BMJ Open

Establishing Anchor-based Minimally Important Differences (MID) with the EORTC Quality of Life Measures: a meta-analysis protocol

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019117
Article Type:	Protocol
Date Submitted by the Author:	11-Aug-2017
Complete List of Authors:	Musoro, Zebedee Jammbe; European Organisation for Research and Treatment of Cancer, Quality of life Hamel, Jean-Francois ; University Hospital of Angers UNAM, Angers, Methodology and Biostatistics Department Ediebah, Divine Ewane; Brain Tumour Center Amsterdam of the VU University Medical Center Cocks, Kim; Adelphi Values King, Madeleine; University of Sydney, PsychoOncology Cooperative Research Group (PoCoG) Groenvold, Mogens; University of Copenhagen, Public Health and Bispebjerg Hospital Sprangers, Mirjam; University of Amsterdam Brandberg, Yvonne; Karolinska Institutet, Learning, Informatics, Management and Ethics Velikova, Galina; University of Leeds Maringwa, John; Quantitative Solutions BV Flechtner, Hans-Henning; University of Magdeburg, 11Clinic for Child and Adolescent Psychiatry and Psychotherapy Bottomley, Andrew; European Organisation for Research and Treatment of Cancer, Quality of Life Department Coens, Corneel; European Organisation for Research and Treatment of Cancer, Quality of Life Department
Primary Subject Heading:	Oncology
Secondary Subject Heading:	Research methods
Keywords:	EORTC-QLQ C30, Anchor-based methods, Minimal important difference (MID), Protocol, Cancer clinical trials

Establishing Anchor-based Minimally Important Differences (MID) with the EORTC Quality of Life Measures: a meta-analysis protocol

Jammbe Z Musoro[1], Jean-Francois Hamel[2], Divine E Ediebah[3], Kim Cocks[4], Madeleine T King[5], Mogens Groenvold[6], Mirjam AG Sprangers[7], Yvonne Brandberg[8], Galina Velikova[9], John Maringwa[10], Hans-Henning Flechtner[11], Andrew Bottomley[1], Corneel Coens[1] on Behalf of the EORTC Quality of Life Group and EORTC Quality of Life Department

[1]European Organisation for Research and Treatment of Cancer (EORTC), Brussels, Belgium
[2]Methodology and Biostatistics Department, University Hospital of Angers UNAM, Angers, France
[3]Brain Tumour Center Amsterdam of the VU University Medical Center, Amsterdam, The Netherlands
[4] Adelphi Values, Bollington, Cheshire, UK
[5]School of Psychology and Sydney Medical School, University of Sydney, Sydney, NSW, Australia
[6]Department of Public Health and Bispebjerg Hospital, University of Copenhagen, Copenhagen, Denmark
[7]Department of Medical Psychology, Academic Medical Center, University of Amsterdam, Amsterdam The Netherlands.
[8]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden.
[9]Leeds Institute of Cancer and Pathology, University of Leeds, St James's Hospital, Leeds, UK.
[10]Quantitative Solutions BV, Breda, The Netherlands
[11]Clinic for Child and Adolescent Psychiatry and Psychotherapy, University of Magdeburg, Magdeburg, Germany

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.be

ABSTRACT

Introduction: As patient assessment of health-related quality of life (HRQOL) in cancer clinical trials has increased over the years, so has the need to attach meaningful interpretations to differences in HRQOL scores between groups and changes within groups. Determining what represents a minimally important difference (MID) in HRQOL scores is useful to clinicians, patients and researchers, and can be used as a benchmark for assessing the success of a health care intervention (e.g. a new treatment). Our objective is to provide an evidence-based protocol to determine MIDs for the European Organisation for Research and Treatment for Cancer (EORTC) core HRQOL questionnaire, QLQ-C30.

Methods and analysis: Data will be derived from published Phase II and III EORTC clinical trials that used the QLQ-C30 instrument, covering melanoma, lung, colorectal, brain, head and neck, prostate, breast, testis, ovarian, pancreas, and esophageal cancer. We will use individual patient data to estimate MIDs for different cancer sites separately. Focus is on anchor-based methods. Anchors will be selected per disease site from available data. A disease-oriented and methodological panel will provide independent guidance on anchor selection. We aim to construct multiple clinical anchors per QLQ-C30 scale and also to compare several anchor-based methods as recommended in the literature. The effects of covariates e.g., gender, age, disease stage, trial etc, will also be investigated. We will examine how our estimated MIDs compare to previously published guidelines, hence further contributing to robust MID guidelines for the EORTC QLQ-C30.

Ethics and dissemination: Ethical approval is not required for this project as it is based on secondary data analysis. Our findings will be presented at scientific conferences, disseminated via peer-reviewed publications, and also compiled in a MID "blue book" which will be made available online on the EORTC Quality of Life Group (QLG) web site as a free guideline document.

Strengths and limitations of this study

- Several anchor-based methods will be applied and compared.
- Multiple clinical anchors will be constructed per QLQ-C30 scale.
- A library of MIDs will be established on the EORTC QLQ-C30 across various patient populations, according to cancer site.
- Will supplement previously published guidelines, hence establishing more robust MID guidelines.
- MIDs can only be estimated for QLQ-C30 scales for which a suitable anchor are available in our database.

1. INTRODUCTION

Patient assessment of health-related quality of life (HRQOL) in cancer clinical trials has increased over the years [1]. Consequently, there is greater need to attach meaningful interpretations to aggregated HRQOL scores, whether differences in HRQOL scores between groups or within-patient changes in HRQOL over time. Determining what represents a minimally important difference (MID) in HRQOL scores is useful to clinicians, patients and researchers, and can be used as a benchmark for assessing the success of a health care intervention (e.g., a new treatment) or the design of future clinical trials (e.g. determining sample sizes).

Minimally important difference (MID) has been 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, 3, 4]. It is important to note that there is a wide-range of terminology for "clinical meaningful change" in the literature. Notable distinctions in terminology have been made when referring to either group-level difference/change or individual-level change. A valuable and comprehensive critique on this topic and relevant references are given by King [5]. For simplicity, the most commonly used term "minimal important difference (MID)" will be applied in this manuscript to denote threshold of clinical relevance. When necessary, we will make distinctions between *(i)* group-level difference: cross-sectional differences in HRQOL scores between clinically-defined groups at a given time point, *(ii)* group-level change: change in HRQOL scores within a group over time and *(iii)* individual-level change: within-patient change in HRQOL scores over time. MIDs that are based on *(i)* and *(ii)* are useful for interpreting group-based trial results, while MIDs for individual-level change can be useful in trials as thresholds to define 'responders' i.e. patients who improved (or conversely patients who deteriorated) by a certain amount. In the literature MIDs are often applied to both group-level and individual-level results interchangeably. It is, however, unclear whether this is appropriate.

There are two broad methods for estimating MIDs; the anchor-based and distribution-based methods. The anchor-based approach has received much attention in the literature [6, 7, 8, 9, 10, 11, 12]. This approach expresses differences or change in HRQOL scores by linking particular HRQOL domains either to

known variables, which have clinical relevance, or to patient/physician-derived ratings of change in the particular domain [3, 6, 12, 13]. In this approach, it is crucial to evaluate the appropriateness of anchors. The usefulness of a MID estimate will depend on the anchor selected, how adjacent groups are defined within that anchor, and the strength of the relationship (conceptually and empirically) between the anchor and the target HRQOL domain [5]. On the other hand, distribution-based methods rely solely on the statistical distribution of HRQOL scores (do not consider patients'/clinicians' perspective) [14, 15], and have been recommended to be used as supportive evidence to anchor based estimates [13].

In this project, we focus on the anchor-based approach, particularly in a setting where both the anchors and HRQOL scores are collected longitudinally. The data will be derived from published Phases II and III EORTC trials which assessed HRQOL using the European Organisation for Research and Treatment for Cancer (EORTC) core HRQOL questionnaire, QLQ-C30. The aim of the project is to provide an evidence-based approach to determine MIDs for HRQOL scores of the EORTC QLQ-C30. Specifically, the appropriateness of particular clinical anchors in determining MIDs will be empirically evaluated. In addition, a library of MIDs will be established on the EORTC QLQ-C30 across various patient populations, according to cancer site (melanoma, lung, brain etc.) as well as stage of disease.

Osoba et al. [6] provided recommendations for small (5 to 10 points), moderate (10 to 20 points) and large changes (>20) for interpreting HRQOL scores of the EORTC QLQ-C30. This was based on individual data from patients with breast and small-cell lung cancers and included four of the EORTC QLQ-C30 scales (physical, emotional, social and global health). A global patient rating of change was the anchor. Similar findings were reported by King [10] based on comparing group differences, from multiple cancer sites, using published study results. More recent guidelines by Cocks *et al*. [16,17] using anchor-based methods highlighted that previous guidelines may be too simplistic in that they do not differentiate between the QLQ-C30 scales as well as between direction of change (improvement vs deterioration). These evidence-based guidelines further recommended using the lower bound as a minimal relevant threshold, arguing that large effect sizes were not always realistically achievable in all settings.

In contrast to Osoba et al. [6] this project will utilize multiple clinical anchors using clinical variables tailored to the specific cancer disease sites that are available in our database. The guidelines of King [10] and Cocks *et al*. [16,17] were based on meta-analyses of published studies, pooling across cancer sites, whereas we will use individual patient data to estimate MIDs for different cancer sites separately. Therefore this project presents an opportunity to add to previously published MID guidelines for the EORTC QLQ-C30 scales e.g. [6, 8, 9, 10, 16, 17] and compare these to estimated MIDs from our study.

