



Interpreting Within-Patient Changes on the EORTC QLQ-C30 and EORTC QLQ-LC13

Cheryl D. Coon¹ · Michael Schlichting² · Xinke Zhang³

Accepted: 26 April 2022 / Published online: 30 June 2022
© The Author(s) 2022

Abstract

Introduction When determining if changes on patient-reported outcome (PRO) scores in clinical trials convey a meaningful treatment benefit, statistical significance tests alone may not communicate the patient perspective. Appraising within-patient changes on PRO scores against established thresholds can determine if improvements or deteriorations experienced by individuals are meaningful. To evaluate the appropriateness of thresholds for interpreting meaningful improvements and deterioration within individuals on the European Organisation for Research and Treatment of Cancer (EORTC) 30-item core instrument (QLQ-C30) and 13-item lung cancer module (QLQ-LC13), a series of psychometric methods were applied to data from a phase III randomized controlled clinical trial in non-small cell lung cancer.

Methods Anchor-based methods of empirical cumulative distribution functions and classification statistics were employed using change scores from Baseline to Week 7 using changes on the QLQ-C30 Global Health Status item as an anchor. Distribution-based methods of one-half standard deviation and standard error of measurement identified the minimum amount of change each domain score can reliably measure.

Results While the correlations between the domain scores and the anchor item were modest in size (i.e., $r \geq 0.30$ for only 5 of 24 domains), consideration of multiple methods along with the magnitude of possible step changes on the score allowed for patterns to emerge. The triangulation process planned *a priori* resulted in different methods being the source for different domain scores. Absolute values of the proposed thresholds ranged from 11.11 to 33.33, and all resulted in the same classifications for all EORTC domains, except QLQ-C30 Fatigue, as would the 10-point threshold that is traditionally used.

Conclusion This study confirms the appropriateness of the 10-point EORTC score threshold generally used by the field for interpreting within-patient changes, but the thresholds proposed from this study enhance interpretability by corresponding to only observable locations along the domain score scale.

✉ Cheryl D. Coon
ccoone@outcometrix.com

¹ Outcometrix, Salida, CO, USA

² Merck Healthcare KGaA, Darmstadt, Germany

³ EMD Serono Research & Development Institute, Inc. (an affiliate of Merck KGaA), Billerica, MA, USA

Key Points for Decision Makers

This study corroborates the appropriateness of 10-point thresholds for identifying meaningful within-patient change on European Organisation for Research and Treatment of Cancer (EORTC) scores in lung cancer.

By accounting for the amount of change that is possible on each score, proposed thresholds for interpreting change on EORTC scores in lung cancer range from 11.11 to 33.33.

To enhance ease of interpretation, researchers should consider applying the thresholds proposed here for identifying individuals who have responded to treatment or who have experienced deterioration in future lung cancer studies.

1 Introduction

Patient-reported outcome (PRO) instruments are commonly used in oncology clinical trials to measure the effect of therapy on functioning, cancer symptoms, and adverse effects of treatment. Patient-experience data, such as those from PRO instruments, can be valuable in risk-benefit assessment, with PRO data possibly being included in drug labeling when analyses are prespecified and alpha-controlled [1, 2] (i.e., included in the clinical trial analysis plan as primary or secondary endpoints). In some cases, PRO data may be included in drug labels descriptively when the evidence is strong and the information complements clinical data, such as the Xalkori label for non-small cell lung cancer (NSCLC) describing a benefit in time to development or worsening of dyspnea [3].

In oncology, PRO instruments developed by the European Organisation for Research and Treatment of Cancer (EORTC) are commonly administered in clinical trials. This includes a 30-item core instrument (QLQ-C30) and cancer type-specific modules, such as the QLQ-LC13 developed for use in NSCLC and small cell lung cancer (SCLC) [4, 5]. When evaluating changes in EORTC scores for individuals over the course of therapy, a 10-point threshold has been used to assist regulators and health authorities interpret the meaningfulness of within-patient score changes [6]. This threshold is also often applied in time-to-deterioration analyses. Despite the widespread use of this 10-point threshold, its derivation from a mixed cancer sample (i.e., breast cancer and SCLC) on scores from only four QLQ-C30 domains in a study intended to evaluate a novel anchor of

‘subjective significance’ renders its generalized use debatable [7]. Evidence to support the use of a 10-point threshold and expand it to other EORTC scores has been generated by other researchers, but the source data still represented a mix of cancer types [8]. Score interpretation work that has been conducted specifically using patients with NSCLC has been limited to QLQ-C30 scores for the purpose of group difference interpretation (i.e., minimal important difference) rather than within-patient change [9, 10].

Recognizing the need for evidence-based, lung cancer-specific thresholds for interpreting meaningful within-patient improvement and deterioration on the QLQ-C30 and QLQ-LC13, this research applied multiple statistical approaches to a phase III randomized controlled trial dataset. While the analysis was applied to all domain scores relevant in a clinical trial setting (i.e., excluding the QLQ-C30 Financial Difficulties score), the manuscript focuses on the key domains of interest to patients, physicians, and regulators (i.e., Physical Functioning, Coughing, Dyspnea, Chest Pain) [2, 11–13], with the results for the remaining domains available in the electronic supplementary files.

2 Methods

2.1 Data Source

Data used for these psychometric analyses came from a multicenter, international, randomized, open-label, phase III trial comparing a programmed death-ligand 1 (PD-L1) immune checkpoint inhibitor with a taxane chemotherapy (NCT02395172). Participants were aged 18 years or older with histologically confirmed stage IIIB, IV, or recurrent NSCLC that progressed after previous platinum doublet therapy. Additional inclusion and exclusion criteria and details about the study design are available in the study by Barlesi et al. [14]. Participants were randomized in a 1:1 ratio, stratified by PD-L1 assay status and NSCLC histology, to receive either the PD-L1 immune checkpoint inhibitor once every 2 weeks or the taxane chemotherapy once every 3 weeks.

PRO questionnaires were completed by participants during site visits using an electronic tablet or site pad at screening, at randomization prior to first administration of the trial treatment, and at selected study visits prior to administration of trial treatment and any trial-related procedures. PRO assessments corresponded to in-clinic treatment visits, resulting in different assessment time points between the two treatment arms. For the checkpoint inhibitor arm, post-randomization PRO assessments occurred at Weeks 1, 3, 7, and 13, and thereafter every 6 weeks while receiving treatment. For the chemotherapy arm, post-randomization

PRO assessments occurred at Weeks 1, 4, 7, and 13, and thereafter every 6 weeks while receiving treatment.

