ORIGINAL ARTICLE

Non-Small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ): Measurement Properties and Estimated Clinically Meaningful Thresholds From a Phase 3 Study

Paul Williams, MPH,[a,*] Thomas Burke, PharmD, PhD,[b] Josephine M. Norquist, MS,[c] Christina Daskalopoulou, PhD,[d] Rebecca M. Speck, PhD, MPH,[e] Ayman Samkari, MD,[f] Sonya Eremenco, MA,[e] Stephen Joel Coons, PhD[e]

[a]*Patient-Centered Solutions, IQVIA, Paris, France*
[b]*Center for Observational and Real World Evidence, Merck & Co., Inc., Kenilworth, New Jersey*
[c]*Patient-Centered Endpoints & Strategy, Merck & Co., Inc., Kenilworth, New Jersey*
[d]*Patient-Centered Solutions, IQVIA, Athens, Greece*
[e]*Clinical Outcome Assessment Program, Critical Path Institute, Tucson, Arizona*
[f]*Clinical Research, Merck & Co., Inc., Kenilworth, New Jersey*

Received 7 January 2022; revised 11 February 2022; accepted 13 February 2022
Available online - 17 February 2022

ABSTRACT

Introduction: The NSCLC Symptom Assessment Questionnaire (NSCLC-SAQ) was developed to assess NSCLC symptom severity in accordance with Food and Drug Administration evidentiary expectations leading to Food and Drug Administration qualification in 2018. This study evaluated the NSCLC-SAQ's measurement properties within a clinical trial.

Methods: The KEYNOTE-598 phase 3 study of participants with stage IV metastatic NSCLC with programmed death-ligand 1 tumor proportion score greater than or equal to 50% was used to assess the NSCLC-SAQ's reliability, construct validity, responsiveness, and estimate clinically meaningful within-person change. Other patient-reported outcome measures included patient global impression items of severity and change in lung cancer symptoms, and the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire core 30 and lung cancer module, LC13.

Results: Participants (N = 560) were mostly men (70%), had a mean age of 64 years, and had Eastern Cooperative Oncology Group performance status of 1 (64%) or 0 (36%). Internal consistency at baseline (Cronbach's $\alpha = 0.74$) and test-retest reliability after 3 weeks (intraclass correlation coefficient = 0.79) were satisfactory. NSCLC-SAQ items, domains, and total score correlated moderately to highly with patient-reported outcome measures capturing similar content, and the total score differentiated among patient global impression of severity groups ($p < 0.001$). The total score detected improvement over time and the estimated clinically meaningful within-person change threshold for improvement ranged from three to five points on the 0 to 20 scale. Few participants exhibited symptom worsening (n = 38), limiting inferences in this group.

Conclusions: The NSCLC-SAQ was found to be reliable, valid, responsive, and interpretable for assessing symptom improvement in NSCLC. Further evaluation is recommended in trial participants whose symptoms worsen over time.

***Corresponding author.**

Disclosure: Mr. Williams and Mrs. Daskalopoulou are employees of IQVIA, the company commissioned by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, New Jersey to conduct this research. Drs. Burke, Norquist, and Samkari are employees of Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, New Jersey, which funded this research study. Dr. Coons, Mrs. Eremenco, and Dr. Speck are employees of the nonprofit Critical Path Institute, which holds the copyright for the NSCLC Symptom Assessment Questionnaire.

Address for correspondence: Paul Williams, MPH, Patient-Centered Solutions, IQVIA, Tour D2, 17 bis place des Reflets, TSA 64567, 92099 La Defense Cedex, France. E-mail: paul.williams@iqvia.com

Cite this article as: Williams P, Burke T, Norquist JM, et al. NSCLC symptom assessment questionnaire: measurement properties and estimated clinically meaningful thresholds from a phase 3 study. *JTO Clin Res Rep*. 2022;3:100298.

ISSN: 2666-3643

https://doi.org/10.1016/j.jtocrr.2022.100298

Keywords: Patient-reported outcome; Non–small cell lung carcinoma; Psychometric properties; Symptom assessment; Clinically meaningful within-person change

Introduction

Lung cancer is the third most common cancer and the leading cause of cancer-related mortality in the United States.[1] People with NSCLC may experience substantial symptom burden and impact on their daily lives, particularly at advanced stages of the disease. The most frequently reported symptoms associated with NSCLC to include cough, pain, shortness of breath (dyspnea), appetite loss, and fatigue.[2] Improvement of symptom severity is considered among the most important factors by persons with advanced NSCLC when making treatment decisions.[3]

Treatment goals for advanced NSCLC focus on prolonging survival and relieving symptoms.[4] Despite this, findings from patient-reported outcome (PRO) assessments, considered the most appropriate method for assessing symptom experience, rarely make their way into U.S. Food and Drug Administration (FDA) drug labeling for oncology drugs.[5] Considering that drug labeling constitutes a formal basis for communicating drug efficacy and safety, this absence limits a holistic understanding of the benefit-risk of a treatment, hindering fully informed treatment decisions. Methodologic challenges that cast doubt on the interpretability of patient-reported data within clinical studies constitute the main barrier, for example, the use of PRO measures for which evidence of adequate validation is lacking, or lack of prespecification or multiplicity adjustment.[6]

After the passage of the 21st Century Cures Act of 2016, FDA has issued guidance aimed to advance the collection of patient experience data for regulatory decision making.[7] Following this and previous guidance,[8] and the processes outlined by FDA's Clinical Outcome Assessment (COA) Qualification Program, the NSCLC Symptom Assessment Questionnaire (NSCLC-SAQ) was developed by the PRO Consortium's NSCLC Working Group for use as an endpoint measure in clinical trials to assess patient-reported NSCLC symptom severity. The PRO Consortium was formed by Critical Path Institute, FDA, and the pharmaceutical industry to support precompetitive, multistakeholder collaboration to advance the FDA qualification of COAs for use in clinical trials in which COA-based endpoints could be used to support product labeling claims (www.c-path.org). Development of the NSCLC-SAQ included qualitative research[2] and cross-sectional quantitative research.[9] However, evidence of longitudinal measurement properties including the interpretation of clinically meaningful within-person change in score is lacking. This evidence is needed to support researchers when considering the appropriateness of including the NSCLC-SAQ as an endpoint measure in clinical trials.

