Article:
Musoro, JZ, Bottomley, A, Coens, C et al. (8 more authors) (2018) 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, 104. pp. 169-181. ISSN 0959-8049

https://doi.org/10.1016/j.ejca.2018.09.005

Interpreting EORTC QLQ-C30 scores as minimally importantly different (MID) for patients with malignant melanoma

Jammbe Z Musoro[1], Andrew Bottomley[1], Corneel Coens[1,] Alexander MM Eggermont[2], Madeleine T King[3,], Kim Cocks[4,5], Mirjam AG Sprangers[6] Mogens Groenvold[7], Galina Velikova[8], Hans-Henning Flechtner[9], Yvonne Brandberg[10] on behalf of the EORTC Melanoma Group and EORTC Quality of Life Group

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

[2] Gustave Roussy Cancer Instiute and University Paris-Sud, Villejuif/Paris-Sud, France

[3]School of Psychology and Sydney Medical School, University of Sydney, Sydney, NSW, Australia

[4]York Trials Unit, Department of Health Sciences, University of York, York, UK

[5]Adelphi Values, Bollington, Cheshire, UK

[6]Department of Medical Psychology, Academic Medical Center, University of Amsterdam, Amsterdam The Netherlands.

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

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

[9]Clinic for Child and Adolescent Psychiatry and Psychotherapy, University of Magdeburg, Magdeburg, Germany

[10]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden.

Corresponding Author:

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

ABSTRACT

Introduction: Health-related quality of life (HRQOL) is increasingly recognized as an important endpoint in cancer clinical trials. The concept of minimally important difference (MID) enables interpreting differences and changes in HRQOL scores in terms of clinical meaningfulness. We aimed to estimate MIDs for interpreting group-level change of EORTC QLQ-C30 scores in patients with malignant melanoma.

Methods: Data was pooled from three published melanoma Phase III trials. Anchors relying on clinician's ratings, e.g. performance status, were selected using correlation strength and clinical plausibility of associating the anchor/EORTC QLQ-C30 scale pair. HRQOL change was evaluated between time periods that were common to all trials: start of treatment to end of treatment, and end of treatment to end of follow-up. Three change-status groups were formed: deteriorated by one anchor category, improved by one anchor category, and no change. Patients with greater anchor change were excluded. The mean change method and linear regression were used to estimate MIDs for change in HRQOL scores within-group and between-groups of patients respectively.

Results: MIDs varied according to QLQ-C30 scale, direction (improvement vs deterioration), anchor and period. MIDs for within-group change ranged from 4 to 18 points (improvement) and -16 to -4 points (deterioration), and MIDs for between-group change; 3 to 16 points and -16 to -3 points. MIDs for most of QLQ-C30 scales ranged from 5 to 10 points in absolute values.

Conclusions: These results are useful for interpreting changes in EORTC QLQ-C30 scores over time, and for performing more accurate sample size calculations in adjuvant melanoma settings.

Keywords: Health-related quality of life (HRQOL); EORTC QLQ-C30; Minimally important difference (MID); Malignant melanoma;

1. INTRODUCTION

Health-related quality of life (HRQOL) is increasingly recognized as an important endpoint in cancer clinical trials [1]. Understanding the amount of change in HRQOL scores that is clinically relevant is crucial for interpretation. The concept of minimally important difference (MID) enables the interpretation of differences between groups and changes over time in HRQOL scores in terms of clinical meaningfulness [[2], [3], [4], [5], [6]]. MID is defined as: 'the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and which would lead the patient or clinician to consider a change in the management'[2]. MIDs are commonly determined via anchor-based and distribution-based methods [7]. Anchor-based methods express differences or change in HRQOL scores by linking specific HRQOL domains to clinical variables which have known clinical relevance [[3], [8], [9], [10]] or to patient/physician-derived ratings of change in the specific domain [[4], [5], [6]]. The usefulness of anchor-based MIDs is reliant on the anchor selected, how discriminant groups are defined with respect to that anchor, and the strength of the relationship (conceptually and empirically) between the anchor and the target HRQOL domain [11]. Distribution-based methods rely on the statistical distribution of HRQOL scores, e.g. standard deviation (SD) criteria or the standard error of measurement, SEM [[12], [13]]. Since distribution-based methods do not consider patients'/clinicians' perspective, they have been recommended to be used as supportive evidence to anchor-based methods [7].

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 [14]. Osoba et al. [4] published guidelines for interpreting small (5 to 10 points), moderate (10 to 20 points) and large changes (> 20 points) in EORTC QLQ-C30 scores using a global patient rating of change as anchor, in patients with breast and small-cell lung cancer. In an early application of clinical anchors, King [3] compiled 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 HRQOL scales. More recent guidelines by Cocks et al. [[5], [6]] highlighted the need not only to differentiate between the EORTC QLQ-C30 scales, but also between direction of change (improvement vs deterioration) and between clinical settings. This implies that a global rule for MIDs applicable to all situations is highly unlikely [[7], [11], [15]].

This study aims to provide MID estimates for EORTC QLQ-C30 scales in patients with malignant melanoma who undergo adjuvant treatment. We focused on examining MIDs for group-level change (both within and between groups) in HRQOL scores over time [[16]]. There are currently no MID guidelines for the EORTC QLQ-C30 specific to malignant melanoma. In contrast to Osoba et al, [4] we used multiple clinical anchors that were available in our 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 we used individual patient data from archived EORTC melanoma trials.

2. METHODS

Data description

Data were pooled from three published adjuvant melanoma Phase III EORTC trials. Trial 1 assessed the effect of two regimens of interferon of intermediate dose versus observation alone in patients with stage IIb/III melanoma after surgery and enrolled 1388 patients [**Error! Reference source not found.**]. Trial 2 compared adjuvant immunotherapy with anti-CTLA-4 monoclonal antibody (ipilimumab) versus placebo after complete resection of high-risk Stage III melanoma and enrolled 951 patients [[18], [19]]. Trial 3 compared the effect of adjuvant therapy with PEG-Intron to observation after adequate dissection of the regional lymph in AJCC Stage III melanoma and enrolled 1256 patients [17]. All three trials assessed HRQOL using the EORTC QLQ-C30 at baseline, during treatment and on several follow-up time points after end of treatment. When pooling, three key time points were identified that were common across all three trials: (i) Start of treatment (T1); time point before or on first day of treatment administration. If no treatment was administered then T1 was the time point before or on date of randomization. (ii) End of treatment (T2); last day of protocol treatment administration. Patients who were under observation alone did not contribute data at T2. (iii) End of follow-up (T3); the last day of the protocol follow-up period. For patients under observation, T3 was the last day after baseline.

The EORTC QLQ-C30

The EORTC QLQ-C30 comprises 30 items, 24 of which are aggregated into 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: dyspnea (DY), appetite loss (AP), sleep disturbance (SL), constipation (CO), diarrhea (DI) and financial impact (FI).

