Article:
Musoro, JZ, Coens, C, Greimel, E et al. (10 more authors) (2020) Minimally important differences for interpreting European Organisation for Research and Treatment of Cancer (EORTC) Quality of life Questionnaire core 30 scores in patients with ovarian cancer. Gynecologic Oncology, 159 (2). pp. 515-521. ISSN 0090-8258

https://doi.org/10.1016/j.ygyno.2020.09.007

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

Jammbe Z Musoro[1], Corneel Coens[1], Elfriede Greimel[2], Madeleine T King[3], Mirjam AG Sprangers[4], Andy Nordin[5], Eleonora BL van Dorst[6], Mogens Groenvold[7], Kim Cocks[8], Galina Velikova[9], Hans-Henning Flechtner[10] Andrew Bottomley[1] on behalf of the EORTC Gynecological and Quality of Life Groups

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

2Medical University Graz, Graz, Austria

3University of Sydney, Faculty of Science, School of Psychology, Sydney, NSW, Australia

4Department of Medical Psychology, Amsterdam University Medical Centers, location Academic Medical Center, University of Amsterdam, Cancer Center Amsterdam, The Netherlands.

5East Kent Gynaecological Oncology Centre, Queen Elizabeth the Queen Mother Hospital

6Department of Obstetrics and Gynecology, Academic Hospital Utrecht, Utrecht, the Netherlands

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

8Adelphi Values, Bollington, Cheshire, UK

9Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK.

10Clinic for Child and Adolescent Psychiatry and Psychotherapy, University of Magdeburg, Magdeburg, Germany

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

ABSTRACT

Introduction: Minimal important differences (MIDs) are useful for interpreting changes or differences in health-related quality of life scores in terms of clinical importance. There are currently no MID guidelines for the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire core 30 (EORTC QLQ-C30) specific to ovarian cancer. This study aims to estimate MIDs for interpreting group-level change of EORTC QLQ-C30 scores in ovarian cancer.

Methods: Data were derived from four EORTC published trials. Clinical anchors for each EORTC QLQ-C30 scale were selected using correlation strength and clinical plausibility. MIDs for within-group change and between-group differences in change over time were estimated via mean change method and linear regression respectively. For each EORTC QLQ-C30 scale, MID estimates from multiple anchors were summarized via weighted-correlation. Distribution-based MIDs were also examined as supportive evidence.

Results: Anchor-based MIDs were determined for deterioration in **7** of the 14 EORTC QLQ-C30 scales assessed, and in 11 scales for improvement. Anchor-based MIDs for within-group change ranged from 4 to 19 (improvement) and -9 to -4 (deterioration). Between-group MIDs ranged from 3 to 13 (improvement) and -11 to -4 (deterioration). Generally, absolute anchor-based MIDs for most scales ranged from 4 to 10 points.

Conclusions: Our findings will aid interpretation of EORTC QLQ-C30 scores in ovarian cancer and inform sample size calculations in future ovarian cancer trials with endpoints that are based on EORTC QLQ-C30 scales.

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

1. INTRODUCTION

Assessing patient reported outcomes such as health-related quality of life (HRQOL) has gained substantial importance in cancer clinical trials [1]. Consequently, there is increasing need to interpret HRQOL scores in a manner that is clinically relevant. Clinical interpretation of HRQOL scores is facilitated by the notion of minimally important difference (MID) [2, 3, 4, 5], which is defined as the smallest change in a HRQOL score that is perceived as "important" by a patient or by a third party (eg a clinician), which may indicate a change in the patient's management[3]. As HRQOL is often a secondary endpoint in clinical trials without associated power calculation, statistical significance of the observed treatment effects may not reflex clinical relevance. Knowledge on MID is useful to clinicians, patients and researchers as it can serve as a benchmark for assessing treatment efficacy, assisting with treatment decision-making, and informing sample size calculations in future clinical trials when HRQOL is the primary or co-primary endpoint. MIDs are commonly estimated via anchor-based [6, 7] and distribution-based methods [9]. Anchor-based methods rely on variables that have clinical relevance, or on patient/physician-derived ratings. For example, a worsening CTCAE pain grade can function as an anchor for patient's self-reported pain rating. Distribution-based methods use statistical features of the data, eg the 0.5 Standard deviation (SD) is commonly used to approximate a MID [5].

In ovarian cancer trials, it is acknowledged that assessing HRQOL as a secondary or co-primary end-point is critical for supporting the commonly used progression free survival outcome [10]. The European Organisation for Research and Treatment of Cancer Quality of life Questionnaire core 30 (EORTC QLQ-C30) is commonly used in ovarian cancer trials [10]. There are currently no MID guidelines for the EORTC QLQ-C30 specific to ovarian cancer.

Mean differences ≥10 points are widely considered as clinically meaningful when interpreting EORTC QLQ-C30 scales in clinical trials [11]. This rule mainly stems from overlaps in initial guidelines by King [2] and Osoba et al. [3] for interpreting the EORTC QLQ-C30. However, there is increasing evidence that MIDs can vary by EORTC QLQ-C30 scales, direction of

change (improvement/deterioration) and clinical settings [4, 6, 7, 12, 14]. This implies that a global rule for MIDs applicable to all settings is highly unlikely [13]. This study aims to estimate anchor-based MIDs for the EORTC QLQ-C30 scales specifically in ovarian cancer. Focus is on establishing MIDs for interpreting HRQOL change scores over time in groups of ovarian cancer patients. This work is part of the EORTC MID project that seeks to gather empirical evidence on patterns of MID estimates across EORTC QLQ-C30 scales and disease sites [6, 12, 14, 15, 16]. Our study differs from Osoba et al. [2] in that we used clinical anchors whereas Osoba used patient-derived ratings. Additionally, as opposed to guidelines by King [3] and Cocks et al. [4] that were based on meta-analyses of published studies, pooling across cancer sites, our study used individual patient data from published trials.