While change from Baseline to Week 13 was planned for these analyses for comparability with PRO change score analyses conducted on other checkpoint inhibitors [15, 16], the sample size of participants experiencing meaningful change on the anchor item at Week 13 (i.e., meaningful improvement, $n = 24$; meaningful deterioration, $n = 29$) did not meet the a priori threshold of $n \geq 30$. Instead, change from Baseline to Week 7 was used (i.e., meaningful improvement, $n = 39$; meaningful deterioration, $n = 43$). Baseline was defined as the last measurement prior to randomization or, if missing, the last measurement prior to the first trial drug administration.

Analysis was conducted on participants who were randomized to study treatment with evaluable PRO data (i.e., completed one baseline PRO assessment and at least one post-baseline PRO questionnaire). Analysis was pooled across treatment groups.

2.2 Patient-Reported Outcome Instruments

The EORTC QLQ-C30 assesses general aspects of health-related quality of life (QoL) of patients with cancer, including five functional scales (i.e., Physical, Role, Cognitive, Emotional, Social), three symptom scales (i.e., Fatigue, Pain, Nausea and Vomiting), a Global Health Status/QoL scale, a number of single items assessing additional common symptoms of cancer (i.e., Dyspnea, Loss of Appetite, Insomnia, Constipation, Diarrhea), and a single item measuring the financial impact of disease. Most items are reported on a 1-to-4 verbal response scale with response options of 'not at all', 'a little', 'quite a bit', and 'very much', while the two Global Health Status/QoL items are reported on a 1-to-7 numeric response scale anchored at the ends with 'very poor' and 'excellent'. All responses are linearly transformed to produce domain scores on a 0-to-100 range, where higher scores represent a greater amount of the concept being measured.

The EORTC QLQ-LC13 comprises 13 questions that measure lung cancer-related symptoms (i.e., Coughing, Hemoptysis, Dyspnea) and treatment-related adverse effects (i.e., Sore Mouth or Tongue, Dysphagia, Hair Loss, Tingling Hands or Feet, Chest Pain, Arm or Shoulder Pain, Other Pain, Usefulness of Pain Medication). The QLQ-LC13 items use the same 1-to-4 verbal response scale as the QLQ-C30 items, and domain scores are also transformed to a 0-to-100 metric.

The primary method by which score interpretation is determined relies on external criteria as anchors to classify participants according to the degree of change in the concept being measured. While an anchor that measures the concept being measured by each PRO score is desirable, in this

study concept-specific anchors were not available. Instead, the Global Health Status item from the QLQ-C30 was used to define change in participants' overall health. This item has been previously used as an anchor, where a two-category increase indicated improvement and a two-category decrease indicated deterioration [17], and these two levels of improvement and deterioration were chosen as the 'target' anchor categories for defining meaningful change for this study.

2.3 Psychometric Methods

Prior to conducting anchor-based analyses, the strength of the relationship between change on the anchor and change on each EORTC score was evaluated to ensure that the anchor is sufficiently related to the instrument. Ideally, correlations should be at least 0.3 to indicate a viable anchor, but results can be interpreted cautiously when the anchor correlation does not reach that threshold [18].

The primary anchor-based approach was the presentation of empirical cumulative distribution functions (ECDFs) [19]. ECDFs plotted the cumulative proportion of participants reporting change scores at each level of the EORTC score range; separate curves were presented for each level of change reported on the Global Health Status item. The location at which the curve corresponding to the target anchor category (i.e., two-category increase for improvement, two-category decrease for deterioration) reaches 50% on the y-axis is a possible location for an interpretable threshold, provided that this curve is distinct from the adjacent category curves at this location [20]. ECDFs considered each level of change on the anchor as a separate curve so that the distance between anchor categories could be compared, but categories towards the tails of the distribution above and below the target anchor categories were collapsed because sample sizes were below 10.

A supportive anchor-based approach examined a series of classification statistics (i.e., sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV]) along each point of the EORTC score scale. These statistics were reported separately for predicting improvement (i.e., two-category or greater increase) and deterioration (i.e., two-category or greater decrease) on each EORTC score. The location where sensitivity and specificity are simultaneously at their highest levels is one possible threshold location, as is the location where PPV and NPV are simultaneously at their highest levels [20].

A supplementary analysis relied on distribution-based methods. While unable to link the meaning of estimates back to the patient perspective as recommended [1], distribution-based methods ensure that proposed thresholds are sufficiently large given the variability of the scores and the reliability of the instrument. The one-half standard deviation (SD) method used data from Baseline to identify the amount

of change on an instrument that would equate to a moderate amount of change [21]. The standard error of measurement (SEM) adjusted the magnitude of the one-half SD estimate for the reliability of the scale (i.e., more reliable scores can measure smaller amounts of change) [22]. The SEM reliability adjustment for each domain was based on the test–retest reliability between the screening and Week 1 administrations for stable participants on the Global Health Status item [23].

Thresholds for identifying improvement and deterioration were proposed by triangulating the results of the anchor-based and distribution-based methods. ECDFs were given preference over classification statistics, assuming sensitivity and specificity were adequate. The chosen anchor-based estimate must also have exceeded the SEM, ensuring that the proposed amount of change can be reliably measured by the scale. The proposed thresholds were considered against the minimum amount of change that can be observed on each domain given the number of items and their four-category response scale (Online Resource 1).

3 Results

3.1 Descriptive Statistics

Data were available for 681 participants randomized to study treatment with evaluable PRO data. Select baseline characteristics of the sample are presented in Table 1. The analysis sample was predominantly male (69%), white (64%), and spread across geographic regions of the world. The average age was 63.1 years, ranging from 29 to 89 years. The majority of participants in the analysis sample had Eastern Cooperative Oncology Group scores of 1 (64%) and non-squamous histology of disease (69%).

Scores at Baseline and Week 7 and the change from Baseline to Week 7 for key domains and the anchor item are presented in Table 2 for participants with evaluable scores at both time points; scores for the remaining domains are reported in Online Resource 2. Although the EORTC scores showed minimal mean change from Baseline to Week 7, the large SDs and wide range of change scores reflected some improvement and deterioration in the sample. For the Global Health Status anchor item, Baseline scores were towards the middle of the score range (mean = 66.0), with responses observed across the whole scale range. While the mean score

Table 1 Descriptive statistics for demographic and clinical characteristics for the study cohort as analyzed