Trial 1 used Version 2 of the QLQ-C30 while Trial 2 and 3 used Version 3. The two versions differ only in the response categories of questions 1 to 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 [14], with the means of the raw scores for each scale transformed to fall between 0 and 100. For consistency in signs of the 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 anchors

Anchors were constructed using clinical data from physician examinations, common terminology criteria for adverse events (CTCAE) and laboratory results that were available in the trial data sets. Anchors were initially selected based on the strength of correlation with the corresponding QLQ-C30 scale. We prioritized anchors with correlations of $\geq|0.30|$ as proposed by Revicki et al. [7] and where achievable, anchors with stronger correlations were targeted [21]. The selected anchors were further verified for clinical plausibility by a panel of melanoma and HRQOL experts to avoid spurious findings. This panel was also tasked to identify clinically relevant changes for each of the selected anchors. For each QLQ-C30 scale, multiple anchors could be selected. Details on the anchor selection procedures have been described by Musoro et al. [16]. The retained anchors comprised WHO performance status (PS) and 7 CTCAEs (gastrointestinal disorder, anorexia, pain, fatigue, immune disorder, diarrhea and nervous system disorder). The 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).

Definition of clinical change groups

Three clinical change status groups (CCG) were defined after 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 considered to be above the "minimal" expected change, and so were excluded from datasets used to estimate mean change and MIDs.

Data analysis

Individual-level change scores of the EORTC QLQ-C30 scales and their corresponding anchors were computed between T1 and T2, and between T2 and T3. Only subjects with both EORTC QLQ-C30 and anchor data available for a given pair of time points contributed to calculation of change scores.

Two anchor-based methods were then used to estimate MIDs for improvements and deterioration for each EORTC QLQ-C30 scale and its corresponding anchors. The primary method involved calculating the mean HRQOL change score for the improvement and deterioration CCGs, respectively. This is applicable for interpreting change within a group of patients and it is analogous to the mean HRQOL change score over time for a single treatment

group in a trial. Effect sizes (ES) were computed by dividing the mean change HRQOL score between adjacent time points (e.g. T1 and T2) by the SD of the HRQOL scores at the earlier time point (T1). Only mean change scores with an ES of > 0.2 or ≤ 0.8 were considered appropriate for inclusion as MIDs. This was based on Cohen's [13] recommendations that ES of 0.2 are small, 0.5 are moderate and ≥ 0.8 are large. The rationale here was that observed effect sizes < 0.2 reflected changes that were clinically unimportant, and those ≥ 0.8 were clearly more than minimally important. We also compared the difference in change scores between the improvement (or deterioration) CCG and no change CCG using ANOVA.

The secondary method involved linear regression applied to compare change scores for subjects in the improvement (or deterioration) CCGs versus the stable CCG. For a given EORTC QLQ-C30 scale/anchor pair, separate models were fitted for improving and deteriorating scores. 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, and 'stable'=0 and 'deterioration'=1 when modelling deterioration. The resulting slope parameters correspond to the mean change score for improvement and deterioration respectively. This is useful for interpreting changes between groups of patients, and it is analogous to comparing the mean HRQOL change score in a target treatment group to a control group in a trial. For a given HRQOL scale, the anchor-based estimates from multiple anchors were triangulated to a single value via a correlation-based weighted average.

Distribution-based techniques were used as supportive methods by estimating the 0.2 SD, 0.3 SD, 0.5 SD and SEM separately at T1, T2 and T3. These techniques have previously been used in the literature to estimate MIDs [7]. However, since these estimates rely solely on the statistical distribution of the HRQOL scores, and do not include an inherent valuation of clinical relevance, they are used to give context to our derived anchor-based estimates. Test-retest reliability estimates to compute SEM for the QLQ-C30 were obtained from Hjermstad et al. [22]. All statistical analyses were performed using the SAS software [23]. An in-depth description of the statistical methodology, including the anchor selection process, has previously been published [16].

3. RESULTS

The baseline demographic and clinical characteristics of the study population are presented in Tables 1a and 1b. The characteristics of the patients across the 3 trials were similar. In Table 2, the descriptive statistics of the QLQ-C30 scale scores at T1, T2 and T3 are summarised. The distribution of the various scale scores were similar across the different time points. The time period (in months) between T1 and T2 ranged from 0.1 to 24.2 with a mean of 10.4 (SD=6.1) for Trial 1, from 0 to 38.4 with a mean of 12.3 (SD=12.8) for Trial 2, and from 0.1 to 57 with

a mean of 23.7 (SD=16.6) for Trial 3. The period between T2 and T3 ranged from 0 to 31.3 with a mean of 8.9 (SD=6.4) for Trial 1, from 0 to 64.4 with a mean of 11.2 (SD=11) for Trial 2, and from 0.5 to 64.4 with a mean of 27.5 (SD=19.7) for Trial 3.

Cross-sectional correlations of the QLQ-C30 scale scores with their corresponding selected anchors (at T1, T2 and T3), and correlations between their change scores (between T2-T1 and T3-T2) are presented in Table 3. At least one anchor was constructed for each QLQ-C30 scale, except for the constipation scale where no suitable anchors were found. The cross-sectional correlations ranged from 0.16 to 0.76 in absolute value, with over 90% of the correlation coefficients being above the 0.3 threshold [7]. Much lower correlations (range: 0.1 to 0.53) were observed between the change scores.

The distribution of patients across the different anchor categories is summarised in Table A.1. According to the anchors, most patients remained stable (63% to 88%), for both periods between T2 & T1 and T3 & T2. Relatively low proportions of patients either improved (4% to 20%) or deteriorated (2% to 11%).

Table 4 presents the range of estimated MID values from the mean change method and the linear regression for each HRQOL scale, across multiple anchors and over time (change between T2 & T1 vs T3 & T2). MID estimates are only presented for scales with at least one appropriate anchor or where CCG has an ES of > 0.2 or ≤ 0.8. Detailed results on the estimates per anchor from the mean change method and the linear regression are presented in Tables A.2 and A.3 respectively. Generally, the MID estimates varied by scale, direction of change scores (improvement vs deterioration), selected anchor and time point. This is illustrated in Figure 1, where estimates from the mean change method in Table 4 are plotted along with their 95% confidence intervals (CI). Though the MID estimates for change between T1 and T2 were comparable to those for change between T2 and T3, relatively wider CIs were observed in the latter time period, reflecting the relatively smaller sample size. The MID estimates were always in the expected direction according to the anchor, i.e. positive vs negative change scores within the improvement vs. deterioration CCG respectively. Based on ANOVA, the difference in change scores between the improvement (or deterioration) CCG and no change CCG for most of the EORTC QLQ-C30 scales were statistically significant (p-value <0.05). Non-significant differences were mostly observed amongst the CCGs with an ES of < 0.2. As shown in Table 4, generally the MIDs for interpreting within-group change in HRQOL scores (estimated using the mean change method) ranged from 4 to 18 points and -16 to -4 points for improvement and deterioration respectively. MIDs for between-group change (estimated using the linear regression) ranged from 3 to 16 points and -16 to -3 points for improvement and deterioration respectively. For the majority of the QLQ-C30 scales, the estimated MIDs ranged from 5 to 10 points in absolute values.

