Title

Estimation of minimally important differences and responder definitions for EORTC QLQ-MY20 scores in multiple myeloma patients

Authors

Sully, Kate
Trigg, Andrew
Bonner, Nicola
et al.

ORIGINAL ARTICLE

European Journal of Haematology WILEY

Estimation of minimally important differences and responder definitions for EORTC QLQ-MY20 scores in multiple myeloma patients

Kate Sully[1] | Andrew Trigg[1] (iD) | Nicola Bonner[1] (iD) | Alejandro Moreno-Koehler[1] |
Claire Trennery[1] | Nina Shah[2,3] (iD) | Emre Yucel[2] (iD) | Sumeet Panjabi[2] | Kim Cocks[1] (iD)

[1]Adelphi Values, Cheshire, UK

[2]Amgen Inc, Thousand Oaks, CA, USA

[3]University of California, San Francisco, CA, USA

Correspondence
Kim Cocks, Adelphi Values, Adelphi Mill, Grimshaw Lane, Bollington, Cheshire, SK10 5JB, UK.
Email: Kim.cocks@adelphivalues.com

Funding information
This project was funded by Amgen, who commissioned Adelphi Values, a health outcomes agency by whom Kate Sully, Andrew Trigg, Nicola Bonner, Alejandro Moreno-Koehler, Claire Trennery and Kim Cocks are employed. Emre Yucel and Sumeet Panjabi are employed by Amgen, and hold Amgen stock. Nina Shah received consultancy fees from Amgen for her involvement in the project.

Abstract

Objectives: Thresholds for the minimally important difference (MID) or responder definition (RD) in health-related quality-of-life (HRQoL) scores are required to interpret the impact of an intervention or change in the trajectory of the condition which is meaningful to patients. This study aimed to establish MID and RD for the European Organisation for Research and Treatment of Cancer Quality of Life Multiple Myeloma questionnaire (EORTC QLQ-MY20).

Methods: A novel mixed-methods approach was applied by utilizing both existing clinical trial data and prospective patient interviews. Anchor-based, distribution-based, and qualitative-based estimates of meaningful change were triangulated to form recommended RDs for each scale of the EORTC QLQ-MY20. Anchor-based MIDs were summarized using weighted correlation.

Results: Recommended MIDs were as follows: Disease Symptoms (DS 10 points), Side Effects of Treatment (SE 10 points), Body Image (BI 13 points), and Future Perspective (FP 9 points). Recommended RDs were as follows: DS (16 improvement; 11 worsening), SE (6 improvement; 9 worsening), BI (33 improvement; 33 worsening), and FP (11 improvement; 11 worsening).

Conclusions: The study generated estimates of the MID and RD for each scale of the EORTC QLQ-MY20. Published estimates will enable investigators and clinicians to adopt these as standard for interpretation and for hypothesis testing. Consequently, analyses from trials of different interventions can be more comparable.

KEYWORDS
EORTC QLQ-MY20, interpretation guidelines, meaningful change, minimally important difference, mixed methods, multiple myeloma, qualitative interviews, quality of life, responder definition

Kate Sully and Andrew Trigg are contributed equally to this work.

1 | INTRODUCTION

Multiple myeloma (MM) is the third most common hematologic malignancy worldwide.[1] Patients often report significant impairment in health-related quality of life (HRQoL) due to disease-related symptoms like fatigue, pain, and reduced physical function, as well as treatment-related toxicities such as neuropathy.[2-4] Therefore, in addition to efficacy, HRQoL should be evaluated as a key endpoint within trials assessing novel treatments for MM. When HRQoL is evaluated, the European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire (EORTC QLQ-C30), in conjunction with its MM-specific module (QLQ-MY20), is among the most commonly employed patient-reported outcome (PRO) measures.[5]

A key challenge in the field of HRQoL measurement is the need to interpret the meaning of differences between treatment groups. Statistical significance does not necessarily reflect clinical relevance, so both are needed for interpretation. Here, we refer to the smallest difference in mean score between groups which could be considered clinically meaningful, as the minimally important difference (MID).[6] The threshold of within-patient change deemed meaningful, used to define a patient as a responder if their change in score exceeds this threshold, is referred to as the responder definition (RD).[7]

While score interpretation guidelines have been published for the EORTC QLQ-C30, including estimates specific to an MM population,[8] no published or accepted guidelines exist for EORTC QLQ-MY20 scores.[9] In light of this, three carfilzomib Phase III MM trials (ASPIRE[2] [NCT01080391], CLARION[10] [NCT018018752], and ENDEAVOR [NCT01568866]) used a distribution-based estimate, the standard error of measurement (SEM),[11] to estimate MIDs since this had been used previously in the literature.[12] The smallest possible score change was used for a RD[12-14] with the next larger score change as a sensitivity analysis. Ideally, MID and RD recommendations should come from triangulating multiple estimates from anchor- and distribution-based approaches,[15,16] but where these are not published, studies tend to rely on the distribution-based estimates alone. Distribution-based estimates do not directly consider what is meaningful to patients or clinicians, and will vary across samples, impeding consistent interpretation. Qualitative patient interviews are an emerging approach to additionally explore score interpretation and contextualize estimates in terms of how a patient feels and functions.[17,18]

This study aimed to establish MID and RD for the EORTC QLQ-MY20 utilizing both existing clinical trial data and prospective patient interviews. The aim of this study was thus to recommend MIDs and RDs for each scale of the EORTC QLQ-MY20 for use in MM patients. This was achieved through triangulation of anchor- and distribution-based analyses of data pooled across the ASPIRE, CLARION and ENDEAVOR trials and prospective qualitative interviews of patients with MM. It is hoped that the publication of such guidelines will standardize the interpretation across future studies using these instruments.

