
Article:
Musoro, JZ, Coens, C, Singer, S et al. (12 more authors) (2020) Minimally important differences for interpreting European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 scores in patients with head and neck cancer. Head & Neck, 42 (11). pp. 3141-3152. ISSN 1043-3074

https://doi.org/10.1002/hed.26363

Minimally important differences for interpreting European Organisation for Research and Treatment for Cancer (EORTC) Quality of life Questionnaire Core 30 scores in patients with head and neck cancer

Jammbe Z Musoro[1] PhD, Corneel Coens[1] MSc, Susanne Singer [2, 3] PhD, Silke Tribius [4] MD, PhD, Sjoukje F Oosting[5] MD, PhD, Mogens Groenvold[6] MD, PhD, Christian Simon[7] MD, Jean-Pascal Machiels[8] MD, PhD, Vincent Grégoire[9] MD, PhD, Galina Velikova[10] MD, PhD, Kim Cocks[11,12] PhD, Mirjam AG Sprangers[13] PhD, Madeleine T King[14] PhD, Andrew Bottomley[1] PhD on behalf of the EORTC Head and Neck and Quality of Life Groups

[1]European Organisation for Research and Treatment of Cancer (EORTC), Brussels, Belgium

[2]Institute of Medical Biostatistics, Epidemiology and Informatics, Division of Epidemiology and Health Services Research University Medical Centre Mainz, Germany

[3]University Cancer Centre Mainz, Germany

[4]Department of Radiation Oncology, Asklepios Hospital St. Georg, Hamburg, Germany

[5]University Medical Center Groningen, Department of Medical Oncology, University of Groningen, the Netherlands

[6]Department of Public Health, University of Copenhagen, and Bispebjerg Hospital, Copenhagen, Denmark

[7]Centre Hospitalier Universitaire Vaudois - Lausanne, Switzerland

[8] Institut Roi Albert II, Service d'oncologie médicale, Cliniques universitaires Saint-Luc and Institut de Recherche Clinique et Expérimentale, U C Louvain, Brussels, Belgium

[9]Radiation Oncology Department, Centre Léon Bérard, Lyon, France.

[10]Leeds Institute of Cancer and Pathology, University of Leeds, St James's Hospital, Leeds, UK.

[11]Department of Health Sciences, University of York, York, UK

[12]Adelphi Values, Bollington, Cheshire, UK

[13]Department of Medical Psychology, Amsterdam University Medical Centers, location Academic Medical Center, University of Amsterdam, Cancer Center Amsterdam, The Netherlands.

[14]University of Sydney, Faculty of Science, School of Psychology, Sydney, NSW, Australia

Corresponding Author:
Jammbe Musoro, Ph.D., European Organization for Research and Treatment of Cancer, EORTC Headquarters, 83/11 Avenue E. Mounier, 1200 Brussels, Belgium; Tel: +32 (0) 2 774 15 39; jammbe.musoro@eortc.org

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

ABSTRACT

Background: We aimed to estimate minimally important difference (MID) for interpreting group-level change over time for European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire core 30 (EORTC QLQ-C30) scores in head and neck cancer.

Methods: Data were derived retrospectively from two published EORTC trials. Clinical anchors were selected using correlation strength and clinical plausibility of the given anchor/QLQ-C30 scale pair. MIDs for within-group and between-group change were estimated via the mean change method and linear regression respectively. Distribution-based MIDs were also examined. MIDs for 2 of the scales, dyspnea and nausea/vomiting, are more uncertain considering their low correlations with the anchors.

Results: Anchor-based MIDs could be determined for deterioration in **7** of the 14 QLQ-C30 scales assessed, and in 3 scales for improvement. MIDs varied by scale, direction of change and anchor. Absolute MIDs values ranged from 5 to 15 points for within-group change and 4 to 12 for between-group change. Most MIDs were within 4 to 10 points.

Conclusions: Our findings, if confirmed, will aid interpreting changes in selected QLQ-C30 scale scores over time and inform sample size calculations in future clinical trials in head and neck cancer.

Keywords: Health-related quality of life (HRQOL), EORTC QLQ-C30, Minimally important difference (MID), Head and neck cancer

1. INTRODUCTION

Health-related quality of life (HRQOL) is now commonly included as an important endpoint in cancer clinical trials [1, 2]. It is acknowledged that interpreting HRQOL data merely through statistical significance might be misleading since small mean differences can be statistically significant, even when the differences are not clinically relevant. Thus, the minimally important difference (MID) approach is important for interpreting HRQOL data as clinically meaningful [3, 4, 5, 6, 7, 8]. 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 (e.g. an informed proxy or a clinician), which may indicate a change in the patient's management[3].

MIDs can be estimated via anchor-based and distribution-based methods [9]. Anchor-based methods express differences or change in HRQOL scores in terms of external variables that have clinical relevance [4, 8, 10, 11, 12] or in relation to patient/physician-derived ratings of change in the specific domain [5, 6, 7]. Distribution-based methods using for example certain proportions of the standard deviation (SD) and standard error of measurement (SEM) [13, 14], depend on the statistical distribution of HRQOL scores and are recommended by some investigators as supportive evidence to anchor-based methods [9].

Assessing the quality of life of patients with head and neck cancer is relevant for understanding the impact of the disease and its treatment on patients and patients' daily life [15]. The European Organisation for Research and Treatment for Cancer Quality of life Questionnaire core 30 (EORTC QLQ-C30) is widely used to assess HRQOL in cancer patients. There are currently no MID guidelines for the EORTC QLQ-C30 specific to head and neck cancer. This study aims to interpret the EORTC QLQ-C30 scales in patients with head and neck cancer. It is important to highlight that the QLQ-C30 assesses generic aspects of HRQOL and not head-and-neck specific dimensions. Addressing MIDs for the head and neck disease-specific module (EORTC HN43 [16]) is out of the scope of this paper.