The results in Table 4 were further summarised to single MID values per scale in Table 5 by taking a correlation-weighted average across multiple anchors. This facilitates the selection of MIDs for per QLQ-C30 scales for use in practice. Furthermore, in Table 5, we also compared the anchor-based MIDs to estimates from commonly used distribution-based approaches in the literature. The distribution-based estimates for each QLQ-C30 scale were very similar across T1, T2 and T3. For a particular distribution-based approach, the estimates across the different time points were mostly within a < 1 point range for a given QLQ-C30 scale. Therefore, only results at T1 are reported in Table 5. The anchor-based MID estimates tended to be larger than the 0.2 SD and smaller than the 0.5 SD. Most of the anchor-based estimates were closer to both the 0.3 SD and the 1 SEM.

4. DISCUSSION

Our study determined MIDs for group-level change of the EORTC QLQ-C30 scores over time, using individual patient data pooled across three published international randomized EORTC adjuvant melanoma clinical trials. Anchors for each QLQ-C30 scale were selected based on both the statistical correlation and clinical plausibility. Multiple anchors were selected for most QLQ-C30 scales. The cross-sectional correlations between the anchors and their corresponding scales were usually greater than the recommended 0.3 correlation threshold [7]. However, lower correlations were observed when considering the changes over time, which may be attributed to cumulative measurement error.

The use of multiple anchors per scale provided some reassurance about the plausibility of the estimated MIDs. Despite the modest correlation between the anchors/scales change scores, the estimated MIDs were often within a small range (generally < 5 points range) and were also in the expected direction of change according to the anchor.

Similar to recent findings on MIDs for the QLQ-C30 by Cocks et al [[5], [6]] and Maringwa et al [[8], [9]], we observed that MIDs vary by scales as well as by the direction of change (improvement vs deterioration). Furthermore, akin to Maringwa et al [[8], [9]] there were no systematic differences in the magnitude of change between deteriorating and improving scores. This is in contrast to Cocks et al [6] and other studies that assessed MIDs for the Functional Assessment of Cancer Therapy (FACT) questionnaires [[24], [25]], where estimates for deterioration tended to be larger than those for improvement. However, we noted that the latter studies used a patient- or clinician-rated global rating of change as anchors, whereas our study and those of Maringwa et al applied clinical anchors. It will be interesting to further examine this observation in other studies.

Our MID estimates across many scales were somewhat within the suggested 5 to 10 points range suggested by Osoba et al. [4], as shown in Table 4. Cocks et al [[5], [6]] and Maringwa et al [8, 9] also made similar observations, which is reassuring. However, as pointed out by Cocks et al [[5], [6]] we also observed that the thresholds for some scales could be much lower. For example the MIDs for the EF and CF scales could be as low as 3 points. On the other hand, much bigger thresholds were observed for scales like RF and AP, where MIDs for the AP scale could be as high as 18 points. 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.

For any given QLQ-C30 scale, no remarkable differences were observed amongst MIDs for change scores between T1 and T2, and between T2 and T3. This is probably because the patients' HRQOL in these adjuvant melanoma studies were relatively stable over time as shown by the mean scores at T1, T2 and T3 in Table 2. Furthermore, according to the anchors, the majority of the patients remained stable over time, or changed by only one category (Table A.1). Comparable estimates (results not shown) were also obtained from applying the mean change method to the merged data of all possible pairwise time point differences of HRQOL scores (where a subject can contribute multiple change scores that are calculated across different pairs of time points). We also made a distinction between MIDs for interpreting within-group changes; obtained from the mean change method, and MIDs for interpreting changes between groups; obtained from the linear regression. Estimates from both approaches were often in the same range.

While clinicians and researchers seeking MID would often like simple guidance, results such as those presented in this paper are often complex, as a consequence of there being numerous anchors, various distribution-based criteria, and various HRQOL scales. In Table 4, we represented this complexity as the range of MIDs generated by the various anchors. However, we appreciate end-users may find such a range of options confusing, wondering which they should use. So in order to provide a single MID value per QLQ-C30 scale, we further simplified by calculating a correlation-weighted average across multiple anchors. End-users can choose to work with either the ranges provided in Table 4 or the single values provided in Table 5, whichever they feel most comfortable with.

A limitation of our study is that anchor-based MIDs could only be estimated for QLQ-C30 scales for which a suitable anchor was available in the database. For example, no suitable anchors were found for the constipation (CO) scale. Different anchors also represent different categorizations of clinical relevance that may or may not exceed a 'true' MID. Furthermore, the available anchors relied exclusively on clinical observations or interpretations. The potentially inflated MID estimates for scales like RF and AP, may be due to an underestimation

of their relevance by the physician rated anchors (such as performance status or CTCAE grades) compared to the patient self-reported assessment. However, given that our data set is limited, it will be interesting to further examine this observation in future studies. Anchors related to mental health/distress of patients were not available in our study, which is a notable lack since these are important aspects of HRQOL. Additionally, anchors that are based on the patient's perspective of change (e.g., subjective significance questionnaires) were not available. Nonetheless, it is reassuring to notice the considerable overlap between our findings and that of Osoba et al. [4] which was based on using individual patients' ratings of change as anchor. One out of the three trials that were pooled in this study used Version 2 of the EORTC QLQ-C30. 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. It will be interesting to further investigate in a larger sample if these difference may affect MID estimates. Another limitation is that our data originates from three controlled clinical trials, each with specific selection and treatment criteria. Although results are consistent among the three trials, extrapolation beyond their specific setting remains unverified.

In conclusion, our findings can help clinicians and researchers to interpret the clinical relevance of group-level change of QLQ-C30 scores over time, in patients with malignant melanoma. We have provided MID estimates for interpreting changes in HRQOL scores over time for both within-group and between-groups of patients. Our results will also aid to perform more accurate sample size calculations when primary outcomes are based on EORTC QLQ-C30 scales.

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

Role of the Funding Source: The sponsor provided financial support for study conduction, reviewed and approved the final version of the manuscript.

Conflict of interest statement: None declared.

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

5. REFERENCES

1. Bottomley A, et al. Health related quality of life outcomes in cancer clinical trials. Eur J Cancer. 2005; 41: 1697-1709.
2. Schünemann HJ, Guyatt GH. Goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res. 2005; 40: 593-597.
3. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res. 1996; 5: 555-567.
4. Osoba D et al. Interpreting the significance of changes in health related quality-of-life scores. J Clin Oncol. 1998; 16: 139-144.
5. Cocks K, 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.
6. Cocks K, 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.
7. 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
8. Maringwa JT, 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.
9. Maringwa J, 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.
10. Cella D 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.
11. 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.
12. 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.
13. Cohen J. Statistical Power Analysis for the Behavioural Sciences (2nd Edition). Lawrence Erlbaum Associates, NJ, USA (1988).