2 | METHODS

2.1 | EORTC QLQ-MY20

The EORTC QLQ-MY20 contains three multi-item scales (Disease Symptoms [6 items], Side Effects of Treatment [10 items], Future Perspective [3 items]), and a single-item scale (Body Image).[19,20] All transformed scale scores range from 0 to 100 with higher scores indicating worse symptoms (Disease Symptoms and Side Effects of Treatment) or better support/functioning (Future Perspective and Body Image). The reliability and validity of these scores has been previously documented in patients with MM.[20] The EORTC QLQ-MY20 is administered in conjunction with the QLQ-C30.

2.2 | Clinical trial data

Data were pooled across three clinical trials: ASPIRE, ENDEAVOR, and CLARION. Each was a Phase III, randomized, open-label study comparing carfilzomib-based regimens, with a primary end point of progression-free survival (PFS). ASPIRE randomized patients with relapsed or refractory MM to receive carfilzomib, lenalidomide, and dexamethasone versus lenalidomide and dexamethasone in 28-day cycles. ENDEAVOR randomized patients with relapsed or refractory MM to receive carfilzomib and dexamethasone in 28-day cycles versus bortezomib and dexamethasone in 21-day cycles. CLARION randomized patients with newly diagnosed MM ineligible for transplant to receive carfilzomib, melphalan, and prednisone versus bortezomib, melphalan, and prednisone in 42-day cycles. Further details of each study including ethical approval are described elsewhere.[2,10,21,22]

Patients were eligible for entry into the pooled sample if they had completed the EORTC QLQ-C30 or QLQ-MY20 at baseline plus at least one other of the following time points: mid-treatment (MT), or end of treatment (EOT). MT was defined as Week 8-12, as while ENDEAVOR and CLARION both had PRO assessments common to both treatment arms at Week 12, the closest PRO assessment common to both arms in ASPIRE was at Week 8. EOT was defined as 30 days after the last administration of treatment in all three studies.

In anchor-based approaches, a criterion "anchor measure" is used to identify patients who have experienced a meaningful change on the concept being measured.[15] The anchor should be sufficiently related to the PRO score to map one onto the other. It also needs to be interpretable on its own. Potential anchors were identified by the authors reviewing the clinical trial protocols to identify measures available across the trials at mid-treatment (MT) and end of treatment (EOT). Anchors were selected via review of the protocols and case report forms. Potential anchors were chosen if they were deemed to have conceptual overlap with any of the EORTC QLQ-MY20 scales. Clinical input was also sought to confirm the clinical relevance and feasibility of the anchors.

Prior to their implementation, polyserial correlations between anchor classification and changes on each score were calculated to ensure sufficient correlation (≥|0.3|) of the proposed anchor with the EORTC QLQ-MY20 scales.[15]

2.3 | Anchor-based analyses

Two anchor-based methods were applied, mirroring an approach in another pooled study of cancer patients.[23] Anchor-based MIDs were estimated by calculating the mean change score of patients classified as improved and deteriorated according to anchor definitions. Linear regression models were also fitted with EORTC QLQ-MY20 score change as the outcome and a binary indicator of stable vs improved/worsened according to the anchor as a predictor, where the coefficient of this indicator was the MID estimate (incorporating the change score of stable patients). Diagnosis (newly diagnosed/relapsed) was accounted for in the model.

Anchor-based RDs were estimated by plotting ROC curves for each anchor-scale combination to discriminate between patients who had changed or remained stable, where the "optimal" score change was determined by minimizing the sums of squares of 1-sensitivity and 1-specificity.[24] Sensitivity and specificity values of at least 0.750 for the optimal score change have been previously recommended for application to individual patients.[25]

2.4 | Distribution-based analyses

Distribution-based estimates of half a standard deviation (0.5 SD) at baseline and one SEM at baseline (using Cronbach's alpha at baseline; multi-item scales only) were calculated.

2.5 | Qualitative data

In this mixed-methods study, the patient interviews were conducted alongside an analysis of existing trial data. To ensure emerging estimates from the clinical trials did not influence the conduct of the patient interviews, the two components of the study were conducted independently until all analyses had been completed.

Semi-structured, qualitative interviews were conducted with adults with MM (newly diagnosed and relapsed/refractory). The aim was to understand and explore what constitutes a meaningful change in concepts assessed by the EORTC QLQ-MY20 from a patient perspective, focusing on RD estimates.

The study aimed to recruit 20 patients: 10 from the UK (5 newly diagnosed, 5 relapsed/refractory) and 10 from the US (5 newly diagnosed, 5 relapsed/refractory). A sample of 20 patients was judged likely to sufficiently explore the topics of interest.[26] Patients were recruited via clinician referral and eligible for inclusion if they were ≥18 years of age, a native English speaker, willing and able to participate in a 60-minute telephone interview, and a clinician confirmed diagnosis of relapsed/refractory or newly diagnosed MM. Relapsed/refractory MM patients were required to have received at least second-line therapy, while newly diagnosed patients were required to be currently receiving or recently completed first-line therapy. Patients with significant hearing, reading or speaking difficulties, or other conditions which in the clinician's judgment would render the patient unable to participate, were excluded. Patients were also recruited according to quotas to provide balance in terms of age, gender, race, education, and transplant eligibility. Ethical approval was obtained (Copernicus IRB approval code ADE1-17-491). Patients read and signed informed consent before participation and could withdraw from the study at any time. Potentially eligible patients were recruited through clinicians, who recorded clinical information relevant to eligibility criteria. Once eligibility was confirmed, demographic information was collected, and an interview scheduled.

Prior to interview, patients were provided with a copy of the EORTC QLQ-MY20. A semi-structured interview guide was designed following a recommended stepwise approach. As a first step patients' understanding of the underlying scale was confirmed, including the concept of interest being assessed, the anchors of the response scale and the direction of the response scale. Meaningful change was then discussed but at the item level (1-4 scale). Next, in the context of multi-item scales of the EORTC QLQ-MY20, the aggregation of items into scores including the 0-100 scaling was discussed, to assess patients' understanding of score change at the scale level. Meaningful change at the score level was then explored in an open-ended manner, supplemented with probing on specific levels of change.