The objectives of this study were to assess cross-sectional and longitudinal measurement properties of the NSCLC-SAQ, including responsiveness and assessment of clinically meaningful within-person change in score, using data from a phase 3 clinical study within the measure's context of use.

Materials and Methods

Data Source

Prespecified analyses were conducted using the interim analysis 1 data cutoff of KEYNOTE-598—a randomized, double-blind, placebo-controlled phase 3 study to evaluate the efficacy of pembrolizumab plus ipilimumab versus pembrolizumab plus placebo for first-line treatment of participants with stage IV, metastatic NSCLC whose tumors are programmed death-ligand 1–positive (tumor proportion score $\geq$50%) and lack EGFR-sensitizing mutations and *ALK* translocations.

Eligible participants were 18 years of age or older, with no previous systemic therapy for metastatic NSCLC, had an Eastern Cooperative Oncology Group (ECOG) performance status score of 0 or 1, and had at least one lesion measurable per Response Evaluation Criteria in Solid Tumors, version 1.1.

The study was conducted in accordance with Good Clinical Practice and all study materials were approved by the appropriate ethics body at each participating center. Further details and findings from KEYNOTE-598 have been reported elsewhere.[10]

PRO Assessments

Under the supervision of trained site personnel, participants completed all PRO measures on tablet computers before all other study procedures at baseline, at each 3-weekly treatment cycle until week 18 (cycle 7), and subsequently on a less frequent basis.

NSCLC-SAQ consists of seven items assessing 5 NSCLC symptom concepts: cough, pain, dyspnea, fatigue, and poor appetite.[2,11] All items have a recall period of past 7 days and a 5-point response scale ranging from 0: "No <symptom> at All" to 4: "Very severe <symptom>" or from 0: "Never" to 4: "Always" to measure attributes of symptom intensity or frequency, respectively. Two

pain items form a single "Pain" domain representing the most severe response of the two items, and two fatigue items form a single "Fatigue" domain by calculating their mean. The NSCLC-SAQ total score is computed as the sum of the 5 domains and ranges between 0 and 20. Higher scores indicate more severe symptomatology.

Other PRO measures used included the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire core 30 (QLQ-C30) with the Lung Cancer Module (QLQ-LC13), EQ-5D-5L, patient global impression of severity-lung cancer symptoms (PGIS-LC), and patient global impression of change-lung cancer symptoms (PGIC-LC). After FDA feedback during the study, modifications were made to the PGIS-LC and PGIC-LC resulting in new versions administered to participants at subsequent visits. PGIS-LC version 2.0 was a single item "How would you rate your symptoms of your lung cancer at this time?" with five response options from "No symptoms" to "very severe"; PGIS-LC version 1.0 included similar but not identical response options. PGIC-LC version 2.0 was a single item "Compared to your first study visit, how would you describe the symptoms of your lung cancer today?" with seven response options from "much better" to "much worse"; PGIC-LC version 1.0 was not used in these analyses as the recall period specified "last study visit," thus preventing comparison to baseline. PGIS-LC and PGIC-LC were completed up to week 18 only.

Analyses

Analyses were conducted on blinded pooled treatment arms data of interim analysis 1 data cutoff using the PRO full analysis set (PROFAS), defined as all participants who had at least 1 PRO assessment available and received at least one dose of study medication.

Demographic and clinical variables at baseline were described using mean, SD, median, minimum, and maximum for quantitative variables, and percentage and frequency for categorical variables.

NSCLC-SAQ compliance rate (i.e., number of participants with available scores/number of participants expected to have available scores) and completion rate (i.e., number of participants with available scores/number of participants in PROFAS) were calculated from baseline to week 18.

NSCLC-SAQ items, domains, and total score at baseline were described using mean, SD, median, first and third quartile, minimum, and maximum; item responses were described using frequencies and percentages. Mean scores were presented graphically from baseline to week 18.

Interitem correlations at baseline were assessed using Spearman rank-order correlation coefficients, ranging from 0 to 1 with higher values indicating greater monotonicity.

Internal consistency reliability at baseline and week 18 was assessed using Cronbach's α coefficients with 95% confidence intervals, ranging from 0 to 1 with values between 0.7 and 0.9 considered to represent acceptable to high reliability.[12] In addition, item-total correlations were calculated using Pearson correlations of the domain with the total score (after removing the domain from the total score).

Test-retest reliability was assessed among stable participants, defined as those with no change in PGIS-LC responses of the same version between baseline and week 3, using 2-way mixed, absolute agreement, single measure intraclass correlation coefficients (ICCs) between the two assessments.[13] Values between 0.5 and 0.9 represent moderate to good reliability.[14]

Construct validity was assessed using convergent, discriminant, and known groups validity approaches. Convergent validity, referring to how well constructs that theoretically should be related to each other are observed to be related, was assessed using Spearman correlations among the NSCLC-SAQ items, domains, and total score against concepts from the other PRO measures (QLQ-C30, QLQ-LC13, EQ-5D-5L, and the PGIS-LC) at baseline. Correlation coefficients of greater than or equal to 0.4 were considered as evidence of convergent validity[15] and greater than or equal to 0.6 considered evidence for concepts expected to be highly related. Discriminant validity, viewed as the counterpart to convergent validity, was assessed using the same method but evaluating correlations between less related concepts from the other PRO measures, hypothesized to be less strongly correlated.