14. Fayers P, Aaronson NK, Bjordal K, Curran D and Groenvold M on behalf of the EORTC Quality of Life Study Group. EORTC QLQ-C30 Scoring Manual (Third edition). Brussels, EORTC Quality of Life Group, 2001.
15. 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):207–21.
16. 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
17. Eggermont AM, Suciu S, MacKie R, et al, for the EORTC Melanoma Group. Post-surgery adjuvant therapy with intermediate doses of interferon alfa 2b versus observation in patients with stage IIb/III melanoma (EORTC 18952): randomised controlled trial. Lancet 2005; 366: 1189–96.
18. Eggermont AM, Chiarion-Sileni V, Grob JJ, et al. Adjuvant ipilimumab versus placebo after complete resection of high-risk stage III melanoma (EORTC 18071): a randomised, double-blind, phase 3 trial. Lancet Oncol 2015; 16:522-30.
19. Eggermont AM, Chiarion-Sileni V, Grob JJ, et al. Prolonged survival with Ipilimumab as adjuvant in stage III melanoma. New Engl J Med 2016; 375:1845-1855.
20. Eggermont AM, Suciu S, Santinami M, et al: Adjuvant therapy with pegylated interferon alfa-2b versus observation alone in resected stage III melanoma: Final results of EORTC 18991, a randomised phase III trial. Lancet 372:117-126, 2008
21. 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.
22. 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: 1249–1254
23. Institute Inc. 2013. Base SAS® 9.4 Procedures Guide. Cary, NC: SAS Institute Inc.
24. 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.
25. Ringash J, O'Sullivan B, Bejzak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. Cancer 2007; 110:196–202.

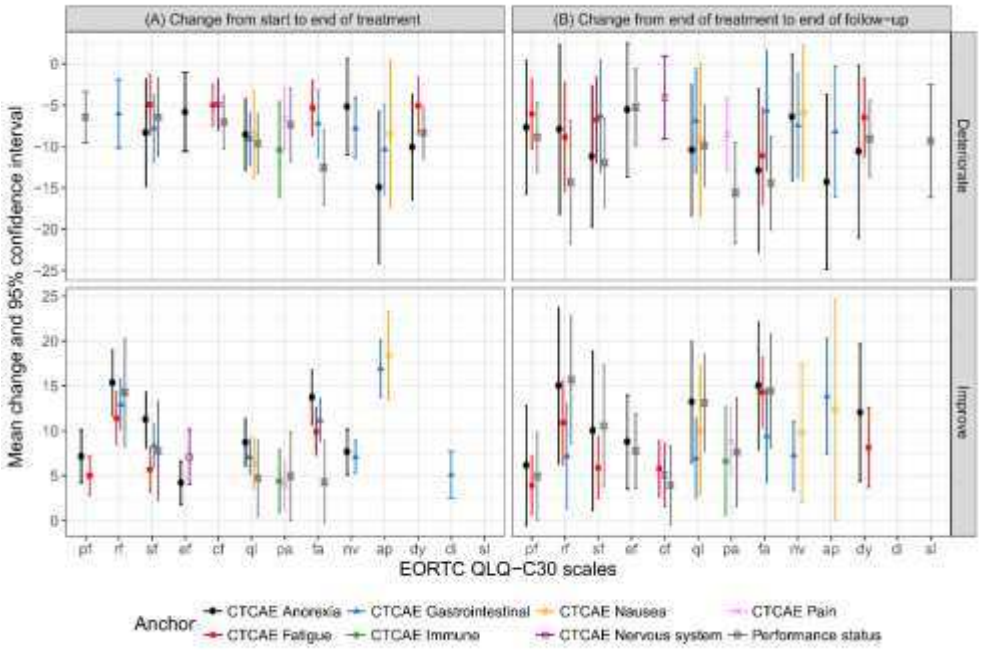

Figure 1: Mean change and 95% confidence interval for improvement and deterioration EORTC QLQ-C30 scales, across multiple anchors and at different time periods.

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; CF, cognitive functioning; CO, constipation; DI, diarrhea; DY, dyspnea; EF, emotional functioning; FA, fatigue; NV, nausea/vomiting; PA, pain; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; SL, sleep disturbance
Deterioration =worsened by 1 anchor category, no change =no change in anchor category and improvement = improved by 1 category

Table 1a: Selected baseline demographic and clinical characteristics of the patients by study

	Study 18952 (N=1388)	**Study 18071 (N=951)**	**Study 18991 (N=1256)**	**Total (N=3595)**
Gender- N (%)				
Male	771 (55.5)	589 (61.9)	731 (58.2)	2091 (58.2)
Female	616 (44.4)	362 (38.1)	525 (41.8)	1503 (41.8)
Missing	1 (0.1)	0 (0.0)	0 (0.0)	1 (0.0)
Country- N (%)				
United Kingdom	142 (10.2)	36 (3.8)	327 (26.0)	505 (14.0)
Italy	91 (6.6)	144 (15.1)	229 (18.2)	464 (12.9)
Netherlands	261 (18.8)	23 (2.4)	152 (12.1)	436 (12.1)
France	181 (13.0)	144 (15.1)	106 (8.4)	431 (12.0)
Germany	140 (10.1)	76 (8.0)	103 (8.2)	319 (8.9)
United States	0 (0.0)	213 (22.4)	0 (0.0)	213 (5.9)
Poland	167 (12.0)	11 (1.2)	28 (2.2)	206 (5.7)
Belgium	116 (8.4)	16 (1.7)	68 (5.4)	200 (5.6)
Switzerland	46 (3.3)	41 (4.3)	44 (3.5)	131 (3.6)
Bulgaria	46 (3.3)	0 (0.0)	29 (2.3)	75 (2.1)

Russian	21 (1.5)	54 (5.7)	0 (0.0)	75 (2.1)
Australia	0 (0.0)	37 (3.9)	36 (2.9)	73 (2.0)
Spain	27 (1.9)	11 (1.2)	35 (2.8)	73 (2.0)
Portugal	38 (2.7)	0 (0.0)	33 (2.6)	71 (2.0)
Denmark	0 (0.0)	59 (6.2)	0 (0.0)	59 (1.6)
Croatia	14 (1.0)	0 (0.0)	31 (2.5)	45 (1.3)
Others*	98 (7.1)	86 (9.0)	35 (2.8)	219 (6.1)
Number of positive lymph nodes- N (%)				
0-1	712 (51.3)	437 (46.0)	678 (54.0)	1827 (50.8)
2-4	389 (28.0)	321 (33.8)	423 (33.7)	1133 (31.5)
>=5	151 (10.9)	192 (20.2)	148 (11.8)	491 (13.7)
Unknown/missing	136 (9.8)	1 (0.1)	7 (0.6)	144 (4.0)
Performance status- N (%)				
0	1199 (86.4)	890 (93.6)	1061 (84.5)	3150 (87.6)
1	180 (13.0)	61 (6.4)	195 (15.5)	436 (12.1)
Missing	9 (0.6)	0 (0.0)	0 (0.0)	9 (0.3)
Age				
Mean (SD)	48.55 (13.46)	51.10 (12.86)	48.80 (12.35)	-
Interquartile	39.0 - 59.0	42.0 - 61.0	40.0 - 58.0	-
Weight (kg)				
Mean (SD)	75.47 (15.15)	82.43 (18.20)	77.08 (14.99)	-
Interquartile	65.0 - 84.5	70.0 - 92.0	66.0 - 85.6	-