Variable	Statistic	Baseline [<i>N</i> = 681]
Age	<i>n</i>	681
	Missing (%)	0 (0.0)
	Mean ± SD (years)	63.1 ± 9.6
	Min–max (years)	29.0–89.0
Sex	Male	468 (68.7)
	Female	213 (31.3)
Race	White	438 (64.3)
	Asian	207 (30.4)
	Black or African American	6 (0.9)
	American Indian or Alaska Native	1 (0.1)
	Native Hawaiian or other Pacific Islander	1 (0.1)
	Other	5 (0.7)
	Missing	23 (3.4)
Region	Asia	204 (30.0%)
	US and Western Europe	171 (25.1%)
	Rest of the world	161 (23.6%)
	Eastern Europe	145 (21.3%)
ECOG	0	244 (35.8)
	1	437 (64.2)
NSCLC histology	Squamous cell	209 (30.7)
	Non-squamous	472 (69.3)

Data are expressed as *n* (%) unless otherwise specified

ECOG Eastern Cooperative Oncology Group, *Min* minimum, *Max* maximum, NSCLC non-small cell lung cancer, *SD* standard deviation

Table 2 Descriptive statistics for participants with patient-reported outcome data at Baseline and Week 7 for key EORTC domain and Spearman correlations between key EORTC domain scores and the Global Health Status item for change from Baseline to Week 7

EORTC domain score	n, Missing (%), mean \pm SD, min–max			<i>r</i> (<i>p</i> -value) Between EORTC domain and Global Health Status for Δ from Baseline to Week 7
	Baseline [<i>n</i> = 483]	Week 7 [<i>n</i> = 483]	Change from Baseline to Week 7 [<i>n</i> = 483]	
EORTC QLQ-C30 Physical Functioning	483 0 (0.0) 77.8 \pm 18.4 6.7–100.0	483 0 (0.0) 74.9 \pm 20.4 6.7–100.0	483 0 (0.0) –2.9 \pm 16.5 –60.0 to 60.0	0.33 (<0.0001)
EORTC QLQ-LC13 Dyspnea	482 1 (0.2) 24.9 \pm 21.0 0.0–100.0	482 1 (0.2) 26.2 \pm 21.2 0.0–100.0	482 1 (0.2) 1.4 \pm 18.1 –66.7 to 66.7	–0.29 (<0.0001)
EORTC QLQ-LC13 Coughing	482 1 (0.2) 34.7 \pm 27.3 0.0–100.0	482 1 (0.2) 31.4 \pm 26.6 0.0–100.0	482 1 (0.2) –3.3 \pm 26.4 –100.0 to 100.0	–0.14 (0.0028)
EORTC QLQ-LC13 Pain in Chest	482 1 (0.2) 17.7 \pm 23.5 0.0–100.0	482 1 (0.2) 14.6 \pm 22.0 0.0–100.0	482 1 (0.2) –3.1 \pm 23.9 –100.0 to 66.7	–0.20 (<0.0001)
EORTC Global Health Status	482 1 (0.2) 66.0 \pm 19.9 0.0–100.0	482 1 (0.2) 65.0 \pm 19.9 0.0–100.0	482 1 (0.2) –1.0 \pm 19.5 –66.7 to 83.3	NA

EORTC European Organisation for Research and Treatment of Cancer, *min* minimum, *max* maximum, QLQ-C30 30-item core instrument, QLQ-LC13 13-item lung cancer module, SD standard deviation, NA not applicable

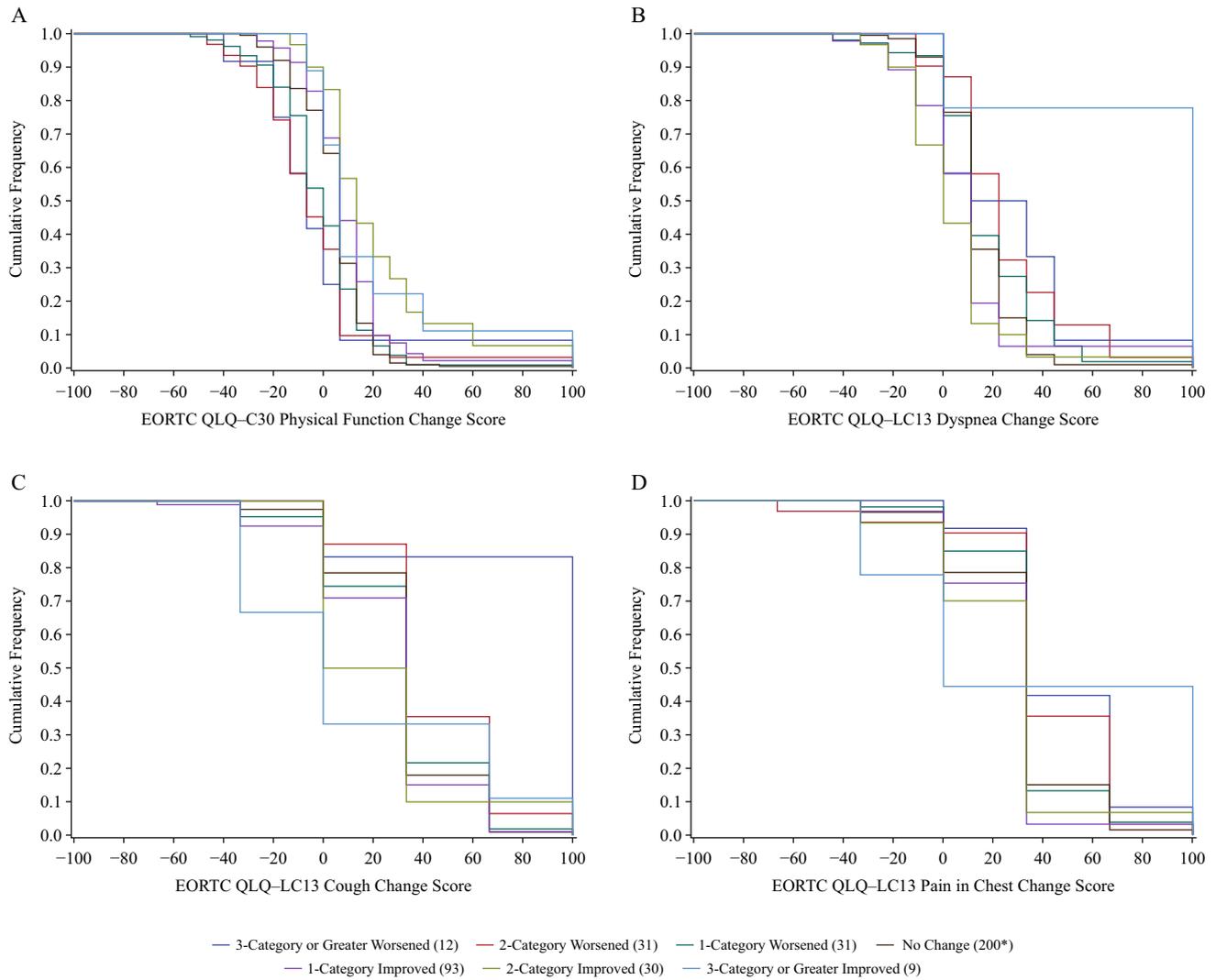
did not change at Week 7 (i.e., 65.0), changes as large as a five-category improvement (i.e., 83.3) and a four-category deterioration (i.e., –66.7) were observed.