2. MATERIALS AND METHODS

Data description

The data were derived from four published phase III trials. Trial 1, carried out between October 2005 and February 2008, evaluated the efficacy of maintenance erlotinib after first-line chemotherapy (enrolled 835 patients [17]. Trial 2 assessed the benefits of early treatment on the basis of increased CA125 concentrations compared with delayed treatment on the basis of clinical recurrence (298 patients enrolled by the EORTC), and was conducted between May 1999, and August 2005 [18]. Trial 3, conducted between April 1994 and December 2009, compared a combination of Taxol-platinum and a combination of cyclophosphamide-platinum chemotherapy in the treatment of advanced epithelial ovarian cancer (231 patients enrolled by the EORTC) [19]. Trial 4 compared upfront debulking surgery versus neo-adjuvant chemotherapy in patients with Stage IIIc or IV epithelial ovarian carcinoma (670 patients enrolled by the EORTC), and was conducted between September 1998 and January 2009 [20]. All trials assessed EORTC QLQ-C30 scores at baseline, during treatment, and at multiple follow-up time points after the end of treatment. All patients with at least a valid HRQOL form were included in the study. No adjustment for missing data was made.

The EORTC QLQ-C30

This questionnaire comprises 30 items that are aggregated into 9 multi-item scales; global health status, physical functioning, role functioning, emotional functioning, cognitive functioning, social functioning, fatigue, pain, nausea and vomiting, and 6 single-item scales; dyspnoea, sleep disturbance, appetite loss, constipation, diarrhoea, and financial impact. Trial 1, 2 and 4 used version 3 of the EORTC QLQ-C30, whereas trial 3 used EORTC QLQ-C30 (+3), which was converted to version 2 according to the scoring manual [21]. The EORTC QLQ-C30 version 2 and 3 differ only in the response categories of questions 1 to 5 (in the physical functioning domain), coded as yes/no in version 2, while 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 scoring manual [21], with means of the raw scores transformed to fall between 0 and 100. For consistency in signs, the symptom scores were reversed to follow the functioning scales' interpretation; 0 represents the worst possible score and 100, the best possible score. The financial impact scale was omitted because suitable anchors were not available.

Clinical anchor

Anchors were selected from available clinical data such as physician examinations, common terminology criteria for adverse events (CTCAE) and laboratory results. For each HRQOL scale, multiple anchors could be selected based on correlation strength. Revicki et al [5] suggested a correlation of $\geq|0.30|$ as acceptable. Depending on the distribution of the HRQOL scale/anchor pair, a polyserial or polychoric correlation was estimated. Retained anchors were verified for clinical plausibility by a panel of five ovarian cancer / HRQOL experts (among whom were practicing gynaecological oncologists and a clinical psychologist), to avoid spurious findings [16].

Definition of clinical change groups

Three clinical change groups (CCG) were defined: deterioration (worsened by 1 anchor category), stable (no change in anchor category), and improvement (improved by 1 anchor category). To avoid overestimating the MIDs, change scores ≥ 2-points were excluded from datasets used to estimate 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 example, patients with HRQOL assessments at time points t_a, t_b and t_c, contributed change scores computed between t_a & t_b, t_a & t_c and t_b & t_c. Hence a patient could contribute multiple change scores, and given their change scores, patients could contribute to multiple CCGs. 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.

A linear regression was used to estimate MIDs for differences between groups in change over time. For each 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 and 'stable' = 0 and 'deterioration' = 1 when modelling deterioration. Since patients could contribute change scores to multiple CCGs, and more than one change score to a particular CCG, we corrected for the association between multiple change scores within subjects via the generalized estimating equations approach [22]. The resulting slope parameters for the 'improved' and 'deteriorated' covariates correspond to the MID for improvement and deterioration respectively. These MIDs are useful for interpreting changes

over time between two distinct groups of patients. Anchor-based MIDs from multiple anchors per scale were summarized to a single value using weighted correlation. The summarized MIDs were computed separately for within-group change and between-group difference in change.

To assess whether MIDs varied between patient populations that started versus successfully completed first-line treatment, a binary treatment indicator was created where trial 1 and 2 were classified as 'post first-line' and trial 3 and 4 as 'pre first-line'. This treatment indicator and its interaction term with the anchor variable were included in a regression model. If an interaction term was statistically significant, MIDs were re-estimated separately for the two subgroups. To account for multiple testing, p-values below 0.001 were considered to be statistically significant.

Distribution-based methods

The 0.3 SD, 0.5 SD and standard error of measurement (SEM) were estimated at the following time points: (i) Start of treatment/observation (t1); time point before or on the first day of treatment administration (or observation for trial 2) and (ii) end of treatment (t2); last day of protocol treatment administration (or last follow-up date for trial 2).

An effect size (ES) was computed within each CCG by dividing the mean of the HRQOL change scores by the SD of the change scores over all time points. The computed ESs were used to screen anchor-based MIDs. Only mean changes with an ES of ≥ 0.2 and <0.8 were considered appropriate for inclusion as anchor-based MIDs. This was based on Cohen's [9] recommendations that an ES of 0.2 is small, 0.5 is moderate and ≥ 0.8 is large. The reason here was that observed ESs <0.2 reflected changes that were clinically unimportant, and those ≥ 0.8 were obviously more than minimally important. All statistical analyses were performed using SAS software [23].