2.6 | Qualitative analysis

All interviews were audio recorded and transcribed verbatim for analysis within ATLAS.ti v8 qualitative data analysis software. Transcripts were analyzed by sorting quotes into concepts using methods derived from thematic analysis. This identifies recurring themes provided by individual patients, ensuring that the study findings directly reflect how patient's think about and describe meaningful change on the concepts assessed by the EORTC QLQ-MY20. Quotes were organized in a data extraction table at the item- and scale-level. Numeric estimates were extracted, combined with the reasoning detailing the impact of the score change on a patient's condition.

2.7 | Triangulation

The anchor-based, distribution-based and qualitative-based analyses lead to multiple derived estimates, which need combining to form recommended thresholds for MIDs and RDs, a process known as triangulation.[15] MID estimates from each anchor were presented on a forest plot to identify if there was overlap and convergence around a small range of values. A weighted average was also calculated for each score to summarize MIDs across anchors, where estimates were weighted by the correlations between change in anchor and PRO score as follows:

$$\text{MID}_{\text{weighted}} = \frac{\sum_{i=1}^{n} |r_i|\,|x_i|}{\sum_{i=1}^{n} |r_i|}$$

where x denotes each [absolute] estimate and r denotes the [absolute] correlation coefficient of each anchor-scale combination, for each i of n total estimates.[27]

TABLE 1 Clinical characteristics of the pooled clinical trial sample and independent qualitative interview sample

Characteristic	Pooled trial sample (N = 2147) n (%)	Independent qualitative sample (N = 20) n (%)
Age category		
<65	742 (34.6)	11 (55.0)
≥65-<74	1013 (47.2)	5 (25.0)
≥75	392 (18.3)	4 (20.0)
Sex		
Male	1111 (51.7)	11 (55.0)
Female	1036 (48.3)	9 (45.0)
Race		
White	1719 (80.1)	16 (80.0)
Black or African American	35 (1.6)	3 (15.0)
Asian	285 (13.3)	1 (5.0)
American Indian or Alaska Native	4 (0.2)	-
Other/ Unknown/ Multiple	30 (1.4)	-
Missing	74 (3.4)	-
Education		
Some high school	-	1 (5.0)
High school diploma	-	4 (20.0)
Some years at college	-	5 (25.0)
College or university degree	-	8 (40.0)
Masters degree	-	1 (5.0)
Doctorate degree	-	1 (5.0)
Region		
Europe	1464 (68.2)	2 (10.0)
North America	236 (11.0)	18 (90.0)
Rest of World	447 (20.8)	
MM type		
Relapsed/ refractory	1421 (66.2)	10 (50.0)
Newly diagnosed	726 (33.8)	10 (50.0)

The distribution-based estimates were also added to the forest plot and were generally considered as the smallest desirable thresholds. All RD estimates, including qualitative input, were presented on separate forest plots alongside the distribution-based estimates. The possible increments in EORTC QLQ-MY20 scores were marked on the x-axis to ensure the chosen threshold could be translated onto a possible score change (eg, the Body Image scale is one item and can only change by increments of 33.3 points for an individual patient). For a given estimate from the clinical trial data, qualitative feedback on the corresponding level of score change was sought to contextualize and guide their meaning, in addition to exploring what smaller or larger changes meant. Anchor-based estimates were prioritized and distribution-based considered supportive to these;[15] qualitative-based estimates were used to further support and narrow the range of values.

3 | RESULTS

3.1 | Anchor-based analyses

The pooled sample comprised 2147 patients with an EORTC QLQ-C30 or QLQ-MY20 assessment at baseline plus MT or EOT; patient and disease characteristics are presented in (Table 1).

Two PRO measures and three clinical measures were identified as potential anchors. Patient-reported anchors were the EORTC QLQ-C30 Global Health Status/Quality of Life (GHS/QOL) scale and Functional Assessment of Cancer Therapy-Gynecologic Oncology Group Neurotoxicity (FACT-GOG-Ntx) Additional Concerns scale). Clinical anchor measures were the Eastern Cooperative Oncology Group Performance Status (ECOG PS), matched adverse events (AEs) and peripheral neuropathy-related AEs). Changes from baseline to MT/EOT were used to identify groups of patients who had remained stable, improved and deteriorated on the anchor measure (see [Table 2] for definitions). Matched AEs were those that were deemed to be potentially related to the EORTC QLQ-MY20 scales; Disease Symptoms (Chest pain, Back pain, Arthralgia, Bone pain, Pain in extremity, Musculoskeletal chest pain, Myalgia, Neuralgia); Side Effects of Treatment (Fatigue, Paresthesia, Neuropathy peripheral, Polyneuropathy, Insomnia, Anxiety, Conjunctivitis, Chest pain, Dyspepsia); Future Perspective (Anxiety, Insomnia). Of the five proposed anchor measures, the FACT-GOG-Ntx Additional Concerns scale and matched AEs (Side Effects of Treatment scale only) were sufficiently correlated for further analysis (see [Table 2]).

Mean EORTC QLQ-MY20 score changes on each scale followed expected trends within each anchor group, where changes indicating better health were observed in the improvement group and vice versa. All mean change estimates and those from linear regression are presented in (Figure 1). Mean changes for deterioration were considerably smaller in magnitude than improvement for Disease Symptoms, where sizable change scores were also observed in the stable groups (−8.02 at MT and −6.74 at EOT according to the FACT-GOG-Ntx anchor). Results from linear regression analyses were generally more consistent across different anchor measures. Estimates for improvement and deterioration were closer in magnitude compared to using the mean change.