Correlations between change scores on the key EORTC domains and the anchor item are also presented in Table 2 and Online Resource 2. The correlation for QLQ-C30 Physical Functioning ($r = 0.33$) reached the a priori threshold of 0.30, while those for QLQ-LC13 Dyspnea ($r = -0.29$), Coughing ($r = -0.14$), and Chest Pain ($r = -0.20$) did not; however, these three QLQ-LC13 score correlations all reached statistical significance ($p < 0.01$). Thus, Global Health Status is sufficiently related to the QLQ-C30 Physical Functioning and may provide some insight into the interpretation of the three QLQ-LC13 scores, but interpretation estimates for the latter may be deflated (i.e., biased towards 0) due to weak correlations.

3.2 Empirical Cumulative Distribution Functions

The ECDFs for the Global Health Status change groups (i.e., 3+ category improved through 3+ category worsened) are presented for the four key EORTC domains in Fig. 1; plots

for the remaining domains are reported in Online Resource 3. The locations where the target anchor categories (i.e., two-category improved and two-category worsened) cross 50% on the *y*-axis (i.e., median) were considered possible thresholds for identifying improvement and deterioration on the domain score. For QLQ-C30 Physical Functioning, the median for the two-category improved group was +13.33, and this line was distinct from the adjacent groups at 50%. The median for the two-category worsened group was –6.67, and this line was distinct from one-category worsened. For QLQ-LC13 Dyspnea, the median for the two-category improved group was 0, which is not a viable location because this score indicates no change. The median for the two-category worsened group was +22.22, which is in the direction of deterioration for symptom domains, and this line was distinct from the adjacent groups at 50%. For QLQ-LC13 Coughing, the two-category improved group also provided an uninterpretable estimate of 0. The two-category worsened group crossed 50% at +33.33, but this line was not distinct from one-category worsened, also limiting the suitability of this estimate. For QLQ-LC13 Chest Pain, the two-category improved group provided an



*201 for EORTC QLQ-C30 Physical Function Change Score

Fig. 1 ECDFs of change on key EORTC domain scores and change on Global Health Status from Baseline to Week 7 (anchor group sample sizes indicated in parentheses). *ECDF* empirical cumulative dis-

tribution function, *EORTC* European Organisation for Research and Treatment of Cancer, *QLQ-C30* 30-item core instrument, *QLQ-LC13* 13-item lung cancer module

uninterpretable estimate of +33.33 in the direction of deterioration, while the median for the two-category worsened group was +33.33, which was not distinct from the one-category worsened group, thus limiting the usefulness of this estimate.

3.3 Classification Statistics

The classification statistics plots for participants with meaningful improvement on Global Health Status (i.e., two-category or greater increase) are presented in Fig. 2 for the four key EORTC domains; plots for the remaining domains are reported in Online Resource 4. The location where the four classification statistics are simultaneously maximized

was considered a possible threshold for identifying improvement on the domain score. However, PPV was generally very low in the range where sensitivity and specificity were simultaneously maximized, therefore priority was given to maximizing sensitivity and specificity. This location was +6.67 for QLQ-C30 Physical Functioning, -11.11 for QLQ-LC13 Dyspnea, and -33.33 for QLQ-LC13 Coughing and QLQ-LC13 Chest Pain. The classification statistics plots for participants with meaningful deterioration on Global Health Status (i.e., two-category or greater decrease) are presented in Fig. 3 for the four key EORTC domains; plots for the remaining domains are reported in Online Resource 5. The location where sensitivity and specificity were maximized when considering deterioration was -6.67 for QLQ-C30