3. RESULTS

A summary of patients characteristics at baseline per trial are presented in Table 1.

The median follow-up time (in months) for quality of life data was 6.1 (SD = 8.7) for trial 1, 2.6 (SD=8.4) for trial 2, 21.5 (SD=23) for trial 3 and 3.9 (SD=8) for trial 4. An overview of patient inclusion is presented in Figure S1.

Twenty two possible clinical anchors were initially evaluated for the EORTC QLQ-C30 scales. After prioritizing on cross-sectional correlation, 69 anchor and HRQOL scale pairs (comprising 3 to 6 anchors per scale) were preselected for review by the clinical panel. A total of 51 anchor and HRQOL scale pairs were further excluded for lack of clinical plausibility e.g., performance status versus cognitive functioning scale and CTCAE constipation versus diarrhoea scale. Table 2 presents the final list of retained anchors comprising WHO performance status (PS) and 7 CTCAEs (anorexia, pain, fatigue, nausea and vomiting, diarrhoea, gastrointestinal and constipation). 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 11 scales. Table 2 also presents the cross-sectional correlations between HRQOL scales and anchors ranging from 0.3 to 0.7 in absolute value, and the correlations between their change scores ranging from 0.1 to 0.4. The distribution of patients and the number of change observations across the anchor categories is presented in Table S1.

Table 3 summarises the anchor-based MIDs 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). An MID range is presented for scales with multiple anchors. Detailed results are presented in Table S2.

Anchor-based MIDs were determined for deterioration in 7 of the 14 scales assessed, and in 11 scales for improvement. MIDs varied according to HRQOL scale, direction of change scores (improvement versus deterioration) and selected anchor. This is illustrated in Figure 1, where estimates from the mean change method are plotted along with their 95% confidence intervals. All MIDs were always in the expected direction according to the anchor change group, i.e.

positive versus negative change scores within the improvement versus deterioration CCG respectively.

MIDs for within-group change ranged from 4 to 19 points for improvement and -9 to -4 points for deterioration, while MIDs for between-group change ranged from 3 to 13 for improvement and -11 to -4 for deterioration. Table 3 also presents a single value summary of MIDs based on a correlation-weighted average. Generally, MIDs for most EORTC QLQ-C30 scales ranged from 4 to 10 points in absolute values for both anchor-based methods.

The interaction effect between the anchor and treatment indicator was statistically significant (p value < 0.001) for diarrhoea (DI) /CTCAE gastrointestinal and fatigue (FA)/PS pairs, for both improvement and deterioration, suggesting significant different MIDs between the two treatment subgroups. For the pre first-line ovarian cancer subgroup, no MIDs were available since the resulting ESs were clinically unimportant (<0.2) or obviously more than minimally important (≥0.8). MIDs for improvement (deterioration) within the post first-line ovarian cancer subgroup were: DI; 9.17 (-12.46) for within-group and 9.83 (-11.07) for between-groups, and FA; 7.13 (-5.34) for within-group and 5.96 (-6.31) for between-groups. The estimates for DI were relatively higher compared to those obtained for the DI in Table 3, while those for FA were mostly lower compared to the corresponding weighted MIDs.

Table S3 presents distribution-based MIDs at t1 for 14 EORTC QLQ-C30 scales that were considered in this study. The distribution-based estimates at t2 for each HRQOL scale were similar to t1, mostly within < 1 point range. Anchor-based MIDs for improvement for most scales were similar to 0.5 SD, whereas those for deterioration tended to range from 0.2 to 0.3 SD.

4. DISCUSSION

Our study is the first to investigate MIDs for interpreting group-level change of EORTC QLQ-C30 scores over time in patients with ovarian cancer. Generally, MIDs for most scales were within the range of 4-10 points, which is similar to the 5-10 point range reported by Osoba et al.[3] in breast and small-cell lung cancer. A similar range was also observed by Cocks et al. [4] in pooled data across multiple cancer sites, by Musoro et al [6,12,14,15] in malignant melanoma, advanced breast cancer and head and neck cancer respectively, and by Maringwa et al. [7, 8] in lung and brain cancer respectively. We observed that MIDs for deterioration were lower than those for improvement, except for the diarrhoea scale. This is in contrast with Cocks et al. [4] where the estimates for deterioration tended to be larger than those for improvement. Note that other studies have also reported no systematic differences between MIDs for improving and deteriorating scores [6, 7, 8]. We also observed that the thresholds for some scales were much lower (MID for fatigue scale = 3 points) or larger (MID the role functioning scale =19 points), which supports earlier claims by Cocks et al. [4] that the 5-10 points rule may not be applicable for all settings. These increasing robust guidelines reiterates the importance of selecting scale specific MIDs with caution, taking into account the direction change (improvement versus deterioration) and disease setting.

We merged data from trials that used either version 2 or 3 of the EORTC QLQ-C30. Even though scales of the two questionnaire version were transformed to have values between 0 and 100, we recognize that physical functioning scale of version 2 can only take a limited range of values compared to version 3. However, our results showed that the MIDs for physical functioning scale did not depend on the EORTC QLQ-C30 version (i.e. version 2 versus version 3). Similar results have been previously reported in other cancer sites [14].

MIDs are often varied as a consequence of there being numerous anchors, various distribution-based criteria, and multiple HRQOL scales. We acknowledge that researchers and clinicians may find such a range of options confusing. So, to provide a single MID value per scale (when multiple anchors were used), we calculated a correlation-weighted average. We understand that

researchers and clinicians can choose to work with either the ranges or the single values provided in Table 3.

