JNCI Cancer Spectrum (2019) 3(3): pkz037

doi: 10.1093/jncics/pkz037
First published online June 4, 2019
Article

ARTICLE

Minimally Important Differences for Interpreting EORTC QLQ-C30 Scores in Patients With Advanced Breast Cancer

Jammbe Z. Musoro*, Corneel Coens, Frederic Fiteni, Pogoda Katarzyna, Fatima Cardoso, Nicola S. Russell, Madeleine T. King, Kim Cocks, Mirjam Ag Sprangers, Mogens Groenvold, Galina Velikova, Hans-Henning Flechtner, Andrew Bottomley; on behalf of the EORTC Breast and Quality of Life Groups

See the Notes section for the full list of authors' affiliations.
*Correspondence to: Jammbe Musoro, PhD, European Organization for Research and Treatment of Cancer, EORTC Headquarters, 83/11 Avenue E. Mounier, 1200 Brussels, Belgium (e-mail: jammbe.musoro@eortc.org).

Abstract

Background: We aimed to estimate the minimally important difference (MID) for interpreting group-level change over time, both within a group and between groups, for the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire core 30 (EORTC QLQ-C30) scores in patients with advanced breast cancer.
Methods: Data were derived from two published EORTC trials. Clinical anchors (eg, performance status [PS]) were selected using correlation strength and clinical plausibility of their association with a particular QLQ-C30 scale. Three change status groups were formed: deteriorated by one anchor category, improved by one anchor category, and no change. Patients with greater anchor changes were excluded. The mean change method was used to estimate MIDs for within-group change, and linear regression was used to estimate MIDs for between-group differences in change over time. For a given QLQ-C30 scale, MID estimates from multiple anchors were triangulated to a single value via a correlation-based weighted average.
Results: MIDs varied by QLQ-C30 scale, direction (improvement vs deterioration), and anchor. MIDs for within-group change ranged from 5 to 14 points (improvement) and −14 to −4 points (deterioration), and MIDs for between-group change over time ranged from 4 to 11 points and from −18 to −4 points. Correlation-weighted MIDs for most QLQ-C30 scales ranged from 4 to 10 points in absolute values.
Conclusions: Our findings aid interpretation of changes in EORTC QLQ-C30 scores over time, both within and between groups, and for performing more accurate sample size calculations for clinical trials in advanced breast cancer.

Patient-reported outcomes such as health-related quality of life (HRQOL) are increasingly assessed as important endpoints in cancer clinical trials. As a result, there is growing interest to improve the interpretation of HRQOL data in cancer clinical trials (1). It is recognized that interpreting HRQOL scores merely via statistical significance might be misleading because small differences in mean scores can be statistically significant, even when clinical relevance is absent. The minimally important difference (MID) approach aids interpreting differences and changes in HRQOL scores as clinically meaningful (2–7). MID can be defined as the smallest change in a HRQOL score that is perceived as "important" by a patient or by a third party (eg, a

Received: March 5, 2019; **Revised:** April 24, 2019; **Accepted:** May 20, 2019

clinician), which may indicate a change in the patient's management (2).

MIDs are commonly estimated using anchor-based and distribution-based methods (7). Anchor-based methods express differences or change in HRQOL scores using other familiar variables that have clinical relevance (3,7–9) or to patient and/or physician-derived ratings of change in the specific domain (4–6). Distribution-based methods use the statistical distribution of HRQOL scores (eg, SD criteria or SEM) and are considered as supportive evidence to anchor-based methods (10).

This study focused on interpreting the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire core 30 (EORTC QLQ-C30) in patients with advanced breast cancer. Guidelines for interpreting the QLQ-C30 were initially published by King (3) and Osoba et al. (4). King (3) evaluated published evidence about differences in QLQ-C30 scores between groups for multiple cancer sites and clinical anchors and found that the score range for small, moderate, and large effects differed between the scales of the QLQ-C30. Osoba et al. (4) provided thresholds for interpreting small (5 to 10 points), moderate (10 to 20 points), and large changes (>20 points) in QLQ-C30 scores using a global rating of change in metastatic breast and small-cell lung cancer patients. Based on King (3) and Osoba et al. (4), mean differences no less than 10 points are widely regarded as clinically meaningful for the QLQ-C30 in randomized clinical trials (11). However, recent guidelines revealed that MIDs can differ by QLQ-C30 scale, direction of change (improvement vs deterioration), and settings (5, 6), rendering a widely applicable rule for MIDs highly unlikely. We therefore need to gather further empirical evidence on patterns of MIDs across QLQ-C30 scales and disease sites (12).