MID guidelines for interpreting the EORTC QLQ-C30 were initially published by King [4] and Osoba et al. [5]. King [4] assessed published evidence about differences in EORTC 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 EORTC QLQ-C30. Osoba et al. [5] provided estimates for interpreting small (5 to 10 points), moderate (10 to 20 points) and large changes (> 20 points) in EORTC QLQ-C30 scores in patients with breast and small-cell lung cancer, using individual patient's ratings about the importance of changes in HRQOL. Based on King [4] and Osoba et al. [5], mean differences ≥10 points are commonly considered as clinically meaningful when interpreting EORTC QLQ-C30 scales in randomized clinical trials [16]. Nevertheless, recent guidelines show that MIDs can differ not only by the

particular EORTC QLQ-C30 scale, but also by direction of change (improvement versus deterioration) and clinical settings [6, 7]. This implies that a global rule for MIDs applicable to all settings is highly unlikely [9, 12, 18]; Therefore there is a need to gather further empirical evidence on patterns of MID estimates across EORTC QLQ-C30 scales and disease sites [19]. This study focuses on assessing MIDs for group-level change in HRQOL scores over time, both within a group and between groups, and differs from Osoba *et al.* [4] in that we used clinical anchors. Additionally, as opposed to the guidelines by King [5] and Cocks *et al.* [6, 7] that were based on meta-analyses of published studies, pooling across cancer sites, this study used individual patient data from archived EORTC trials. We also examined how anchor-based MIDS compared to MIDs that are based on commonly used distribution-based methods.

2. MATERIALS AND METHODS

Data description

The data were derived from two published EORTC trials in head and neck cancer. Trial 1 (EORTC 24954) compared sequential induction chemotherapy and radiotherapy versus alternating chemo-radiotherapy for patients with resectable cancers of the hypopharynx and larynx and enrolled 450 patients [20]. Trial 2 (EORTC 24971) compared neoadjuvant docetaxel (Taxotere) plus cisplatin plus 5-fluorouracil versus neoadjuvant cisplatin plus 5-fluorouracil in patients with locally advanced inoperable squamous cell carcinoma of the head and neck (n=358) [19]. Both trials assessed HRQOL using the EORTC QLQ-C30 at baseline, during treatment and on several follow-up time points after the end of treatment.

The EORTC QLQ-C30

The EORTC QLQ-C30 comprises 30 items that are aggregated into fifteen scales; nine multi-item scales, i.e. five functioning scales: physical (PF), role (RF), cognitive (CF), emotional (EF), and social (SF), three symptom scales: fatigue (FA), pain (PA), and nausea/vomiting (NV) and one global health status (QL) scale. The remaining six single items assess symptoms: dyspnoea (DY), appetite loss (AP), sleep disturbance (SL), constipation (CO), diarrhoea (DI) and financial impact (FI). Trial 1 used version 2 of the EORTC QLQ-C30, whereas trial 2 used version 3. The two versions differ only in the response categories of questions 1–5 (in the PF domain), coded as yes/no in version 2, whereas version 3 uses a four-point Likert scale ranging from ‘not at all’ to ‘very much’. The scoring of the EORTC QLQ-C30 scales was done according to the EORTC QLQ-C30 scoring manual [22], with the means of the raw scores for each scale transformed to fall between 0 and 100. For consistency in signs of the HRQOL change scores across the various scales, the symptom scores were reversed to follow the

functioning scales' interpretation, i.e. all scales were scored such that 0 represents the worst possible score and 100, the best possible score. The FI scale was omitted from the analysis because suitable anchors were not available.

Clinical anchor

Clinical anchors were selected from clinical data that were available in the two trial data sets, e.g. physician examinations, clinician-rated common terminology criteria for adverse events (CTCAE), clinician-rated performance status and laboratory results. Anchors were selected for each QLQ-C30 scale based on correlation strength. Depending on the distribution of the HRQOL scale/anchor pair, a Spearman's rank, polyserial or polychoric correlation was estimated. Anchors with correlations of $\geq|0.30|$ [9] were prioritized and where achievable, anchors with much stronger correlations were targeted [23]. For scales where the majority of the anchors did not reach 0.3 threshold, we selected anchors with a mixture of weak (<0.3) to optimal correlations. We also aimed for multiple anchors per HRQOL scale to provide some reassurance about the plausibility of the MID estimates

The retained anchors were verified for clinical plausibility by a panel of eight head and neck cancer / HRQOL experts to avoid spurious findings. The clinicians could suggest or request any excluded anchors. The final list of retained anchors were based on a consensus between the methodological and clinical panel.. Details on the anchor selection procedures have been described by Musoro et al. [19].

Definition of clinical change groups

Three clinical change status groups (CCGs) were defined after a systematic consultation with our panel of clinical experts: deterioration (worsened by 1 anchor category), stable (no change in anchor category) and improvement (improved by 1 anchor category). Patients who changed by 2 or more categories of an anchor were excluded from data sets used to estimate mean change and MIDs because they were considered to be above the 'minimal' expected change.

Data analysis

Anchor-based methods

HRQOL and anchor change scores were computed across all pairwise time points and then combined into one dataset to provide sufficient data for examining clinically important changes. For instance, if a subject is measured at time points t_a, t_b and t_c, change scores were computed between t_a & t_b, t_a & t_c and t_b & t_c. This implies that 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.

The mean change method was used to estimate MIDs for within-group change over time. With this approach, 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 group of patients, and it is similar to the mean HRQOL change score over time for a treatment group in a trial. We also compared the difference in change scores between the improvement (or deterioration) CCG and no change CCG using analysis of variance (ANOVA).