Known groups validity, referring to how well the NSCLC-SAQ total score distinguishes among groups of participants known to differ in health status, was evaluated using analysis of variance in groups defined by PGIS-LC response categories, QLQ-C30 Global Health Status/Quality of Life score dichotomized by the NSCLC population norm (i.e., low <58.8 and high ≥ 58.8 health status),[16] and ECOG performance status 0 versus 1.

Responsiveness and Clinically Meaningful Within-Person Change

First, the appropriateness of both PGIS-LC and PGIC-LC anchors was assessed using Spearman correlations by correlating NSCLC-SAQ total score change from baseline with the anchors, with correlation coefficients greater than 0.3 considered desirable.[17]

Responsiveness, reflecting the ability of a measure to detect changes in groups of participants who have changed in the measured concept, was assessed between

baseline and week 18. NSCLC-SAQ total score change from baseline was evaluated within groups of participants who had a change in the PGIS-LC category (for participants who completed the same PGIS-LC version) using one-sample *t* tests to evaluate change from 0 and reporting the magnitude of change using effect sizes (ESs, i.e., mean change from baseline/$SD_{baseline}$) and standardized response means (SRMs, i.e., mean change from baseline/$SD_{change\ from\ baseline}$), and across groups using analysis of covariance adjusted for baseline score.

To generate meaningful within-person change estimates in NSCLC-SAQ total score between baseline and week 18, anchor-based methods were used and supported by distribution-based methods and visual displays of the change score using empirical cumulative distribution function (eCDF) and probability density function (PDF) curves. Anchor-based methods included parametric and non-parametric descriptive statistics of the NSCLC-SAQ total score change from baseline computed for anchor groups on the basis of PGIS-LC change from baseline to week 18 (in participants who completed the same PGIS-LC version). Distribution-based estimates included the half SD of the baseline score and standard error of measurement (SEM, i.e. $SD_{baseline} \times \sqrt{[1 - \text{test-retest reliability coefficient}]}$).

Meaningful change thresholds were evaluated separately for participants improving and worsening over time. The estimates of interest for meaningful improvement (or worsening) were a change from baseline in PGIS-LC of 1 and 2 response categories indicating improvement (or worsening). Estimates from these anchor-based analyses were considered alongside visual inspection of the eCDF curves and distribution-based estimates to provide a range of plausible meaningful within-person change thresholds.

As a prespecified sensitivity analysis, the PGIC-LC (version 2.0 only) response at week 18 was considered as an anchor, acknowledging the limitation that this nonstatic anchor would likely be susceptible to recall bias.[18]

Table 1. Demographic and Clinical Characteristics at Baseline

Parameter/Category	PROFAS (N = 560)
Age (y)	
n	560
Mean (SD)	64.0 (9.06)
Median (min-max)	65 (35-85)
Age group, n (%)	
<65 y	279 (49.8)
≥65 y	281 (50.2)
Sex, n (%)	
Male	389 (69.6)
Female	170 (30.4)
Undetermined (nondisclosed)	1 (<0.1)
Region, n (%)[a]	
East Asia	63 (11.3)
Non-East Asia	497 (88.8)
ECOG performance status, n (%)	
0	203 (36.3)
1	357 (63.8)
Histology, n (%)	
Squamous	157 (28.0)
Nonsquamous	403 (72.0)
PGIS-LC version 1.0, n (%)[b]	
Not severe	89 (25.9)
Mildly severe	100 (29.2)
Moderately severe	108 (31.5)
Very Severe	39 (11.4)
Extremely severe	7 (2.0)
Missing	217
PGIS-LC version 2.0, n (%)[b]	
No symptoms	38 (19.8)
Mild	66 (34.4)
Moderate	57 (29.7)
Severe	25 (13.0)
Very severe	6 (3.1)
Missing	368

[a]Countries included were as follows: East Asia region (Taiwan, Thailand, South Korea); Non-East Asia region (Argentina, Brazil, Canada, Chile, Colombia, France, Germany, Hungary, Ireland, Italy, Latvia, Mexico, Peru, Poland, South Africa, Spain, Turkey, Ukraine, United Kingdom, and the United States).
[b]Percentages calculated from participants in PROFAS who received this PGIS-LC version. Missing values owing to the PGIS-LC version change whereby some participants completed version 1.0 and some others version 2.0.
ECOG, Eastern Cooperative Oncology Group; max, maximum; min, minimum; PGIS-LC, Patient Global Impression of Severity-Lung Cancer Symptoms; PROFAS, patient-reported outcome full analysis set.

Results

Study Population

A total of 568 participants were randomly allocated to treatment arms from 171 sites across 24 countries and were included in the KEYNOTE-598 study, of which 560 (98.6%) were included in the PROFAS analysis set.

At baseline, participants' mean age was 64 years, most were men (69.6%), enrolled in non-East Asia study sites (88.8%), had an ECOG performance status 1 (63.8%), and nonsquamous histology (72.0%) (Table 1). Just over half of participants reported "not severe" or "mildly severe" symptoms at baseline for PGIS-LC version 1.0 (55.1%) or "no symptoms" or "mild" for PGIS-LC version 2.0 (54.2%). At baseline, only a few participants reported the highest severity of symptoms using version 1.0 ("very severe": 11.4%; "extremely severe": 2.0%) or version 2.0 ("severe": 13.0%; "very severe": 3.1%).

NSCLC-SAQ Compliance, Completion, and Distribution

NSCLC-SAQ compliance was high at baseline (95.7%) through week 18 (92.6%) (Supplementary Table 1, compliance and completion rates). As expected,

completion rates continued to decrease through week 18 (66.8%) as participants discontinued the study owing to disease progression, physician decision, adverse events, or death.