Given the rapid increase in using HRQOL scores for managing individual patients, our MIDs can be a useful starting point for defining cut-offs for individual-level change that are clinically meaningful for ovarian cancer patients [24]. For example in clinical trials, patients who change by the MID or more can be considered 'responders' and the proportion of responders can be compared between treatments. In clinical practice, our MIDs can serve as screening thresholds for identifying patients with clinically important problems. However, two important caveats apply to setting thresholds for use at individual level [13, 24]. First, not all MID values will translate into a score that is achievable for an individual because every HRQOL scale has a limited number of observable values. For example, any single-item scale from the EORTC QLQ-C30 has only four possible values, resulting in discrete number of change scores, while the multi-item scales have many more possible values and hence change scores. For single-item scales in particular, it may be necessary to select values on either side of the MID for individual thresholds, with selection of either the higher or lower value depending on clinical context. The second caveat is that individual thresholds must be set above bounds of measurement error to avoid false positive changes that might trigger unjustified clinical actions [24].

Although we focused on obtaining anchor-based MIDs, distribution-based estimates were also provided. The anchor-based and distribution-based approaches have their pros and cons. For instance, while anchor-based methods are often preferred because they consider what a meaningful change is to patients or clinicians, they do not account for the measurement precision of the used instrument or anchor. On the other hand, the distribution-based approach accounts for measurement precision but lacks information about the clinical relevance of observed changes, and are much more sample dependent. It is recommended that distribution-based methods be used as supportive evidence to anchor-based methods [5]. Our results showed that anchor-based MIDs for improvement for most scales were close to 0.5 SD, while those for deterioration mainly ranged from 0.2 to 0.3 SD, supporting the plausibility of these estimates.

As limitation, anchor-based MIDs were only available for EORTC QLQ-C30 scales for which suitable anchors were available in the database. No suitable anchors were found for 3 of the 14 scales assessed; cognitive functioning, dyspnoea and sleep disturbance. Furthermore, the anchors relied exclusively on clinical interpretations. Although cross-sectional correlations between the HRQOL scale/anchor pairs were mostly greater than the recommended 0.3 threshold, the correlations between their change scores were mostly suboptimal (<0.3). The low correlations may be due to the subjective nature of the anchors, particularly the CTCAEs, which can be misrepresented by the different physicians compared to patients' ratings as already reported by Basch et al [25], and also because a change variable is intrinsically more variable than a cross-sectional observation due to the compounding of measurement error at both time points. Although it is reassuring that our MID estimates are comparable to previously published MIDs [3, 6, 7, 8, 12], it is still interesting to compare these MIDs to those in future studies that use anchors with stronger correlations. Another limitation was the lack of anchors that are based on patients' perspective of change. Such anchors are particularly important since we are dealing with patient reported outcomes. However, patients' self-rating of change in various QLQ-C30 scales are not always available in retrospective databases, and would need to be planned as future work to supplement the current results. In addition, it is reassuring that there was considerable overlap between our findings and those of Osoba et al.[3], which was based on patients' ratings of change as anchor.

It is important to note that our data is limited to four clinical trials, each with specific selection and treatment criteria. Trial 1 was in the maintenance setting after first line chemotherapy, while trial 2 evaluated patients who had completed front line therapies. Trials 3 and 4 on the other hand included patients undergoing initial treatment for ovarian cancer. Unfortunately, there were no available trials involving patients with recurrent ovarian cancer or targeting women with platinum-resistant ovarian cancer. Thus, extrapolation beyond the specific settings of the trials used in this study should be done with caution, especially for the palliative setting which is not represented in our sample. In addition, clinical trial populations may not represent

adequately the overall cancer population in terms of prognosis, treatment options, ethnicity, education or affluence. Even within our four ovarian cancer trials, significant differences in MIDs were found between the two treatment categories. Our data are also limited by the lower prevalence of high PS or CTCAE toxicity grades, as patient mainly reported grade 0, 1 or 2 during the trial. These results form part of a larger overarching project that aims to develop an evidence-based MID catalogue that is more refined than the single value rule-of-thumb currently still in use. However, an overly granular approach would be too data driven and impractical. Therefore we aim to further undertake a comprehensive synthesis of MID estimates to identify plausible ranges based on patterns across multiple sources, beyond retrospective clinical anchors.

In conclusion, our findings will help clinicians and researchers to interpret the clinical relevance of group-level change of selected EORTC QLQ-C30 scores over time in ovarian cancer. The provided estimates can be a useful benchmark for assessing the effectiveness of an intervention and for computing sample size in future ovarian cancer trials with endpoints that are based on EORTC QLQ-C30 scales.

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. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res. 1996; 5: 555-567.
3. Osoba D et al. Interpreting the significance of changes in health related quality-of-life scores. J Clin Oncol. 1998; 16: 139-144.
4. 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.
5. 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
6. 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
7. 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.
8. 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.