2. METHODS AND ANALYSIS

2.1 Datasets and definition

Databases used for the analysis:

The data will be derived from published Phase II and III EORTC clinical trials that collected HRQOL data at baseline and follow-up using the EORTC QLQ-C30 and supplementary EORTC questionnaire modules. Cancer types include melanoma, lung, colorectal, brain, head and neck, prostate, breast, testis, ovarian, pancreas, and oesophageal cancer. Data from more recent EORTC studies, completed during this project, will also be included as well as non-EORTC data when available.

Data will be pooled within each cancer site separately using study time (defined as days since randomization) as the common temporal scale per patient. MIDs will be established per cancer site, with attention to robustness across the different subpopulations.

The EORTC QLQ-C30:

The focus of the analysis is on the EORTC QLQ-C30, a self-administered questionnaire designed for use in cancer clinical trials. The EORTC QLQ-C30 comprises 30 items, 24 of which are aggregated into nine multi-item scales, i.e. five functioning scales (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, and nausea/vomiting) and one global health status scale. The remaining six single-item (dyspnea, appetite loss, sleep disturbance, constipation, diarrhea and the financial impact) scales assess symptoms. The financial impact scale will be omitted from the analysis because suitable anchors are unlikely to exist.

Scoring of the EORTC QLQ-C30 scales will follow the standard procedures (see EORTC QLQ-C30 Scoring Manual [19]). For consistency in signs of the change scores across the various EORTC QLQ-C30 scales, the symptom scores will be reversed to follow the functioning scales interpretation; i.e. all scales will be scored such that 0 represents the worst possible score and 100 the best possible score. All versions of the EORTC QLQ-C30 will be used [19].

2.2 Anchor selection

We hope to identify at least one suitable clinical anchor for each EORTC QLQ-C30 scale from among potential clinical factors (e.g. laboratory measures, physiological measures, clinician ratings) that are available in the databases. Since the QLQ-C30 yields 15 scales measuring a wide range of symptoms and functioning, the suitability of an anchor must be considered relative to specific HRQOL domains. A suitable anchor for any particular QLQ-C30 scale should fulfil several criteria. Most notably the anchor should be relevant for the disease indication, should have clear medical interpretation and clinicians should be familiar with it. Also there should be a conceptual and empirical relationship between the anchor and its patient-reported counterpart [13].

Anchors will be selected per cancer site. This exercise will be guided by a panel of five to six clinical experts (per disease site) who are familiar with the specific trials, as well as with the structure of the EORTC QLQ-C30. These experts will primarily be recruited from the EORTC QOL group and from the panel of investigators involved in the included studies.

Clinical anchors will be pre-selected based on availability (i.e. the total that can be successfully matched to existing QLQ-C30 assessments), strength of correlation with the corresponding EORTC QLQ-C30 scale and finally clinical plausibility. A clinical anchor will be matched to an EORTC QLQ-C30 form if their respective assessment dates are within a predefined window. This time window will be determined on a per trial basis to ensure that the underlying true associations in the data are preserved. First, a candidate list of relevant clinical variables will be assembled based on the availability within each disease site. The acceptable compliance rate (i.e. availability of complete information) will depend on both relative and absolute available numbers. We aim for compliance rates ≥50% and an effective sample size of at least 200 patients with repeated observations.

Thereafter, we will evaluate how well the anchors correlate with the corresponding QLQ-C30 scale at various time-points of interest. Either a Spearman's rank, polyserial or polychoric correlation will be used, depending on the distribution of the pair of variables. The correlation between their change scores will also be checked. Revicki et al. [13] suggested a correlation of ≥ |0.30| as a measure of an acceptable association. Where achievable, however, anchors with much stronger correlations will be prioritized as suggested by recent simulation studies [20]. The list of retained anchors will be independently scrutinized for clinical relevance by the clinical experts, who will help to define clinical relevant cut-offs points in the anchor. Multiple anchors will be constructed for each QLQ-C30 scale where possible. If no suitable anchors can be identified for a given scale, no anchor-based MID will be estimated and reported for that scale.

Availability of anchors

When an anchor is only available for a subset of trials, only that subset will be used. A table will be constructed to summarise the availability of each anchor in the set of trials, and the QLQ-C30 scales to which each anchor is related (conceptually, clinically and empirically).

2.3 Preliminary analyses

a. Descriptive tabulation of the distribution of anchors and the EORTC QLQ-C30 scales will be made by trial, and pooled across trials. If insufficient variation is present or missing data is substantial in any anchor or scale, its inclusion in further analyses will be re-evaluated.

b. As a first step to establish the validity of an anchor, correlations between the anchors and their corresponding QLQ-C30 scales will be calculated using all matched anchor/ HRQOL scale pairs, regardless of time point. Scatterplots of the correlations will be inspected to gain greater understanding of bivariate distributions. The correlations will be calculated taking potential confounding factors into account (e.g. treatment, gender, age, disease stage, country, trial etc.,), to investigate the robustness of the associations in the overall population. Anchor/HRQOL scale pairs that fail to correlate at least 0.30 in at least one subgroup will be excluded from further consideration. Subgroups with associations < 0.30 may be excluded from further analysis, after discussion with the clinical experts. Similarly, we will investigate the correlation between change

scores of the anchor and HRQOL scale over time. Priority will be given to anchor/HRQOL scale pairs with correlations of at least 0.30 when MIDs for change scores are to be calculated.

c. The HRQOL score will be presented descriptively (e.g. mean, median, range, SD) at every time point of interest, within various subgroups (e.g. treatment, gender, age group, disease stage, country, trial etc.,), as well as in the overall population.

2.4 Handling of missing data

Missing HRQOL data:

We will cross-check compliance with the protocol schedule and verify the reasons for missing data. A cross-tabulation of the clinical anchors with HRQOL compliance will be made. We will evaluate the proportion of missing HRQOL forms per category of the anchor, and also check if subjects with missing HRQOL forms differ systematically from those with complete HRQOL data. If systematic differences are found, a panel of methodological experts will be consulted to suggest appropriate sensitivity analyses (e.g. imputation techniques) to check the robustness of the MID estimate.

Missing clinical anchor data:

Clinical anchors will be selected in such a way that missing data is minimized. For each EORCT QLQ-C30 scale the subset of anchors with the least amount of missing data will be prioritized. Similar to the handling of missing HRQOL data, we will also explore the anchor data to identify patterns as well as reasons for missingness.

2.5 Statistical Analysis

Cross-Sectional analysis of HRQOL scores

Cross-sectional differences of HRQOL scores between clinically relevant known-groups of patients will be investigated at selected time points. The groups will include the constructed anchor categories (e.g. grouping patients by their performance status or CTCAE score), extent of the disease or disease progression and other clinical factors which have distinct ordered categories with clear clinical relevance. We will compare between groups by examining cross-sectional differences for specific time-points of interest e.g. at baseline, at the end of treatment and at the end of follow-up. For each

HRQOL scale, the mean differences in HRQOL for each pair of adjacent group categories at the time-point of interest will be calculated. In addition, we will calculate effect sizes for these groups by dividing the difference of the mean HRQOL score from both groups by the standard deviation between patients in either group [10].

Anchor-based method for change scores:

The focus will be on examining both group-level and individual-level change over time. We will compute all possible pairwise time point differences in HRQOL scores and combine the data. This means that a subject can contribute multiple change scores that are calculated across different pairs of time-points, and the resulting dependency within the data will be accounted for whenever a regression model is applied. We will also consider specific time intervals, e.g. changes in HRQOL scales in the periods between start and end of treatment, and between end of treatment and end of follow-up. Subjects will be assigned to distinct subgroups reflecting various levels of change (e.g. no change, small positive changes, large positive changes, small negative changes, or large negative changes) based on the clinical anchor(s). These groups will be referred to as clinical change groups (CCG) and they are mutually exclusive. For each pair of time-points and for a given anchor, a patient can thus belong to only one CCG category.