At baseline, all NSCLC-SAQ item means were similar (range: 0.86 to 1.66) with a median of 1 (i.e., "mild" or "rarely" for symptom intensity or frequency, respectively) except for the fatigue items with a median of 2 (i.e., "sometimes") (Table 2). Ceiling effects indicating the lowest symptom severity were most notable for items "pain in chest" (47.6%), "pain in areas other than chest" (39.2%), and "poor appetite" (44.2%). No floor effects indicating the highest symptom severity were observed at baseline. The NSCLC-SAQ total score at baseline seemed normally distributed (mean = 6.88, SD = 3.99, median = 6.75 [on a range of 0–20]) equivalent to an average item response of 1.4 (6.88/5) reflecting item responses between "mild" and "moderate" or between "rarely" and "sometimes" for symptom intensity or frequency, respectively.

By week 6, a trend of improvement was seen for all item scores, with continued improvement through week 18 (Fig. 1). Similarly, the NSCLC-SAQ total score revealed a trend of improvement through week 18 (mean = 4.54; SD = 3.44; median = 4.00) equivalent to an average item response of 0.9 (4.54/5) reflecting item responses between "not at all" and "mild" or between "never" and "rarely" for symptom intensity or frequency, respectively.

Reliability

Interitem Spearman correlations were generally weak to moderate in the range of 0.11 to 0.82, with the largest correlation between the two fatigue items "lack of energy" and "tire easily" (Supplementary Table 2, correlations between items and domains).

Internal consistency reliability of the NSCLC-SAQ was above the conventional threshold of 0.70 at both baseline (Cronbach's $\alpha = 0.74$, 95% confidence interval [CI]: 0.70–0.77) and week 18 (Cronbach's $\alpha = 0.78$, 95% CI: 0.74–0.81). All domains were moderately correlated with the total score at baseline (range: 0.37–0.71) and week 18 (range: 0.43–0.72) (Supplementary Table 3, correlations between domains and the total score).

Test-retest reliability of the NSCLC-SAQ between baseline and week 3 was ICC equal to 0.79 (95% CI: 0.74–0.84). This was calculated on the 210 participants who had stable overall symptoms (i.e., participants who did not have a change in PGIS-LC score between baseline and week 3), had completed the same PGIS-LC version, and had nonmissing NSCLC-SAQ data.

Construct Validity

Convergent validity was shown by moderate to high correlations of NSCLC-SAQ items, domains, and total score with concepts hypothesized to be similar from the other PRO measures and especially those with the most similar concepts (Supplementary Table 4, which describes all correlations with the hypothesized correlations of $\geq$0.4 and $\geq$0.6 indicated). Discriminant validity was supported by the observation of lower (low to moderate) correlations with the other PRO measures, which captured less similar concepts.

Known groups validity was shown by the logical trend in NSCLC-SAQ total scores among groups hypothesized to be different at baseline (all $p < 0.001$) (Table 3). Higher mean NSCLC-SAQ total scores, indicating more severe symptoms, were seen in groups defined using PGIS-LC response (higher responses indicating more severe lung cancer symptom severity), QLQ-C30 Global Health Status/Quality of Life (lower values indicating worse overall health), and ECOG performance status (higher score indicating a lower level of functionality).

Responsiveness

In the assessment of the appropriateness of the anchors, the correlation between change from baseline to week 18 for the NSCLC-SAQ total score and PGIS-LC was r equal to 0.46, supporting the adequacy of the PGIS-LC as an anchor for responsiveness analyses and determining meaningful within-person change thresholds. The correlation with PGIC-LC was weak ($r = 0.25$), and, thus, did not support the use of PGIC-LC as an anchor in these analyses.

Responsiveness of the NSCLC-SAQ total score was supported by the logical differences in change from baseline to week 18 among PGIS-LC change score groups ($p < 0.001$) (Supplementary Table 5, responsiveness analysis). Mean changes were statistically significantly different from "0" for participants improving by 1 PGIS-LC response category (mean change = −2.06; 95% CI: −2.77 to −1.36) and participants improving by greater than one PGIS-LC response category (mean change = −4.08; 95% CI: −5.01 to −3.15); ES and SRM were moderate to large. Results were inconclusive for detecting worsening owing to the small magnitude of NSCLC-SAQ total score change from baseline in the small sample (n = 38) of participants who worsened on the basis of the PGIS-LC. For this group, ES and SRM were small, and no evidence was found that the mean change was statistically significantly different from "0."

Interpretation of Meaningful Within-Person Change in Score

NSCLC-SAQ change scores from baseline to week 18 by PGIS-LC anchor groups are presented in Table 4. Within the group of 68 participants who improved by 1

Table 2. Distribution of NSCLC-SAQ Items, Domains, and Total Score at Baseline

Items, Domains, Total Score	Response Category, n (%)					Mean (SD)	Median (Q1-Q3)	Minimum-maximum
	0, n (%)	1, n (%)	2, n (%)	3, n (%)	4, n (%)			
Item 1. Cough	117 (21.8)	210 (39.2)	142 (26.5)	49 (9.1)	18 (3.4)	1.33 (1.02)	1.00 (1.00-2.00)	
Item 2. Pain in chest	255 (47.6)	147 (27.4)	97 (18.1)	27 (5.0)	10 (1.9)	0.86 (1.00)	1.00 (0.00-1.50)	
Item 3. Pain in areas other than chest	210 (39.2)	155 (28.9)	103 (19.2)	51 (9.5)	17 (3.2)	1.09 (1.12)	1.00 (0.00-2.00)	
Item 4. Shortness of breath	159 (29.7)	146 (27.2)	129 (24.1)	77 (14.4)	25 (4.7)	1.37 (1.18)	1.00 (0.00-2.00)	
Item 5. Lack of energy	113 (21.1)	143 (26.7)	158 (29.5)	95 (17.7)	27 (5.0)	1.59 (1.15)	2.00 (1.00-2.00)	
Item 6. Tire easily	98 (18.3)	147 (27.4)	154 (28.7)	111 (20.7)	26 (4.9)	1.66 (1.14)	2.00 (1.00-3.00)	
Item 7. Poor appetite	237 (44.2)	99 (18.5)	100 (18.7)	63 (11.8)	37 (6.9)	1.19 (1.30)	1.00 (0.00-2.00)	
Derived pain domain[a]						1.36 (1.11)	1.00 (0.00-2.00)	0.00-4.00
Derived fatigue domain[a]						1.63 (1.09)	1.50 (1.00-2.50)	0.00-4.00
Total score[b]						6.88 (3.99)	6.75 (3.50-9.50)	0.00-18.00