This study examined MIDs for group-level change in HRQOL scores over time. In contrast to Osoba et al. (4), we used available clinical anchors in the database. Furthermore, the guidelines of King (3) and Cocks et al. (5, 6) were based on meta-analyses of published studies, pooling across cancer sites, whereas this study used individual patient data from archived trials.

Methods

Data Description

Data were derived from two published phase III EORTC trials. Trial 1 assessed the clinical benefit of a dose-intensive anthracycline-based regimen compared with standard treatment in women with locally advanced breast cancer and enrolled 448 patients (13). Trial 2 compared a combination of doxorubicin and paclitaxel vs doxorubicin and cyclophosphamide as first-line chemotherapy in advanced (metastatic) breast cancer and enrolled 275 patients (14). Both trials assessed HRQOL using the QLQ-C30 at baseline, during treatment, and at several follow-up time points after the end of treatment.

The EORTC QLQ-C30

The EORTC QLQ-C30 comprises 30 items, 24 of which are aggregated into 9 multi-item scales: 5 functioning scales (physical, role, cognitive [CF], emotional, and social); 3 symptom scales (fatigue, pain, and nausea and/or vomiting); and 1 global health-status scale. The remaining six single items assess symptoms of dyspnea, appetite loss (AP), sleep disturbance, constipation, diarrhea, and financial impact. Both trials used version 2 of the QLQ-C30, with standard scoring applied to the scales (15). For consistency in signs, all scales were scored such that 0 represents the worst possible score and 100, the best possible score. Financial impact was omitted from the analysis because suitable anchors were not available.

Clinical Anchor

Anchors were selected from variables that were available in the trial datasets (eg, physician examinations and common terminology criteria for adverse events [CTCAE]). Anchors were selected for each HRQOL scale based on correlation strength. Spearman rank, polyserial, or polychoric correlation was used, depending on the distribution of the pair of variables. Anchors with correlations no less than 0.30 were prioritized (10), and where achievable, anchors with much stronger correlations were targeted. The retained anchors were further verified for clinical plausibility by a panel of six breast cancer and/or HRQOL experts to avoid spurious findings. Multiple anchors could be selected for each HRQOL scale (12).

For trial 1, the retained anchors comprised 1) World Health Organization PS, scored between 0 (completely active with no limitations) and 4 (bedbound); and 2) four CTCAEs (nausea, vomiting, fatigue, and alopecia), graded between 0 (no toxicity) to 4 (life-threatening). The only anchor retained for trial 2 was the PS.

Definition of Clinical Change Groups

Three clinical change status groups (CCG) were defined: deterioration (worsened by 1 anchor category), stable (no change in anchor category), and improvement (improved by 1 anchor category). In order not to overestimate the MIDs, change scores no less than a 2-point change in anchor categories were excluded from datasets used to estimate MIDs because they were considered to be above the "minimal" expected change.

Statistical Analysis

Anchor-Based Methods

Change scores of HRQOL scale and anchor pairs were computed across all pairwise time points and combined to provide sufficient data for examining clinically important changes. For example, for a subject measured at time points t_a, t_b, and t_c, change scores were computed between t_a and t_b; t_a and t_c; and t_b and t_c. Hence, a subject can contribute multiple change scores, and given their change scores, subjects can contribute to multiple CCGs. Only subjects with HRQOL and anchor data for a given pair of time points contributed to the calculation of change scores. Data from the two trials were pooled to estimate MIDs.