Change in HRQOL score between two time-points is commonly expressed as a simple difference. We will explore other ways to express this change e.g. using relative differences that correct for the scores at baseline or another previous time-point. Table 1 presents a list of alternative summary scores for expressing change in HRQOL scores that will be explored. For each CCG, the summary scores will be presented descriptively (e.g. mean, median, range, SD). Differences in HRQOL summary scores between adjacent CCGs will be evaluated using primarily non-parametric techniques.

a. **Mean change method:** For a given HRQOL scale and its corresponding anchor, the MID for improvement is equal to the mean summary score of the "small positive change" CCG and the MID for deterioration is equal to "small negative change" CCG. The mean summary scores of the "small change" CCGs and that of the "no change" CCG will be compared. If the mean

summary score for "no change" CCG is similar to any of the two "small change" CCGs, the estimated MID is doubtful [18].

b. **Linear regression:** The estimate of the numerical change in HRQOL summary scores (see table 1) that is associated with the transition between adjacent CCG categories will be determined using a linear regression. Separate models will be fitted for improving and deteriorating scores based on the anchor. The outcome variable is the summary score, and the covariate is a binary anchor variable; coded as "no change" = 0, and "small positive change" = 1 for model on improvement, and "no change" = 0 and "small negative change" = 1 for model on deterioration. The resulting β's (i.e. slope parameters) correspond to the MIDs for improvement and deterioration respectively. This approach can be extended to correct for other covariates that could possibly affect the MID estimates [21].

c. **Receiver operating characteristic (ROC) curves:** For each summary score, the ROC analysis will be used to estimate thresholds that optimally discriminate between "minimally importantly changed" and "not minimally importantly changed" individuals, based on the anchor. Changes in different directions will be examined separately, i.e. no change group versus small positive CCG, and no change group versus small negative CCG. Different approaches will be used to calculate threshold values, e.g. by; (i) minimizing the gap between sensitivity and specificity, (ii) minimizing the sum of 1- sensitivity and 1- specificity and (iii) minimizing the sum of squares of 1- sensitivity and 1- specificity [22]. The various estimates will be compared and triangulation considered in order to establish robust guidelines. The assurance with which an estimated threshold can be used will depend on their corresponding sensitivity and specificity values. It is commonly not recommended to apply thresholds to individual patients when sensitivity and specificity are less than 75% [23].

d. **Empirical cumulative distribution function**: For each possible value of a given summary score (see Table 1) expressing change in the EORCT QLQ-C30 domains over time, the percentage of patients achieving at least that amount of change will be plotted separately for each CCG, and also separately for improving and deteriorating scores. The benefit of this approach is that the separation between CCGs may be visually compared across all values of

the summary scores, thus offering a range of possible individual-level thresholds for clinical relevance that can be considered simultaneously [24, 25].

The estimated MIDs across these methods will be compared, and the percentage of patients with improved or deteriorated HRQOL scores based on these MIDs will be reported. Recommendation for using a MID estimate for classifying individual patients will be based on whether the probability of misclassification is low [23] or whether the MID values exceed the measurement error level by comparing the MID to the minimum detectable change (MDC) [14, 5, 23, 26]. The MDC $= 1.96 * \sqrt{2} * SEM$ [26] represents the smallest change that can be considered to be above the measurement error. Usually if the MDC is > MID then the measure is insufficiently precise to monitor individual patients.

Distribution-based methods

We will examine the distribution-based approaches based on the standard deviation criteria, e.g. 0.2 SD, 0.3 SD, 0.5 SD, and the standard error of mean, SEM [14]. The SDs and SEM will be calculated on the summary scores (see Table 1) yielding MIDs corresponding to the rules above. Since this approach requires that the data are normally distributed, those summary scores that violate this assumption (based on standard testing techniques) will not be considered.

Effect sizes (ES) [15] will be calculated by dividing the summary scores in Table 1 by the pooled standard deviation of subjects at baseline (i.e., before treatment). This will be done for any two adjacent time points, e.g. depending on whether the level of compliance is acceptable. As a variation we will also calculate the effect sizes between adjacent time points by using the standard deviations of subjects at the previous time point [27].

3. VALIDATION AND SENSITIVITY ANALYSIS

3.1 Stability of the estimated MIDs

Internal validation:

The stability of the estimated MIDs across different patient groups (e.g. sex, age, disease stage, country, etc.) will be investigated by including the grouping factors (one at a time) and an interaction

term between the grouping factor and the anchor in a regression model. We will include as many socio-demographic and clinically relevant covariates as are available from the study database and that can be evaluated by the available sample size.

External validation:

For each cancer site in order to perform external validation, we will examine external (i.e. non-EORTC) studies having comparable data. This is subject to the availability of such data.

3.2 Handling the boundaries (floor and ceiling) effects

We will check for the proportion of patients with boundary (extreme) scores. For those patients where the later time-point was a boundary score, the change over time may be incorrectly estimated by simple subtraction. The change in clinical anchor for these patients at the boundaries will be used to estimate the magnitude of the problem. The proportion of patients with a change in clinical anchor that is not reflected in the HRQOL change due to the boundary constraints would be an indication of a limiting boundary problem. As a sensitivity check, we will investigate how much the MIDs are affected if we include or exclude these patients.

4. ETHICS AND DISSEMINATION

Ethical approval is not required for this project that is based on secondary data analysis. Our findings will be presented at scientific conferences, disseminated via peer-reviewed publications, and also compiled in a MID "blue book" which will be made available online on the EORTC Quality of Life Group (QLG) web site as a free guideline document.

5. CONCLUSION

In this project we will determine minimally important differences (MID) for HRQOL scores of the EORTC QLQ-C30, using empirical individual patient data. The main focus is on the anchor-based approach. We aim to construct multiple anchors per QLQ-C30 multiple-item or single-item scale and apply and compare results from several anchor-based methods as recommended in the literature [13, 18]. Hopefully the resulting MID estimates can triangulate to one value or a small range of values.

It is important to highlight that there are diverse opinions in the literature on whether or not it is plausible to use the same methods for interpreting individual-level change versus group-level differences/change. For instance, the mean change method and the ROC curve method have been labelled to be appropriate for comparing group-level and individual-level change respectively [21, 28]. On the other hand, both methods have been recommended to be useful for estimating MIDs that are useful for interpreting either group-level or individual-level change as long as the anchor is available at the individual level [23, 29, 30]. There is a need for a consensus on this matter. We will compare and contrast MID estimates from the different methods to provide empirical evidence, and assess whether it is possible to apply a simplified guideline to between group differences/change and individual-level change.

A strength of our research is its integral combination of both clinical and methodological expertise. The findings will ultimately improve the interpretation of the QLQ-C30 scale scores in clinical trials by providing empirical guidelines for relevant improvements and deteriorations.

Each year there are over 5000 newly registered downloads of the EORTC quality of life measures. The information from our research will be of added value to all its users (e.g. pharma and academic) since a frequent issue raised by regulators and trial sponsors is an understanding of MID.

Overall, this project will supplement previously published research by using individual patient data to estimate MIDs for different cancer sites separately, hence further providing evidence to robust and practical MID guidelines for the EORTC QLQ-C30.

6. References

1. Bottomley A, et al. Health related quality of life outcomes in cancer clinical trials. Eur J Cancer. 2005; 41: 1697-1709.
2. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. Controlled Clinical trials. 1989; 10: 407-415.
3. Schünemann HJ, Guyatt GH. Goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res. 2005; 40: 593-597.
4. Schünemann HJ et al. Measurement properties and interpretability of the Chronic Respiratory Disease Questionnaire (CRQ*)*. COPD. 2005; 2: 81-89.
5. 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.
6. Osoba D et al. Interpreting the significance of changes in health related quality-of-life scores. J Clin Oncol. 1998; 16: 139-144.
7. Lydick F, Epstein RS. Interpretation of quality of life changes. Qual Life Res. 1993; 2: 221-226.
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. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res. 1996; 5: 555-567.
11. Cella D, et al. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol. 2002; 55: 285-295.
12. 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.
13. 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
14. 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.
15. Cohen J. Statistical Power Analysis for the Behavioural Sciences (2nd Edition). Lawrence Erlbaum Associates, NJ, USA (1988).

16. 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.
17. 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.
18. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. COPD 2(1), 63–67 (2005).
19. 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
20. 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.
21. Angst F., Aeschlimanna A. and Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol. 2016; 82: 128-136.
22. Froud R, Abel G. Using ROC Curves to Choose Minimally Important Change Thresholds when Sensitivity and Specificity Are Valued Equally: The Forgotten Lesson of Pythagoras. Theoretical Considerations and an Example Application of Change in Health Status. Caylà JA, ed. PLoS ONE. 2014;9(12):e114468. doi:10.1371/journal.pone.0114468
23. de Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW, Hendriks EJ, Bouter LM, Terwee CB. Three ways to quantify uncertainty in individually applied "minimally important change" values. J Clin Epidemiol. 2010 Jan;63(1):37-45. doi: 10.1016/j.jclinepi.2009.03.011.
24. Farrar JT, Dworkin RH, Max MB. Use of the cumulative proportion of responders analysis graph to present pain data over a range of cutoff points: making clinical trial data more understandable. J. Pain Symptom Manage. 2006; 31(4):369–377.
25. McLeod LD, Coon CD, Martin SA, Fehnel SE, and Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharmacoecon Outcomes Res. 2011; 11(2): 163–169. doi:10.1586/erp.11.12.
26. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual. Life Outcomes. 2006; 4, 54.
27. Yost, Kathleen J. et al. Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) instrument. Value in Health , 2005; 8, 117 – 127

28. Wells GA, Beaton D, Shea B, Boers M, Simon L, Strand V, et al. Minimal clinically important differences: review of methods. J Rheumatol. 2001; 28:406-12.
29. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, the Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002; 77:371-83.
30. 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.

Acknowledgements: We thank the various EORTC disease group members and their clinical investigators, and all the patients who participated in the trials that we shall be analyzing.

Contributors: AB, CC, DEE, MK, MG, MAGS, GV and HJ contributed to the conception and design of the study. MG, MAGS, YB, GV, HJ revised the proposed methodology for clinical plausibility. JZM, HJ, CC, CK, DEE, JM and MK provided critical input on the proposed statistical analysis. JZM and CC drafted the protocol. All the authors read and corrected the drafts and approved the final version.

,

Funding: This study was funded by an unrestricted academic grant from the EORTC Quality of Life Group.

Competing interests: The authors declare no conflicts of interest.

Data Sharing Statement: Please visit http://www.eortc.org/data-sharing/ for details about the data sharing policy of the EORTC.