Others* Comprise country with total percentage < 1% (Canada, Czech Republic, Austria, Turkey, Finland, Sweden, Estonia, Norway, Slovenia, Israel, Serbia, Hungary, Slovakia)

Table 1b: Distribution of patient by baseline disease stage

Study 18952 (N=1388)		**Study 18071 (N=951)**		**Study 18991 (N=1256)**	
Tumor stage	**N (%)**	**Tumor stage**	**N (%)**	**Tumor stage**	**N (%)**
TXN2M0	749 (54.0)	Stage III B	420 (44.2)	TanyN2MO	743 (59.2)
T4N0M0	355 (25.6)	Stage III C (>= 4LN+)	193 (20.3)	TanyN1M0	508 (40.4)
TXN1M0	283 (20.4)	Stage III A	186 (19.6)	TxN0M0	5 (0.4)
Missing	1 (0.1)	Stage III C (1-3 LN+)	152 (16.0)		

Table 2: Summary statistics of the EORTC QLQ-C30 scale scores at T1, T2 and T3

	Scale (n=1575 - 1840)													
	PF	**RF**	**SF**	**EF**	**CF**	**QL**	**PA**	**FA**	**NV**	**AP**	**DY**	**DI**	**SL**	**CO**
T1														
Median	93.3	83.3	100.0	83.3	100.0	75.0	100.0	66.7	100.0	100.0	100.0	100.0	100.0	100.0
Mean (SD)	86.1 (18.7)	76.5 (27.3)	83.7 (22.7)	79.5 (20.5)	88.3 (17.9)	70.2 (20.0)	84.1 (21.9)	69.6 (26.0)	92.4 (14.3)	82.1 (27.9)	88.6 (21.4)	89.5 (19.5)	81.5 (25.8)	93.3 (16.7)
T2														
Median	100.0	100.0	100.0	83.3	100.0	75.0	100.0	77.8	100.0	100.0	100.0	100.0	100.0	100.0
Mean (SD)	87.8 (17.7)	80.8 (24.8)	85.4 (21.8)	79.7 (21.0)	85.9 (20.2)	71.3 (20.3)	84.0 (22.8)	72.2 (24.0)	94.0 (13.6)	86.2 (24.1)	87.2 (21.8)	89.2 (20.5)	79.2 (26.6)	92.0 (18.8)
T3														
Median	100.0	100.0	100.0	83.3	100.0	75.0	100.0	77.8	100.0	100.0	100.0	100.0	100.0	100.0

Table 2: Summary statistics of the EORTC QLQ-C30 scale scores at T1, T2 and T3

	Scale (n=1575 - 1840)													
	PF	**RF**	**SF**	**EF**	**CF**	**QL**	**PA**	**FA**	**NV**	**AP**	**DY**	**DI**	**SL**	**CO**
Mean	86.7	81.1	85.6	80.2	86.5	72.0	84.1	75.5	94.9	90.1	87.9	92.9	79.4	92.3
(SD)	(19.7)	(27.2)	(23.8)	(21.5)	(20.2)	(22.1)	(23.9)	(24.9)	(14.3)	(21.7)	(21.8)	(17.4)	(27.2)	(17.8)

T1, T2 and T3 are time points for start of treatment, end of treatment and end of follow-up respectively. AP, appetite loss; CF, cognitive functioning; CO, constipation; DI, diarrhea; DY, dyspnea; EF, emotional functioning; FA, fatigue; NV, nausea/vomiting; PA, pain; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; SL, sleep disturbance; SD, standard deviation.

Table 3: Cross-sectional correlations of the EORTC QLQ-C30 scale scores with anchors, and correlations between their change scores

		Cross-sectional			**Change scores**	
Scale	**Anchor**	**T1**	**T2**	**T3**	**T2-T1**	**T3-T2**
PF	Performance status	-0.39	-0.41	-0.35	-0.23	-0.28
	CTCAE Anorexia	-0.41	-0.34	-0.35	-0.18	-0.26
	CTCAE Fatigue	-0.38	-0.36	-0.29	-0.19	-0.16
RF	Performance status	-0.36	-0.51	-0.48	-0.22	-0.35
	CTCAE Gastrointestinal	-0.44	-0.38	-0.39	-0.26	-0.23
	CTCAE Anorexia	-0.45	-0.44	-0.45	-0.24	-0.3
	CTCAE Fatigue	-0.53	-0.46	-0.4	-0.24	-0.24
SF	Performance status	-0.35	-0.50	-0.47	-0.22	-0.32
	CTCAE Gastrointestinal	-0.43	-0.41	-0.37	-0.25	-0.24
	CTCAE Anorexia	-0.46	-0.44	-0.4	-0.26	-0.26
	CTCAE Fatigue	-0.45	-0.44	-0.4	-0.23	-0.2
EF	Performance status	-0.18	-0.37	-0.3	-0.12	-0.24
	CTCAE Nervous system	-0.44	-0.48	-0.4	-0.21	-0.17
	CTCAE Anorexia	-0.34	-0.41	-0.3	-0.14	-0.21
CF	Performance status	-0.28	-0.36	-0.3	-0.14	-0.18
	CTCAE Nervous system	-0.39	-0.36	-0.32	-0.12	-0.19
	CTCAE Fatigue	-0.39	-0.37	-0.3	-0.11	-0.13
QL	Performance status	-0.38	-0.48	-0.44	-0.26	-0.36
	CTCAE Nausea	-0.45	-0.42	-0.38	-0.24	-0.23
	CTCAE Gastrointestinal	-0.48	-0.39	-0.38	-0.3	-0.3
	CTCAE Anorexia	-0.49	-0.46	-0.46	-0.26	-0.32
PA	Performance status	-0.26	-0.43	-0.44	-0.18	-0.3
	CTCAE Pain	-0.39	-0.43	-0.35	-0.19	-0.28
	CTCAE Immune	-0.38	-0.49	-0.44	-0.26	-0.24
FA	Performance status	-0.39	-0.49	-0.46	-0.26	-0.36
	CTCAE Gastrointestinal	-0.53	-0.42	-0.45	-0.3	-0.3
	CTCAE Anorexia	-0.55	-0.49	-0.5	-0.26	-0.34
	CTCAE Fatigue	-0.63	-0.55	-0.45	-0.3	-0.3
NV	CTCAE Nausea	-0.67	-0.73	-0.58	-0.43	-0.31
	CTCAE Gastrointestinal	-0.61	-0.57	-0.5	-0.33	-0.35
	CTCAE Anorexia	-0.51	-0.53	-0.49	-0.21	-0.4
AP	CTCAE Nausea	-0.6	-0.53	-0.54	-0.3	-0.24
	CTCAE Gastrointestinal	-0.71	-0.60	-0.62	-0.41	-0.38

	CTCAE Anorexia	-0.76	-0.72	-0.67	-0.43	-0.52
DY	Performance status	-0.37	-0.35	-0.34	-0.15	-0.2
	CTCAE Fatigue	-0.38	-0.42	-0.33	-0.15	-0.21
DI	CTCAE Gastrointestinal	-0.38	-0.38	-0.34	-0.26	-0.37
	CTCAE Diarrhea	-0.76	-0.68	-0.56	-0.5	-0.53
SL	Performance status	-0.16	-0.28	-0.3	-0.09	-0.18