A linear regression approach was used to estimate MIDs for differences between groups in change over time. For a given HRQOL scale/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 modelling improvement (deteriorated observations were excluded) and 'stable' = 0 and 'deterioration' = 1 when modelling deterioration (improved observations were excluded). Since some patients contributed change scores to multiple CCGs, and more than one change score to a particular CCG, we corrected for the association between multiple change scores contributed by some patients (i.e. within subjects correlation) by specifying a suitable covariance structure using generalized estimating equations (GEE) [24, 25]. The resulting 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. Hence these MIDs are useful for interpreting changes over time between two distinct groups of patients. For a given HRQOL scale, the anchor-based MID estimates from multiple anchors were triangulated to a single value via a correlation-based weighted average.

In order to assess whether MIDs varied by age, gender, disease stage (based on the N classification) and study (trial 1 versus trial 2), we included these factors (one at a time) and their interaction term with the anchor variable in a regression model. Separate models were fitted for improving and deteriorating HRQOL scores. The model for improving scores excluded deteriorated observations and vice versa.

Distribution-based methods

For this approach, 0.3 SD, 0.5 SD and SEM were estimated at two time points that were common in both trials: *(i)* Start of treatment (t1); time point before or on the first day of treatment administration and *(ii)* end of treatment (t2); last day of protocol treatment

administration. The resulting estimates are compared with those from the anchor-based approach.

The effect size (ES) within each CCG was computed by dividing the mean of the HRQOL change scores (derived from all the pairwise time point differences) by the standard deviation (SD) of the HRQOL change scores over all time points. Only mean changes with an ES of ≥ 0.2 and <0.8 were considered appropriate for inclusion as MIDs. This was based on Cohen's[14] recommendations that an ES of 0.2 is small, 0.5 is moderate and ≥0.8 is large. The rationale here was that an observed ES <0.2 reflects changes that were clinically unimportant, and ESs ≥0.8 were obviously more than minimally important. All statistical analyses were performed using SAS software [26].

3. RESULTS

Table 1 presents a summary of demographic and clinical characteristics of patients at baseline per trial. The median follow-up time (in months) for HRQOL was 6.1 (SD = 14.2) for trial 1 and 1.6 (SD=4.9) for trial 2. An overview of the flow of patients through this study is presented on Figure 1.

A total of 35 potential clinical anchors were initially assessed for the QLQ-C30 scales. After prioritization on cross-sectional correlation, 5 to 7 anchors were preselected per HRQOL scale for review by the clinical panel. The majority of anchors that were considered implausible by the clinical panel had cross-sectional correlations of < 0.2. There were a few cases, for example performance status versus the pain scale, where the 0.3 correlation criteria was met but were excluded by the clinical panel. Table 2 presents the final list of retained anchors comprising WHO performance status (PS) and 4 CTCAEs (nausea, weight loss, gastrointestinal and pulmonary). PS was scored between 0 (no symptoms of cancer) and 4 (bedbound) while the CTCAEs were graded between 0 (no toxicity) to 4 (life-threatening). At least one clinical anchor was retained for 7 of the 14 scales (PF, RF, SF, QL, FA, NV and DY). Table 2 also provides estimates of cross-sectional correlations between the EORTC QLQ-C30 scale scores with their selected anchors (over all time points) and correlations between their change scores. The cross-sectional correlations between HRQOL scales and anchors ranged from 0.2 to 0.4 in absolute value, while the correlations between their change scores ranged from 0.1 to 0.3. The correlations (cross-sectional and change scores) between the NV and DY scales and their respective retained anchors were less than the 0.3 threshold.

The distribution of patients and the number of change observations across the categories of suitable anchors is presented in Table A.1. According to the anchors, there were relatively more patients who remained stable compared to patients who either improved or deteriorated. Table

3 shows results from the mean change method (for interpreting within-group change over time) and the linear regression (for interpreting between-group differences in change over time) for each HRQOL scale, along with the estimated ES within the various CCGs. The ES for most CCGs improvement across the various anchors were < 0.2 which was too small to meet our minimum requirement.

Results in Table 3 are further summarised in Table 4, where MID estimates are presented for only those scales for which the CCG has an ES of ≥ 0.2 and <0.8. Anchor-based MIDs were determined for deterioration in **7** of the 14 QLQ-C30 scales assessed, and in **3** scales for improvement. The MID estimates varied according to the scale, direction of change scores (improvement versus deterioration) and anchor. This is illustrated in Figure 2, in which estimates from the mean change method in Table 4 are plotted along with their 95% confidence intervals (CIs). The final anchor-based estimates (Table 4) were always in the expected direction according to the anchor, i.e. positive versus negative change scores within the improvement versus deterioration CCG, respectively. Except for the nausea and vomiting scale, statistically significant differences (ANOVA p-value <0.05) were observed between the HRQOL change scores for the improvement or deterioration CCGs (with ES ≥ 0.2 and < 0.8) and no change CCG.

Anchor-based MIDs for within-group change (based on the mean-change method) ranged from 5 to 15 points in absolute values and MIDs for between-group change (based on the linear regression) ranged from 4 to 12 points (Table 4). For the nausea and vomiting scale, the estimated MID values from two different anchors were summarised in to a single value by taking a correlation-weighted average across the two anchors. Generally, the estimated MIDs ranged from 4 to 10 points for the majority of the HRQOL scales. The interaction effects between the anchor and age, gender, disease stage and study respectively showed no statistically significant differences (results not shown). This suggests that the MIDs estimates did not depend on these factors. Table 4 also compares the anchor-based estimates to those from some commonly used distribution-based methods. Except for the role functioning and nausea and vomiting scales, the distribution-based estimates at t1 and t2 were very similar, often within a <1 point range. The majority of the anchor-based estimates were > 0.2 SD (Table 5). Estimates for the social functioning, global quality of life and nausea/vomiting scales tended to range between 0.3 SD and 0.5 SD. Estimates for the physical and role functioning scales were closer to 0.5 SD, while those for the dyspnoea scale were closer to 0.3 SD. The distribution-based estimates for all 14 scales are presented in Table 5.