The mean change method was used to estimate MIDs for within-group change over time. MIDs for improvement and deterioration were computed as the mean HRQOL change scores for the improvement and deterioration CCGs, respectively. This is relevant for interpreting change within a single group of patients, and it is similar to the mean HRQOL change score over time for a treatment group in a trial. Effect sizes (ESs) were computed within each CCG by dividing the mean of the HRQOL change scores (derived from all the pairwise time point differences) by the SD of the HRQOL change scores over all time points. Only mean changes with an ES no less than 0.2 and less than 0.8 were considered appropriate for inclusion as MIDs. This was based on Cohen's (16) recommendations that an ES of 0.2 is small, 0.5 is moderate, and no less than 0.8 is large. The

Table 1. Baseline demographic and clinical characteristics of the patients by study (all patients had advanced breast cancer)

Characteristic	Study 10921 No. (%) (N = 448)	Study 10961 No. (%) (N = 275)	Total (N = 723)
Performance status			
0	394 (87.9)	119 (43.3)	513 (71.0)
1	54 (12.1)	133 (48.4)	187 (25.9)
2	0 (0.0)	22 (8.0)	22 (3.0)
Unknown	0 (0.0)	1 (0.4)	1 (0.1)
Number of positive nodes			
N_0–N_1	250 (55.8)	144 (52.4)	394 (54.5)
N_2	176 (39.3)	26 (9.5)	202 (27.9)
N_{4+}	0 (0.0)	51 (18.5)	51 (7.1)
N_x	9 (2.0)	41 (14.9)	50 (6.9)
N_3	13 (2.9)	13 (4.7)	26 (3.6)
Country			
France	97 (21.7)	41 (14.9)	138 (19.1)
Netherlands	41 (9.2)	42 (15.3)	83 (11.5)
United Kingdom	11 (2.5)	68 (24.7)	79 (10.9)
Poland	78 (17.4)	0 (0.0)	78 (10.8)
Belgium	48 (10.7)	29 (10.5)	77 (10.7)
Canada	68 (15.2)	0 (0.0)	68 (9.4)
Slovenia	22 (4.9)	26 (9.5)	48 (6.6)
Switzerland	28 (6.3)	8 (2.9)	36 (5.0)
Russia	27 (6.0)	0 (0.0)	27 (3.7)
Italy	0 (0.0)	18 (6.5)	18 (2.5)
Israel	0 (0.0)	16 (5.8)	16 (2.2)
South Africa	3 (0.7)	12 (4.4)	15 (2.1)
Portugal	13 (2.9)	0 (0.0)	13 (1.8)
Czech Republic	12 (2.7)	0 (0.0)	12 (1.7)
Spain	0 (0.0)	9 (3.3)	9 (1.2)
Austria	0 (0.0)	6 (2.2)	6 (0.8)
Age, y			
Mean (SD)	50.07 (9.68)	52.27 (9.61)	—
Range	26.0–79.0	28.0–70.0	—

rationale was that ESs less than 0.2 reflect changes that are clinically unimportant, and those no less than 0.8 are obviously more than minimally important. The difference in change scores between the improvement (or deterioration) CCG and no change CCG was compared using analysis of variance (ANOVA).

A linear regression was used to estimate MIDs for differences between groups in change over time. For a given HRQOL scale and anchor pair, the outcome variable was the HRQOL change score, and the covariate was a binary anchor variable (coded as stable = 0 and improvement = 1 when modeling improvement [deteriorated observations were excluded], and stable = 0 and deterioration = 1 when modeling deterioration [improved observations were excluded]).

Because change scores were computed across all pairwise time points, some patients contributed change scores to more than one CCG and more than one change score to a particular CCG. We corrected for the association between multiple change scores contributed by the same patients by specifying a suitable covariance structure using the generalized estimating equations (17). The slope parameters for the "improved" and "deteriorated" covariates correspond to the MID for improvement and deterioration, respectively. This approach is similar to comparing the mean HRQOL change score over time in a treatment group to a control group in a trial, which is why these MIDs are useful for interpreting changes over time between two distinct groups of patients. Furthermore, we compared the two trials by adding a "trial" effect in a linear regression model, separately for improving and deteriorating HRQOL scores. This was based on the data with PS as the anchor.

Both within-group and between-group MID estimates for a given HRQOL scale, from multiple anchors, were triangulated to a single value via a correlation-based weighted average.

Distribution-Based Methods

The SEM, 0.2 SD, 0.3 SD, and 0.5 SD were applied to HRQOL scores at two time points common to both trials: 1) start of treatment (t1), time point before or on the first day of treatment, and 2) end of treatment (t2), last day of protocol treatment. Test–retest reliability estimates to compute SEM for the QLQ-C30 were based on Hjermstad et al. (18). All analyses were performed using the SAS software (19).