The change scores that optimally discriminated between anchor-based change and stability, as identified in ROC curves and forming RD estimates, are presented in (Table 3) and (Figures S1-S4). The

TABLE 2 Anchor-scale correlations based on change from baseline to mid-treatment and change from baseline to end of treatment

Anchor measure	Definition (Improvement/Stable/Deterioration) (Stable/Deterioration for AE anchors)	Number improved/stable/deteriorated at MT	Correlation MT	Number improved/stable/deteriorated at EOT	Correlation EOT
GHS/QOL	+8.3/ 0/ −8.3[32]	Improvement: 209 Stable: 465 Deterioration: 220	DS: −0.076 SE: −0.082 BI: 0.031 FP: 0.081	Improvement: 95 Stable: 227 Deterioration: 98	DS: −0.036 SE: −0.082 BI: 0.095 FP: 0.114
FACT-GOG-Ntx Additional Concerns	≥3.3/ 0/ ≤−3.3[33]	Improvement: 222 Stable: 789 Deterioration: 333	DS: **−0.410** SE: **−0.545** BI: 0.299 FP: 0.227	Improvement: 85 Stable: 343 Deterioration: 250	DS: **−0.416** SE: **−0.607** BI: **0.336** FP: **0.314**
ECOG PS	−1/ 0/ +1[34]	Improvement: 210 Stable: 952 Deterioration: 299	DS: −0.135 SE: −0.124 BI: 0.011 FP: 0.030	Improvement: 87 Stable: 616 Deterioration: 231	DS: −0.171 SE: −0.226 BI: 0.101 FP: 0.149
Matched grade 2 + AEs[a]	None within MT/EOT window/ ≥1 within MT/EOT window[35]	Stable (DS): 1941 Deterioration (DS): 79 Stable (FP): 1992 Deterioration (FP): 28 Stable (SE): 1865 Deterioration (SE): 155	DS: −0.188 SE: −0.160 FP: 0.040	Stable (DS): 1025 Deterioration (DS): 27 Stable (FP): 1045 Deterioration (FP): 7 Stable (SE): 1016 Deterioration (SE): 36	DS: −0.196 SE: **−0.322** FP: 0.044
Peripheral Neuropathy grade 2 + AEs	None within MT/EOT window/ ≥1 within MT/EOT window[35]	Stable: 1261 Deterioration: 84	DS: −0.029 SE: −0.271 BI: 0.096 FP: 0.035	Stable: 662 Deterioration: 16	DS: 0.124 SE: −0.235 BI: −0.042 FP: −0.060

Note: Correlations in bold are ≥0.3 and thus retained for anchor-based analyses on the scale of interest.

Abbreviations: AE, Adverse event; BI, Body image; DS, Disease Symptoms; ECOG PS, Eastern Cooperative Oncology Group Performance Status; EOT, End of treatment; FP, Future perspective; GHS/QOL, Global Health Status/Quality of Life; MT, mid-treatment; SE, Side Effects of treatment

[a]Matched AEs for each scale are as follows: Disease Symptoms (DS: Chest pain, Back pain, Arthralgia, Bone pain, Pain in extremity, Musculoskeletal chest pain, Myalgia, Neuralgia); Future Perspective (FP: Anxiety, Insomnia); Side Effects of Treatment (SE: Fatigue, Paresthesia, Neuropathy peripheral, Polyneuropathy, Insomnia, Anxiety, Conjunctivitis, Chest pain, Dyspepsia). No AEs corresponding to the Body Image scale were identified.

AUC of all ROC curves had a lower 95% confidence interval >0.5; therefore, all were able to discriminate between anchor groupings better than chance alone. However, no sensitivity and specificity values of any estimate met a threshold of ≥0.750 recommended for application to individual patients.

3.2 | Distribution-based analyses

Distribution-based estimates of 0.5 SD at baseline were as follows: Disease Symptoms (10.6), Side Effects (7.0), Body Image (13.7), and Future Perspective (12.6). SEM estimates were as follows: Disease Symptoms (9.1), Side Effects (9.5), Future Perspective (10.7).

3.3 | Qualitative analyses

Twenty patients were interviewed; 18 (90%) recruited from the US and 2 (10%) from the UK. While it was originally intended that patients would be recruited from both the US (n = 10) and the UK (n = 10), UK recruitment was not as rapid. Demographic and clinical characteristics are presented in (Table 1). As per sample targets, 50% (n = 10) of the sample had relapsed or refractory MM and 50% (n = 10) were newly diagnosed with MM. A range of education levels

was also represented, where 50% (n = 10) patients had less than a college degree.

Patients were able to understand the task and provided change estimates for all of the EORTC QLQ-MY20 scales in terms of both improvement and deterioration.

> "*Important meaning something that would really make a difference for me, um, and improve my overall condition and, you know, physically and mentally.*"
>
> (US-02)

> "*When it gets worse again, then it makes me worry more, and it reminds me more that I'm sick all the time. So, then it makes me worry about the future.*"
>
> (US-07)