4. DISCUSSION

This study investigated MIDs for interpreting group-level change of EORTC QLQ-C30 scores over time in patients with head and neck cancer. Anchor-based MIDs could be determined for deterioration in **7** of the 14 QLQ-C30 scales assessed, and in 3 scales for improvement. Similar to recent findings [6, 7, 8, 10, 11] in patients with other types of cancer, the estimated anchor-based MIDs varied according to EORTC QLQ-C30 scale and direction of change (improvement versus deterioration). In agreement with Cocks *et al.* [7], the estimates for deterioration tended to be larger than those for improvement. Cella *et al.*[27] and Ringash *et al.*[27, 28] observed the same pattern when examining MIDs for the Functional Assessment of Cancer Therapy questionnaires. However, other studies have reported no systematic differences in the magnitude of change between deteriorating and improving scores [8, 10, 11]

We differentiated between MIDs for interpreting within-group changes; obtained from the mean change method, and for interpreting the differences between groups (that is the stable CCG versus the improvement or deterioration CCGs) in within-group change; obtained from the linear regression. The estimates from both approaches were often in the same range. MID estimates for most scales were within the range of 5-10 points that was suggested by Osoba *et al.*[5] in patients with breast and small-cell lung cancer and also observed by Cocks *et al.* [6,7] in pooled data across multiple cancer sites, Musoro et al [8] in patients with malignant melanoma and Maringwa *et al.* [10, 11] in patients with lung and brain cancer respectively. However, it is important to note that Cocks *et al.* [6, 7] also highlighted that the thresholds for some scales could be lower in some settings. For example, Musoro *et al.* [8] reported MIDs that were as low as 3 points for the cognitive functioning scale in patients with malignant melanoma. We also observed much bigger threshold for deterioration for some scales, e.g. MIDs for fatigue and role functioning scales were around 15 points. A similar threshold was reported for the role functioning scale in patients who received adjuvant treatment for melanoma [8]. 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.

As a limitation, suitable clinical anchors were not always available in our study datasets, hence anchor-based MIDs could not be estimated for eight of the EORTC QLQ-C30 scales which were omitted in this study. Although we aimed for multiple anchors per scale for reassurance about the plausibility of our MID estimates, only one suitable anchor was retained for most scales, which was often WHO performance status (PS). PS is widely used in oncological trials to assess patients' general physical functioning, and has previously been shown to be correlated with HRQOL [4, 8, 10, 11]. Furthermore, the anchors that were used in our study relied exclusively on clinical observations or interpretations, and were not necessarily suitable in all situations.

For instance, although CTCAE fatigue met the requirements of a plausible clinical relationship with the EORTC QLQ-C30 fatigue scale, the resulting correlation between their change scores was 0.05, which was too low to be retained. The low correlation can be explained by the discrete nature of the CTCAE scale where only few high-grade events were scored. Moreover, due to the subjective nature of 'fatigue' there is likely also misrepresentation by the different physicians compared to patients' ratings as already reported by Basch *et al* [28].

Clinical variables that measure swallowing ability, such as the adverse event dysphagia, were suggested by the clinical experts as potentially good anchors for head and neck cancer patients. However, such variables were often collected in a time limited period or were reported in just a few patients, hence could not be used because of insufficient data.

Generally, we recognize that it is often challenging to obtain suitable clinical anchors from retrospective clinical trial data. Even when potentially suitable anchors are identified, their correlation with HRQOL scales are often undesirably low, with a majority of the patients often remaining in the stable clinical change group as seen in Table A.1 and also previously reported by others [8, 10, 11]. This often limits the data needed to calculate MIDs. Furthermore, as shown in Table 2, correlations between anchor and HRQOL change scores are often lower compared to cross-sectional scores, probably because the change variables are intrinsically more varying due to the compounding of measurement error. We also acknowledge that the low correlations , particularly for the dyspnea and nausea/vomiting scales, raise concerns about the plausibility of the selected anchor as well as the reliability of the estimated MIDs. We recognized that our data are limited and thus it is imperative to further compare our MID estimates, especially for dyspnea and nausea/vomiting scales, to those in future studies that use anchors with much stronger correlations.

Given these limitations in the ancho-based approach, it is informative to use distribution-based estimates as an independent way to confirm the plausibility of the numerical range of anchor-based MID estimates [9]. Most of the anchor-based MIDs in our study for were either close to or in the range of 0.3 SD and 0.5 SD which have been used to define MIDs in the literature [30]. In addition, anchors that are based on the patient's perspective of change (e.g. 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.*[5], which was based on using individual patients' ratings of change as anchor. Patients' self-assessed rating 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. It is important to note that our data are limited to two controlled clinical trials, each

with specific selection and treatment criteria. Thus, extrapolation beyond their specific setting should be done with caution.

This study combined data from two trials that used different versions of the EORTC QLQ-C30; trial 1 used version 2 and trial 2 used version 3. Although the scales were transformed to have values between 0 and 100, the PF scale of version 2 can only take a limited range of values compared to version 3. However, our findings suggested that the MIDs for PF in our study did not depend on the questionnaire version. It will be interesting to further investigate in a larger sample if these differences may affect MID estimates.

It is important to highlight that a number of articles are available that provide general guidelines for selecting MIDs for the EORTC QLQ-C30 scales [6, 7, 16]. Cocks *et al.* [7] provided MIDs for interpreting EORTC QLQ-C30 change scores over time for all 15 scales based on meta-analyses of published studies, pooling across multiple cancer sites. For the seven scales considered in this study, the estimated MID values were often within the threshold ranges presented by Cocks *et al.* [7]. These increasingly robust guidelines advocate a more nuanced approach to clinical relevance beyond a single threshold.