9. Cohen J. Statistical Power Analysis for the Behavioural Sciences (2nd Edition). Lawrence Erlbaum Associates, NJ, USA (1988).
10. Joly F, Hilpert F, Okamoto A et al. Fifth Ovarian Cancer Consensus Conference of the Gynecologic Cancer InterGroup: Recommendations on incorporating patient-reported outcomes in clinical trials in epithelial ovarian cancer. Eur J Cancer 2017;78:133–138
11. Cocks K, King MT, Velikova G, et al: Quality, interpretation and presentation of European Organisation for Research and Treatment of Cancer quality of life questionnaire core 30 data in randomised controlled trials. Eur J Cancer (2008) 44:1793-1798.
12. Musoro J, Coens C, Fiteni F, et al. EORTC Breast and Quality of Life Groups, Minimally Important Differences for Interpreting EORTC QLQ-C30 Scores in Patients With Advanced Breast Cancer, JNCI Cancer Spectrum, Volume 3, Issue 3, September 2019, pkz037, https://doi.org/10.1093/jncics/pkz037
13. 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.
14. Musoro JZ, Coens C, Singer S, et al. 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 [published online ahead of print, 2020 Jul 6]. Head Neck. 2020;10.1002/hed.26363. doi:10.1002/hed.26363.
15. Musoro, J.Z., Sodergren, S.C., Coens, C., Pochesci, A., Terada, M., King, M.T., Sprangers, M.A.G., Groenvold, M., Cocks, K., Velikova, G., Flechtner, H.-H., Bottomley, A. and (2020), Minimally important differences for interpreting the EORTC QLQ-C30 in patients with advanced colorectal cancer treated with chemotherapy. Colorectal Dis. doi:10.1111/codi.15295.
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. Vergote IB, Jimeno A, Joly F, eta al. Randomized phase III study of erlotinib versus observation in patients with no evidence of disease progression after first-line platin-based

chemotherapy for ovarian carcinoma: a European Organisation for Research and Treatment of Cancer-Gynaecological Cancer Group, and Gynecologic Cancer Intergroup study. J Clin Oncol. 2014; 32(4):320-6. doi: 10.1200/JCO.2013.50.5669.

18. Rustin GJ1, van der Burg ME, Griffin CL, et al. Early versus delayed treatment of relapsed ovarian cancer (MRC OV05/EORTC 55955): a randomised trial. Lancet. 2010; 376(9747):1155-63. doi: 10.1016/S0140-6736(10)61268-8.
19. Piccart MJ, Bertelsen K, James K,, et al. Randomized intergroup trial of cisplatin-paclitaxel versus cisplatin–cyclophosphamide in women with advanced epithelial ovarian cancer: three-year results. J Natl Cancer Inst, 92 (2000), pp. 699-708.
20. Vergote I, Coens C, Nankivell M, eta al. Neoadjuvant chemotherapy versus debulking surgery in advanced tubo-ovarian cancers: pooled analysis of individual patient data from the EORTC 55971 and CHORUS trials. Lancet Oncol. 2018 ;19(12):1680-1687. doi: 10.1016/S1470-2045(18)30566-7.
21. 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.
22. Liang KY, Zeger SL. Regression analysis for correlated data. *Annu. Rev.* Pub Health. 1993;14:43–68.
23. Institute Inc. 2013. Base SAS® 9.4 Procedures Guide. Cary, NC: SAS Institute Inc.
24. 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.
25. 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.

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

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

Conflict of interest statement: None declared.

Author contributors: JZM, CC, EG, MTK, MAGS, MG, KC, GV, H-HF and AB: contributed to the conception and design of the study. EG, MTK, MAGS, AN, EBLVD, MG, KC and GV: reviewed the proposed methodology for clinical plausibility. ZJM, CC, KC, and MTK: provided critical input on the statistical analysis. ZJM: Performed the statistical analysis and drafted the manuscript. All the authors read and corrected the drafts and approved the final version.

Tables

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

	Study				
	55041 (N=835)	**55955 (N=298)**	**55931 (N=231)**	**55971 (N=670)**	**Total (N=2034)**
	N (%)	**N (%)**	**N (%)**	**N (%)**	**N (%)**
Performance status					
0	559 (66.9)	199 (66.8)	121 (52.4)	300 (44.8)	1179 (58.0)
1	276 (33.1)	96 (32.2)	83 (35.9)	284 (42.4)	739 (36.3)
2	0 (0.0)	3 (1.0)	25 (10.8)	84 (12.5)	112 (5.5)
3	0 (0.0)	0 (0.0)	2 (0.9)	0 (0.0)	2 (0.1)
Unknown/Missing	0 (0.0)	0 (0.0)	0 (0.0)	2 (0.3)	2 (0.1)
International Federation of Gynecology and Obstetrics (FIGO) Stage					
I	57 (6.8)	41 (13.8)	0 (0.0)	0 (0.0)	98 (4.8)
II	62 (7.4)	35 (11.7)	10 (4.3)	0 (0.0)	107 (5.3)
III	563 (67.4)	179 (60.1)	165 (71.4)	511 (76.3)	1418 (69.7)
IV	152 (18.2)	43 (14.4)	56 (24.2)	157 (23.4)	408 (20.1)
Unknown/Missing	1 (0.1)	0 (0.0)	0 (0.0)	2 (0.3)	3 (0.1)
Country					
France	328 (39.3)	42 (14.1)	9 (3.9)	11 (1.6)	390 (19.2)
Belgium	98 (11.7)	17 (5.7)	31 (13.4)	133 (19.9)	279 (13.7)
Netherlands	46 (5.5)	87 (29.2)	31 (13.4)	104 (15.5)	268 (13.2)
Spain	63 (7.5)	111 (37.2)	5 (2.2)	62 (9.3)	241 (11.8)