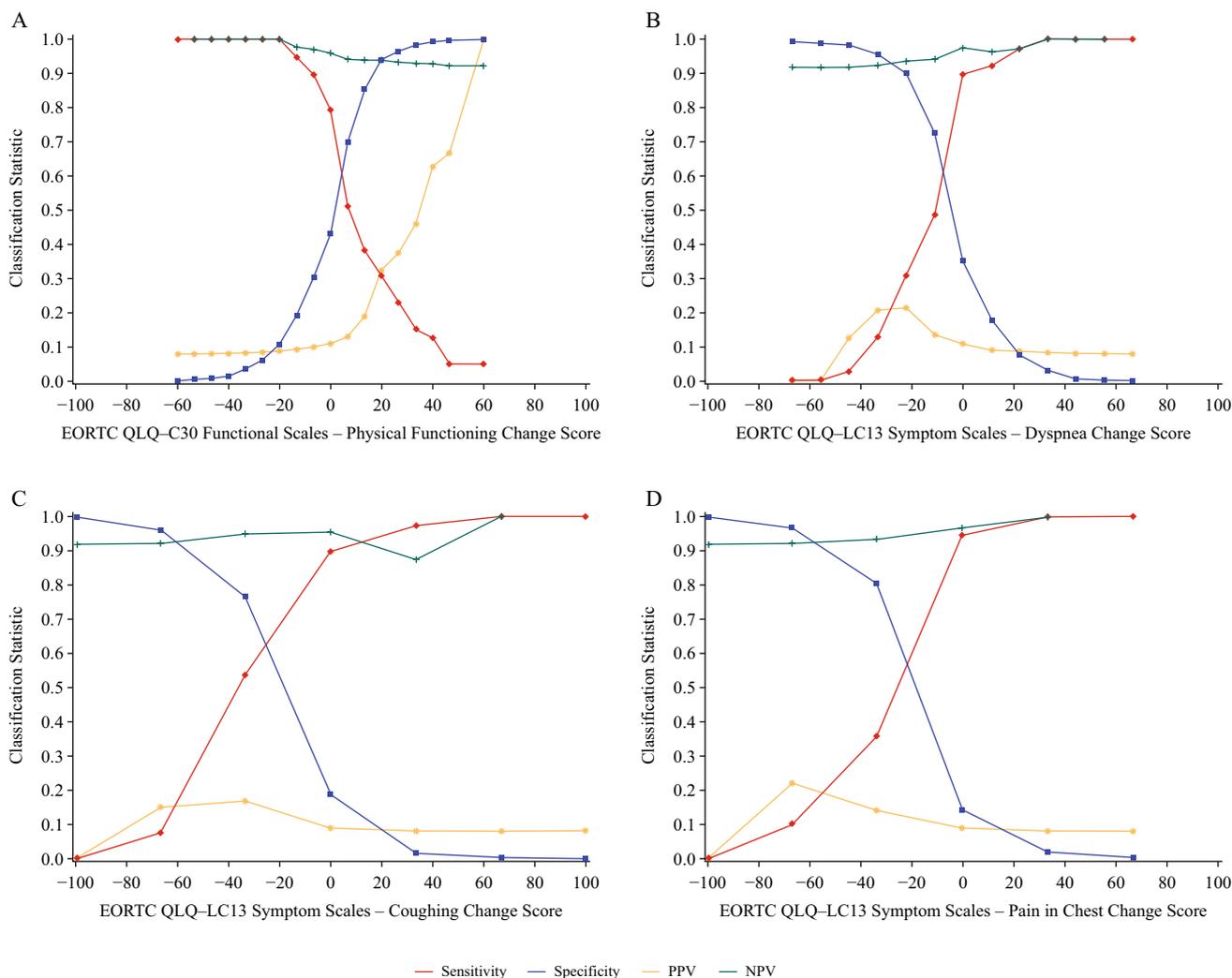


Fig. 2 Classification statistics for improvement on key EORTC domain scores anchored on a two-category improvement on Global Health Status from Baseline to Week 7. *EORTC* European Organisa-

tion for Research and Treatment of Cancer, *NPV* negative predictive value, *PPV* positive predictive value, *QLQ-C30* 30-item core instrument, *QLQ-LC13* 13-item lung cancer module

Physical Functioning, +11.11 for QLQ-LC13 Dyspnea, and +33.33 for QLQ-LC13 Coughing and QLQ-LC13 Chest Pain.

10.29, and 13.42, respectively. For QLQ-LC13 Chest Pain, the test–retest reliability was smaller, therefore the SEM estimate of 14.95 reflected the scale’s more modest reliability.

3.4 Distribution-Based Methods

3.5 Triangulation

Results for the distribution-based approaches are presented in Table 3 for the four key EORTC domains; values for the remaining domains are reported in Online Resource 6. Because distribution-based estimates do not account for the meaningfulness of the magnitude and are always positive, their values were applicable for both improvement and deterioration. For QLQ-C30 Physical Functioning, QLQ-LC13 Dyspnea, and QLQ-LC13 Coughing, the test–retest reliability value was large enough to render the SEM estimate smaller than the 0.5 SD estimate. These estimates were 8.62,

To determine appropriate thresholds for identifying meaningful improvement and deterioration, the estimates from the anchor-based and distribution-based methods for the four key EORTC domains were graphically examined (Online Resource 7) and the proposed thresholds are presented in Table 4. Proposed thresholds for the remaining domains are reported in Online Resource 8.

For QLQ-C30 Physical Functioning, while the estimates for improvement were in the same range, they differed by a score step, with an ECDF estimate of +13.33 and a

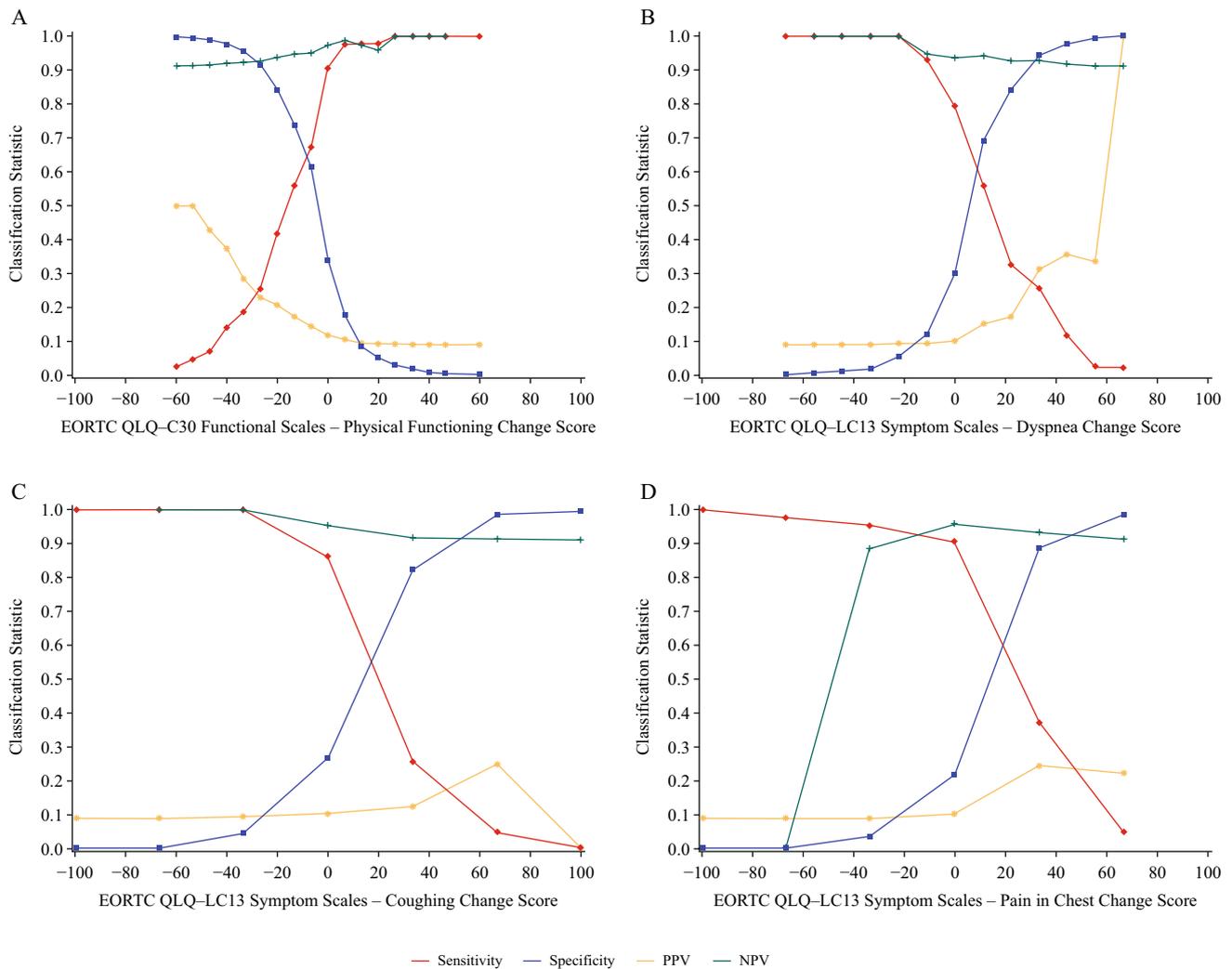


Fig. 3 Classification statistics for deterioration on key EORTC domain scores anchored on a two-category deterioration on Global Health Status from Baseline to Week 7. *EORTC* European Organisa-

tion for Research and Treatment of Cancer, *NPV* negative predictive value, *PPV* positive predictive value, *QLQ-C30* 30-item core instrument, *QLQ-LC13* 13-item lung cancer module