There is emerging interest in using HRQOL scores in monitoring and managing individual patients. Our MID estimates can be a useful starting point for defining cut-offs for individual-level change that are clinically meaningful for head and neck cancer patients. For example in a clinical trial, patients who change by the MID or more can be considered 'responders' and the proportion of responders can be compared between treatment arms, while in clinical practice, our MIDs can serve as screening thresholds for identifying patients with clinically important problems. However, two caveats apply to setting thresholds for use at individual level. First, the actual threshold for application to individuals needs to be chosen with knowledge of the underlying scores for each HRQOL scale, since not all MID values will translate into a plausible score for an individual to achieve. Any scale has a limited number of observable values; the values either side of the MID may be good candidates for individual thresholds, with selection of either the higher or lower scale value determined by study investigators depending on clinical context. Second, individual thresholds must be set above bounds of measurement error to, avoid false positive changes that might trigger unjustified clinical actions [18, 29]. Giesinger et al [32] have developed clinical thresholds for physical functioning, emotional functioning, fatigue and pain scales of the EORTC QLQ-C30 to aid individual-level and group-level interpretation in clinical practice. Their data comprised patients with diverse types of malignancies. Instead of change scores over time, these thresholds apply to values observed at singular visits.

In conclusion, our findings can help clinicians and researchers to interpret the clinical relevance of group-level change of selected EORTC QLQ-C30 scale scores over time in patients with head and neck cancer. We have provided MID estimates for interpreting changes in HRQOL

scores over time, particularly with respect to deterioration, for both within and between groups of patients. The MIDs for the dyspnea and nausea/vomiting scales in particular are more uncertain and require further empirical scrutiny. These findings, if confirmed, will allow more accurate sample size calculations for clinical trials in head and neck cancer with endpoints that are based on EORTC QLQ-C30 scales.

5. REFERENCES

1. Bottomley A, Flechtner H, Efficace F et al. Health related quality of life outcomes in cancer clinical trials. Eur J Cancer. 2005; 41: 1697-1709.
2. Zikos E, Coens C, Quinten C, et al. The Added Value of Analyzing Pooled Health-Related Quality of Life Data: A Review of the EORTC PROBE Initiative. J Natl Cancer Inst 2016;108(5):djv391
3. Schünemann HJ, Guyatt GH. Goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res. 2005; 40: 593-597.
4. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res. 1996; 5: 555-567.
5. 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: 139-144.
6. 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 2010; 29(1): 89–96.

7. 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. European Journal of Cancer (2012) 48, 1713– 1721.
8. 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. European Journal of Cancer (2018) 104, 169-181. doi.org/10.1016/j.ejca.2018.09.005
9. Revicki D, Hays RD, Cella D, Sloan J Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008; 61:102–109
10. Maringwa JT, Quinten C , King M, et al. on behalf of the EORTC PROBE project and the Lung Cancer Group. 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 Nov; 19(11):1753-60.
11. 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 Sep; 22(9):2107-12.
12. Cella D, Eton DT, Lai JS, et al. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of cancer therapy (FACT) Anemia and Fatigue scales. J Pain Symptom Manage. 2002; 24:547-561.
13. 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), 861–873.
14. Cohen J. Statistical Power Analysis for the Behavioural Sciences (2nd Edition). Lawrence Erlbaum Associates, NJ, USA (1988).
15. D. Clasen, J. Keszte, A. Dietz, et al. Quality of life during the first year after partial laryngectomy: longitudinal study Head Neck, 40 (6) (2018), pp. 1185-1195
16. Singer S, Amdal CD, Hammerlid E, et al. International validation of the revised European Organisation for Research and Treatment of Cancer Head and Neck Cancer Module, the EORTC QLQ-HN43: Phase IV. Head Neck. 2019; 41(6):1725-1737.
17. 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) 44:1793-1798.
18. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcome Res. 2011 Apr; 11(2):171-84.

19. 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 2017; 7:e019117. doi:10.1136/bmjopen-2017-019117
20. Lefebvre JL, Rolland F, Tesselaar M, et al.; EORTC Head and Neck Cancer Cooperative Group; EORTC Radiation Oncology Group. Phase 3 randomized trial on larynx preservation comparing sequential vs alternating chemotherapy and radiotherapy.J Natl Cancer Inst. 2009 Feb 4; 101(3):142-52. doi: 10.1093/jnci/djn460. Epub 2009 Jan 27.
21. Vermorken JB, Remenar E, van Herpen C, et al.; EORTC 24971/TAX 323 Study Group. Cisplatin, fluorouracil, and docetaxel in unresectable head and neck cancer. N Engl J Med. 2007 Oct 25; 357(17):1695-704.
22. Fayers P, Aaronson, NK, Bjordal K, Groenvold M, Curran D, and Bottomley A on behalf of the EORTC Quality of Life Study Group. EORTC QLQ-C30 Scoring Manual (Third edition). Brussels, EORTC Quality of Life Group, 2001.
23. Coon CD. Empirical Telling the Interpretation Story: The Case for Strong Anchors and Multiple Methods. 23rd Annual Conference of the International Society for Quality of Life Research, Copenhagen, Denmark, October 2016. Qual Life Res 25, 1, ab2, p: 1-2.
24. Liang KY, Zeger SL. Regression analysis for correlated data. *Annu. Rev.* Pub Health. 1993; 14:43–68.
25. Ying GS, Maguire MG, Glynn R, Rosner B. Tutorial on Biostatistics: Linear Regression Analysis of Continuous Correlated Eye Data. Ophthalmic Epidemiol. 2017; 24(2):130-140.
26. Institute Inc. 2013. Base SAS® 9.4 Procedures Guide. Cary, NC: SAS Institute Inc.
27. Cella D, Bullinger M, Scott C, Barofsky I, Clinical Consensus Meeting Group. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. Mayo Clin Proc. 2002; 77:384-92.
28. Ringash J, O'Sullivan B, Bejzak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. Cancer 2007; 110:196–202.
29. 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.
30. 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.
31. King MT, Dueck AC, Revicki DA. Can methods developed for interpreting group-level patient-reported outcome data be applied to individual patient management? Med Care. 2019;57(supp 1):S38–S45.