Results

Table 1 summarizes the demographic and clinical characteristics of patients at baseline. The median follow-up time (in months) for HRQOL was 5.3 (16.9) for trial 1 and 1.6 (2.8) for trial 2. An overview of the flow of patients through this study is presented in Supplementary Figure 1 (available online). Cross-sectional correlations ranged from 0.20 to 0.62 in absolute value, with a majority of the correlation coefficients being above the 0.30 threshold (7) (Table 2). Correlations between the change scores ranged from 0.14 to 0.51. At least one suitable anchor was constructed for 8 of the 14 QLQ-C30 scales that were considered for this study. The distribution of patients and the number of change observations across the categories of suitable anchors are summarized in Supplementary Table 1 (available online).

Table 2. Correlations over all time points of the EORTC QLQ-C30 scale scores with suitable anchors, and correlations between change scores of the EORTC QLQ-C30 scales and anchors

Scale	Anchor	Scores		Change scores	
		n_1 (n_{1R})*	Correlation	n_2 (n_{2R})*	Correlation
PF	Performance status	587 (2922)	−0.52	548 (8508)	−0.30
	CTCAE fatigue	355 (2658)	−0.30	343 (11102)	−0.20
	CTCAE vomiting	355 (2656)	−0.30	343 (11077)	−0.25
RF	Performance status	587 (2922)	−0.54	547 (8520)	−0.20
SF	Performance status	594 (2890)	−0.34	545 (8390)	−0.20
	CTCAE fatigue	355 (2630)	−0.21	340 (10984)	−0.15
	CTCAE vomiting	355 (2628)	−0.25	340 (10959)	−0.20
CF	CTCAE fatigue	355 (2638)	−0.20	342 (11032)	−0.14
QL	CTCAE vomiting	355 (2628)	−0.39	341 (10892)	−0.30
	CTCAE nausea	355 (2628)	−0.39	341 (10892)	−0.30
	CTCAE alopecia	355 (2629)	−0.39	341 (10914)	−0.35
	Performance status	585 (2893)	−0.32	547 (8351)	−0.25
FA	Performance status	587 (2915)	−0.40	546 (8476)	−0.23
	CTCAE nausea	355 (2644)	−0.21	341 (11 014)	−0.15
	CTCAE vomiting	355 (2644)	−0.22	341 (11 014)	−0.16
NV	CTCAE nausea	355 (2654)	−0.60	343 (11 050)	−0.51
	CTCAE vomiting	355 (2654)	−0.62	343 (11 050)	−0.48
AP	CTCAE nausea	355 (2621)	−0.58	343 (10 816)	−0.44
	CTCAE vomiting	355 (2621)	−0.59	343 (10 816)	−0.48

*n_1 (n_{1R}) and n_2 (n_{2R}) can vary by anchor and EORTC QLQ-C30 scale. AP = appetite loss; CF = cognitive functioning; EORTC QLQ-C30 = European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire core 30; FA = fatigue; n_1 = number of patients with at least 1 matched EORTC QLQ-C30 and an anchor form; n_{1R} = number of repeated anchor and HRQOL matched forms across all subjects; n_2 = number of patients with at least 2 matched EORTC QLQ-C30 and an anchor form (at least 2 forms are needed to compute change scores); n_{2R} = number of repeated EORTC QLQ-C30 scale and anchor change scores across all subjects; NV = nausea and/or vomiting; PF = physical functioning; QL = global quality of life; RF = role functioning; SF = social functioning.

Table 3. Range of anchor-based MID estimates from the mean change method and linear regression

Scale	Mean change method*		Linear regression†	
	Improvement	Deterioration	Improvement	Deterioration
PF	7 to 10	−11 to −10	7 to 9	−10 to −8
RF	No MID	−6	No MID	−4
SF	7 to 9	−9 to −5	6 to 7	−11 to −5
CF	5	−4	4	−4
QL	10 to 14	−11 to −5	8 to 11	− 13 to −6
FA	8	−9 to −7	8	−8 to −6
NV	No MID	−12	No MID	−14
AP	No MID	−14	No MID	−18

*The mean change method is useful for interpreting within-group change over time. The symptom scores were reversed to follow the functioning scales interpretation (ie, 0 represents the worst possible score and 100 the best possible score); "no MID" is used where no MID estimate is available either because of the absence of a suitable anchor or ES was either <0.2 or ≥0.8. All of the ESs for the no change group were <0.2. AP = appetite loss; CF = cognitive functioning; ES = effect size; FA = fatigue; MID = minimally important difference; NV = nausea and/or vomiting; PF = physical functioning; QL = global quality of life; RF = role functioning; SF = social functioning.