Table 3 Distribution-based estimates for key EORTC domain scores (*N* = 681)

EORTC domain	Test-retest reliability ^a	One-half standard deviation ^b	Standard error of measurement
EORTC QLQ-C30 Physical Functioning	0.82	10.12	8.62
EORTC QLQ-LC13 Dyspnea	0.77	10.81	10.29
EORTC QLQ-LC13 Coughing	0.76	13.67	13.42
EORTC QLQ-LC13 Pain in Chest	0.62	12.06	14.95

EORTC European Organisation for Research and Treatment of Cancer, *QLQ-C30* 30-item core instrument, *QLQ-LC13* 13-item lung cancer module

^aTest-retest reliability is the intraclass correlation coefficient using a two-way mixed effects regression model based on absolute agreement with screening data as the ‘test’ administration and Week 1 data as the ‘retest’ administration, using only participants who report the same value on the Global Health Status item at both administrations

^b Based on Baseline scores

Table 4 Proposed thresholds (and their key sources) for interpreting meaningful improvement and deterioration on key EORTC domain scores

EORTC domain score	Number of items	Minimum score step	Improvement	Deterioration
EORTC QLQ-C30 Physical Functioning	5	6.67	+13.33 (ECDF)	-13.33 (SEM)
EORTC QLQ-LC13 Dyspnea	3	11.11	-11.11 (classification statistics)	+11.11 (classification statistics)
EORTC QLQ-LC13 Coughing	1	33.33	-33.33 (classification statistics)	+33.33 (ECDF and classification statistics)
EORTC QLQ-LC13 Pain in Chest	1	33.33	-33.33 (classification statistics)	+33.33 (ECDF and classification statistics)

ECDF empirical cumulative distribution function, *EORTC* European Organisation for Research and Treatment of Cancer, *QLQ-C30* 30-item core instrument, *QLQ-LC13* 13-item lung cancer module, *SEM* standard error of measurement

classification statistics estimate of +6.67. Because +6.67 is smaller than the SEM of 8.62, it was not a reasonable estimate. Therefore, the ECDF estimate of +13.33 (i.e., a two-category improvement on one of the five Physical Functioning items, or a one-category improvement on two of the five Physical Functioning items) was the proposed threshold for improvement. For deterioration, both the ECDF and classification statistics estimates were -6.67; however, because these estimates were smaller than the SEM, instead the proposed threshold for deterioration of -13.33 (i.e., a two-category deterioration on one of the five Physical Functioning items, or a one-category deterioration on two of the five Physical Functioning items) was based on the lowest possible score that exceeds the SEM.

For QLQ-LC13 Dyspnea improvement, the ECDF produced an uninterpretable estimate of 0.00, while the classification statistics produced a viable estimate of -11.11. This estimate exceeded the SEM of 10.29, therefore -11.11 (i.e., a one-category improvement on one of the three Dyspnea items) was the proposed threshold for improvement. For QLQ-LC13 Dyspnea deterioration, the ECDF estimate was +22.22, while the classification statistics estimate was +11.11; both estimates exceeded the SEM. At a threshold of +11.11, sensitivity was below 50% and specificity above 70%; increasing the threshold to the ECDF estimate of +22.22 produced a sensitivity of around 30% and specificity approaching 90% (Fig. 3b). This divergence of classification statistics is undesirable because it would produce many false negative results, therefore a threshold of +11.11 would be more appropriate. In the ECDF plot (Fig. 1b), the two-category worsened group had a cumulative frequency of around 60% at +11.11, indicating that about 60% of that target anchor group had QLQ-LC13 Dyspnea change scores of +11.11 or worse. The one-category worsened group had a cumulative frequency below 40% at +11.11, showing ample separation between the two-category worsened target anchor group and the one-category worsened group showing less change on the anchor. Therefore, +11.11 (i.e., a one-category deterioration on one of the three Dyspnea items) was an appropriate location for a proposed deterioration threshold for QLQ-LC13 Dyspnea.

For QLQ-LC13 Coughing, the ECDF for improvement produced an uninterpretable estimate of 0.00, while the classification statistics for improvement produced a viable estimate of -33.33. This estimate exceeded the SEM of 13.42, therefore -33.33 (i.e., a one-category improvement on the single Coughing item) was the proposed improvement threshold. For deterioration, both the ECDF and classification statistics produced estimates of +33.33, which also exceeded the SEM. Therefore, +33.33 (i.e., a one-category deterioration on the single Coughing item) was the proposed deterioration threshold.

For QLQ-LC13 Chest Pain improvement, the ECDF produced an uninterpretable estimate of +33.33 (i.e., in the direction of deterioration), but the classification statistics estimate was -33.33 (i.e., in the correct direction). This estimate exceeded the SEM of 14.95, therefore -33.33 (i.e., a one-category improvement on the single Chest Pain item) was the proposed threshold for improvement. For QLQ-LC13 Chest Pain deterioration, both the ECDF and classification statistics estimates were +33.33, exceeding the SEM. Thus, the proposed threshold for deterioration was +33.33 (i.e., a one-category deterioration on the single Chest Pain item).

4 Discussion

Estimates from anchor-based and distribution-based methods and knowledge of the possible score steps were used to propose data-supported thresholds for domain scores on the QLQ-C30 and QLQ-LC13 using a sample of NSCLC clinical trial participants. Absolute values of proposed thresholds ranged from 11.11 to 33.33 on the EORTC scales. For all but three of the domains, the thresholds corresponded to the minimum score step location (i.e., the smallest observable change on the scale), and this pattern is consistent with recent research on QLQ-C30 and QLQ-LC13 scores in NSCLC [24]. For each domain, the proposed thresholds for improvement and deterioration are equidistant from 0, although this methodology considered the possibility that a greater amount of change might be needed to be meaningful

in one direction. The key sources for the proposed thresholds varied across the domains and the direction for interpretation, with some relying on ECDFs alone, some relying on classification statistics alone, some relying on both, and for some the SEM was the only viable estimate. While different domains relied on different sources, this practice of judging results across multiple sources is the triangulation process and was established a priori, thus there can be confidence in the appropriateness of the proposed thresholds regardless of the source. While the correlation with the anchor did not reach 0.30 for most domains, patterns emerged when considering the target and adjacent anchor groups in the ECDFs, the distance between sensitivity and specificity at possible threshold locations, and the measurement variability of the domain score.