32. Giesinger JM, Kuijpers W, Young T, et al. Thresholds for clinical importance for four key domains of the EORTC QLQ-C30: physical functioning, emotional functioning, fatigue and pain. Health Qual Life Outcomes. 2016;14:87. Published 2016 Jun 7. doi:10.1186/s12955-016-0489-4

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

Competing interests: The authors declare no conflicts of interest.

Table 1: Baseline demographic and clinical characteristics of the patients by study

	Study		
	Trial 1 (N=450)	**Trial 2 (N=358)**	**Total (N=808)**
	No. of patients (%)	**No. of patients (%)**	**No. of patients (%)**
Gender			
Male	402 (89.3)	320 (89.4)	722 (89.4)
Female	48 (10.7)	38 (10.6)	86 (10.6)
Performance status			
0	301 (66.9)	181 (50.6)	482 (59.7)
1	145 (32.2)	176 (49.2)	321 (39.7)
2	4 (0.9)	1 (0.3)	5 (0.6)
N classification			
N0	160 (35.6)	42 (11.7)	202 (25.0)
N1	117 (26.0)	56 (15.6)	173 (21.4)
N2	173 (38.4)	205 (57.3)	378 (46.8)
N3	0 (0.0)	52 (14.5)	52 (6.4)
Nx	0 (0.0)	3 (0.8)	3 (0.4)
T classification			
T1	0 (0.0)	4 (1.1)	4 (0.5)
T2	61 (13.6)	25 (7.0)	86 (10.6)
T3	252 (56.0)	77 (21.5)	329 (40.7)
T4	137 (30.4)	252 (70.4)	389 (48.1)
Country			
France	222 (49.3)	67 (18.7)	289 (35.8)
Netherlands	143 (31.8)	30 (8.4)	173 (21.4)
Italy	56 (12.4)	5 (1.4)	61 (7.5)
Belgium	16 (3.6)	41 (11.5)	57 (7.1)
Hungary	0 (0.0)	45 (12.6)	45 (5.6)
Spain	0 (0.0)	32 (8.9)	32 (4.0)
Germany	0 (0.0)	25 (7.0)	25 (3.1)
United Kingdom	0 (0.0)	23 (6.4)	23 (2.8)

Table 1: Baseline demographic and clinical characteristics of the patients by study

		Study		
		Trial 1 (N=450)	**Trial 2 (N=358)**	**Total (N=808)**
		No. of patients (%)	**No. of patients (%)**	**No. of patients (%)**
	Austria	0 (0.0)	21 (5.9)	21 (2.6)
	Czech Republic	0 (0.0)	19 (5.3)	19 (2.4)
	Switzerland	10 (2.2)	9 (2.5)	19 (2.4)
	Serbia	0 (0.0)	17 (4.7)	17 (2.1)
	Poland	0 (0.0)	8 (2.2)	8 (1.0)
	Slovakia	0 (0.0)	8 (2.2)	8 (1.0)
	Turkey	0 (0.0)	8 (2.2)	8 (1.0)
	Israel	3 (0.7)	0 (0.0)	3 (0.4)
Age				
	Mean (SD)	56.27 (8.43)	53.08 (8.08)	54.86 (8.42)
	Interquartile	50.0 - 62.0	48.0 - 58.0	49.0 – 51.0

Table 2: Cross-sectional correlations of the EORTC QLQ-C30 scales with anchors, and correlations between their change scores

		Cross-sectional		**Change scores**	
Scale	**Anchor**	**No. of patients (No. of observations)**	**Correlation**	**No. of patients (No. of observations)**	**Correlation**
PF	Performance status	752 (2502)	-0.31	641 (3720)	-0.21
RF	Performance status	738 (2459)	-0.40	629 (3663)	-0.30
SF	Performance status	753 (2498)	-0.30	642 (3697)	-0.27
QL	Performance status	751 (2473)	-0.30	635 (3632)	-0.23
FA	Performance status	752 (2506)	-0.32	642 (3730)	-0.30
NV	CTCAE Nausea	353 (1135)	-0.20	309 (1604)	-0.12
	CTCAE Gastrointestinal	353 (1284)	-0.20	314 (2214)	-0.14
DY	Performance status	753 (2501)	-0.21	641 (3718)	-0.10
	CTCAE pulmonary	353 (1134)	-0.24	308 (1607)	-0.11

Abbreviations: FA, fatigue; NV, nausea/vomiting; DY, dyspnoea; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; CTCAE, common terminology criteria for adverse events

Table 3: Means (effect sizes) of HRQOL change score in three clinical change groups that are based on anchors for selected EORTC QLQ-C30 scales and mean change scores based on the linear regression

	Mean change method[1]				**Linear regression**[2]	
Scale	**Anchor**	**Improvement (ES)**	**Stable (ES)**	**Deterioration (ES)**	**Improvement**	**Deterioration**
PF	Performance status	-1.95 (-0.11)[a]	-2.43 (-0.14)	-10.67 (-0.65)	0.11 [a]	-7.31
RF	Performance status	3.62 (0.14) [a]	-1.26 (-0.05)	-15.11 (-0.61)	4.99 [a]	-12.17
SF	Performance status	6.41 (0.26)	1.08 (-0.05)	-7.72 (-0.37)	4.94	-7.74
QL	Performance status	8.64 (0.42)	2.73 (0.14)	-4.71 (-0.23)	5.41	-6.53
FA	Performance status	1.94 (0.08)[a]	-2.15 (-0.09)	-15.36 (-0.68)	3.03 [a]	-11.92