†The linear regression is useful for interpreting between-group differences in change over time.

Table 3 shows the range of MIDs from the mean change method (useful for interpreting within-group change over time) and the linear regression (useful for interpreting between-group differences in change over time) for each HRQOL scale, across multiple anchors. Detailed results are presented in Supplementary Table 2 (available online).

The MIDs varied according to the scale, direction of change scores (improvement vs deterioration), and anchor (Figure 1). Estimates were always in the expected direction according to the anchor (ie, positive vs negative change scores within the improvement vs deterioration CCG, respectively). Statistically significant differences (ANOVA $P < .05$) were observed between the HRQOL change scores for all improvement and deterioration CCGs vs no change CCG.

MIDs for within-group change (based on the mean-change method) ranged from 5 to 14 points (improvement) and −14 to −4 points (deterioration), and MIDs for between-group change (based on the linear regression) ranged from 4 to 11 points and from −18 to −4 points (Table 3). For the majority of the QLQ-C30 scales, the estimated MIDs ranged from 4 to 10 points in absolute values. Adding a trial effect to the regression models showed no statistically significant differences in change scores between the two trials, hence, supporting the combination of the two trials.

The MIDs in Table 3 are summarized to single MID values per scale in Table 4 and ranged from 4 to 10 points in absolute values for most HRQOL scales. Table 4 also compares the anchor-based estimates to the distribution-based estimates at t1. The distribution-based estimates at t2 for each HRQOL scale were similar to t1, mostly within a less-than-1-point range. All anchor-based estimates were no less than 0.2 SD, with most estimates being less than 0.5 SD. The anchor-based estimates tended to be closer to both the 0.3 SD and the 1 SEM.

Discussion

This study examined MIDs for interpreting group-level change of EORTC QLQ-C30 scores over time in patients with advanced breast cancer. Anchors for each HRQOL scale were selected