T1, T2 and T3 are time points for start of treatment, end of treatment and end of follow-up respectively. AP, appetite loss; CF, cognitive functioning; CO, constipation; DI, diarrhea; DY, dyspnea; EF, emotional functioning; FA, fatigue; NV, nausea/vomiting; PA, pain; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; SL, sleep disturbance.
Example of cross-sectional correlations: PF at T1 vs. Performance status at T1 = -0.39, PF at T2 vs. Performance status at T2 =-0.41 and PF at T3 vs. Performance status at T3 =-0.35.
Example of change score correlations: (PF at T2 - PF at T1) vs. (Performance status at T2 – Performance status at T1)= -0.23 and (PF at T3 - PF at T2) vs. (Performance status at T3 – Performance status at T2)= -0.28

Table 4: Range of anchor-based MID estimates from the mean change method and linear regression

	Mean change method		**Linear regression**	
	T2-T1	**T3-T2**	**T2-T1**	**T3-T2**
Scale	**Improvement (Deterioration)**	**Improvement (Deterioration)**	**Improvement (Deterioration)**	**Improvement (Deterioration)**
PF	5 to 7 (-6)	4 to 6 (-9 to -6)	4 to 6 (-6)	5 to 7 (-8 to -5)
RF	11 to 15 (-6)	7 to 16 (-14 to -8)	9 to 12 (-8)	7 to 14 (-16 to -9)
SF	6 to 11 (-8 to -5)	6 to 11 (-12 to -6)	4 to 11 (-9 to -7)	6 to 10 (-13 to -6)
EF	4 to 7 (-6)	4 to 9 (-6 to -5)	4 to 7 (-6)	3 to 8 (-7 to -6)
CF	nM (-7 to -5)	4 to 6 (-4)	nM (-4 to -3)	5 to 7 (-3)
QL	5 to 9 (-10 to -9)	7 to 13 (-10 to -7)	5 to 9 (-9)	7 to 12 (-11 to -7)
PA	4 to 5 (-10 to -7)	7 to 9 (-16 to -9)	4 (-11 to -7)	4 to 9 (-15 to -8)
FA	4 to 14 (-13 to -5)	9 to 15 (-14 to -6)	6 to 13 (-11 to -6)	9 to 15 (-14 to -6)
NV	7 to 8 (-8 to -5)	7 to 10 (-7 to -6)	6 to 7 (-8 to -6)	8 to 10 (-7 to -6)
AP	17 to 18 (-15 to -9)	12 to 14 (-14 to -8)	16 (-11 to -16)	12 to 14 (-14 to -8)
DY	nM (-8 to -5)	8 (-9 to -7)	nM (-6 to -4)	9(-8 to -5)
DI	5 (nM)	nM (nM)	5 (nM)	nM (nM)
SL	nM (nM)	nM (-9)	nM (nM)	nM (-9)

MIDs from the mean change method and the linear regression are useful for interpreting within-group and between-groups change respectively
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 (nM) is used where no MID estimate is available; either due to the absent of a suitable anchor or ES were either <0.2 or ≥ 0.8
Abbreviations: T1, T2 and T3 are time points for start of treatment, end of treatment and end of follow-up respectively. AP, appetite loss; CF, cognitive functioning; CO, constipation; DI, diarrhea; DY, dyspnea; EF, emotional functioning; FA, fatigue; NV, nausea/vomiting; PA, pain; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; SL, sleep disturbance

Table 5: Summary of anchor-based MIDs for within and between group change compared with distribution-based MID estimates

Scale	Anchor-based MID for within-group change		Anchor-based MID for between-groups change		Distribution-based: QOL scores at T1 (n=1829 -1839)			
	T2-T1	T3-T2	T2-T1	T3-T2				
	Improvement (Deterioration)	Improvement (Deterioration)	Improvement (Deterioration)	Improvement (Deterioration)	0.2 SD	0.3 SD	0.5 SD	1 SEM
PF	6 (-6)	5 (-8)	5 (-6)	6 (-7)	3.7	5.6	9.4	5.6
RF	13 (-6)	12 (-10)	11 (-9)	11 (-11)	5.5	8.2	13.6	11.6
SF	8 (-7)	9 (-9)	7 (-8)	8 (-9)	4.5	6.8	11.3	8.2
EF	6 (-6)	8 (-5)	6 (-6)	6 (-6)	4.1	6.1	10.2	7.6
CF	nM (-6)	5 (-4)	nM (-3)	6 (-3)	3.6	5.4	8.9	7.6
QL	7 (-9)	11 (-9)	7 (-9)	10 (-10)	4.0	6.0	10.0	8.5
PA	4 (-8)	8 (-12)	4 (-9)	7 (-12)	4.4	6.6	11.0	8.2
FA	10 (-8)	13 (-11)	10 (-8)	13 (-11)	5.2	7.8	13.0	10.7
NV	7 (-7)	9 (-7)	7 (-8)	9 (-6)	2.9	4.3	7.2	8.7
AP	18 (-12)	13 (-11)	16 (-13)	13 (-11)	5.6	8.4	14.0	12.8
DY	nM (-7)	8 (-8)	nM (-5)	9 (-7)	4.3	6.4	10.7	8.8
DI	5 (nM)	nM (nM)	5 (nM)	nM (nM)	3.9	5.9	9.8	10.3
SL	nM (-4)	nM (-9)	nM (-4)	nM (-9)	5.2	7.8	12.9	11.3
CO	nM (nM)	nM (nM)	nM (nM)	nM (nM)	3.4	5.0	8.4	6.9

The within group MIDs (from the mean change method) and the between groups MIDs (from the linear regression) were summarized via weighted averages based on scale/anchor pair 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 (nM) is used where no MID estimate is available; either due to the absent of a suitable anchor or ES were either <0.2 or ≥ 0.8
Abbreviations: T1, T2 and T3 are time points for start of treatment, end of treatment and end of follow-up respectively. AP, appetite loss; CF, cognitive functioning; CO, constipation; DI, diarrhea; DY, dyspnea; EF, emotional functioning; FA, fatigue; NV, nausea/vomiting; PA, pain; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; SL, sleep disturbance
Note: no suitable anchors were found for constipation.