NV	CTCAE Nausea	-1.71 (-0.12) [a]	-0.65 (-0.05)	-5.56 (-0.31)	-1.25 [a]	-4.77
	CTCAE Gastrointestinal	0.00 (0.00) [a]	-0.42 (-0.01)	-5.83 (-0.45)	0.79 [a]	-3.68
DY	Performance status	-2.51(-0.12) [a]	-1.23 (-0.06)	-6.71 (-0.31)	-1.78 [a]	-3.86
	AE pulmonary	6.02(0.25)	-1.04 (-0.05)	0.00 (0.00) [a]	6.71	0.56 [a]

[1]The mean change method is useful for interpreting within-group change over time

[2]The linear regression is useful for interpreting between-group differences in change over time

[a] These estimated change scores were not considered to summarise the MID estimate because their ES were either <0.2 or ≥0.8

The symptom scores were reversed to follow the functioning scales' interpretation; i.e. 0 represents the worst possible score and 100 the best possible score

Abbreviations: FA, fatigue; NV, nausea/vomiting; DY, dyspnoea; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; ES, effect size; CTCAE, common terminology criteria for adverse events

Table 4: Summary of Anchor-based MIDs for within and between-group change over time compared with distribution-based estimates.

	Anchor-based MID for within-group change		**Anchor-based MID for between-group difference in change**		**Distribution-based QOL scores at t1 (t2) (No. of patients = 543 to 575)**		
Scale	**Improvement**	**Deterioration**	**Improvement**	**Deterioration**	**0.3SD**	**0.5SD**	**1SEM**
PF	no MID	-10.67	no MID	-7.31	5.11 (6.01)	8.52 (10.01)	5.11 (6.01)
RF	no MID	-15.11	no MID	-12.17	6.83 (7.99)	11.38 (13.32)	9.66 (11.30)
SF	6.41	-7.72	4.94	-7.74	6.31 (6.67)	10.51 (11.12)	7.58 (8.02)
QL	8.64	-4.71	5.41	-6.53	5.93 (6.39)	9.88 (10.66)	8.39 (9.04)
FA	no MID	-15.36	no MID	-11.90	6.58 (7.64)	10.96 (12.73)	9.04 (10.50)
NV	no MID	-5.83 to -5.56 (-5.71 [w])	no MID	-4.77 to -3.68 (-4.18 [w])	3.43 (6.25)	5.72 (10.41)	6.96 (12.67)
DY	6.02	-6.71	6.71	-3.86	7.02 (6.27)	11.69 (10.44)	9.64 (8.61)

The within-group MIDs are derived from the mean change method and the between-group MIDs from the linear regression

[w] represents weighted average based on scale/anchor pair change score correlation.

The symptom scores were reversed to follow the functioning scales' interpretation, i.e. 0 represents the worst possible score and 100, the best possible score; 'no MID' is used where no MID estimate is available either due to the absence of a suitable anchor or effect size <0.2 or ≥0.8

Abbreviations: t1 is the time point for the start of treatment; t2 is the time point for the end of treatment; FA, fatigue; NV, nausea/vomiting; DY, dyspnoea; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning.

Table 5: Distribution-based estimates

	Distribution-based HRQOL scores at t1 (t2) (No. of patients = 541 to 575)			
Scale	**0.2 SD**	**0.3 SD**	**0.5 SD**	**1 SEM**
PF	3.41 (4.00)	5.11 (6.01)	8.52 (10.01)	5.11 (6.01)
RF	4.55 (5.33)	6.83 (7.99)	11.38 (13.32)	9.66 (11.30)
SF	4.21 (4.45)	6.31 (6.67)	10.51 (11.12)	7.58 (8.02)
CF	3.18 (3.75)	4.78 (5.62)	7.96 (9.37)	6.75 (7.95)
EF	4.49 (4.05)	6.74 (6.07)	11.23 (10.12)	8.40 (7.57)
QL	3.95 (4.26)	5.93 (6.39)	9.88 (10.66)	8.39 (9.04)
FA	4.39 (5.09)	6.58 (7.64)	10.96 (12.73)	9.04 (10.50)
PA	4.59 (4.82)	6.89 (7.23)	11.48 (12.04)	8.59 (9.01)
DY	4.68 (4.18)	7.02 (6.27)	11.69 (10.44)	9.64 (8.61)
NV	2.29 (4.17)	3.43 (6.25)	5.72 (10.41)	6.96 (12.67)
AP	5.23 (6.62)	7.84 (9.93)	13.07 (16.55)	11.98 (15.17)
SL	5.86 (5.47)	8.79 (8.21)	14.65 (13.68)	12.78 (11.92)
CO	4.25 (5.52)	6.38 (8.28)	10.64 (13.80)	8.77 (11.38)
DI	3.04 (3.91)	4.56 (5.86)	7.61 (9.77)	8.05 (10.34)

Abbreviations: t1 is the time point for the start of treatment; t2 is the time point for the end of treatment; PF, physical functioning; RF, role functioning; CF, cognitive functioning; EF, emotional functioning; SF, social functioning; FA, fatigue; PA, pain; NV, nausea/vomiting; QL, global health status; DY, dyspnoea; AP, appetite loss; SL; sleep disturbance; CO, constipation; DI, diarrhoea