Table 1: Description of various summary scores for expressing changes in HRQOL scores over time

Name	Formula	Description
Absolute difference (AD)	$Q_2 - Q_1$	Simple difference in the HRQOL between two time-points. The mean of the AD values from all subjects corresponds to the Osoba's MID (Osoba et al., 1998).
Piecewise absolute difference (pAD)	$Q_{2c} - Q_{1c}$	Applying AD per subgroup c = 1 to 4, where c is based on baseline QOL values grouped as; *0-25, 26-50, 51-75* and *76-100.*
Relative difference (RD_1) – ordered	$\frac{(Q_2 - Q_1)}{Q_1} \times 100$	% change from previous value. Two variations based on direction Note: Q_1 and Q_2 are based on the Raw Scores (RS). RS is the mean of the component items for a particular scale that takes values from 1to 4 (instead of 0 to100).
Relative difference (RD_2) – sum	$\frac{(Q_2 - Q_1)}{(Q_2 + Q_1)}$	% change compared to sum. Convention: if $Q_2 = Q_1 = 0$ then RD2 = 0.
Relative difference (RD_3) – baseline	$\frac{(Q_2 - Q_1)}{Q_B}$	Q_B = baseline value Note: Q_1, Q_2 and Q_B are based on RS.
Fixed value	Q_2	MID is dependent on the observed value. This will result in a fixed threshold, not dependent on change in HRQOL score from previous value.
Slope	$\frac{(Q_2 - Q_1)}{(T_2 - T_1)}$	The absolute difference between two time points is weighted by the corresponding time period

Q_x = HRQOL outcome at timepoint T_x

BMJ Open

Establishing Anchor-based Minimally Important Differences (MID) with the EORTC Quality of Life Measures: a meta-analysis protocol

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019117.R1
Article Type:	Protocol
Date Submitted by the Author:	13-Oct-2017
Complete List of Authors:	Musoro, Zebedee Jammbe; European Organisation for Research and Treatment of Cancer, Quality of life Hamel, Jean-Francois ; University Hospital of Angers UNAM, Angers, Methodology and Biostatistics Department Ediebah, Divine Ewane; Brain Tumour Center Amsterdam of the VU University Medical Center Cocks, Kim; Adelphi Values King, Madeleine; University of Sydney, PsychoOncology Cooperative Research Group (PoCoG) Groenvold, Mogens; University of Copenhagen, Public Health and Bispebjerg Hospital Sprangers, Mirjam; University of Amsterdam Brandberg, Yvonne; Karolinska Institutet, Learning, Informatics, Management and Ethics Velikova, Galina; University of Leeds Maringwa, John; Quantitative Solutions BV Flechtner, Hans-Henning; University of Magdeburg, 11Clinic for Child and Adolescent Psychiatry and Psychotherapy Bottomley, Andrew; European Organisation for Research and Treatment of Cancer, Quality of Life Department Coens, Corneel; European Organisation for Research and Treatment of Cancer, Quality of Life Department
Primary Subject Heading:	Oncology
Secondary Subject Heading:	Research methods
Keywords:	EORTC-QLQ C30, Anchor-based methods, Minimal important difference (MID), Protocol, Cancer clinical trials, Responder threshold

Establishing Anchor-based Minimally Important Differences (MID) with the EORTC Quality of Life Measures: a meta-analysis protocol

Jammbe Z Musoro[1], Jean-Francois Hamel[2], Divine E Ediebah[3], Kim Cocks[4], Madeleine T King[5], Mogens Groenvold[6], Mirjam AG Sprangers[7], Yvonne Brandberg[8], Galina Velikova[9], John Maringwa[10], Hans-Henning Flechtner[11], Andrew Bottomley[1], Corneel Coens[1] on Behalf of the EORTC Quality of Life Group and EORTC Quality of Life Department

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

[2]Methodology and Biostatistics Department, University Hospital of Angers UNAM, Angers, France

[3]Brain Tumour Center Amsterdam of the VU University Medical Center, Amsterdam, The Netherlands

[4] Adelphi Values, Bollington, Cheshire, UK

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

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

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

[8]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden.

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

[10]Quantitative Solutions BV, Breda, The Netherlands

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

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.be

ABSTRACT

Introduction: As patient assessment of health-related quality of life (HRQOL) in cancer clinical trials has increased over the years, so has the need to attach meaningful interpretations to differences in HRQOL scores between groups and changes within groups. Determining what represents a minimally important difference (MID) in HRQOL scores is useful to clinicians, patients and researchers, and can be used as a benchmark for assessing the success of a health care intervention. Our objective is to provide an evidence-based protocol to determine MIDs for the European Organisation for Research and Treatment for Cancer (EORTC) core HRQOL questionnaire, QLQ-C30. We will mainly focus on MID estimation for group-level comparisons. Responder thresholds (RT) for individual-level change will also be estimated.

Methods and analysis: Data will be derived from published Phase II and III EORTC trials that used the QLQ-C30 instrument, covering several cancer sites. We will use individual patient data to estimate MIDs for different cancer sites separately. Focus is on anchor-based methods. Anchors will be selected per disease site from available data. A disease-oriented and methodological panel will provide independent guidance on anchor selection. We aim to construct multiple clinical anchors per QLQ-C30 scale and also to compare several anchor-based methods. The effects of covariates e.g., gender, age, disease stage etc., will also be investigated. We will examine how our estimated MIDs compare to previously published guidelines, hence further contributing to robust MID guidelines for the EORTC QLQ-C30.

Ethics and dissemination: All patient data originate from completed clinical trials with mandatory written informed consent, approved by local ethical committees. Our findings will be presented at scientific conferences, disseminated via peer-reviewed publications, and also compiled in a MID "blue book" which will be made available online on the EORTC Quality of Life Group (QLG) web site as a free guideline document.

Strengths and limitations of this study

- Several anchor-based methods will be applied and compared.
- Multiple clinical anchors will be constructed per QLQ-C30 scale.
- A library of MIDs will be established on the EORTC QLQ-C30 across various patient populations, according to cancer site.
- Will supplement previously published guidelines, hence establishing more robust MID guidelines.
- MIDs can only be estimated for QLQ-C30 scales for which a suitable anchor are available in our database.

1. INTRODUCTION

Patient assessment of health-related quality of life (HRQOL) in cancer clinical trials has increased over the years [1]. Consequently, there is greater need to attach meaningful interpretations to aggregated HRQOL scores, whether differences in HRQOL scores between groups or within-patient changes in HRQOL over time. Determining what represents a minimally important difference (MID) in HRQOL scores is useful to clinicians, patients and researchers, and can be used as a benchmark for assessing the success of a health care intervention (e.g., a new treatment) or the design of future clinical trials (e.g. determining sample sizes).

Minimally important difference (MID) has been 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' (p.594)[2]. It is important to note that there is a wide-range of terminology for "clinical meaningful change" in the literature. Notable distinctions in terminology have been made when referring to either group-level difference/change or individual-level change. A valuable and comprehensive critique on this topic and relevant references are given by King [3]. In this manuscript, we shall use the term "minimal important difference (MID)" to refer to group-level thresholds and "responder threshold (RT)" to refer to individual-level change. This project will mainly focus on MID estimation and will make distinctions between *(i)* group-level difference: cross-sectional differences in HRQOL scores between clinically-defined groups at a given time point, *(ii)* group-level change: change in HRQOL scores within a group over time and *(iii)* individual-level change: within-patient change in HRQOL scores over time. MIDs that are based on *(i)* and *(ii)* are useful for interpreting group-based trial results, while RTs for individual-level change can be useful in trials as thresholds to define 'responders' i.e. patients who improved (or conversely patients who deteriorated) by a certain amount.

There are two broad methods for estimating MIDs/RTs; the anchor-based and distribution-based methods. The anchor-based approach has received much attention in the literature [4, 5, 6, 7, 8, 9, 10]. This approach expresses differences or change in HRQOL scores by linking particular HRQOL domains either to known variables, which have clinical relevance, or to patient/physician-derived ratings of change in the particular domain [2, 4, 10, 11]. In this approach, it is crucial to evaluate the appropriateness

of anchors. The usefulness of the estimated threshold will depend on the anchor selected, how adjacent groups are defined within that anchor, and the strength of the relationship (conceptually and empirically) between the anchor and the target HRQOL domain [3]. It is worth noting that the estimated thresholds will depend on a range of factors, including the instrument, patient population, selected anchors, and the methods used. Hence a global rule for MIDs/RTs applicable to all situations is highly unlikely [11, 12]. It is recommended that thresholds be estimated by applying several anchor-based methods and using several types of anchors, and then to triangulate on a single value or small range of values [11]. Also, the literature does not clearly distinguish between the methods for estimating group-level vs individual-level thresholds. This study will offer a great opportunity to compare results across several anchor-based methods.

Distribution-based methods, on the other hand, rely solely on the statistical distribution of HRQOL scores (do not consider patients'/clinicians' perspective) [13, 14], and have been recommended to be used as supportive evidence to anchor based estimates [11].