Note: N equals 536 (response options for items 1 to 3: 0 = not at all; 1 = mild; 2 = moderate; 3 = severe; 4 = very severe; response options for items 4 to 7: 0 = never; 1 = rarely; 2 = sometimes; 3 = often; 4 = always).
[a]Derived domains are presented for comprehensiveness and are not intended to be used as an endpoint measure in clinical trials.
[b]Total score ranges from 0 to 20 with higher scores indicating more severe NSCLC-related symptomatology.
NSCLC-SAQ, NSCLC Symptom Assessment Questionnaire; Q1, first quartile; Q3, third quartile.

PGIS-LC category, that is, the smallest amount of change, the mean change in NSCLC-SAQ total score was −2.65 (95% CI: −3.52 to −1.78). Half of the participants in this group had changed by at least −2.25 (median) and 75% had either improved or remained unchanged (third quartile = 0.00). Within the group of 36 participants who improved by 2 PGIS-LC categories (on the PGIS-LC possible range of 1–5), the mean change in NSCLC-SAQ total score was −5.90 (95% CI: −7.10 to −4.71). Half of the participants in this group had changed by at least −5.50 (median) and 75% had improved by at least −3.75 (third quartile). Combining both of these groups—that is, participants who improved by 1 or 2 PGIS-LC categories—the mean change in NSCLC-SAQ

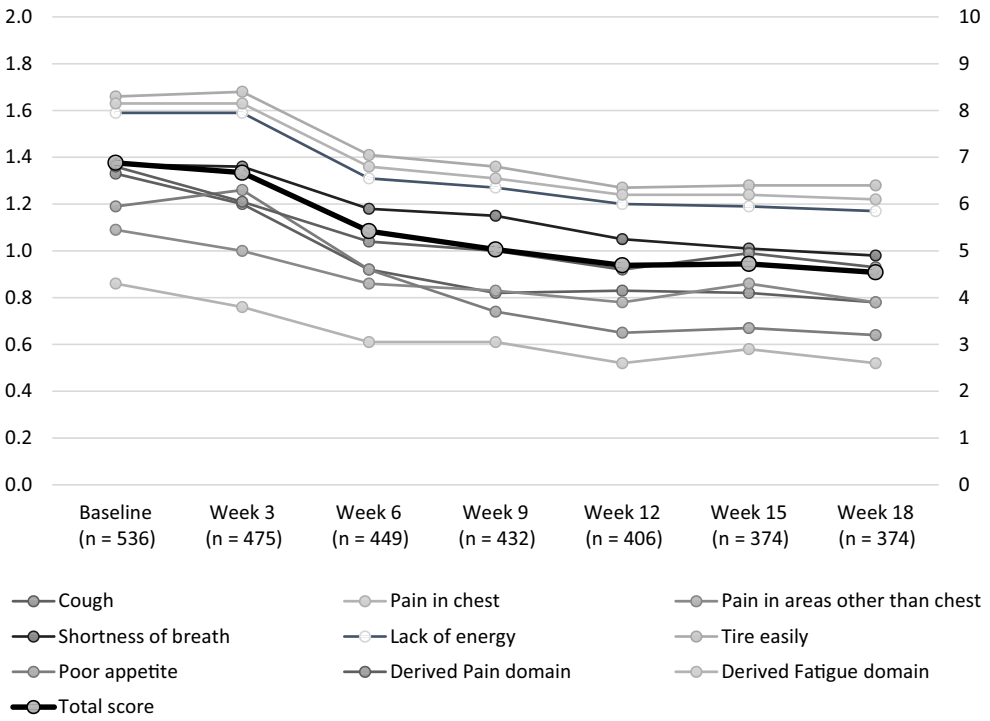

Figure 1. NSCLC-SAQ domain and total score means from baseline through week 18. Total score on the secondary axis. The y axes have been truncated for parsimony, domain scores possible range (0-4), total score possible range (0-20). NSCLC-SAQ, NSCLC Symptom Assessment Questionnaire.

Table 3. Known Groups Validity for NSCLC-SAQ Total Score at Baseline

Known Group Classification	*n*	LSmean (SE)[a]	95% CI	*p* Value
PGIS-LC version 1.0				<0.0001
Not severe	89	4.03 (0.36)	3.31-4.74	
Mildly severe	100	5.48 (0.34)	4.80-6.15	
Moderately severe	108	8.31 (0.33)	7.66-8.95	
Very severe/extremely severe[b]	46	10.70 (0.50)	9.70-11.69	
PGIS-LC version 2.0				
No symptoms	38	4.43 (0.46)	3.52-5.35	<0.0001
Mild	66	5.83 (0.35)	5.13-6.52	
Moderate	57	8.04 (0.38)	7.29-8.78	
Severe/very severe[b]	31	12.06 (0.51)	11.05-13.08	
EORTC QLQ-C30 GHS/QoL[c]				
<58.8	209	9.39 (3.76)	8.88-9.90	<0.0001
≥58.8	326	5.25 (3.24)	4.90-5.61	
ECOG performance status[d]				
0	191	5.74 (3.87)	5.19-6.29	<0.0001
1	345	7.50 (3.93)	7.09-7.92	