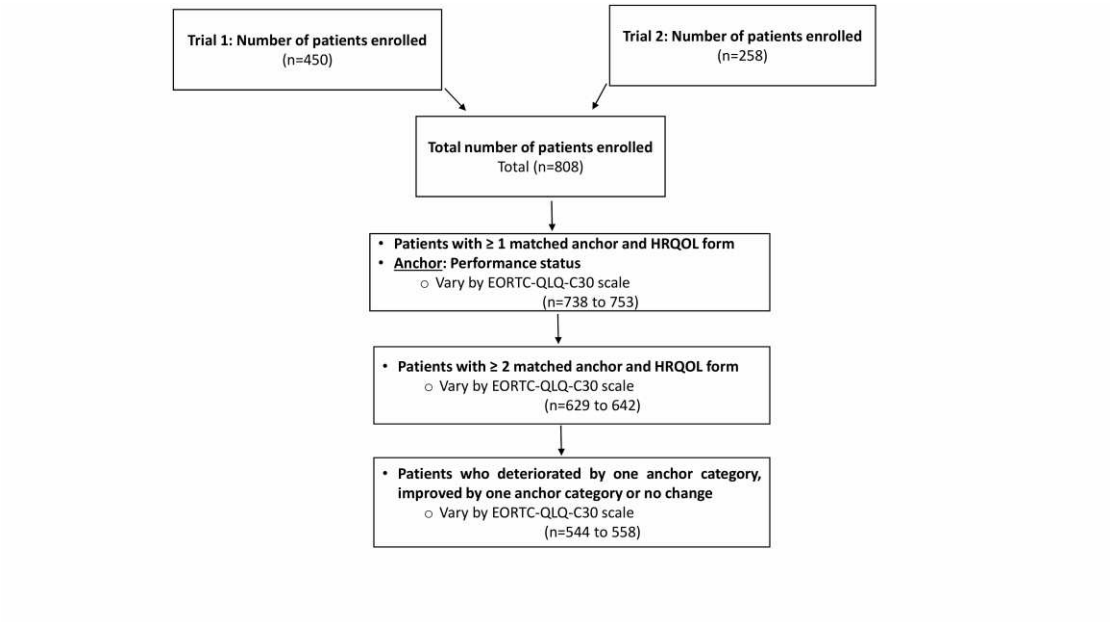

Figure 1: An overview of the flow of patients through the study

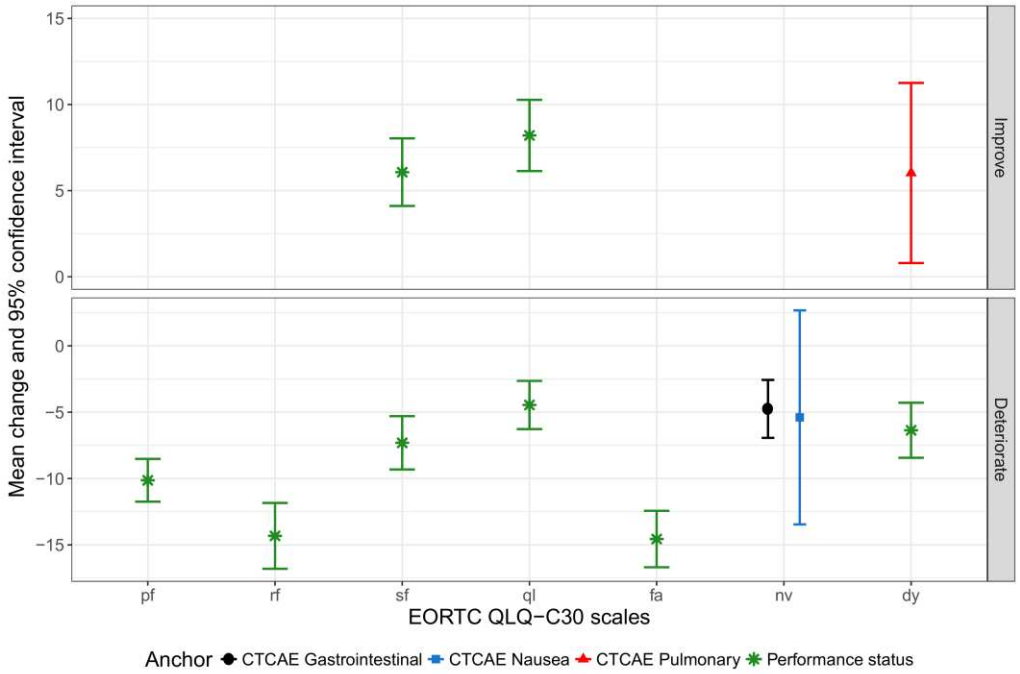

Figure 2: Mean change and 95% confidence interval for improvement and deterioration in EORTC QLQ-C30 scales, across multiple anchors and averaged across different time periods.

Estimates are available only for scales with at least 1 suitable anchor or 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.

Abbreviations: AP, appetite loss; FA, fatigue; NV, nausea/vomiting; DY, dyspnoea; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; CTCAE, common terminology criteria for adverse events.

Deteriorate = worsened by 1 anchor category, no change =no change in anchor category and improve = improved by 1 category

Appendix

Table A.1: Number of patients (number of observations) by change scores of suitable anchors

Anchor change score	CTCAE Nausea	CTCAE Pulmonary	CTCAE Gastrointestinal	Performance status
-4		1 (4)	1(2)	
-3	4 (6)	1 (3)	16 (28)	
-2	37 (94)	5 (8)	81 (214)	3 (9)
-1	64 (167)	37 (83)	132 (489)	181 (538)
0	287 (1285)	301 (1460)	267 (1116)	586 (2453)
1	28 (34)	33 (52)	106 (249)	208 (596)
2	15 (19)	6 (7)	56 (87)	27 (60)
3	3 (3)	1 (1)	19 (30)	1 (1)
4			3 (5)	

Since a patient can have multiple assessments, that patient can contribute to multiple anchor change score category.

Abbreviations: CTCAE, common terminology criteria for adverse events