In this project, we focus on the anchor-based approach, particularly in a setting where both the anchors and HRQOL scores are collected longitudinally. The data will be derived from published Phases II and III EORTC trials which assessed HRQOL using the European Organisation for Research and Treatment for Cancer (EORTC) core HRQOL questionnaire, QLQ-C30. The aim of the project is to provide an evidence-based approach to determine MIDs for HRQOL scores of the EORTC QLQ-C30. Specifically, the appropriateness of particular clinical anchors in determining MIDs will be empirically evaluated. In addition, a library of MIDs will be established on the EORTC QLQ-C30 across various patient populations, according to cancer site (melanoma, lung, brain etc.) as well as stage of disease.

Osoba et al. [4] provided recommendations for small (5 to 10 points), moderate (10 to 20 points) and large changes (>20) for interpreting HRQOL scores of the EORTC QLQ-C30. This was based on individual data from patients with breast and small-cell lung cancers and included four of the EORTC QLQ-C30 scales (physical, emotional, social and global health). A global patient rating of change was the anchor. Similar findings were reported by King [8] based on comparing group differences, from multiple cancer sites, using published study results. More recent guidelines by Cocks *et al*. [15,16] using

anchor-based methods highlighted that previous guidelines may be too simplistic in that they do not differentiate between the QLQ-C30 scales as well as between direction of change (improvement vs deterioration). These evidence-based guidelines further recommended using the lower bound as a minimal relevant threshold, arguing that large effect sizes were not always realistically achievable in all settings.

In contrast to Osoba et al. [4] this project will utilize multiple clinical anchors using clinical variables tailored to the specific cancer disease sites that are available in our database. The guidelines of King [8] and Cocks *et al.* [15,16] were based on meta-analyses of published studies, pooling across cancer sites, whereas we will use individual patient data to estimate MIDs for different cancer sites separately. Therefore this project presents an opportunity to add to previously published MID guidelines for the EORTC QLQ-C30 scales e.g. [4, 6, 7, 8, 15, 16] and compare these to estimated MIDs from our study.

2. METHODS AND ANALYSIS

2.1 Datasets and definition

Databases used for the analysis:

The data will mainly be extracted from published Phase II and III EORTC clinical trials. We will include only studies that collected HRQOL data at baseline and follow-up using the EORTC QLQ-C30 and supplementary EORTC questionnaire modules. Cancer types include melanoma, lung, colorectal, brain, head and neck, prostate, breast, testis, ovarian, pancreas, and oesophageal cancer. Data from more recent EORTC studies, completed during this project, will also be included as well as non-EORTC data when available.

Data will be pooled within each cancer site separately using study time (defined as days since randomization) as the common temporal scale per patient. MIDs will be established per cancer site, with attention to robustness across the different subpopulations.

The EORTC QLQ-C30:

The focus of the analysis is on the EORTC QLQ-C30, a self-administered questionnaire designed for use in cancer clinical trials. The EORTC QLQ-C30 comprises 30 items, 24 of which are aggregated into nine multi-item scales, i.e. five functioning scales (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, and nausea/vomiting) and one global health status scale. The remaining six single-item (dyspnea, appetite loss, sleep disturbance, constipation, diarrhea and the financial impact) scales assess symptoms. The financial impact scale will be omitted from the analysis because suitable anchors are unlikely to exist.

Scoring of the EORTC QLQ-C30 scales will follow the standard procedures (see EORTC QLQ-C30 Scoring Manual [17]). For consistency in signs of the change scores across the various EORTC QLQ-C30 scales, the symptom scores will be reversed to follow the functioning scales interpretation; i.e. all scales will be scored such that 0 represents the worst possible score and 100 the best possible score. All versions of the EORTC QLQ-C30 will be used [17].

2.2 Anchor selection

We hope to identify at least one suitable clinical anchor for each EORTC QLQ-C30 scale from among potential clinical factors (e.g. laboratory measures, physiological measures, clinician ratings) that are available in the databases. No patient ratings of change (e.g. subjective significance questionnaires) are available in the database. HRQOL scores will only be considered as anchors if valid MIDs are known. Since the QLQ-C30 yields 15 scales measuring a wide range of symptoms and functioning, the suitability of an anchor must be considered relative to specific HRQOL domains. A suitable anchor for any particular QLQ-C30 scale should fulfil several criteria. Most notably the anchor should be relevant for the disease indication, should have clear medical interpretation and clinicians should be familiar with it. Also there should be a conceptual and empirical relationship between the anchor and its patient-reported counterpart [11].

Anchors will be selected per cancer site. This exercise will be guided by a panel of five to six clinical experts (per disease site) who are familiar with the specific trials, as well as with the structure of the EORTC QLQ-C30. These experts will primarily be recruited from the EORTC QOL group and from the panel of investigators involved in the included studies.

Clinical anchors will be pre-selected based on availability (i.e. the total that can be successfully matched to existing QLQ-C30 assessments), strength of correlation with the corresponding EORTC QLQ-C30 scale and finally clinical plausibility. A clinical anchor will be matched to an EORTC QLQ-C30 form if their respective assessment dates are within a predefined window. This time window will be determined on a per trial basis to ensure that the underlying true associations in the data are preserved. First, a candidate list of relevant clinical variables will be assembled based on the availability within each disease site. The acceptable compliance rate (i.e. availability of complete information on both the anchor and the HRQOL scale) will depend on both relative and absolute available numbers. We aim for compliance rates ≥50% and an effective sample size of at least 200 patients with repeated observations after pooling data for each cancer site separately. Thereafter, we will evaluate how well the anchors correlate with the corresponding QLQ-C30 scale at various time-points of interest. Either a Spearman's rank, polyserial or polychoric correlation will be used, depending on the distribution of the pair of variables. The correlation between their change scores will also be checked. Revicki et al. [11] suggested a correlation of ≥ |0.30| as a measure of an acceptable association. Where achievable, however, anchors with much stronger correlations will be prioritized as suggested by recent simulation studies [18]. The list of retained anchors will be independently scrutinized for clinical relevance by the clinical experts, who will help to define clinical relevant cut-offs points in the anchor. Multiple anchors will be constructed for each QLQ-C30 scale where possible. If no suitable anchors can be identified for a given scale, no anchor-based MID will be estimated and reported for that scale.

Availability of anchors

When an anchor is only available for a subset of trials, only that subset will be used. A table will be constructed to summarise the availability of each anchor in the set of trials, and the QLQ-C30 scales to which each anchor is related (conceptually, clinically and empirically). For each anchor, we will present how important change will be defined (as prescribed by our panel of clinical experts), along with the estimated correlation with the corresponding QLQ-C30 scale.

2.3 Preliminary analyses

a. Descriptive tabulation of the distribution of anchors and the EORTC QLQ-C30 scales will be made by trial, and pooled across trials. If insufficient variation is present or missing data is substantial in any anchor or scale, its inclusion in further analyses will be re-evaluated.

b. As a first step to establish the validity of an anchor, correlations between the anchors and their corresponding QLQ-C30 scales will be calculated using all matched anchor/ HRQOL scale pairs, regardless of time point. Scatterplots of the correlations will be inspected to gain greater understanding of bivariate distributions. The correlations will be calculated taking potential confounding factors into account (e.g. treatment, gender, age, disease stage, country, trial etc.,), to investigate the robustness of the associations in the overall population. Anchor/HRQOL scale pairs that fail to correlate at least 0.30 in at least one subgroup will be excluded from further consideration. Subgroups with associations < 0.30 may be excluded from further analysis, after discussion with the clinical experts. Similarly, we will investigate the correlation between change scores of the anchor and HRQOL scale over time. Priority will be given to anchor/HRQOL scale pairs with correlations of at least 0.30 when MIDs for change scores are to be calculated.

c. The HRQOL score will be presented descriptively (e.g. mean, median, range, SD) at every time point of interest, within various subgroups (e.g. treatment, gender, age group, disease stage, country, trial etc.,), as well as in the overall population.

2.4 Handling of missing data

Missing HRQOL data:

We will cross-check compliance with the protocol schedule and verify the reasons for missing data. A cross-tabulation of the clinical anchors with HRQOL compliance will be made. We will evaluate the proportion of missing HRQOL forms per category of the anchor, and also check if subjects with missing HRQOL forms differ systematically from those with complete HRQOL data. If systematic differences are found, a panel of methodological experts will be consulted to suggest appropriate sensitivity analyses (e.g. imputation techniques) to check the robustness of the estimated thresholds.

Missing clinical anchor data:

Clinical anchors will be selected in such a way that missing data is minimized. For each EORTC QLQ-C30 scale the subset of anchors with the least amount of missing data will be prioritized. Similar to the handling of missing HRQOL data, we will also explore the anchor data to identify patterns as well as reasons for missingness.

2.5 Statistical Analysis

Cross-Sectional analysis of HRQOL scores

Cross-sectional differences (i.e. at the same time point) of HRQOL scores will be calculated between distinct subgroups of patients, where the grouping has been done on the clinical anchor. The categorization based on the clinical anchor are expected to yield groups that are distinct in health state, as this property is part of the clinical anchor building and evaluation process. For each HRQOL scale, the difference in mean HRQOL between each pair of adjacent group categories will be calculated at specific time-points of interest e.g. at baseline, at the end of treatment and at the end of follow-up. In addition, we will calculate effect sizes for these groups by dividing the difference of the mean HRQOL score from both groups by the standard deviation between patients in either group [8].