[a]LS means were calculated using ANOVA Welch F-test, with $\alpha = 0.05$ level.
[b]Categories were combined owing to low cell counts.
[c]Population reference value for NSCLC from EORTC QLQ-C30 reference values.[16]
[d]ECOG performance status 0 means normal activity (fully active, able to carry on all predisease performance without restriction) whereas 1 means with symptoms but ambulatory (restricted in physically strenuous activity, but ambulatory and able to carry out work of a light or sedentary nature [e.g., light housework, office work]).
CI, confidence interval; ECOG, Eastern Cooperative Oncology Group; EORTC QLQ-C30 GHS/QoL, European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire core 30 Global Health Status/Quality of Life scale; LSmean, least square mean; NSCLC-SAQ, NSCLC Symptom Assessment Questionnaire; PGIS-LC, patient global impression of severity-lung cancer symptoms.

total score was −3.77 (95% CI: −4.53 to −3.02) and the median and third quartile were −3.75 and −1.00, respectively.

It is reasonable to consider that the smallest estimates derived from the 1 PGIS-LC response category improvement group (e.g., −2.25, −2.65) and the largest estimates from the 2 PGIS-LC response category improvement group (e.g., −5.50, −5.90) represent values too extreme for a meaningful within-person change threshold; therefore, a logical threshold most likely lies inside these estimates. Considering further that the meaningful within-person change threshold needs to be an attainable number (i.e., an increment of 0.5 on the 0–20 range of the NSCLC-SAQ total score), these estimates converge toward a score change of three to five points. This range is larger than the estimates derived from distribution-based methods ($^1/_2$ SD = 2.00; SEM = 1.82), which indicates the magnitude of the change in score that is likely to be larger than measurement error. This range is also larger than the distribution of change scores in participants who had no change in PGIS-LC (mean = −1.28, 95% CI: −1.99 to −0.57), which was small and within the distribution-based estimates.

The eCDF curves (Fig. 2) for the NSCLC-SAQ total score by PGIS-LC groups revealed the expected negative change scores for improvement groups, and clear differentiation was evidenced by the consistent separation between curves throughout the range of change scores. PDF curves revealed a similar clear separation among improved, stable, and worsened groups (see Fig. in Supplementary Table 2, NSCLC-SAQ total score PDF curve).

For participants with any amount of worsening in PGIS-LC response category between baseline and week 18, the sample size was small (n = 38) with a large variance around the NSCLC-SAQ total change score as observed by the 95% confidence interval including "0" (95% CI: −0.12 to 2.12), thus, limiting inference in this group. For participants with any worsening in PGIS-LC, a mean increase (i.e., worsening) in NSCLC-SAQ total score of 1.00 was observed (median = 1.00).

Discussion

The NSCLC-SAQ has been developed in accordance with evidentiary expectations detailed in regulatory guidance and the rigorous FDA COA Qualification Program, which led to its qualification in 2018 for use in drug development as a measure of NSCLC symptom severity.[19] To inform meaningful endpoint construction, evaluation of the NSCLC-SAQ in a longitudinal setting was needed and had not been previously reported.

The novel and central contribution of this work were to assess the responsiveness and clinically meaningful within-person change in score for the NSCLC-SAQ within the randomized, double-blind, placebo-controlled phase 3 KEYNOTE-598 study. Results

Table 4. NSCLC-SAQ Total Score Change From Baseline to Week 18 by PGIS-LC Anchor Groups

PGIS-LC Anchor Response Category	n[a]	Distribution of NSCLC-SAQ Change Score[b]	
		Mean (95% CI)	Median (Q1-Q3)
All change response categories			
<−2 (improvement)	5	−3.70 (−8.38 to 0.98)	−2.00 (−6.00 to −2.00)
−2	36	−5.90 (−7.10 to −4.71)	−5.50 (−8.50 to −3.75)
−1	68	−2.65 (−3.52 to −1.78)	−2.25 (−5.00 to 0.00)
0 (no change)	108	−1.28 (−1.99 to −0.57)	−1.00 (−3.00 to 1.00)
1	31	1.05 (−0.28 to 2.38)	1.00 (−1.50 to 3.00)
2	7	0.79 (−1.54 to 3.11)	1.00 (−1.00 to 2.00)
>2 (worsening)	0	(NC)	(NC)
1 or 2 response category improvement/worsening			
−2 or −1 (improvement)	104	−3.77 (−4.53 to −3.02)	−3.75 (−6.50 to −1.00)
1 or 2 (worsening)	38	1.00 (−0.12 to 2.12)	1.00 (−1.50 to 3.00)
Any improvement/ worsening			
≤−1 (improvement)	109	−3.77 (−4.50 to −3.04)	−3.50 (−6.50 to −1.00)
≥+1 (worsening)	38	1.00 (−0.12 to 2.12)	1.00 (−1.50 to 3.00)

[a]The n describes the number of participants from the PROFAS population who completed the same version of PGIS-LC at baseline and week 18.
[b]Change scores are calculated as week 18 score minus baseline score.
CI, confidence interval; NC, not calculated; NSCLC-SAQ, NSCLC Symptom Assessment Questionnaire; PGIS-LC, patient global impression of severity-lung cancer symptoms; Q1, first quartile; Q3, third quartile.

of the clinically meaningful within-person improvement analysis support an estimated range of 3 – 5 points change from baseline in NSCLC-SAQ total score, about 15% to 25% of the 0 to 20 possible score range. This threshold was identified using a sufficiently correlated anchor variable and by comparing estimates from different anchor categories, providing a basis for researchers considering the NSCLC-SAQ total score for an endpoint evaluating improvement in NSCLC symptom severity.

Researchers have suggested that differences may exist between what participants perceive as meaningful in terms of deterioration versus improvement.[20] In our study, insufficient evidence was obtained to support a clear meaningful within-person change threshold for symptom worsening as only a small subset of participants exhibited symptom worsening up to week 18 (n = 38). Although a threshold greater than the distribution-based method estimates (two-points) may provide a starting point, it is recommended that future studies estimate meaningful within-person change for participants whose symptom severity worsens over time.