For all domain scores, the proposed thresholds exceeded the traditional 10-point threshold, therefore an individual who meets these proposed thresholds would also meet the 10-point threshold for interpreting change. For all domains except QLQ-C30 Fatigue, the proposed thresholds were the next score step above a score of 10, therefore an individual who meets the 10-point threshold for interpreting change would also meet the proposed threshold.

The methodology applied here has several advantages over the traditional 10-point threshold. This analysis used an NSCLC sample to establish thresholds for interpreting changes experienced by patients with lung cancer, whereas the 10-point threshold was originally derived using a combined cancer sample. This analysis also applied the methodology to all relevant domain scores from the QLQ-C30 and QLQ-LC13 to establish a unique threshold for each domain, whereas the original 10-point threshold analysis only considered four QLQ-C30 domain scores, the results from which have been generalized to all EORTC scores. Additionally, this analysis considered the steps along the domain score scale where changes can be observed, proposing thresholds at observable locations.

Because the thresholds proposed here generally result in the same classifications as 10-point thresholds, these analyses can be considered confirmation that it is appropriate to apply the traditional 10-point threshold to all QLQ-C30 and QLQ-LC13 functional, symptom, and adverse effect domain scores in patients with NSCLC. However, the thresholds proposed here are easier to communicate and interpret because they correspond to locations along the score scale where change can be observed in an individual.

This research adds to the evidence required by health authorities (e.g., US FDA [1]) for identifying NSCLC clinical trial participants who experienced a clinically relevant amount of change in their EORTC QLQ-C30 and QLQ-LC13 scores. While many health authorities desire such empirical evidence, there are examples of current HTA requirements where universal thresholds are instead

applied, thus necessitating that meaningful change analyses be conducted different ways depending on the health authority reviewing the submission. For example, the German Federal Joint Committee (G-BA), which is a key stakeholder in the EUnetHTA21 methodological workstream, has set a universal 15% of the PRO scale range as a minimum criterion for assessing the clinical relevance of individual changes on PRO scores [25]. However, the thresholds supported by the evidence presented here may provide different classifications than the 15% criterion adopted by the G-BA. For each of the EORTC domain scores, 15% of the scale equates to a 15-point change, and the proposed thresholds for QLQ-C30 Physical Functioning (± 13.33) and QLQ-LC13 Dyspnea (± 11.11) were lower than that criterion. A review of the classification statistics plots reveals that the 15% criterion increased specificity but decreased sensitivity when compared with the proposed thresholds, leading to fewer false positives but greater false negatives. Thus, on these two domains, the use of the 15% criterion would incorrectly classify some individuals who experienced a meaningful change according to this global anchor, which would result in an under-reporting of treatment responders on QLQ-C30 Physical Functioning and QLQ-LC13 Dyspnea to G-BA. However, for the remaining EORTC domains, the proposed thresholds would identify the same individuals who responded to treatment or experienced meaningful deterioration as the 15% criterion. In consequence, depending on whether empirically-derived thresholds or universal thresholds are required, different rates of meaningful change on QLQ-C30 Physical Functioning and QLQ-LC13 Dyspnea scores may be submitted to different health authorities for the same therapy [25].

Despite agreement with previous research, this study has several limitations. While anchors that were directly rated by patients were preferred over clinical anchors rated by someone other than the patient, this study did not include other PRO instruments or global status items, therefore the only suitable anchor was one that measured overall health rather than the specific concept measured by each domain score. This disconnect was reflected in the correlations, which reached the desired level for only five of the QLQ-C30 domains and none of the QLQ-LC13 domains. Additionally, although the overall study sample was robust, the number of participants who fell into either of the target anchor change groups was small (i.e., 30 with two-category improvement, 31 with two-category worsening). The restricted range of observed change on the anchor, with most participants reporting little to no change, may have also been a source of the modest correlations and a different clinical trial sample may produce more change and a stronger relationship to Global Health Status. Later time points post-baseline may also reflect greater change in health status, but because the dropout rate increases steadily as cancer clinical trials

progress, later time points provided smaller samples. Finally, the selection of the two-category improved/worsened target anchor change groups was also not driven based on patient input as is desired [19], although this definition of meaningful change has been previously used and is larger than the minimum change on the anchor.

Given the body of evidence to date, the proposed thresholds should be considered for use in EORTC score interpretation in NSCLC, but additional research would strengthen the confidence in these estimates. Data from other clinical trials should be used to cross-validate these estimates, ideally by applying multiple methods to multiple anchors, preferably ones targeted to the domain scores being evaluated. Patients with NSCLC should also be interviewed using qualitative methods to obtain their direct impressions of changes on these EORTC domain scores, as well as any anchors used for interpretation. Finally, as additional evidence is generated, the results from future studies and this study should be triangulated to refine the thresholds while considering the properties of the source estimates (e.g., methods, anchors, correlation strength, sample size).

5 Conclusion

By considering data from multiple anchor-based methods, the measurement variability of each domain score, and the observable score change steps, this study proposes robust and defensible thresholds for each of the domains on the EORTC QLQ-C30 and QLQ-LC13 in patients with NSCLC. While this study has limitations, the thresholds it proposes are consistent with those that have been traditionally applied to interpreting EORTC scores. The field has questioned the use of a 10-point threshold for identifying meaningful change on EORTC scores and this study corroborates the appropriateness of this threshold for QLQ-C30 and QLQ-LC13 scores in NSCLC. Additionally, these results highlight that score interpretation is facilitated when the amount of change that is possible on the score is considered. Given the consistency with previous estimates and the enhanced interpretability, researchers should consider applying the thresholds proposed here for identifying individuals who have responded to treatment or who have experienced deterioration in future studies of patients with lung cancer.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40271-022-00584-w>.

Acknowledgements The authors thank the clinical trial participants and their families, and the investigators, co-investigators, and study teams at each participating center. They also thank Arman Altincatal (formerly of Evidera) for assistance with data analysis and interpretation, and Vivek Pawar (EMD Serono Research & Development Institute, Inc., Billerica, MA, USA, an affiliate of Merck KGaA) for

assistance with interpretation. Editorial assistance, including redrawing figures and checking final proofs, was provided to the authors by Krati Dixit of Bioscript Stirling (Macclesfield, UK) and funded by Merck (CrossRef Funder ID: 10.13039/100009945).

Declarations

Funding This psychometric analysis was funded by EMD Serono Research & Development Institute, Inc., Billerica, MA, USA, an affiliate of Merck KGaA (CrossRef Funder ID: 10.13039/100004755). The clinical trial from which this analysis was derived was sponsored by Merck (CrossRef Funder ID: 10.13039/100009945), as part of an alliance between Merck and Pfizer.