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

	Study				
	55041 (N=835)	**55955 (N=298)**	**55931 (N=231)**	**55971 (N=670)**	**Total (N=2034)**
	N (%)	**N (%)**	**N (%)**	**N (%)**	**N (%)**
Italy	75 (9.0)	0 (0.0)	107 (46.3)	38 (5.7)	220 (10.8)
United Kingdom	86 (10.3)	0 (0.0)	3 (1.3)	101 (15.1)	190 (9.3)
Austria	93 (11.1)	3 (1.0)	0 (0.0)	11 (1.6)	107 (5.3)
Canada	0 (0.0)	0 (0.0)	0 (0.0)	83 (12.4)	83 (4.1)
Norway	0 (0.0)	0 (0.0)	0 (0.0)	82 (12.2)	82 (4.0)
Australia	40 (4.8)	0 (0.0)	0 (0.0)	0 (0.0)	40 (2.0)
Portugal	4 (0.5)	0 (0.0)	17 (7.4)	7 (1.0)	28 (1.4)
South Africa	0 (0.0)	28 (9.4)	0 (0.0)	0 (0.0)	28 (1.4)
Sweden	0 (0.0)	0 (0.0)	0 (0.0)	23 (3.4)	23 (1.1)
Switzerland	0 (0.0)	0 (0.0)	14 (6.1)	0 (0.0)	14 (0.7)
Denmark	0 (0.0)	0 (0.0)	0 (0.0)	13 (1.9)	13 (0.6)
Israel	0 (0.0)	0 (0.0)	12 (5.2)	0 (0.0)	12 (0.6)
Ireland	0 (0.0)	10 (3.4)	0 (0.0)	1 (0.1)	11 (0.5)
Greece	0 (0.0)	0 (0.0)	2 (0.9)	0 (0.0)	2 (0.1)
New Zealand	2 (0.2)	0 (0.0)	0 (0.0)	0 (0.0)	2 (0.1)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.1)	1 (0.0)
Age					
Mean (SD)	58.93 (10.37)	58.16 (11.34)	56.12 (11.35)	61.46 (10.26)	59.33 (10.73)
Range	19.0 - 85.0	19.0 - 85.0	22.0 - 85.0	25.0 - 86.0	19.0 - 86.0

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

		Scores		Change scores	
Scale	**Anchor**	**n_1 (n_{1R})***	**Correlation**	**n_2 (n_{2R})***	**Correlation**
PF	Performance status	1715 (6397)	-0.41	1226 (14972)	-0.23
RF	Performance status	1714(6391)	-0.47	1226 (14956)	-0.30
SF	Performance status	1713(6377)	-0.38	1226 (14921)	-0.22
EF	Performance status	1715 (6389)	-0.30	1226 (14948)	-0.20
QL	Performance status	1715 (6363)	-0.41	1226 (14828)	-0.25
	CTCAE Anorexia	796 (3333)	-0.33	668 (8606)	-0.12
PA	Performance status	1717(6408)	-0.35	1226 (15028)	-0.20
	CTCAE Pain	947 (4413)	-0.34	813 (13025)	-0.20
FA	Performance status	1717(6400)	-0.45	1226 (14973)	-0.30
	CTCAE Fatigue	796 (3351)	-0.39	668 (8684)	-0.15
	CTCAE Anorexia	796 (3352)	-0.35	668 (8685)	0.10
NV	Performance status	1716(6394)	-0.34	1226 (14961)	-0.20
	CTCAE Nausea & vomiting	1832 (10078)	-0.41	813 (12986)	-0.20

AP	Performance status	1715 (6379)	-0.41	1226 (14907)	-0.21
	CTCAE Anorexia	795 (3343)	-0.51	668 (8664)	-0.25
DI	CTCAE Diarrhoea	1540 (6461)	-0.69	813 (12850)	-0.40
	CTCAE Gastrointestinal	1832 (9991)	-0.38	813 (12850)	-0.23
CO	CTCAE Constipation	948 (4366)	-0.43	813 (12733)	-0.13

* n_1 (n_{1R}) and n_2 (n_{2R}) can vary by anchor and EORTC QLQ-C30 scale.

Abbreviations:

EORTC QLQ-C30 = European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire core 30; n_1 = number of patients with at least 1 matched EORTC QLQ-C30 and an anchor form; n_{1R} = number of repeated anchor and HRQOL matched forms across all subjects; n_2 = number of patients with at least 2 matched EORTC QLQ-C30 and an anchor form (at least 2 forms are needed to compute change scores); n_{2R} = number of repeated EORTC QLQ-C30 scale and anchor change scores across all subjects; PF = physical functioning; RF = role functioning; SF = social functioning; EF = emotional functioning; QL = global quality of life; PA = pain; FA = fatigue; NV = nausea and/or vomiting; AP = appetite loss; DI= Diarrhoea; CO = Constipation; CTCAE, common terminology criteria for adverse events.

.

Table 3: Summary of Anchor-based MIDs for within and between-group change over time.

	Anchor-based MID for within-group change		Anchor-based MID for between-group difference in change	
Scale	**Improvement**	**Deterioration**	**Improvement**	**Deterioration**
PF	9.41	-4.98	6.04	-5.50
RF	18.48	-7.16	12.96	-9.95
SF	14.53	no MID	9.82	no MID
EF	8.45	no MID	5.96	no MID
QL	12.69	-7.94 to -4.72 (-5.76 ʷ)	8.96	-7.90 to -7.25 (-7.46 ʷ)
PA	8.68 to 11.2 (9.94ʷ)	no MID	5.73 to 8.21 (-6.97 ʷ)	no MID
FA	5.71 to 14.51 (11.58ʷ)	-7.62 to -4.76 (-5.47ʷ)	2.46 to 10.34 (7.71ʷ)	-10.77 to -6.59 (-7.63ʷ)
NV	4.33 to 6.54 (5.44 ʷ)	-5.07 to -3.73 (-4.40 ʷ)	3.64 to 5.03 (4.34ʷ)	-5.89 to -4.32 (-5.11 ʷ)
AP	14.74 to 15.71 (15.27 ʷ)	no MID	8.92 to 12.53 (10.88 ʷ)	no MID
DI	5.36	-8.95	5.48	-8.23
CO	11.16	-6.00	6.66	-7.59

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: PF = physical functioning; RF = role functioning; SF = social functioning; EF = emotional functioning; QL = global quality of life; PA = pain; FA = fatigue; NV = nausea and/or vomiting; AP = appetite loss; DI= diarrhoea; CO = Constipation.