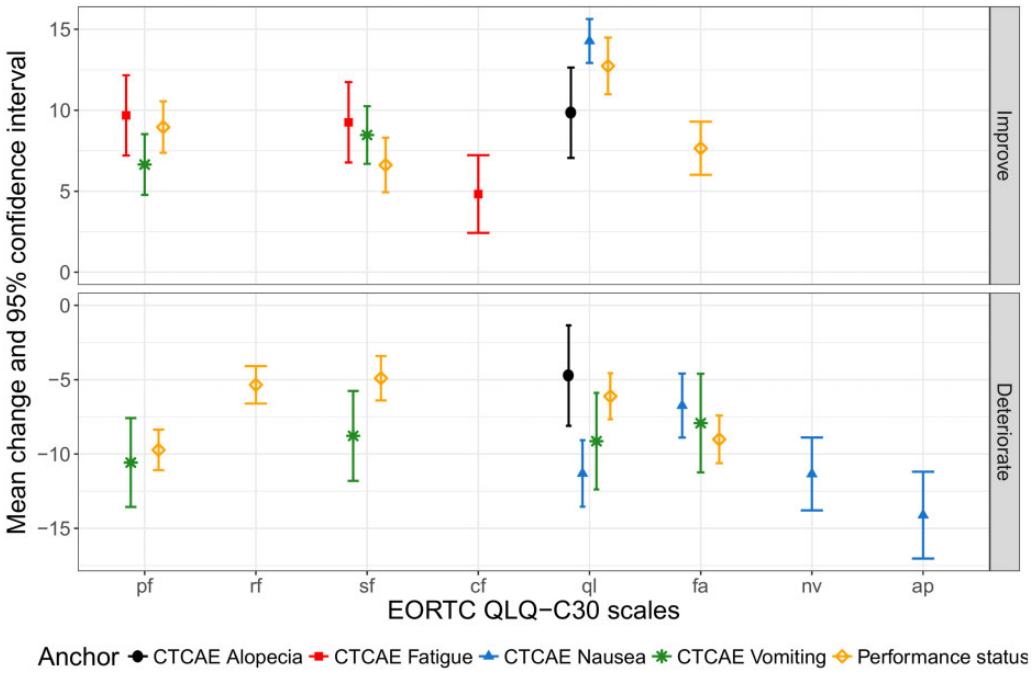

Figure 1. Mean change and 95% confidence interval for improvement and deterioration EORTC QLQ-C30 scales, across multiple anchors and averaged across different time periods. Estimates are available only for scales with at least 1 suitable anchor and with effect size ≥0.2 and <0.8 within the "deteriorate" and "improve" groups, respectively. These mean change scores are useful for interpreting within-group change over time. AP = appetite loss; CF = cognitive functioning; CTCAE = common terminology criteria for adverse events; deteriorate = worsened by 1 anchor category; EORTC QLQ-C30 = European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30; FA = fatigue; Improve = improved by 1 category; NV = nausea and/or vomiting; PF = physical functioning; QL = global quality of life; RF = role functioning; SF = social functioning.

Table 4. Summary of anchor-based MIDs for within- and between-group changes compared with distribution-based estimates

	Anchor-based MID for within-group change*		Anchor-based MID for between-groups difference in change*		Distribution-based QL scores at t_1 (n = 415–425)			
Scale	Improvement	Deterioration	Improvement	Deterioration	0.2 SD	0.3 SD	0.5 SD	1 SEM
PF	9	−10	8	−9	4.7	7.0	11.7	7.0
RF	No MID	−6	No MID	−4	5.1	7.6	12.7	10.7
SF	8	−7	7	−8	5.3	7.9	13.1	9.5
CF	5	−4	4	−4	4.1	6.2	10.3	8.8
QL	12	−8	10	−10	4.9	7.3	12.2	10.3
FA	8	−8	8	−7	4.9	7.3	12.2	10.0
NV	No MID	−11	No MID	−14	3.4	5.1	8.5	10.3
AP	No MID	−14	No MID	−18	5.2	7.8	13.1	12.0

*The within-group MIDs (from the mean change method) and the between-group MIDs (from the linear regression) were summarized via weighted averages based on scale and anchor pair correlation. The symptom scores were reversed to follow the functioning scales interpretation (ie, 0 represents the worst possible score and 100, the best possible score); "no MID" is used where no MID estimate is available either because of the absence of a suitable anchor or ES was either <0.2 or ≥0.8. = appetite loss; CF = cognitive functioning; ES = effect size; FA = fatigue; MID = minimally important difference; n = number of patients; NV = nausea/vomiting; PF = physical functioning; QL = global quality of life; RF = role functioning; SF = social functioning; t1 = time points for the start of treatment.

based on both the correlation strength and the clinical plausibility. When available, multiple anchors were used per HRQOL scale to provide some reassurance about the plausibility of the estimated MIDs. Despite the modest correlation between anchors and scales change scores, most MID estimates from multiple anchors were in a narrow range (often<5 points) and were always in the expected direction according to the anchor change category.

In agreement with recent findings (5–9), our estimates varied by HRQOL scale and direction of change (improvement vs deterioration). Similar to Maringwa et al. (8, 9). and Musoro et al. (7), no systematic differences were observed in the magnitude of change between deteriorating and improving scores. However, other studies reported that estimates for deterioration tended to be larger than those for improvement (6, 20).

We distinguished between MIDs for interpreting the degree of change within a group (obtained from the mean change method) and MIDs for interpreting the degree of differences between groups in within-group change (obtained from linear regression). Interestingly, estimates from both approaches were often in the same range. For many scales, the MIDs were within the range of 5–10 points that was suggested by Osoba et al. (4)

and also observed by Cocks et al. (5, 6), Musoro et al. (7), and Maringwa et al. (8, 9). However, similar to Cocks et al. (5, 6), we noticed that the thresholds for some scales were much lower. For example, MIDs of 4 points were observed for the CF scale. Musoro et al. (7) also reported MIDs that were as low as 3 points for the CF scale in patients with malignant melanoma. On the other hand, similar to Musoro et al. (7), we observed a much bigger threshold of 18 points for the AP scale. This reinforces the evidence that there is no single global standard for clinically meaningful change, and scale-specific MIDs should therefore be selected with more caution.