Conflict of interest Cheryl D. Coon is an employee of Outcometrix and received funding for this psychometric analysis from EMD Serono Research & Development Institute, Inc., Billerica, MA, USA, an affiliate of Merck KGaA (CrossRef Funder ID: 10.13039/100004755). Michael Schlichting is an employee of Merck Healthcare KGaA, Darmstadt, Germany. Xinke Zhang is an employee of EMD Serono Research & Development Institute, Inc., Billerica, MA, USA, an affiliate of Merck KGaA.

Ethics approval The study protocol was approved by Institutional Review Boards and Ethics Committees at each institution. The study was conducted in accordance with the protocol, Good Clinical Practice guidelines, and the Declaration of Helsinki. All study participants provided written informed consent.

Consent to participate Informed consent was obtained from all individual participants included in the original study, whose PRO data were evaluated in the analysis reported herein.

Consent for publication Not applicable (no identifying information is included in this article).

Data sharing For all new products or new indications approved in both the European Union and the US after 1 January 2014, Merck (CrossRef Funder ID: 10.13039/100004755) shares patient-level and study-level data after de-identification, and redacted study protocols and clinical study reports from clinical trials in patients. These data will be shared with qualified scientific and medical researchers, upon the researcher's request, as necessary for legitimate research. Such requests should be submitted in writing to Merck's data sharing portal. When Merck has a co-research, co-development, co-marketing or co-promotion agreement, or when the product has been out-licensed, the responsibility for disclosure might be dependent on the agreement between parties. Under these circumstances, Merck will endeavor to gain agreement to share data in response to requests.

Code availability Not applicable.

Clinical trial registration This psychometric analysis used data from a clinical trial that is registered with ClinicalTrials.gov, identifier NCT02395172.

Author contributions All authors contributed to the design of the psychometric study and were involved in analysis of data, interpretation of results, and drafting and revising the manuscript. All authors read and approved the final manuscript.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction

in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

References

1. US FDA. Guidance for industry on patient-reported outcome measures: Use in medical product development to support labeling claims. *Fed Reg.* 2009;74(235):65132–3.
2. Fiero MH, Roydhouse JK, Vallejo J, King-Kallimanis BL, Kluetz PG, Sridhara R. US Food and Drug Administration review of statistical analysis of patient-reported outcomes in lung cancer clinical trials approved between January, 2008, and December, 2017. *Lancet Oncol.* 2019;20(10):e582–9.
3. Xalkori [package insert]. New York, NY: Pfizer, Inc.; 2021.
4. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organisation for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst.* 1993;85:365–76.
5. Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M. The EORTC QLQ-LC13: a modular supplement to the EORTC core quality of life questionnaire (QLQ-C30) for use in lung cancer clinical trials. *Euro J Cancer.* 1994;30(5):635–42.
6. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol.* 1998;16(1):139–44.
7. Cocks K, King MT, Velikova G, Fayers PM, Brown JM. Quality, interpretation and presentation of European organization for research and treatment of cancer quality of life questionnaire core 30 data in randomized controlled trials. *Eur J Cancer.* 2008;44:1793–8.
8. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual Life Res.* 1996;5:555–67.
9. Maringwa JT, Quinten C, King M, Ringash J, Osoba D, Coens C, et al. Minimal important differences for interpreting health-related quality of life scores from the EORTC QLQ-C30 in lung cancer patients participating in randomized controlled trials. *Support Care Cancer.* 2011;19(11):1753–60.
10. Koller M, Musoro JZ, Tomaszewski K, Coens C, King MT, Sprangers MAG, et al. Minimally important differences of EORTC QLQ-C30 scales in patients with lung cancer or malignant pleural mesothelioma - Interpretation guidance derived from two randomized EORTC trials. *Lung Cancer.* 2022;167:65–72.
11. Kluetz PG, Slagle A, Papadopoulos EJ, Johnson LL, Donoghue M, Kwitkowski VE, et al. Focusing on core patient-reported outcomes in cancer clinical trials: symptomatic adverse events, physical function, and disease-related symptoms. *Clin Cancer Res.* 2016;22:1553–8.
12. Roy UB, King-Kallimanis BL, Kluetz PG, Selig W, Ferris A. Learning from patients: reflections on use of patient-reported outcomes in lung cancer trials. *J Thorac Oncol.* 2018;13(12):1815–7.
13. US FDA. Voice of the patient. A series of reports from the U.S. Food and Drug Administration's (FDA's) Patient-Focused Drug Development Initiative. *Lung Cancer.* 2013. <https://www.fda.gov/media/130344/download>. Accessed 9 Nov 2021.
14. Barlesi F, Vansteenkiste J, Spigel D, Ishii H, Garassino M, de Marinis F, et al. Avelumab versus docetaxel in patients with platinum-treated advanced non-small-cell lung cancer (JAVELIN Lung 200): an open-label, randomised, phase 3 study. *Lancet Oncol.* 2018;19(11):1468–79.
15. Gralla RJ, Coon C, Taylor F, et al. Evaluation of disease-related symptoms in patients with advanced squamous non-small cell lung cancer treated with nivolumab or docetaxel. *J Thorac Oncol.* 2015;10(9):S233–4.
16. Barlesi F, Garon E, Kim DW, et al. Assessment of health-related quality of life (HRQoL) in KEYNOTE-010: a phase 2/3 study of pembrolizumab vs docetaxel in patients with previously treated advanced NSCLC. *Ann Oncol.* 2016;27(Suppl 6):1219P.
17. Bedard G, Zeng L, Zhang L, Lauzon N, Holden L, Tsao M, et al. Minimal important differences in the EORTC QLQ-C30 in patients with advanced cancer. *Asia-Pac J Clin Oncol.* 2014;10(2):109–17.
18. 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(2):102–9.
19. Fda US. Patient-focused drug development: select, develop or modify fit-for-purpose clinical outcome assessments. *Fed Reg.* 2018;83(156):40057–9.
20. Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Qual Life Res.* 2017;27:33–40.
21. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care.* 2003;41(5):582–92.
22. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol.* 1999;52(9):861–73.
23. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15(2):155–63.
24. Beaumont J, Yu H, Lin HM, Goodman E, Hecht S, Le K, et al. Estimation of meaningful change thresholds for the EORTC QLQ-C30 and QLQ-LC13 inpatients with ALK+ non-small cell lung cancer (NSCLC). Poster presented at the 28th Annual Conference of the International Society for Quality of Life Research; October 2021.
25. IQWiG. General Methods Version 6.0 of 5 November 2020. https://www.iqwig.de/methoden/general-methods_version-6-0.pdf. Accessed 27 Jan 2022.