RD estimates were provided for each scale for both improvement and worsening (Table 4). An overall estimate, which represents the most frequently reported change estimate, has been provided for each scale. For details of all change estimates provided for each scale, see Table S1. Where patient's responses were varied, meaning an
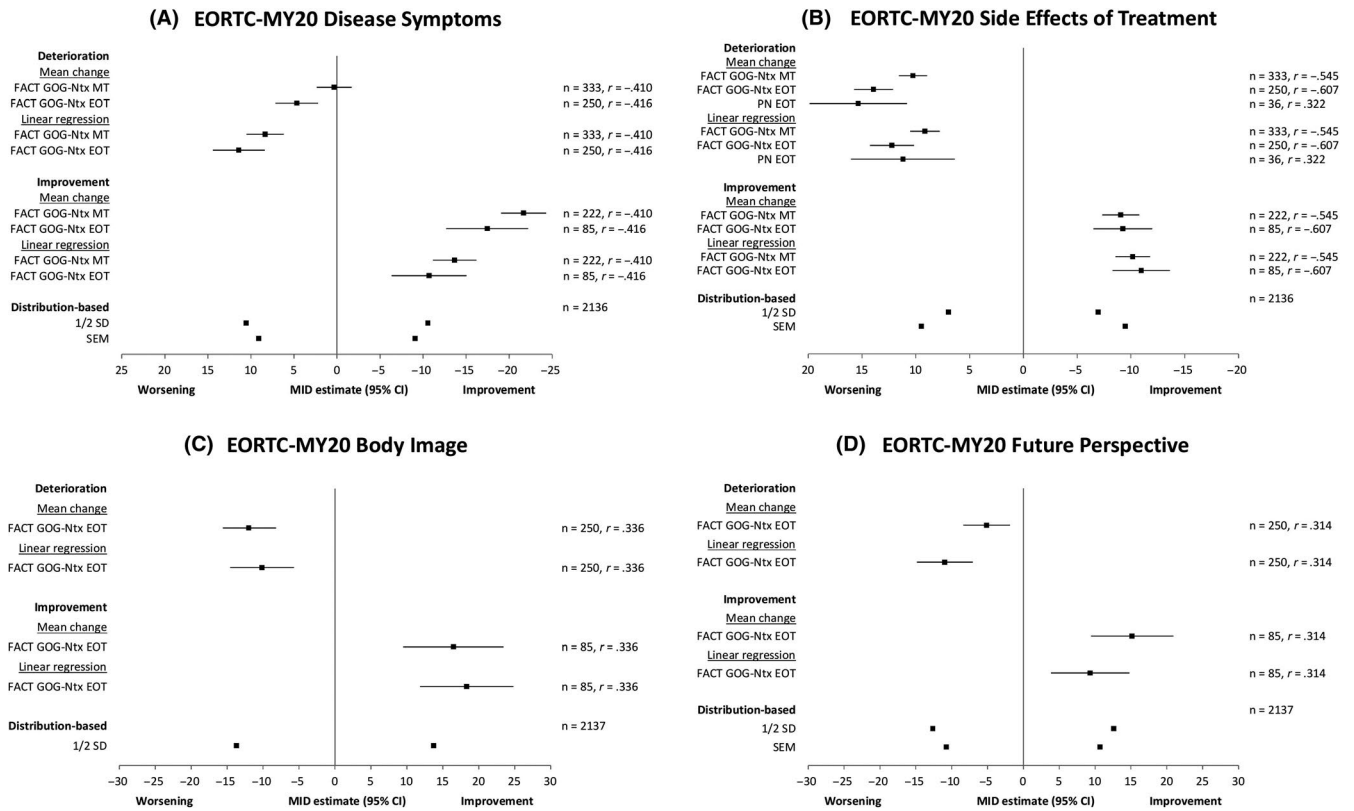
(A) EORTC-MY20 Disease Symptoms

(B) EORTC-MY20 Side Effects of Treatment

(C) EORTC-MY20 Body Image

(D) EORTC-MY20 Future Perspective

FIGURE 1   Results of the MID estimates definitions: anchor-based analyses and distribution-based analyses

TABLE 3   ROC results where anchors have sufficient correlation with PRO scores

Scale	Direction	Anchor	AUC	Optimal threshold	Sensitivity	Specificity
Disease symptoms	Deterioration	FACT-GOG-Ntx MT	0.631	5.6	0.429	0.813
Disease symptoms	Deterioration	FACT-GOG-Ntx EOT	0.662	5.6	0.500	0.751
Disease symptoms	Improvement	FACT-GOG-Ntx MT	0.702	−11.1	0.613	0.699
Disease symptoms	Improvement	FACT-GOG-Ntx EOT	0.644	−11.1	0.506	0.699
Side effects	Deterioration	FACT-GOG-Ntx MT	0.722	6.7	0.652	0.709
Side effects	Deterioration	FACT-GOG-Ntx EOT	0.764	10.0	0.632	0.784
Side effects	Deterioration	Matched AEs	0.711	6.7	0.750	0.571
Side effects	Improvement	FACT-GOG-Ntx MT	0.735	−3.3	0.707	0.670
Side effects	Improvement	FACT-GOG-Ntx EOT	0.743	−6.7	0.624	0.784
Body image	Deterioration	FACT-GOG-Ntx EOT	0.601	−33.3	0.380	0.787
Body image	Improvement	FACT-GOG-Ntx EOT	0.661	33.3	0.435	0.833
Future perspective	Deterioration	FACT-GOG-Ntx EOT	0.595	−11.1	0.268	0.871
Future perspective	Improvement	FACT-GOG-Ntx EOT	0.592	11.1	0.400	0.740

overall estimate was not possible, a range was provided (Side Effects of Treatment and Future Perspective scales), rather than a single value (Disease Symptoms and Body Image). For the Disease Symptoms scale, an improvement of 20 points was considered meaningful, while a 10-point worsening was felt to be important to patients. For the Side Effects of Treatment scale an equal number of patients suggested 10, 20 or 30 points would constitute a meaningful improvement, however, in the opposite direction a deterioration of 5-10 points was considered meaningful. Estimates for the Body Image scale ranged

from 20-points for an improvement to be considered meaningful, to a deterioration of 10-20 points considered important to patients. An improvement of 10-points on the Future Perspective scale was considered important to most patients, while in the direction of deterioration estimates of between 10-20 points were provided.

Findings from the interviews highlighted which scales (Disease Symptoms and Side Effects of Treatment) were most important to patients, as reflected in direct qualitative feedback from patients. While Body Image and Future Perspective were clearly important