Most often, investigators seeking MIDs would desire simple guidelines. However, as shown in this article, results are often varied as a consequence of there being numerous anchors, various distribution-based criteria, and multiple HRQOL scales. Results shown in Figure 1 and Table 3 represent this diversity because the range of MIDs varies by the different anchors. We acknowledge that end users may find such a range of options confusing. So, to provide a single MID value per scale, we further simplified this by calculating a correlation-weighted average across multiple anchors. End users can choose to work with either the ranges provided in Table 3 or the single values provided in Table 4, whichever they feel most comfortable with. Most of the anchor-based estimates were closer to 0.3 SD and 1 SEM compared to the commonly reported 0.5 (21).

A limitation of this study is that suitable anchors were not always available, hence, anchor-based MIDs could not be estimated for seven of the EORTC QLQ-C30 scales, which were omitted in this article Furthermore, the available anchors (PS or CTCAE grades) relied exclusively on clinical observations or interpretations. Because the two trials that were used in this study evaluated chronic delivery of cytotoxic chemotherapy, clinical anchors such as CTCAE nausea, CTCAE vomiting, and CTCAE fatigue were reasonable and relevant. The availability of a pretreatment baseline assessment also allows detecting persistent effects such as alopecia. However, such anchors might not be relevant in other settings, treatments, or subtypes of breast cancer. The available anchors were also not necessarily suitable in all situations. For example, although CTCAE fatigue met the requirements of a plausible clinical relationship with the QLQ-C30 fatigue scale, the resulting correlations were too low (<0.1) to be retained. The low correlation can be explained by the discrete nature of the CTCAE scale where only a few high-grade events were scored. Moreover, because of the subjective nature of fatigue, there is likely also misrepresentation by physicians compared to patients' ratings as already reported by Basch et al. (22).This might also explain the potentially inflated MID estimates for the AP scale. Also, anchors that are based on patients' perspective of change (eg, subjective significance questionnaires) were not available in our study. Nonetheless, it is reassuring to notice the considerable overlap between our findings and those of Osoba et al. (4), which used patients' ratings of change as the anchor. Patients' self-assessed ratings across the different QLQ-C30 scales and across different disease sites are rarely available from retrospective data sources and would need to be planned as future research to complement our findings.

Another limitation is that our data originate from two controlled clinical trials, each with specific selection and treatment criteria. Although results were consistent between the trials, extrapolation beyond their specific setting should be made with caution. A number of articles are available that provide general guidelines for selecting MIDs for the QLQ-C30 scales (5, 6, 11). For instance, Cocks et al. (6) published MIDs for interpreting QLQ-C30 change scores over time for all 15 scales using published results from multiple cancer sites. The MID values obtained for the eight scales considered in this study were comparable to those presented by Cocks et al. (6). These increasingly robust guidelines advocate a more nuanced approach to clinical relevance beyond a single threshold.

In conclusion, our findings can help clinicians and researchers interpret the clinical relevance of group-level change of QLQ-C30 scores over time in patients with advanced breast cancer. The fact that MIDs can vary by QLQ-C30 scale and anchor suggests that we cannot rely on global standards for defining clinically meaningful change. Finally, our results will also inform more accurate sample size calculations for clinical trials in advanced breast cancer with endpoints that are based on EORTC QLQ-C30 scales.

Funding

This study was funded by the EORTC Quality of Life Group.