Anchor-based method for change scores:

The focus will be on examining both group-level and individual-level change over time. We will compute all possible pairwise time point differences in HRQOL scores and combine the data. This means that a subject can contribute multiple change scores that are calculated across different pairs of time-points, and the resulting dependency within the data will be accounted for whenever a regression model is applied. We will consider specific time intervals, namely changes in HRQOL scales in the periods between start and end of treatment, and between end of treatment and end of follow-up as these are often well defined across several studies. Furthermore, depending on the study design and setting, we will consider additional shorter time intervals prior to the end of treatment where feasible. Subjects will be assigned to distinct subgroups reflecting various levels of change (e.g. no change, small positive changes, large positive changes, small negative changes, or large negative

changes) based on the clinical anchor(s). These groups will be referred to as clinical change groups (CCG) and they are mutually exclusive. For each pair of time-points and for a given anchor, a patient can thus belong to only one CCG category.

Change in HRQOL score between two time-points is commonly expressed as a simple difference. We will explore other ways to express this change e.g. using relative differences that correct for the scores at baseline or another previous time-point. Table 1 presents a list of alternative summary scores for expressing change in HRQOL scores that will be explored. For each CCG, the summary scores will be presented descriptively (e.g. mean, median, range, SD). Differences in HRQOL summary scores between adjacent CCGs will be evaluated using primarily non-parametric techniques.

a. **Mean change method:** For a given HRQOL scale and its corresponding anchor, the MID for improvement is equal to the mean summary score of the "small positive change" CCG and the MID for deterioration is equal to "small negative change" CCG. The mean summary scores of the "small change" CCGs and that of the "no change" CCG will be compared. If the mean summary score for "no change" CCG is similar to any of the two "small change" CCGs, the estimated MID is doubtful [19].

b. **Linear regression:** The estimate of the numerical change in HRQOL summary scores (see table 1) that is associated with the transition between adjacent CCG categories will be determined using a linear regression. Separate models will be fitted for improving and deteriorating scores based on the anchor. The outcome variable is the summary score, and the covariate is a binary anchor variable; coded as "no change" = 0, and "small positive change" = 1 for model on improvement, and "no change" = 0 and "small negative change" = 1 for model on deterioration. The resulting β's (i.e. slope parameters) correspond to the MIDs for improvement and deterioration respectively. This approach can be extended to correct for other covariates that could possibly affect the MID estimates [20].

c. **Receiver operating characteristic (ROC) curves:** For each summary score, the ROC analysis will be used to estimate RTs based on an anchor. Changes in different directions will be examined separately. For example, for defining improvement, we will create an "at least minimally important change" group using all CCGs for improvements, i.e. small positive and

large positive CCGs, and a “no minimally important change" group using no change CCG and any level of worsening (i.e. small negative and large negative CCGs). Different approaches will be used to calculate threshold values, e.g. by; (i) minimizing the gap between sensitivity and specificity, (ii) minimizing the sum of 1- sensitivity and 1- specificity and (iii) minimizing the sum of squares of 1- sensitivity and 1- specificity [21]. The various estimates will be compared and triangulation considered in order to establish robust guidelines. The assurance with which an estimated threshold can be used will depend on their corresponding sensitivity and specificity values. It is commonly not recommended to apply thresholds to individual patients when sensitivity and specificity are less than 75% [22].

d. **Empirical cumulative distribution function (ECDF)**: For each possible value of a given summary score (see Table 1) expressing change in the EORTC QLQ-C30 domains over time, the percentage of patients achieving at least that amount of change will be plotted separately for each CCG, and also separately for improving and deteriorating scores. The benefit of this approach is that the separation between CCGs may be visually compared across all values of the summary scores, thus offering a range of possible RTs for clinical relevance that can be considered simultaneously [23, 24].

The estimated thresholds across these methods will be compared, and the percentage of patients with improved or deteriorated HRQOL scores will be reported. Recommendation for using estimated RTs for classifying individual patients will be based on whether the probability of misclassification is low [22] or whether the RT values exceed the measurement error level by comparing the thresholds to the minimum detectable change (MDC) [13, 3, 22, 25]. The MDC $= 1.96 * \sqrt{2} * SEM$ [25] represents the smallest change that can be considered to be above the measurement error. Usually if the MDC is > the RT then the measure is insufficiently precise to monitor individual patients. Furthermore, when setting RTs, especially on domains that are computed based on a single item, we will check that the RTs align with the underlying change levels of the scale scores [26].

Distribution-based methods

We will examine the distribution-based approaches based on the standard deviation criteria, e.g. 0.2 SD, 0.3 SD, 0.5 SD, and the standard error of mean, SEM [13]. The SDs and SEM will be calculated on the summary scores (see Table 1) yielding MIDs corresponding to the rules above. Since this approach requires that the data are normally distributed, those summary scores that violate this assumption (based on standard testing techniques) will not be considered.

Effect sizes (ES) [14] will be calculated by dividing the summary scores in Table 1 by the pooled standard deviation of subjects at baseline (i.e., before treatment). This will be done for any two adjacent time points, e.g. depending on whether the level of compliance is acceptable. As a variation we will also calculate the effect sizes between adjacent time points by using the standard deviations of subjects at the previous time point [27].

3. VALIDATION AND SENSITIVITY ANALYSIS

3.1 Stability of the estimated MIDs

Internal validation:

Characteristics such as age, gender, disease stage, country, etc. typically influence the absolute score outcomes of many HRQOL scales [28]. The stability of the estimated MIDs will therefore be investigated by including these factors (one at a time) and an interaction term with the anchor in a regression model. We will include as many socio-demographic and clinically relevant covariates as are available from the study database and that can be evaluated by the available sample size.

External validation:

For each cancer site in order to perform external validation, we will examine external (i.e. non-EORTC) studies having comparable data. This is subject to the availability of such data.

3.2 Handling the boundaries (floor and ceiling) effects

We will check for the proportion of patients with boundary (extreme) scores. For those patients where the later time-point was a boundary score, the change over time may be incorrectly estimated by simple subtraction. The change in clinical anchor for these patients at the boundaries will be used to estimate the magnitude of the problem. The proportion of patients with a change in clinical anchor

that is not reflected in the HRQOL change due to the boundary constraints would be an indication of a limiting boundary problem. As a sensitivity check, we will investigate how much the MIDs are affected if we include or exclude these patients.

4. ETHICS AND DISSEMINATION

All patient data originate from completed clinical trials with mandatory written informed consent, approved by local ethical committees. Our findings will be presented at scientific conferences, disseminated via peer-reviewed publications, and also compiled in a MID "blue book" which will be made available online on the EORTC Quality of Life Group (QLG) web site as a free guideline document.

5. CONCLUSION

In this project we will determine minimally important differences (MID) for HRQOL scores of the EORTC QLQ-C30, using empirical individual patient data. The main focus is on the anchor-based approach. We aim to construct multiple anchors per QLQ-C30 multiple-item or single-item scale and apply and compare results from several anchor-based methods as recommended in the literature [11, 19]. Figure 1 presents a flow diagram summarizing the key data component, the clinical anchor construction procedure and the main statistical methods which will be applied in this project. Hopefully the resulting MID estimates can triangulate to one value or a small range of values.

It is important to highlight that there are diverse opinions in the literature on whether or not it is plausible to use the same methods for interpreting individual-level change versus group-level differences/change. For instance, the mean change method and the ROC curve method have been labelled to be appropriate for comparing group-level and individual-level change respectively [20, 29]. On the other hand, both methods have been recommended to be useful for estimating MIDs that are useful for interpreting either group-level or individual-level change as long as the anchor is available at the individual level [22, 30, 31]. There is a need for a consensus on this matter. We will compare and contrast MID estimates from the different methods to provide empirical evidence, and assess whether

it is possible to apply a simplified guideline to between group differences/change and individual-level change.

A strength of our research is its integral combination of both clinical and methodological expertise. The findings will ultimately improve the interpretation of the QLQ-C30 scale scores in clinical trials by providing empirical guidelines for relevant improvements and deteriorations.

Each year there are over 5000 newly registered downloads of the EORTC quality of life measures. The information from our research will be of added value to all its users (e.g. pharma and academic) since a frequent issue raised by regulators and trial sponsors is an understanding of MID.

Overall, this project will supplement previously published research by using individual patient data to estimate MIDs for different cancer sites separately, hence further providing evidence to robust and practical MID guidelines for the EORTC QLQ-C30.

6. 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. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcome Res. 2011 Apr;11(2):171-84.
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. Lydick F, Epstein RS. Interpretation of quality of life changes. Qual Life Res. 1993; 2: 221-226.
6. 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.
7. 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.
8. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res. 1996; 5: 555-567.
9. Cella D, et al. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol. 2002; 55: 285-295.
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. 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
12. Nordin A, Taft C, Lundgren-Nilsson A, Dencker A. Minimal important differences for fatigue patient reported outcome measures-a systematic review. BMC Med Res Methodol, 16 (2016), p. 62.

13. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J. Clin. Epidemiol. 1999; 52(9), 861–873.
14. Cohen J. Statistical Power Analysis for the Behavioural Sciences (2nd Edition). Lawrence Erlbaum Associates, NJ, USA (1988).
15. 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.
16. 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.
17. 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
18. 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.
19. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. COPD 2(1), 63–67 (2005).
20. Angst F., Aeschlimanna A. and Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol. 2016; 82: 128-136.
21. Froud R, Abel G. Using ROC Curves to Choose Minimally Important Change Thresholds when Sensitivity and Specificity Are Valued Equally: The Forgotten Lesson of Pythagoras. Theoretical Considerations and an Example Application of Change in Health Status. Caylà JA, ed. PLoS ONE. 2014;9(12):e114468. doi:10.1371/journal.pone.0114468
22. de Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW, Hendriks EJ, Bouter LM, Terwee CB. Three ways to quantify uncertainty in individually applied "minimally important change" values. J Clin Epidemiol. 2010 Jan;63(1):37-45. doi: 10.1016/j.jclinepi.2009.03.011.
23. Farrar JT, Dworkin RH, Max MB. Use of the cumulative proportion of responders analysis graph to present pain data over a range of cutoff points: making clinical trial data more understandable. J. Pain Symptom Manage. 2006; 31(4):369–377.
24. McLeod LD, Coon CD, Martin SA, Fehnel SE, and Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharmacoecon Outcomes Res. 2011; 11(2): 163–169. doi:10.1586/erp.11.12.

25. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual. Life Outcomes. 2006; 4, 54.
26. Wyrwich K. What's the Score? Moving from Items to Scores -Methods, Considerations, and Case Examples. Eighth Annual Patient-Reported Outcome Consortium Workshop. Bethesda, MD, April 2017; 13-28 (https://c-path.org/wp-content/uploads/2017/05/2017_session5_scoringfinal.pdf).
27. Yost, Kathleen J. et al. Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) instrument. Value in Health , 2005; 8, 117 – 127
28. Wan GJ, Counte MA, Cella DF, Hernande L, Deasy S, Shiomoto G: An analysis of the impact of demographic, clinical and social factors on health-related quality of life. Value in Health 1999, 2: 308–318. 10.1046/j.1524-4733.1999.24006.x
29. Wells GA, Beaton D, Shea B, Boers M, Simon L, Strand V, et al. Minimal clinically important differences: review of methods. J Rheumatol. 2001; 28:406-12.
30. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, the Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002; 77:371-83.
31. 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.

Acknowledgements: We thank the various EORTC disease group members and their clinical investigators, and all the patients who participated in the trials that we shall be analyzing.

Contributors: AB, CC, DEE, MTK, MG, MAGS, GV and HF contributed to the conception and design of the study. MG, MAGS, YB, GV, HF revised the proposed methodology for clinical plausibility. JZM, JH, CC, CK, DEE, JM and MTK provided critical input on the proposed statistical analysis. JZM and CC drafted the protocol. All the authors read and corrected the drafts and approved the final version.

Funding: This study was funded by an unrestricted academic grant from the EORTC Quality of Life Group.

Competing interests: The authors declare no conflicts of interest.

Data Sharing Statement: Please visit http://www.eortc.org/data-sharing/ for details about the data sharing policy of the EORTC.

Table 1: Description of various summary scores for expressing changes in HRQOL scores over time

Name	Formula	Description
Absolute difference (AD)	$Q_2 - Q_1$	Simple difference in the HRQOL between two time-points. The mean of the AD values from all subjects corresponds to the Osoba's MID (Osoba et al., 1998).
Piecewise absolute difference (pAD)	$Q_{2c} - Q_{1c}$	Applying AD per subgroup c = 1 to 4, where c is based on baseline QOL values grouped as; *0-25, 26-50, 51-75* and *76-100.*
Relative difference (RD_1) – ordered	$\frac{(Q_2 - Q_1)}{Q_1} \times 100$	% change from previous value. Two variations based on direction Note: Q_1 and Q_2 are based on the Raw Scores (RS). RS is the mean of the component items for a particular scale that takes values from 1to 4 (instead of 0 to100).
Relative difference (RD_2) – sum	$\frac{(Q_2 - Q_1)}{(Q_2 + Q_1)}$	% change compared to sum. Convention: if $Q_2 = Q_1 = 0$ then RD2 = 0.
Relative difference (RD_3) – baseline	$\frac{(Q_2 - Q_1)}{Q_B}$	Q_B = baseline value Note: Q_1, Q_2 and Q_B are based on RS.
Fixed value	Q_2	MID is dependent on the observed value. This will result in a fixed threshold, not dependent on change in HRQOL score from previous value.
Slope	$\frac{(Q_2 - Q_1)}{(T_2 - T_1)}$	The absolute difference between two time points is weighted by the corresponding time period

Q_x = HRQOL outcome at time point T_x

Figure 1: A flow diagram summarizing the data (e.g. the cancer sites, QLQ-C30 scales and types of clinical variables that will be used for anchor construction), the clinical anchor construction step and the main statistical methods which will be applied in this project.

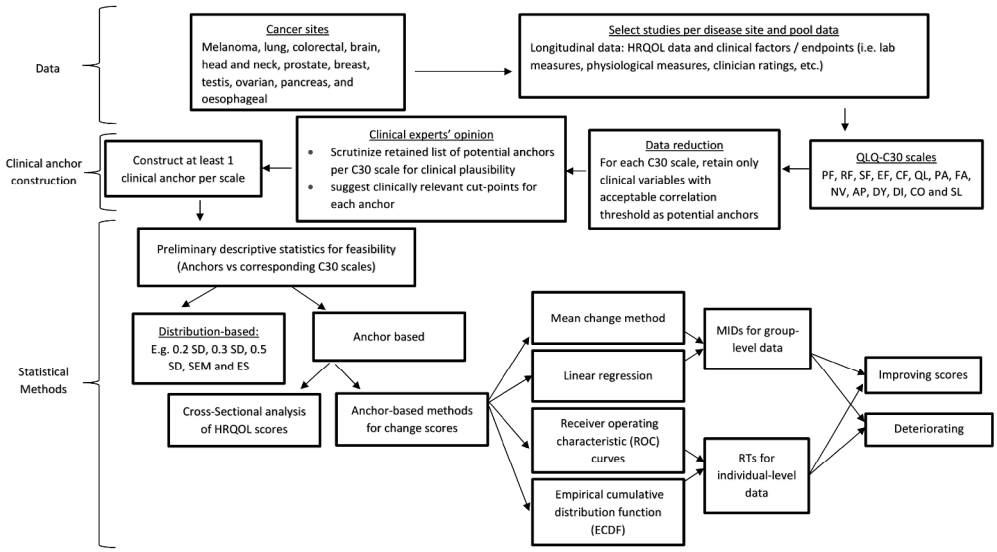

Figure 1: A flow diagram summarizing the data (e.g. the cancer sites, QLQ-C30 scales and types of clinical variables that will be used for anchor construction), the clinical anchor construction step and the main statistical methods which will be applied in this project.

275x153mm (300 x 300 DPI)

BMJ Open

Establishing Anchor-based Minimally Important Differences (MID) with the EORTC Quality of Life Measures: a meta-analysis protocol

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-019117.R2
Article Type:	Protocol
Date Submitted by the Author:	15-Nov-2017
Complete List of Authors:	Musoro, Zebedee Jammbe; European Organisation for Research and Treatment of Cancer, Quality of life Hamel, Jean-Francois ; University Hospital of Angers UNAM, Angers, Methodology and Biostatistics Department Ediebah, Divine Ewane; Brain Tumour Center Amsterdam of the VU University Medical Center Cocks, Kim; Adelphi Values King, Madeleine; University of Sydney, PsychoOncology Cooperative Research Group (PoCoG) Groenvold, Mogens; University of Copenhagen, Public Health and Bispebjerg Hospital Sprangers, Mirjam; University of Amsterdam Brandberg, Yvonne; Karolinska Institutet, Department of Oncology-Pathology Velikova, Galina; University of Leeds Maringwa, John; Quantitative Solutions BV Flechtner, Hans-Henning; University of Magdeburg, 11Clinic for Child and Adolescent Psychiatry and Psychotherapy Bottomley, Andrew; European Organisation for Research and Treatment of Cancer, Quality of Life Department Coens, Corneel; European Organisation for Research and Treatment of Cancer, Quality of Life Department
Primary Subject Heading:	Oncology
Secondary Subject Heading:	Research methods
Keywords:	EORTC-QLQ C30, Anchor-based methods, Minimal important difference (MID), Protocol, Cancer clinical trials, Responder threshold

Establishing Anchor-based Minimally Important Differences (MID) with the EORTC Quality of Life Measures: a meta-analysis protocol

Jammbe Z Musoro[1], Jean-Francois Hamel[2], Divine E Ediebah[3], Kim Cocks[4], Madeleine T King[5], Mogens Groenvold[6], Mirjam AG Sprangers[7], Yvonne Brandberg[8], Galina Velikova[9], John Maringwa[10], Hans-Henning Flechtner[11], Andrew Bottomley[1], Corneel Coens[1] on Behalf of the EORTC Quality of Life Group

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

[2]Methodology and Biostatistics Department, University Hospital of Angers UNAM, Angers, France

[3]Brain Tumour Center Amsterdam of the VU University Medical Center, Amsterdam, The Netherlands

[4] Adelphi Values, Bollington, Cheshire, UK

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

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

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

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

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

[10]Quantitative Solutions BV, Breda, The Netherlands

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

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.be

ABSTRACT

Introduction: As patient assessment of health-related quality of life (HRQOL) in cancer clinical trials has increased over the years, so has the need to attach meaningful interpretations to differences in HRQOL scores between groups and changes within groups. Determining what represents a minimally important difference (MID) in HRQOL scores is useful to clinicians, patients and researchers, and can be used as a benchmark for assessing the success of a health care intervention. Our objective is to provide an evidence-based protocol to determine MIDs for the European Organisation for Research and Treatment for Cancer (EORTC) core HRQOL questionnaire, QLQ-C30. We will mainly focus on MID estimation for group-level comparisons. Responder thresholds (RT) for individual-level change will also be estimated.