TABLE 4 Overview of EORTC-MY20 scale improvement and worsening change scores

Scale	Improvement change score across interview sample	Example quote	Worsening change score across interview sample	Example quote
Disease symptoms	20	*"Well then I could tolerate the pain instead at being at 70 or 90. I can tolerate the pain and do all what I want to do."* (US-09, change from 90-70, 20-point improvement)	10	*"Because it's hard to—I mean it's hard to handle a normal life when it gets, when it gets to that kind of pain, you know."* (US-17, change from 20 to 30, 10-point worsening)
Side effects of treatment	No clear pattern (evenly distributed 10/20/30)	*"It would indicate a significant decrease in the pain I'm experiencing in my hands and feet, uh, my back, uh, and at times other parts of my body."* (US-11, change from 90 to 60, 30-point improvement)	5-10	*"Side effects would cause me some concern, in that they had developed and that could be because my myeloma would be deteriorating."* (UK-02, change from 10 to 20, 10-point worsening)
Body image	20	*"I was thinking of really what would be significant, um, in terms of my body image, you know, how much of a change and I feel that a 20 change would be noticeable and important."* (US-02, change from 80 to 100, 20-point improvement)	10 or 20	*"There would be some gradual degradation over time, um, if we're talking about months, quarters, and years. Uh, and I should be mentally prepared for a physical appearance, you know, changing for the poorer."* (US-14, change from 75 to 65, 10-point worsening)
Future perspective	10	*"Um, maybe just that much a little closer to a longer remission than a 70."* (US-15, change from 70 to 80, 10-point improvement)	10 or 20	*"I think there would be more, more bad days than good days. Um, and, you know, not just physically hindering you but mentally, um, there's definitely times when I can't put my finger on it. And I don't want to say depressed, but there's, there's times where you do feel down, but you don't know why.. Um, including the future."* (US-18, change from 80 to 70, 10-point worsening)

to patients during the development of the EORTC QLQ-MY20, they did not seem to be concepts which patients would expect to change with treatment. This may reflect disease defining impact concepts, that is, core symptoms or impacts (namely Disease Symptoms and Side Effects of Treatment), compared to more distal disease impact concepts, that is, additional symptoms or general impacts (Body Image and Future Perspectives).

3.4 | Triangulation

Results of the MID analyses were triangulated by plotting all estimates on a forest plot (Figure 1); the anchor-based estimates were also summarized using a correlation-weighted average. Inspection of the plots and weighted estimates yielded the following recommendations for MID: Disease Symptoms (10 points), Side Effects of Treatment (10 points), Body Image (13 points) and Future Perspective (9 points).

Results of the RD anchor-based analyses and distribution-based analyses were triangulated with the qualitative interview findings to converge on a recommended value or small range of values for each scale (Figure 2). Recommended meaningful change estimates were

as follows: Disease Symptoms (16 improvement; 11 worsening), Side Effects of Treatment (6 improvement; 9 worsening), Body Image (33 improvement; 33 worsening), Future Perspective (11 improvement; 11 worsening).

4 | DISCUSSION

The novel mixed-methods approach allowed estimation of both MIDs and RDs for each scale of the EORTC QLQ-MY20. Previous studies had relied on distribution-based estimates alone to aid interpretation of scores. The availability of data from three pooled trials to derive multiple anchor and distribution-based estimates is a key strength of this analysis. The data included both newly diagnosed and relapsed myeloma patients and combined estimates from both mid- and end-of-treatment points, aiding the generalizability of these estimates to future studies.

It is acknowledged that there are limitations of the study. All three data sources were from trials with defined eligibility criteria; thus, results may not be fully generalizable to the wider MM population
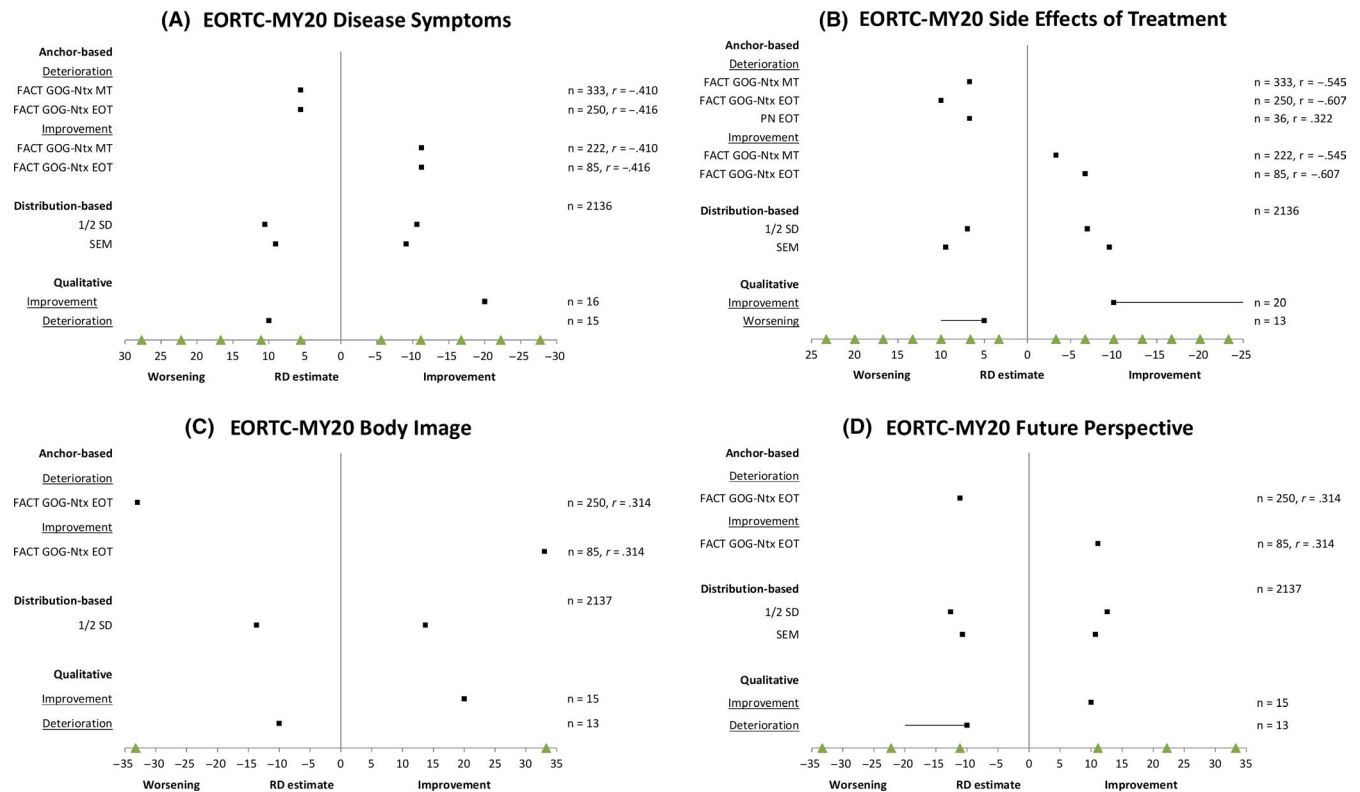
FIGURE 2 Results of the responder definitions: anchor-based analyses, distribution-based analyses, and qualitative-based analyses [Colour figure can be viewed at wileyonlinelibrary.com]