Notes

Affiliations of authors: European Organisation for Research and Treatment of Cancer, Brussels, Belgium (JZM, CC, AB); Department of Medical Oncology, University Hospital of Nîmes, France (FF); Institut de Recherche en Cancérologie de Montpellier, France (FF); University of Montpellier, France (FF); Maria Sklodowska-Curie Institute-Oncology Center, Warsaw, Poland (PK); Breast Unit, Champalimaud Clinical Centre, Champalimaud Foundation, Lisbon, Portugal (FC); Netherlands Cancer Institute, Amsterdam, The Netherlands (NSR); University of Sydney, Faculty of Science, School of Psychology, Sydney, NSW, Australia (MTK); Department of Health Sciences, University of York, York, UK (KC); Adelphi Values, Bollington, Cheshire, UK (KC); Department of Medical Psychology, Amsterdam University Medical Centers, Academic Medical Center, University of Amsterdam, Cancer Center Amsterdam, The Netherlands (MAS); Department of Public Health, University of Copenhagen, and Bispebjerg Hospital, Copenhagen, Denmark (MG); Leeds Institute of Cancer and Pathology, University of Leeds, St James's Hospital, Leeds, UK (GV); Clinic for Child and Adolescent Psychiatry and Psychotherapy, University of Magdeburg, Magdeburg, Germany (H-HF).

We thank the EORTC breast disease group members and their clinical investigators and all the patients who participated in the trials that we used for this analysis.

The authors declare no conflicts of interest.

References

1. Bottomley A, Flechtner H, Efficace F. Health related quality of life outcomes in cancer clinical trials. *Eur J Cancer*. 2005;41(12):1697–1709.
2. Schünemann HJ, Guyatt GH. Goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res*. 2005;40(2):593–597.
3. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual Life Res*. 1996;5(6):555–567.
4. Osoba D, Rodrigues G, Myles J, et al. Interpreting the significance of changes in health related quality-of-life scores. *J Clin Oncol*. 1998;16(1):139–144.
5. Cocks K, King MT, Velikova G, et al. Evidence-based guidelines for determination of sample size and interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. *J Clin Oncol*. 2011;29(1):89–96.
6. Cocks K, King MT, Velikova G, et al. Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. *Eur J Cancer*. 2012;48(11):1713–1721.

7. Musoro ZJ, Bottomley A, Coens C, et al. Interpreting European Organisation for Research and Treatment for Cancer Quality of Life Questionnaire core 30 scores as minimally importantly different for patients with malignant melanoma. *Eur J Cancer*. 2018;104(0):169–181.
8. 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.
9. Maringwa J, Quinten C, King M, et al. Minimal clinically meaningful differences for the EORTC QLQ-C30 and EORTC QLQ-BN20 scales in brain cancer patients. *Ann Oncol*. 2011;22(9):2107–2112.
10. 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.
11. Cocks K, King MT, Velikova G, et al. Quality, interpretation and presentation of European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 data in randomised controlled trials. *Eur J Cancer*. 2008;44(13):1793–1798.
12. 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.
13. Therasse P, Mauriac L, Welnicka-Jaskiewicz M, et al. Final results of a randomized phase III trial comparing cyclophosphamide, epirubicin, and fluorouracil with a dose-intensified epirubicin and cyclophosphamide + filgrastim as neoadjuvant treatment in locally advanced breast cancer: an EORTC-NCIC-SAKK multicenter study. *J Clin Oncol*. 2003;21(5):843–850.
14. Biganzoli L, Cufer T, Bruning P, et al. Doxorubicin and paclitaxel versus doxorubicin and cyclophosphamide as first-line chemotherapy in metastatic breast cancer: the European Organization for Research and Treatment of Cancer 10961 Multicenter Phase III Trial. *J Clin Oncol*. 2002;20(14):3114–3121.
15. Fayers P, Aaronson NK, Bjordal K, et al. *EORTC QLQ-C30 Scoring Manual (Third edition)*. Brussels, Belgium; EORTC Quality of Life Group; 2001.
16. Cohen J. *Statistical Power Analysis for the Behavioural Sciences (2nd Edition)*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
17. Liang KY, Zeger SL. Regression analysis for correlated data. *Annu Rev Public Health*. 1993;14(1):43–68.
18. Hjermstad MJ, Fossa SD, Bjordal K, Kaasa S. Test/retest study of the European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire. *J Clin Oncol*. 1995;13(5):1249–1254.
19. Institute Inc. *Base SAS® 9.4 Procedures Guide*. Cary, NC: SAS Institute Inc; 2013.
20. Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. *Cancer*. 2007;110(1):196–202.
21. 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.
22. Basch E, Dueck AC, Rogak LJ, et al. Feasibility assessment of patient reporting of symptomatic adverse events in multicenter cancer clinical trials. *JAMA Oncol*. 2017;3(8):1043–1050.