APPENDIX

Table A.1: Frequency of patients by change scores of anchors

		CTCAE Nausea	CTCAE Anorexia	CTCAE Nervous system	CTCAE Immune	CTCAE Gastro-intestinal	CTCAE Diarrhea	CTCAE Pain	CTCAE Fatigue	Performance status
	Anchor change score									
T2-T1	**-4**								2 (0.1)	
	-3	5 (0.3)	6 (0.3)	5 (0.3)	6 (0.7)	9 (0.5)		9 (0.5)	8 (0.4)	
	-2	50 (2.7)	47 (2.6)	42 (2.3)	44 (4.9)	68 (3.7)	4 (0.4)	69 (3.8)	72 (3.9)	5 (0.3)
	-1	206 (11.2)	229 (12.5)	179 (9.8)	178 (19.8)	363 (19.8)	43 (4.6)	332 (18.2)	325 (17.7)	91 (5.6)
	0	1491 (81.3)	1459 (79.6)	1404 (76.6)	563 (62.6)	1199 (65.4)	825 (88.3)	1166 (63.8)	1178 (64.3)	1373 (85)
	1	71 (3.9)	77 (4.2)	151 (8.2)	85 (9.5)	158 (8.6)	53 (5.7)	199 (10.9)	201 (11.0)	136 (8.4)
	2	10 (0.6)	13 (0.7)	38 (2.1)	19 (2.1)	30 (1.6)	8 (0.9)	48 (2.6)	41 (2.2)	9 (0.6)
	3		2 (0.1)	13 (0.7)	4 (0.4)	6 (0.3)	1 (0.1)	6 (0.3)	6 (0.3)	1 (0.1)
	4			1 (0.1)						
	Total	1833	1833	1833	899	1833	934	1829	1833	1615
T3-T2	**-4**	1 (0.1)		3 (0.2)		4 (0.3)		1 (0.1)	2 (0.1)	
	-3	3 (0.2)	4 (0.3)	13 (0.8)	10 (1.1)	19 (1.2)	6 (0.9)	9 (0.6)	12 (0.8)	2 (0.2)
	-2	15 (0.9)	19 (1.2)	19 (1.2)	10 (1.1)	46 (2.9)	10 (1.4)	32 (2)	43 (2.7)	11 (0.9)
	-1	63 (4)	61 (3.8)	145 (9.1)	68 (7.7)	145 (9.1)	45 (6.4)	182 (11.5)	185 (11.7)	120 (9.8)
	0	1463 (92.2)	1440 (90.8)	1298 (81.8)	738 (83.7)	1283 (80.9)	638 (90.6)	1223 (77.1)	1208 (72.6)	982 (80.1)
	1	31 (2)	44 (2.8)	78 (4.9)	40 (4.5)	63 (4)	2 (0.3)	108 (6.8)	99 (6.2)	93 (7.6)
	2	7 (0.4)	14 (0.9)	22 (1.4)	13 (1.5)	21 (1.3)	3 (0.4)	27 (1.7)	32 (2)	15 (1.2)
	3	3 (0.2)	4 (0.3)	8 (0.5)	1 (0.1)	4 (0.3)		4 (0.3)	4 (0.3)	3 (0.2)
	4				2 (0.2)	1 (0.1)			1 (0.1)	
	Total	1586	1586	1586	882	1586	704	1586	1586	1226

CTCAE, Common terminology criteria for adverse events

T1, T2 and T3 are time points for start of treatment, end of treatment and end of follow-up respectively.

Anchor change scores: -4 to -1, 0 and 1to 4 represents improvement, no change, and deterioration respectively. Only the -1, 0 and 1 change score categories were used to estimate MIDs. No MIDs for deterioration were calculated for CTCAE Diarrhea between T2 and T3 because only 2 patients experienced a clinically minimal deterioration.

Table A.2: Means (effect sizes) of HRQOL change scores in three clinical change groups that are based on selected anchors per EORTC QLQ-C30 scale

		T2-T1			T3-T2		
Scale	**Anchor**	**Improvement (ES)** n = 43 to 363	**No change** n = 563 to 1491	**Deterioration (ES)** n = 53 to 201	**Improvement (ES)** n = 45 to 185	**No change** n = 638 to 1463	**Deterioration (ES)** n = 31 to 108
PF	Performance status	2.30 (0.16)†	-0.24	-6.42 (-0.45)	4.93 (0.30)	-1.03	-8.86 (-0.53)
	CTCAE Anorexia	7.13 (0.39)	0.80	-1.30 (-0.07)†	6.12 (0.34)	-0.52	-7.69 (-0.43)
	CTCAE Fatigue	5.03 (0.28)	1.46	-1.92 (-0.11)†	3.91 (0.22)	-0.73	-6.05 (-0.34)
RF	CTCAE Gastrointestinal	12.94 (0.48)	2.85	-6.01 (-0.22)	7.18 (0.28)	0.50	-1.32 (-0.05)†
	CTCAE Anorexia	15.36 (0.57)	2.94	-2.38 (-0.09)†	15.03 (0.58)	1.27	-7.95 (-0.31)
	CTCAE Fatigue	11.38 (0.43)	2.84	-0.66 (-0.02)†	10.90 (0.42)	1.11	-8.84 (-0.34)
	Performance status	14.29 (0.56)	2.74	-2.57 (-0.10)†	15.69 (0.59)	1.31	-14.31 (-0.54)
SF	CTCAE Gastrointestinal	8.36 (0.38)	0.88	-7.75 (-0.35)	3.79 (0.17)†	0.00	-6.28 (-0.29)
	CTCAE Anorexia	11.26 (0.51)	0.68	-8.33 (-0.37)	10.00 (0.45)	0.56	-11.24 (-0.51)
	CTCAE Fatigue	5.68 (0.26)	1.67	-4.92 (-0.22)	5.89 (0.27)	0.29	-6.77 (-0.31)
	Performance status	7.78 (0.37)	1.15	-6.42 (-0.30)	10.54 (0.47)	0.69	-11.96 (-0.53)
EF	Performance status	2.44 (0.13)†	0.07	-3.07 (-0.16)†	7.75 (0.36)	0.82	-5.25 (-0.24)
	CTCAE Anorexia	4.20 (0.21)	-0.07	-5.81 (-0.29)	8.75 (0.41)	0.91	-5.56 (-0.26)
	CTCAE Nervous system	7.08 (0.35)	0.19	-2.31 (-0.11)†	4.14 (0.20)	0.95	-3.17 (-0.15)†
CF	Performance status	-2.22 (-0.15)†	-2.79	-7.04 (-0.46)	3.92 (0.21)	-1.13	-2.90 (-0.15)†
	CTCAE Nervous system	-1.13 (-0.06)†	-1.91	-4.92 (-0.28)	5.06 (0.26)	-0.66	-4.06 (-0.21)
	CTCAE Fatigue	-0.77 (-0.04)†	-2.33	-4.98 (-0.28)	5.80 (0.30)	-0.99	-1.20 (-0.06)†
QL	CTCAE Gastrointestinal	7.12 (0.36)	0.19	-9.08 (-0.46)	6.94 (0.33)	0.04	-6.85 (-0.33)
	CTCAE Anorexia	8.70(0.44)	0.10	-8.55 (-0.43)	13.19 (0.63)	0.82	-10.42(0.50)
	CTCAE Nausea	6.40 (0.33)	0.51	-8.57 (-0.44)	10.05 (0.48)	0.69	-9.17 (-0.44)
	Performance status	4.72 (0.25)	-0.31	-9.64 (-0.52)	13.11 (0.61)	0.69	-9.87 (-0.46)
PA	Performance status	4.95 (0.24)	0.68	-7.35 (-0.35)	7.64 (0.33)	-0.27	-15.59 (-0.68)
	CTCAE Pain	3.92 (0.20)	-0.06	-6.62 (-0.31)	8.79 (0.38)	-0.48	-8.49 (-0.37)
	CTCAE Immune	4.40 (0.20)	0.51	-10.39 (-0.28)	6.62 (0.29)	2.28	-2.92 (-0.13)†
FA	CTCAE Gastrointestinal	11.20 (0.44)	0.54	-7.17 (-0.28)	9.43 (0.39)	0.70	-5.64 (-0.24)
	CTCAE Anorexia	13.71 (0.53)	0.53	-4.47 (-0.17)†	15.03(0.62)	1.45	-12.88 (0.53)
	CTCAE Fatigue	9.91 (0.39)	0.95	-5.33 (-0.21)	14.29 (0.59)	0.41	-11.11 (-0.46)
	Performance status	4.27 (0.20)	-1.95	-12.54 (-0.59)	14.44 (0.61)	-0.35	-14.40 (-0.61)
NV	CTCAE Gastrointestinal	7.10 (0.53)	0.78	-7.75 (-0.57)	7.24 (0.59)	-0.48	-7.41 (-0.60)
	CTCAE Anorexia	7.63 (0.55)	0.71	-5.19 (-0.38)	12.02 (0.90)†	-0.15	-6.44 (-0.48)
	CTCAE Nausea	12.50 (0.92)†	0.47	-13.62 (-1.00)†	9.79 (0.77)	-0.18	-5.91 (-0.46)
AP	CTCAE Gastrointestinal	16.94 (0.64)	1.04	-10.34(-0.39)	13.85 (0.64)	-0.13	-8.20 (-0.38)
	CTCAE Anorexia	23.16 (0.88)†	1.10	-14.91 (-0.56)	28.25(1.29)†	0.49	-14.29 (-0.65)
	CTCAE Nausea	18.37 (0.68)	2.36	-8.45 (-0.31)	12.37 (0.56)	0.90	-3.23 (-0.15)†
DY	Performance status	-0.37 (-0.02)†	-2.08	-8.33 (-0.46)	3.33 (0.17)†	-1.02	-9.06 (-0.47)
	CTCAE Fatigue	2.37 (0.11)†	-1.26	-5.03 (-0.23)	8.15 (0.40)	-1.14	-6.46 (-0.32)
DI	CTCAE Gastrointestinal	5.09 (0.27)	-0.28	-14.86 (-0.80)†	16.20 (0.85)†	0.63	0.00 (0.00)†
	CTCAE Diarrhea	20.16 (1.16)†	-1.10	-27.67 (-1.59)†	31.85 (1.30)†	3.16	-
SL	Performance status	2.20 (0.09)†	-0.27	-4.41 (-0.17)†	3.89 (0.15)†	-0.31	-9.32 (-0.35)