seen in clinical practice. Additionally, a number of the potential anchors did not correlate sufficiently with the PRO scores for analysis. Anchor-based estimates were therefore based mainly on the patient-reported neurotoxicity (FACT-GOG-Ntx) and peripheral neuropathy adverse events, which were collected in two out of the three pooled trials. Ideally, broader anchors would have been found but this may largely reflect the nature of what the disease-specific EORTC QLQ-MY20 is measuring. Finally, the trials did not contain any patient-reported rating of their own change which is often utilized to estimate the MID in prospective studies.[28] Often, however, these global ratings of change anchors are not highly correlated with actual score change;[29,30] thus, the availability of patient opinions through prospective interviews alongside the trial data is of importance.

The level at which meaningful change is discussed with patients must map on to the scale in order to perform triangulation. However, exploring meaningful change at the scale level can be challenging, as patients must consider multiple items when generating an estimate. In the current study, a stepwise approach was utilized to facilitate this process. This scaffolding approach (stepwise learning strategy) worked well and helped patients understand and articulate the concept of meaningful change before moving onto the scale level (0-100 scale). It is acknowledged that while it was still challenging for some patients to discuss meaningful change at the scale level this approach led to more confidence in the estimates that were provided.

The methodology used here is novel and challenges remain associated with collecting, analyzing and interpreting qualitative data

related to meaningful change thresholds. Relevance of the EORTC QLQ-MY20 Side Effects of Treatment and Body Image items to patients impacted the meaningful change estimates provided. Patients relied on their own experiences and if the items from the EORTC QLQ-MY20 were not relevant to them (eg, a side effect not experienced) then they felt unable to talk about meaningful change for that scale or provided large estimates (eg, where Body Image was not important to them).

There was variability in the scale level estimates provided by patients, which made it challenging to narrow to a single value or narrow range. For the Side Effects of Treatment scale, an equal number of patients suggested 10, 20 or 30 points would constitute a meaningful improvement. Patient's reasoning was very similar across the different estimates, namely reduction in the frequency and severity of side effects and general improvement in well-being. When it was challenging to draw conclusive evidence from qualitative data, priority was given to the anchor-based and distribution-based estimates when triangulating the estimates. Generally, estimates for the RD from both the patient interviews and from the clinical data had limitations. While patient estimates could be very variable as discussed above, the ROC curves also had a lower than ideal AUC. The recommended threshold was not met; however, this threshold was recommended based on applications to individual patient care (ie, clinical practice).[25] In a clinical trial, where the RD estimates will be used for group-based analysis, it may be that a greater extent of misclassification is acceptable compared to evaluating a single patient. Despite the limitations, the RD estimates provided are still

an improvement on the previously used thresholds which were generally based on a global estimate across scales, without consideration of the underlying scale and whether an individual patient could achieve that level of change. Additionally, variability on RD estimates should not be viewed as a weakness, given current regulatory opinion that a range of appropriate thresholds for meaningful within-patient change can be used in practice.[31]

The SEM estimates originally used to interpret the ASPIRE, CLARION, and ENDEAVOR trials were 9-10 for Disease Symptoms, 7 for the Side Effects and 10-11 for Future Perspective. SEM estimates in this study remained similar except for the Side Effects scale which was slightly higher (9.5). The SEM was not estimated for the Body Image scale as it comprises only one item and Cronbach's alpha could not be computed; future studies assessing test-retest reliability could use this coefficient to derive the SEM.

Significantly more US patients (n = 18) were recruited into the study than UK patients (n = 2). However, it was not anticipated that there would be differences between US and UK patients, in terms of meaningful change estimates.

5 | CONCLUSION

Findings from this study addressed the objective to recommend MIDs and RDs for each scale of the EORTC QLQ-MY20 for use in MM patients. To our knowledge, this study is the first application of a mixed-methods approach to establish meaningful change estimates using both existing clinical trial data and prospective patient interviews. This integration allowed a more thorough exploration of meaningful change than if performing qualitative or quantitative research alone. Published estimates of the MID and RD will enable other users of the EORTC QLQ-MY20 to adopt the same estimates as standard for interpretation, making different study analyses more comparable for all stakeholders including patients, doctors, researchers, sponsors, regulators, and payers.

CONFLICT OF INTEREST

The authors report no other conflicts of interest in this work.

