoms in COPD (E-RS:COPD),[25] may be good candidates for symptom severity measures in the context of clinical trials.

The examination of the measurement properties of the SGRQ using the RMT analyses conducted using these SGRQ data from phase III clinical trials came with some challenges and limitations. First, while pooling data from five large phase III trials created a dataset including thousands of SGRQ assessments which led to precise statistical estimates in the Rasch model, most statistical indices and tests used in this setting are highly sensitive to sample size (eg, for fit to the Rasch model, or for exploration of DIF). Hence, any significance testing had to be considered with extreme caution. Secondly, our results were obtained from five trials with specific characteristics. The patients mostly had moderate to severe COPD and the percentage of male in our sample was slightly greater than in the general COPD population.[26] Whether our results can be generalized, especially to patients with milder or very severe COPD, is an open question. Finally, our analyses focused on the cross-sectional measurement properties of the SGRQ Activity and Symptoms domains. Further analyses would be needed to inform their longitudinal measurement properties (ie, how appropriate is it to use them to characterize a change in activity limitations or symptom severity) and the measurement properties of the third domain composing the SGRQ (Impact).

Conclusion

While the SGRQ could be a potentially relevant summary index of overall HRQoL, its use to specifically target symptom severity or activity limitations in clinical trials in COPD is not warranted. Hence, other relevant PRO instruments should be considered to be used along with

SGRQ to provide a holistic assessment of the health status of COPD patients in clinical research.

Author Contributions

All authors contributed to data analysis, drafting or revising the article, have agreed on the journal to which the article will be submitted, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Disclosure

Angely Loubert, Antoine Regnault, and Juliette Meunier are employees of Modus Outcomes, a patient-centered outcomes consultancy commissioned by Novartis to conduct this study. Florian S. Gutzwiller and Stéphane A. Regnier are employees of Novartis Pharma AG. The authors report no other conflicts of interest in this work.

References

1. Miravitlles M, Ribera A. Understanding the impact of symptoms on the burden of COPD. *Respir Res*. 2017;18(1):67. doi:10.1186/s12931-017-0548-3
2. Food and Drug Administration. Chronic obstructive pulmonary disease: developing drugs for treatment guidance for industry (draft guidance). U.S. Department of Health and Human Services; 2016.
3. Ekstrom M, Sundh J, Larsson K. Patient reported outcome measures in chronic obstructive pulmonary disease: which to use? *Expert Rev Respir Med*. 2016;10(3):351–362. doi:10.1586/17476348.2016.1146595
4. Jones PW, Quirk FH, Baveystock CM, Littlejohns P. A self-complete measure of health status for chronic airflow limitation. The St. George's respiratory questionnaire. *Am Rev Respir Dis*. 1992;145(6):1321–1327. doi:10.1164/ajrccm/145.6.1321
5. Jones P, Miravitlles M, van der Molen T, Kulich K. Beyond FEV1 in COPD: a review of patient-reported outcomes and their measurement. *Int J Chron Obstruct Pulmon Dis*. 2012;7:697. doi:10.2147/COPD.S32675
6. Andrich D, Marais I. *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Springer; 2019.
7. Mahler DA, Kerwin E, Ayers T, et al. FLIGHT1 and FLIGHT2: efficacy and safety of QVA149 (indacaterol/glycopyrrolate) versus its monocomponents and placebo in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2015;192(9):1068–1079. doi:10.1164/rccm.201505-1048OC
8. Zhong N, Wang C, Zhou X, et al. LANTERN: a randomized study of QVA149 versus salmeterol/fluticasone combination in patients with COPD. *Int J Chron Obstruct Pulmon Dis*. 2015;10:1015.
9. Hashimoto S, Ikeuchi H, Murata S, Kitawaki T, Ikeda K, Banerji D. Efficacy and safety of indacaterol/glycopyrronium in Japanese patients with COPD: a subgroup analysis from the SHINE study. *Int J Chron Obstruct Pulmon Dis*. 2016;11:2543. doi:10.2147/COPD.S111408
10. Wedzicha JA, Decramer M, Ficker JH, et al. Analysis of chronic obstructive pulmonary disease exacerbations with the dual bronchodilator QVA149 compared with glycopyrronium and tiotropium (SPARK): a randomised, double-blind, parallel-group study. *Lancet Respir Med*. 2013;1(3):199–209. doi:10.1016/S2213-2600(13)70052-3
11. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. Global initiative for chronic obstructive lung disease (GOLD). 2020. Available from: https://goldcopd.org/. Accessed February 22, 2020.
12. Meguro M, Barley EA, Spencer S, Jones PW. Development and validation of an improved, COPD-specific version of the St. George respiratory questionnaire. *Chest*. 2007;132(2):456–463. doi:10.1378/chest.06-0702
13. Rasch G. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. 1960.
14. Andrich D. *Rasch Models for Measurement*. Newbury Park, CA: Sage; 1988.
15. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess*. 2009;13(12):iii, ix–x, 1–177. doi:10.3310/hta13120
16. Andrich D. Rating scales and Rasch measurement. *Expert Rev Pharmacoecon Outcomes Res*. 2011;11(5):571–585. doi:10.1586/erp.11.59
17. Wright BD, Masters GN. *Rating Scale Analysis*. Mesa Press; 1982.
18. Alrubaiy L, Hutchings HA, Williams JG. Assessing patient reported outcome measures: a practical guide for gastroenterologists. *United European Gastroenterol J*. 2014;2(6):463–470. doi:10.1177/2050640614558345
19. Andrich D. An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Educ Res Perspect*. 1982;9(1):95–104.
20. Nunnally JC, Bernstein IH. *Psychometric Theory*. McGraw-Hill; 1994.
21. United Nations SD. Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings. Available from: https://web.archive.org/web/20021216043228/http://unstats.un.org/unsd/methods/m49/m49regin.htm. Accessed August 12, 2020.
22. Lo C, Liang WM, Hang LW, Wu TC, Chang YJ, Chang CH. A psychometric assessment of the St. George's respiratory questionnaire in patients with COPD using Rasch model analysis. *Health Qual Life Outcomes*. 2015;13:131. doi:10.1186/s12955-015-0320-7
23. Regnault A, Hamel JF, Patrick DL. Pooling of cross-cultural PRO data in multinational clinical trials: how much can poor measurement affect statistical power? *Qual Life Res*. 2015;24(2):273–277. doi:10.1007/s11136-014-0765-x
24. Crawford B, Monz B, Hohlfeld J, et al. Development and validation of a cough and sputum assessment questionnaire. *Respir Med*. 2008;102(11):1545–1555. doi:10.1016/j.rmed.2008.06.009
25. Leidy NK, Murray LT. Patient-reported outcome (PRO) measures for clinical trials of COPD: the EXACT and E-RS. *COPD*. 2013;10(3):393–398. doi:10.3109/15412555.2013.795423
26. Ntritsos G, Franek J, Belbasis L, et al. Gender-specific estimates of COPD prevalence: a systematic review and meta-analysis. *Int J Chron Obstruct Pulmon Dis*. 2018;13:1507–1514. doi:10.2147/COPD.S146390
27. St George's University of London. Available from: http://www.healthstatus.sgul.ac.uk/SGRQ_download/SGRQ%20Manual%20June%202009.pdf. Accessed September 02, 2020