Figure legends

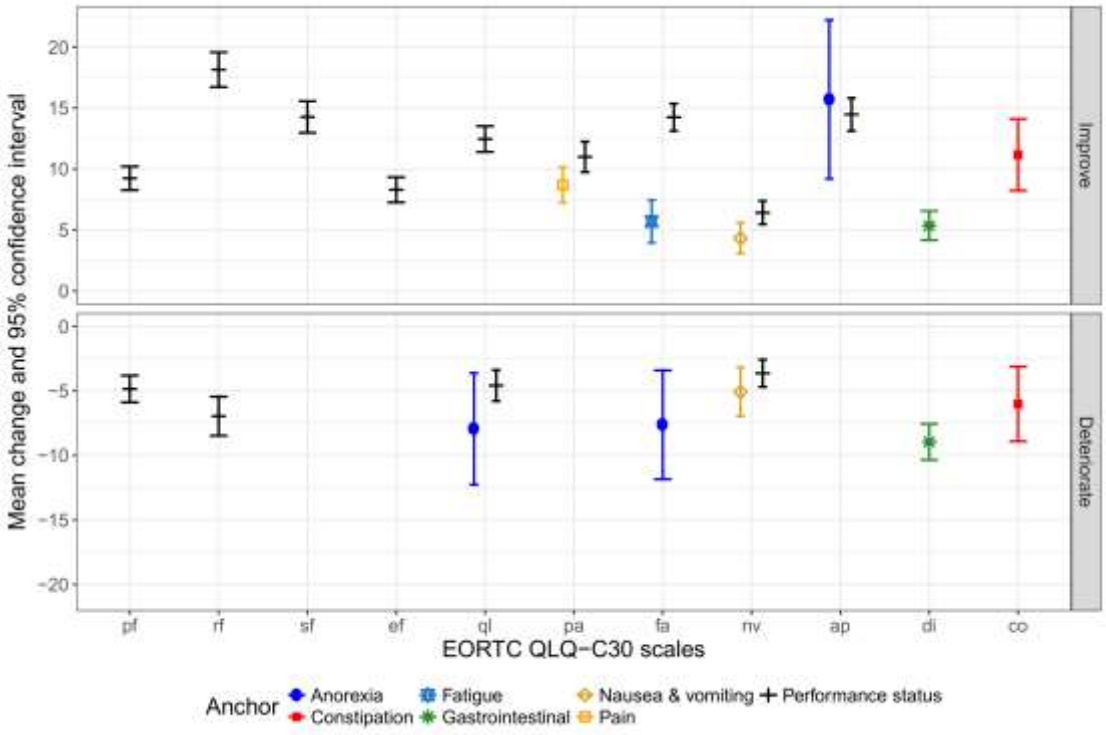

Figure 1: Mean change and 95% confidence interval for improvement and deterioration EORTC QLQ-C30 scales, across multiple anchors and averaged across different time periods.

Estimates are available only for scales with at least 1 suitable anchor 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: PF = physical functioning; RF = role functioning; SF = social functioning; EF = emotional functioning; QL = global quality of life; PA = pain; FA = fatigue; NV = nausea and/or vomiting; AP = appetite loss; DI= Diarrhoea; CO = Constipation; 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

Supplementary materials

Table S1: Number of patients (number of observations) by change scores of suitable anchors

Anchor change score	CTCAE Nausea & vomiting	CTCAE Fatigue	CTCAE Pain	CTCAE Anorexia	CTCAE Diarrhoea	CTCAE Constipation	CTCAE Gastrointestinal	Performance status
-4						2 (21)	3 (16)	
-3	22 (116)	3 (3)	10 (34)		8 (35)	4 (9)	40 (172)	2 (2)
-2	54 (411)	25 (148)	54 (164)	1(1)	46 (140)	20 (101)	112 (604)	39 (112)
-1	134 (1002)	133 (622)	224 (1246)	26(106)	143 (662)	113 (670)	278 (1750)	434 (2098)
0	792 (10760)	622 (7087)	755 (10069)	659(8539)	1238 (15201)	798 (11645)	733 (8463)	1092 (11169)
1	149 (620)	192 (793)	297 (1253)	39(129)	194 (812)	139 (564)	364 (1604)	424 (1598)
2	54 (161)	53 (121)	108 (277)	2(9)	84 (167)	25 (71)	161 (400)	42(90)
3	12 (23)	7 (9)	17 (45)		13 (18)	5 (11)	40 (78)	3(5)
4	1 (1)		1 (6)		1 (1)	1 (2)	5(7)	1(6)

Abbreviations: CTCAE, common terminology criteria for adverse events.