ORCID

Andrew Trigg https://orcid.org/0000-0002-1613-8042

Nicola Bonner https://orcid.org/0000-0002-9810-9893

Nina Shah https://orcid.org/0000-0002-1971-8173

Emre Yucel https://orcid.org/0000-0002-6036-9526

Kim Cocks https://orcid.org/0000-0003-2595-708X

## REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359-E386.
2. Stewart AK, Dimopoulos MA, Masszi T, et al. Health-related quality-of-life results from the open-label, randomized, phase III ASPIRE trial evaluating carfilzomib, lenalidomide, and dexamethasone versus lenalidomide and dexamethasone in patients with relapsed multiple myeloma. *J Clin Oncol*. 2016;34(32):3921-3930.
3. Leleu X, Kyriakou C, Broek IV, et al. Prospective longitudinal study on quality of life in relapsed/refractory multiple myeloma patients receiving second-or third-line lenalidomide or bortezomib treatment. *Blood cancer journal*. 2017;7(3):e543.
4. Mortensen G, Salomo M. Quality of life in patients with multiple myeloma: a qualitative study. *J Cancer Sci Ther*. 2016;8:289-293.
5. Sonneveld P, Verelst S, Lewis P, et al. Review of health-related quality of life data in multiple myeloma patients treated with novel agents. *Leukemia*. 2013;27(10):1959.
6. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10(4):407-415.
7. Uryniak T, Chan I, Fedorov VV, et al. Responder Analyses—A PhRMA Position Paper. *Stat Biopharm Res*. 2011;3(3):476-487.
8. Kvam AK, Fayers P, Wisloff F. What changes in health-related quality of life matter to multiple myeloma patients? A prospective study. *Eur J Haematol*. 2010;84(4):345-353.
9. Ousmen A, Touraine C, Deliu N, et al. Distribution-and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. *Health Qual Life Outcomes*. 2018;16(1):228.
10. Facon T, Lee JH, Moreau P, et al. Phase 3 study (CLARION) of carfilzomib, melphalan, prednisone (KMP) v bortezomib, melphalan, prednisone (VMP) in newly diagnosed multiple myeloma (NDMM). *Clin Lymphoma Myeloma Leuk*. 2017;17(1):e26-e27.
11. Lord FM. *Applications of item response theory to practical testing problems*. London, UK: Routledge; 2012.
12. Dimopoulos MA, Delforge M, Hajek R, et al. Lenalidomide, melphalan, and prednisone, followed by lenalidomide maintenance, improves health-related quality of life in newly diagnosed multiple myeloma patients aged 65 years or older: results of a randomized phase III trial. *Haematologica*. 2013;98(5):784-788.
13. Ludwig H, Moreau P, Dimopoulos MA, et al. Health related quality of life results from the open-label, randomized, phase III endeavor trial evaluating carfilzomib and dexamethasone versus bortezomib and dexamethasone in patients with relapsed or refractory multiple myeloma. *Am Soc Hematology*. 2016;34:3921-3930.
14. Leleu X, Kyriakou C, Broek IV, et al. Continued treatment duration, drug dosing and health-related quality of life (HRQoL) of patients with relapsed/refractory multiple myeloma (RRMM) receiving 2nd and 3rd line treatments: results from a European Multicentre Study. *Am Soc Hematology*; 2013;122:5368.
15. 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):102-109.
16. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res*. 2011;11(2):171-184.
17. Gelhorn HL, Kulke MH, O'Dorisio T, et al. Patient-reported symptom experiences in patients with carcinoid syndrome after participation in a study of telotristat etiprate: a qualitative interview approach. *Clin Ther*. 2016;38(4):759-768.
18. Brigden A, Parslow RM, Gaunt D, Collin SM, Jones A, Crawley E. Defining the minimally clinically important difference of the SF-36 physical function subscale for paediatric CFS/ME:
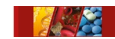
triangulation using three different methods. *Health Qual Life Outcomes*. 2018;16(1):202.

19. Stead M, Brown J, Velikova G, et al. Development of an EORTC questionnaire module to be used in health-related quality-of-life assessment for patients with multiple myeloma. *Br J Haematol*. 1999;104(3):605-611.

20. Cocks K, Cohen D, Wisløff F, et al. An international field study of the reliability and validity of a disease-specific questionnaire module (the QLQ-MY20) in assessing the quality of life of patients with multiple myeloma. *Eur J Cancer*. 2007;43(11):1670-1678.

21. Dimopoulos MA, Moreau P, Palumbo A, et al. Carfilzomib and dexamethasone versus bortezomib and dexamethasone for patients with relapsed or refractory multiple myeloma (ENDEAVOR): a randomised, phase 3, open-label, multicentre study. *Lancet Oncol*. 2016;17(1):27-38.

22. Facon T, Lee JH, Moreau P, et al. Randomized phase 3 study of carfilzomib or bortezomib with melphalan-prednisone for transplant-ineligible, NDMM patients. *Blood*. 2019;133:1953-1963. blood-2018-2009-874396.

23. Musoro ZJ, Hamel J-F, Ediebah DE, et al. Establishing anchor-based minimally important differences (MID) with the EORTC quality-of-life measures: a meta-analysis protocol. *BMJ Open*. 2018;8(1):e019117.

24. Froud R, Abel G. Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: The forgotten lesson of Pythagoras. Theoretical considerations and an example application of change in health status. *PLoS ONE*. 2014;9(12):e114468.

25. de Vet HC, Terluin B, Knol DL, et al. Three ways to quantify uncertainty in individually applied "minimally important change" values. *J Clin Epidemiol*. 2010;63(1):37-45.

26. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. *Value Health*. 2011;14(8):978-988.

27. Harper A, Trennery C, Sully K, Trigg A. *Triangulating estimates of meaningful change or difference in patient-reported outcomes: application of a correlation-based weighting procedure*. Paper presented at: Quality of Life Research; 2018.

28. 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(3):207-221.

29. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *J Clin Epidemiol*. 1997;50(8):869-879.

30. Schmitt J, Di Fabio RP. The validity of prospective and retrospective global change criterion measures. *Arch Phys Med Rehabil*. 2005;86(12):2270-2276.

31. FDA. Document for Patient-Focused Drug Development Public Workshop on Guidance 3: Select, Develop or Modify Fit-For-Purpose. Clinical Outcome Assessments. 2018; https://www.fda.gov/media/116277/download

32. Bedard G, Zeng L, Zhang L, et al. Minimal important differences in the EORTC QLQ-C30 in patients with advanced cancer. *Asia-Pacific J Clin Oncol*. 2014;10(2):109-117.

33. Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. *Eval Health Prof*. 2005;28(2):172-191.

34. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six PROMIS-Cancer scales in advanced-stage cancer patients. *J Clin Epidemiol*. 2011;64(5):507.

35. Maringwa JT, Quinten C, King M, 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):1753-1760.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.