Responsiveness of the NSCLC-SAQ total score was exhibited by detectable differences in NSCLC-SAQ total score change observed over the 18-week period among participants categorized by their global impression of the

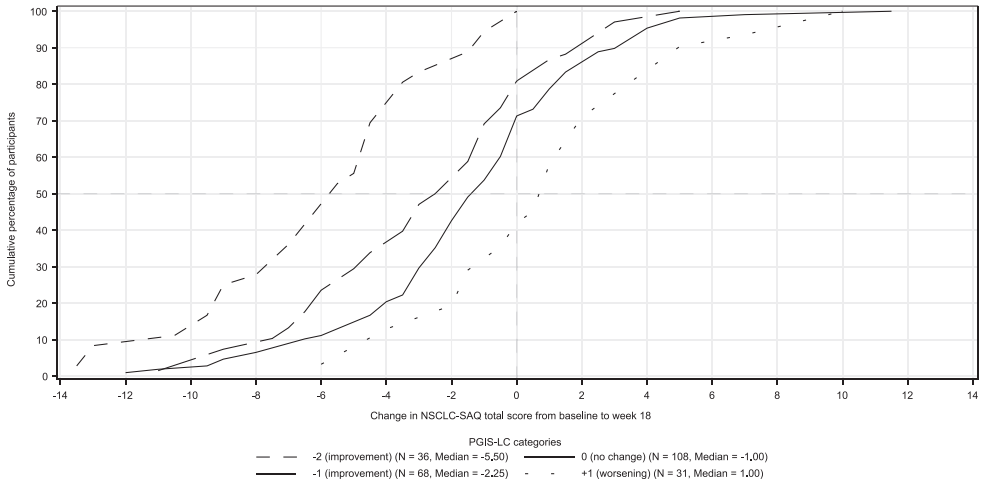

Figure 2. Empirical cumulative distribution function curves for NSCLC-SAQ total score change from baseline to week 18 by PGIS-LC anchor categories; the change in score from baseline to week 18 equals score at week 18 minus the baseline score. Curves toward the left of 0 (negative NSCLC-SAQ total score change) indicate improvement in symptom severity. PGIS change groups with low cell counts (n <10) are not provided for parsimony. NSCLC-SAQ, NSCLC Symptom Assessment Questionnaire; PGIS-LC, patient global impression of severity-lung cancer symptoms.

severity of their lung cancer symptoms. Score changes differed statistically significantly from "0" for participants who improved (i.e., mean change for those with >1 category improvement = −4.08, 95% CI: −5.01 to −3.15), and by a larger amount than for participants who stayed at the same severity on the PGIS-LC (mean change = −1.94, 95% CI: −2.51 to −1.38). For participants who worsened, a small statistically nonsignificant change was observed (mean change = 0.16, 95% CI: −0.78 to 1.10). Again, this was likely owing to the small subset of participants whose symptom severity worsened over the study period.

Participants in KEYNOTE-598 reported a low level of NSCLC symptom severity at baseline as seen by low item means and ceiling effects indicating "not at all" or "never" experiencing certain symptoms in the preceding 7 days. This was similar to previous findings reported elsewhere.[9] Furthermore, the overall symptom severity measured by the NSCLC-SAQ total score was toward the lower end of the score range (mean at baseline = 6.88, SD = 3.99), which revealed a trend of improvement by week 18 (mean at week 18 = 4.54, SD = 3.44).

This study also evaluated the cross-sectional measurement properties and test-retest reliability of the NSCLC-SAQ, and our findings were similar to those previously reported.[9] Specifically, the interitem correlation between the two fatigue items was large ($r = 0.82$), thus, confirming the close relationship between participant reports of "tiring easily" and having "lack of energy" and providing support for their averaging in the derived fatigue domain. The NSCLC-SAQ total score revealed acceptable internal consistency reliability (i.e., Cronbach's α at baseline = 0.74 and at week 18 = 0.78) and test-retest reliability (i.e., ICC = 0.79). Although test-retest reliability was slightly lower than previously reported (cf., ICC = 0.87), this may be a consequence of the longer assessment period in our study (3 weeks at the first treatment cycle after baseline) compared with the 7- to 10-day period previously reported. Finally, construct validity was well revealed by consistent patterns of moderate to high correlations with similar concepts captured by the other PRO measures and by the magnitude of differences between groups expected to differ in NSCLC symptom severity.

Strengths of this study include the ability to observe a change in NSCLC symptom severity by means of a blinded interventional study in a multinational setting, with PRO measures administered electronically at study sites before all other study procedures, thus minimizing the possibility of experimental processes influencing participants' responses. In addition, it is noteworthy that the performance of the measure's responsiveness to improvement, including the ability to quantify meaningful within-person change, was observed despite a sample with generally low symptom severity at baseline. This may be of particular interest to researchers considering using this measure in the first-line setting.

The main limitation of this study was the inclusion of relatively few participants with high symptom severity at baseline, and the small subset of participants who exhibited symptom worsening up to week 18. As a result, it was not possible to support a clear within-person meaningful change threshold for symptom worsening. Inclusion of participants with greater functional limitations (e.g., ECOG performance status 2) or investigating symptom worsening in participants with progressive disease may be a way to investigate this further. Another limitation included the version change in PGIS-LC and PGIC-LC during the study. To address this, participants completing different versions were excluded from the analyses assessing change in score. In addition, the weak correlation between the NSCLC-SAQ change score and the PGIC-LC at week 18 suggested an inadequacy of the PGIC-LC to detect change over the 18-week interval. This aligns with previously reported recommendations that nonstatic anchors (i.e., requiring participants to recall their change, often over long periods) may be more susceptible to recall bias,[18] and researchers should consider this carefully for future studies.