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

All the ESs for the no change group were < 0.2

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: T1, T2 and T3 are time points for start of treatment, end of treatment and end of follow-up respectively. AP, appetite loss; CF, cognitive functioning; CO, constipation; DI, diarrhea; DY, dyspnea; EF, emotional functioning; FA, fatigue; NV, nausea/vomiting; PA, pain; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; SL, sleep disturbance

No results are presented for deterioration in DI scale based on CTCAE Diarrhea between T2 and T3 because only 2 patients experienced a clinically minimal deterioration.

Table A.3: Mean change scores based on the linear regression

		T2-T1		T3-T2	
Scale	**Anchor**	**Improvement**	**Deterioration**	**Improvement**	**Deterioration**
PF	Performance status	2.54†	-6.18	5.96	-7.84
	CTCAE Anorexia	6.32	-2.10†	6.64	-7.17
	CTCAE Fatigue	3.57	-3.38†	4.64	-5.32
RF	CTCAE Gastrointestinal	10.09	-8.86	6.68	-1.82†
	CTCAE Anorexia	12.42	-5.32†	13.76	-9.22
	CTCAE Fatigue	8.55	-3.50†	9.79	-9.95
	Performance status	11.55	-5.31†	14.38	-15.63
SF	CTCAE Gastrointestinal	7.47	-8.63	3.79†	-6.28
	CTCAE Anorexia	10.58	-9.01	9.44	-11.80
	CTCAE Fatigue	4.01	-6.58	5.59	-7.06
	Performance status	6.63	-7.57	9.86	-12.64
EF	Performance status	2.37†	-3.13†	6.93	-6.07
	CTCAE Anorexia	4.27	-5.74	7.84	-6.47
	CTCAE Nervous system	6.90	-2.50†	3.19	-4.12†
CF	Performance status	0.57†	-4.24	5.05	-1.77†
	CTCAE Nervous system	0.78†	-3.01	5.72	-3.40
	CTCAE Fatigue	1.55†	-2.65	6.78	-0.21†
QL	CTCAE Gastrointestinal	6.93	-9.27	6.90	-6.89
	CTCAE Anorexia	8.60	-8.65	12.37	-11.24
	CTCAE Nausea	5.90	-9.08	9.37	-9.85
	Performance status	5.03	-9.33	12.42	-10.56
PA	Performance status	4.26	-8.03	7.91	-15.32
	CTCAE Pain	3.97	-6.56	9.27	-8.01
	CTCAE Immune	3.90	-10.90	4.34	-5.20†
FA	CTCAE Gastrointestinal	10.66	-7.72	8.73	-6.34
	CTCAE Anorexia	13.18	-5.00†	13.57	-14.33
	CTCAE Fatigue	8.96	-6.29	13.88	-11.52
	Performance status	6.22	-10.59	14.79	-14.05
NV	CTCAE Gastrointestinal	6.32	-8.53	7.72	-6.92
	CTCAE Anorexia	6.92	-5.90	12.17†	-6.29
	CTCAE Nausea	12.03†	-14.09	9.97	-5.73
AP	CTCAE Gastrointestinal	15.91	-11.37	13.98	-8.07
	CTCAE Anorexia	22.05†	-16.02	27.76	-14.77
	CTCAE Nausea	16.01	-10.82	11.47	-4.12†
DY	Performance status	1.72†	-6.25	4.36†	-8.03
	CTCAE Fatigue	3.63†	-3.77	9.29	-5.33
DI	CTCAE Gastrointestinal	5.37	-14.58†	15.57	-0.63†

	CTCAE Diarrhea	21.26†	-26.57†	28.69	-
SL	Performance status	2.47†	-4.14†	4.20†	-9.01

Separate regression models were fitted for each scale/anchor pair: Outcome = HRQOL change score, covariate = binary anchor variable; coded as 'stable'=0 and 'improvement'=1 or 'deterioration'=1 for models on improvement and deterioration respectively. The mean change scores = slope parameters. No results are presented for deterioration in DI scale based on CTCAE Diarrhea between T2 and T3 because only 2 patients experienced a clinically minimal deterioration.

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

Abbreviations: T1, T2 and T3 are time points for start of treatment, end of treatment and end of follow-up respectively. AP, appetite loss; CF, cognitive functioning; CO, constipation; DI, diarrhea; DY, dyspnea; EF, emotional functioning; FA, fatigue; NV, nausea/vomiting; PA, pain; PF, physical functioning, QL, global quality of life; RF, role functioning; SF, social functioning; SL, sleep disturbance