Table S2: Means (effect sizes) of HRQOL change score in three clinical change groups that are based on selected anchors per EORTC QLQ-C30 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	9.41 (0.44)	2.57 (0.12)	-4.98 (-0.23)	6.04	-5.50
RF	Performance status	18.48 (0.62)	3.8 (0.13)	-7.16 (-0.24)	12.96	-9.95
SF	Performance status	14.53 (0.53)	3.41 (0.12)	-2.44 (-0.09) [a]	9.82	-5.26 [a]
EF	Performance status	8.45 (0.36)	1.45 (0.06)	-2.2 (-0.09) [a]	5.96	-3.52 [a]
QL	Performance status	12.69 (0.58)	2.82 (0.13)	-4.72 (-0.21)	8.96	-7.25
	CTCAE Anorexia	0.64 (0.03) [a]	1.23 (0.06)	-7.94 (-0.41)	-0.08 [a]	-7.90
PA	Performance status	11.2 (0.46)	2.02 (0.08)	-3.49 (-0.14) [a]	8.21	-5.37 [a]
	CTCAE Pain	8.68 (0.40)	2.29 (0.10)	-3.61 (-0.17) [a]	5.73	-5.06 [a]
FA	Performance status	14.51 (0.59)	3.15 (0.13)	-4.76 (-0.20)	10.34	-6.59
	CTCAE Fatigue	5.71 (0.26)	3.39 (0.15)	-3.52 (-0.16) [a]	2.46	-5.06 [a]
	CTCAE Anorexia	2.1 (0.09) [a]	3.06 (0.14)	-7.62 (-0.34)	-1.92 [a]	-10.77
NV	Performance status	6.54 (0.39)	0.79 (0.05)	-3.73 (-0.22)	5.03	-4.32
	CTCAE Nausea & vomiting	4.33 (0.29)	0.34 (0.02)	-5.07 (-0.34)	3.64	-5.89
AP	Performance status	14.74 (0.55)	3.96 (0.15)	-0.9 (-0.03) [a]	8.92	-6.17 [a]
	CTCAE Anorexia	15.71 (0.74)	0.5 (0.02)	-17.99 (-0.85) [a]	12.53	-18.27 [a]
DI	CTCAE Diarrhoea	16.33 (0.80) [a]	0.04 (0.0)	-20.53 (-0.96) [a]	16.28 [a]	-18.99 [a]
	CTCAE Gastrointestinal	5.36 (0.25)	-0.16 (-0.01)	-8.95 (-0.42)	5.48	-8.23
CO	CTCAE Constipation	11.16 (0.41)	1.92 (0.07)	-6 (-0.22)	6.66	-7.59

[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:

PF = physical functioning; RF = role functioning; SF = social functioning; EF = emotional functioning; QL = global quality of life; PA = pain; FA = fatigue; NV = nausea and/or vomiting; AP = appetite loss; DI= diarrhoea; CO = Constipation; CTCAE, common terminology criteria for adverse events.

Table S3: Distribution-based estimates

	Distribution-based HRQOL scores at t1 (No. of patients = 1462 to 1479)			
Scale	**0.2 SD**	**0.3 SD**	**0.5 SD**	**1 SEM**
PF	4.34	6.51	10.84	6.51
RF	6.48	9.72	16.19	13.74
SF	5.80	8.71	14.51	10.46
EF	4.61	6.91	11.52	9.77
CF	5.36	8.04	13.40	11.05
QL	5.33	8.00	13.33	9.97
FA	4.34	6.51	10.84	9.20
PA	4.79	7.19	11.99	8.97
NV	3.85	5.78	9.63	11.72
AP	6.40	9.60	16.00	14.66
DY	5.37	8.05	13.41	11.06
DI	6.22	9.33	15.54	12.82
CO	4.17	6.26	10.43	11.04
SL	6.30	9.45	15.75	13.73

Abbreviations: t1 is the time point for the start of treatment; HRQOL= health-related quality of life; SD = standard deviation; SEM = standard error of measurement; 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

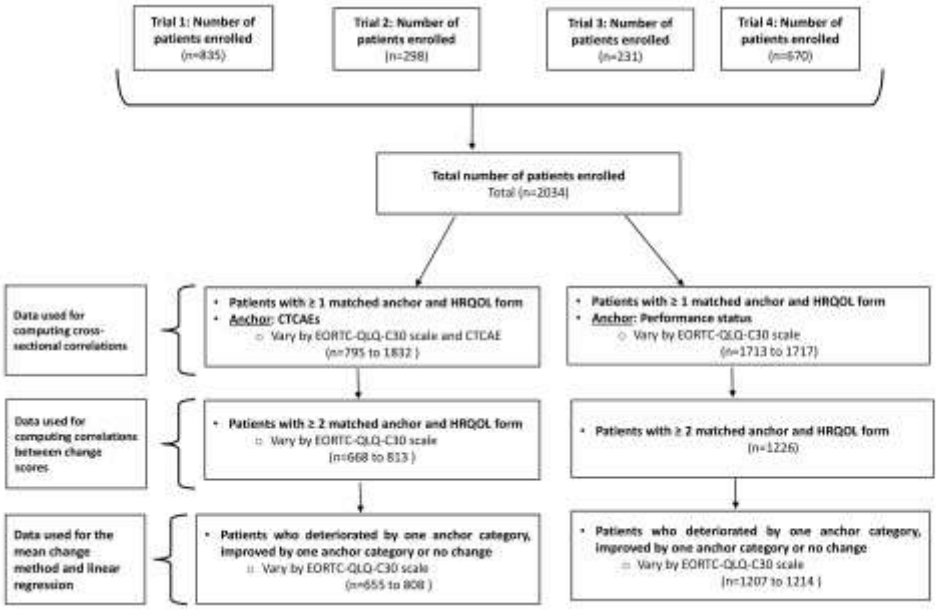

Figure S1: An overview of patient inclusion.

Abbreviations: CTCAE; common terminology criteria for adverse events, HRQOL; health-related quality of life