In conclusion, this study provides a rationale for consideration of the NSCLC-SAQ total score to assess NSCLC symptom severity as an endpoint measure in clinical trials, including trials involving participants who are mildly symptomatic at baseline. It is recommended that future studies be conducted to better understand the performance of the NSCLC-SAQ in trials enrolling participants likely to have greater lung cancer symptom severity at baseline or more likely to experience symptom worsening over time, such as those who have failed previous line(s) of therapy for their metastatic disease or ECOG performance status 2.

CRediT Authorship Contribution Statement

Paul Williams: Methodology, Visualization, Writing - review & editing.

Thomas Burke: Conceptualization, Methodology, Supervision, Writing – review & editing.

Josephine M. Norquist: Conceptualization, Methodology, Writing – review & editing.

Christina Daskalopoulou: Methodology, Formal analysis, Writing – original draft.

Rebecca M. Speck: Writing – review & editing.

Ayman Samkari: Conceptualization, Investigation.

Sonya Eremenco, Stephen Joel Coons: Writing – review & editing.

Supplementary Data

Note: To access the supplementary material accompanying this article, visit the online version of the *JTO Clinical and Research Reports* at www.jtocrr.org and at https://doi.org/10.1016/j.jtocrr.2022.100298.

Acknowledgments

The preparation of this article was supported by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, New Jersey, who sponsored the KEYNOTE-598 clinical trial and the analyses reported here. The sponsor was involved with the study design, collection, and interpretation of data, in the writing of the report and article, and in the decision to submit the article for publication. Critical Path Institute is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) and is 54.2% funded by the FDA / HHS, totaling $13,239,950, and 45.8% funded by nongovernment sources, totaling $11,196,634. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS or the U.S. Government. The authors thank Cristina Ivanescu (IQVIA, The Netherlands) for support with the analysis of data for this study.

References

1. Division of Cancer Prevention and Control, Centers for Disease Control and Prevention. Lung cancer statistics 2020. https://www.cdc.gov/cancer/lung/statistics/index.htm. Accessed March 8, 2022.
2. McCarrier KP, Atkinson TM, DeBusk KP, et al. Qualitative development and content validity of the Non-Small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ), a patient-reported outcome instrument. *Clin Ther*. 2016;38:794–810.
3. Bridges JF, Mohamed AF, Finnern HW, Woehl A, Hauber AB. Patients' preferences for treatment outcomes for advanced non-small cell lung cancer: a conjoint analysis. *Lung Cancer*. 2012;77:224–231.
4. American Cancer Society. Treatment choices for non-small cell lung cancer, by stage. https://www.cancer.org/cancer/lung-cancer/treating-non-small-cell/. Accessed March 8, 2022.
5. Gnanasakthy A, Barrett A, Evans E, D'Alessio D, Romano CD. A review of patient-reported outcomes labeling for oncology drugs approved by the FDA and the EMA (2012-2016). *Value Health*. 2019;22:203–209.
6. Basch E, Geoghegan C, Coons SJ, et al. Patient-reported outcomes in cancer drug development and US regulatory review: perspectives from industry, the Food and Drug Administration, and the patient. *JAMA Oncol*. 2015;1:375–379.
7. Food and Drug Administration (FDA). Plan for issuance of patient-focused drug development guidance. https://www.fda.gov/files/about%20fda/published/Plan-for-Issuance-of-Patient%E2%80%90Focused-Drug-Development-Guidance.pdf. Accessed March 8, 2022.
8. Food and Drug Administration (FDA). Guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims. Accessed March 8, 2022.
9. Bushnell DM, Atkinson TM, McCarrier KP, et al. Non-small cell lung cancer symptom assessment questionnaire: psychometric performance and regulatory qualification of a novel patient-reported symptom measure. *Curr Ther Res Clin Exp*. 2021;95:100642.
10. Boyer M, Sendur MAN, Rodriguez-Abreu D, et al. Pembrolizumab plus ipilimumab or placebo for metastatic non-small-cell lung cancer with PD-L1 tumor proportion score $\geq$ 50%: randomized, double-blind phase III KEYNOTE-598 study. *J Clin Oncol*. 2021;39:2327–2338.
11. *Critical Path Institute*. Non-small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ) User Manual. Version 1.0: January 16, 2018. Critical Path Institute. Accessed March 8, 2022.
12. Nunnally JC, Bernstein IH. *Psychometric Theory*. 3rd ed. New York, NY: McGraw-Hill; 1994.
13. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428.
14. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163.
15. Messick S. Standards of validity and the validity of standards in performance Asessment. *Educ Meas Issues Pract*. 1995;14:5–8.
16. Scott NW, Fayers PM, Aaronson NK, et al. EORTC QLQ-C30 reference values. EORTC Quality of Life Group. https://www.eortc.org/app/uploads/sites/2/2018/02/reference_values_manual2008.pdf. Accessed March 8, 2022.
17. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61:102–109.
18. Food and Drug Administration (FDA). Patient-focused drug development guidance public workshop. methods to identify what is important to patients & select, develop or modify fit-for-purpose clinical outcomes assessments. https://www.fda.gov/drugs/news-events-human-drugs/patient-focused-drug-development-guidance-methods-identify-what-important-patients-and-select. Accessed March 8, 2022.
19. Food and Drug Administration (FDA). Qualification of non-small cell lung cancer symptom assessment questionnaire (NSCLC-SAQ) – a patient-reported outcome instrument. https://www.fda.gov/files/drugs/published/Qualifications-of-Non-Small-Cell-Lung-Cancer-Symptom-Assessment-Questionnaire-%28NSCLC-SAQ%29—A-Patient-Reported-Outcome-Instrument.pdf. Accessed March 8, 2022.
20. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Qual Life Res*. 2002;11:207–221.