Methods and analysis: Data will be derived from published Phase II and III EORTC trials that used the QLQ-C30 instrument, covering several cancer sites. We will use individual patient data to estimate MIDs for different cancer sites separately. Focus is on anchor-based methods. Anchors will be selected per disease site from available data. A disease-oriented and methodological panel will provide independent guidance on anchor selection. We aim to construct multiple clinical anchors per QLQ-C30 scale and also to compare several anchor-based methods. The effects of covariates e.g., gender, age, disease stage etc., will also be investigated. We will examine how our estimated MIDs compare to previously published guidelines, hence further contributing to robust MID guidelines for the EORTC QLQ-C30.

Ethics and dissemination: All patient data originate from completed clinical trials with mandatory written informed consent, approved by local ethical committees. Our findings will be presented at scientific conferences, disseminated via peer-reviewed publications, and also compiled in a MID "blue book" which will be made available online on the EORTC Quality of Life Group (QLG) web site as a free guideline document.

Strengths and limitations of this study

- Several anchor-based methods will be applied and compared.
- Multiple clinical anchors will be constructed per QLQ-C30 scale.
- A library of MIDs will be established on the EORTC QLQ-C30 across various patient populations, according to cancer site.
- Will supplement previously published MID guidelines, hence establishing more robust MID guidelines.
- Anchor-based MIDs can only be estimated for QLQ-C30 scales for which a suitable anchor are available in our database.

1. INTRODUCTION

Patient assessment of health-related quality of life (HRQOL) in cancer clinical trials has increased over the years [1]. Consequently, there is greater need to attach meaningful interpretations to aggregated HRQOL scores, whether differences in HRQOL scores between groups or within-patient changes in HRQOL over time. Determining what represents a minimally important difference (MID) in HRQOL scores is useful to clinicians, patients and researchers, and can be used as a benchmark for assessing the success of a health care intervention (e.g., a new treatment) or the design of future clinical trials (e.g. determining sample sizes).

Minimally important difference (MID) has been 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' (p.594)[2]. It is important to note that there is a wide-range of terminology for "clinical meaningful change" in the literature. Notable distinctions in terminology have been made when referring to either group-level difference/change or individual-level change. A valuable and comprehensive critique on this topic and relevant references are given by King [3]. In this manuscript, we shall use the term "minimal important difference (MID)" to refer to group-level thresholds and "responder threshold (RT)" to refer to individual-level change. This project will mainly focus on MID estimation and will make distinctions between *(i)* group-level difference: cross-sectional differences in HRQOL scores between clinically-defined groups at a given time point, *(ii)* group-level change: change in HRQOL scores within a group over time and *(iii)* individual-level change: within-patient change in HRQOL scores over time. MIDs that are based on *(i)* and *(ii)* are useful for interpreting group-based trial results, while RTs for individual-level change can be useful in trials as thresholds to define 'responders' i.e. patients who improved (or conversely patients who deteriorated) by a certain amount.

There are two broad methods for estimating MIDs/RTs; the anchor-based and distribution-based methods. The anchor-based approach has received much attention in the literature [4, 5, 6, 7, 8, 9, 10]. This approach expresses differences or change in HRQOL scores by linking particular HRQOL domains either to known variables, which have clinical relevance, or to patient/physician-derived ratings of change in the particular domain [2, 4, 10, 11]. In this approach, it is crucial to evaluate the appropriateness

of anchors. The usefulness of the estimated threshold will depend on the anchor selected, how adjacent groups are defined within that anchor, and the strength of the relationship (conceptually and empirically) between the anchor and the target HRQOL domain [3]. It is worth noting that the estimated thresholds will depend on a range of factors, including the instrument, patient population, selected anchors, and the methods used. Hence a global rule for MIDs/RTs applicable to all situations is highly unlikely [11, 12]. It is recommended that thresholds be estimated by applying several anchor-based methods and using several types of anchors, and then to triangulate on a single value or small range of values [11]. Also, the literature does not clearly distinguish between the methods for estimating group-level vs individual-level thresholds. This study will offer a great opportunity to compare results across several anchor-based methods.

Distribution-based methods, on the other hand, rely solely on the statistical distribution of HRQOL scores (do not consider patients'/clinicians' perspective) [13, 14], and have been recommended to be used as supportive evidence to anchor based estimates [11].

In this project, we focus on the anchor-based approach, particularly in a setting where both the anchors and HRQOL scores are collected longitudinally. The data will be derived from published Phases II and III EORTC trials which assessed HRQOL using the European Organisation for Research and Treatment for Cancer (EORTC) core HRQOL questionnaire, QLQ-C30. The aim of the project is to provide an evidence-based approach to determine MIDs for HRQOL scores of the EORTC QLQ-C30. Specifically, the appropriateness of particular clinical anchors in determining MIDs will be empirically evaluated. In addition, a library of MIDs will be established on the EORTC QLQ-C30 across various patient populations, according to cancer site (melanoma, lung, brain etc.) as well as stage of disease.

Osoba et al. [4] provided recommendations for small (5 to 10 points), moderate (10 to 20 points) and large changes (>20) for interpreting HRQOL scores of the EORTC QLQ-C30. This was based on individual data from patients with breast and small-cell lung cancers and included four of the EORTC QLQ-C30 scales (physical, emotional, social and global health). A global patient rating of change was the anchor. Similar findings were reported by King [8] based on comparing group differences, from multiple cancer sites, using published study results. More recent guidelines by Cocks *et al*. [15,16] using

anchor-based methods highlighted that previous guidelines may be too simplistic in that they do not differentiate between the QLQ-C30 scales as well as between direction of change (improvement vs deterioration). These evidence-based guidelines further recommended using the lower bound as a minimal relevant threshold, arguing that large effect sizes were not always realistically achievable in all settings.

In contrast to Osoba et al. [4] this project will utilize multiple clinical anchors using clinical variables tailored to the specific cancer disease sites that are available in our database. The guidelines of King [8] and Cocks *et al.* [15,16] were based on meta-analyses of published studies, pooling across cancer sites, whereas we will use individual patient data to estimate MIDs for different cancer sites separately. Therefore this project presents an opportunity to add to previously published MID guidelines for the EORTC QLQ-C30 scales e.g. [4, 6, 7, 8, 15, 16] and compare these to estimated MIDs from our study.

2. METHODS AND ANALYSIS

2.1 Datasets and definition

Databases used for the analysis:

The data will mainly be extracted from published Phase II and III EORTC clinical trials. We will include only studies that collected HRQOL data at baseline and follow-up using the EORTC QLQ-C30 and supplementary EORTC questionnaire modules. Cancer types include melanoma, lung, colorectal, brain, head and neck, prostate, breast, testis, ovarian, pancreas, and oesophageal cancer. Data from more recent EORTC studies, completed during this project, will also be included as well as data from non-EORTC clinical trials when available.

Data will be pooled within each cancer site separately using study time (defined as days since randomization) as the common temporal scale per patient. MIDs will be established per cancer site, with attention to robustness across the different subpopulations.

The EORTC QLQ-C30:

The focus of the analysis is on the EORTC QLQ-C30, a self-administered questionnaire designed for use in cancer clinical trials. The EORTC QLQ-C30 comprises 30 items, 24 of which are aggregated into nine multi-item scales, i.e. five functioning scales (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, and nausea/vomiting) and one global health status scale. The remaining six single-item (dyspnea, appetite loss, sleep disturbance, constipation, diarrhea and the financial impact) scales assess symptoms. The financial impact scale will be omitted from the analysis because suitable anchors are unlikely to exist.

Scoring of the EORTC QLQ-C30 scales will follow the standard procedures (see EORTC QLQ-C30 Scoring Manual [17]). For consistency in signs of the change scores across the various EORTC QLQ-C30 scales, the symptom scores will be reversed to follow the functioning scales interpretation; i.e. all scales will be scored such that 0 represents the worst possible score and 100 the best possible score. All versions of the EORTC QLQ-C30 will be used [17].

2.2 Anchor selection

We hope to identify at least one suitable clinical anchor for each EORTC QLQ-C30 scale from among potential clinical factors (e.g. laboratory measures, physiological measures, clinician ratings) that are available in the databases. No patient ratings of change (e.g. subjective significance questionnaires) are available in the database. HRQOL scores will only be considered as anchors if valid MIDs are known. Since the QLQ-C30 yields 15 scales measuring a wide range of symptoms and functioning, the suitability of an anchor must be considered relative to specific HRQOL domains. A suitable anchor for any particular QLQ-C30 scale should fulfil several criteria. Most notably the anchor should be relevant for the disease indication, should have clear medical interpretation and clinicians should be familiar with it. Also there should be a conceptual and empirical relationship between the anchor and its patient-reported counterpart [11].

Anchors will be selected per cancer site. This exercise will be guided by a panel of five to six clinical experts (per disease site) who are familiar with the specific trials, as well as with the structure of the EORTC QLQ-C30. These experts will primarily be recruited from the EORTC QOL group and from the panel of investigators involved in the included studies. The clinical experts will be briefed on the

purpose of the project and the importance of selecting anchors that are clinically related to the corresponding HRQOL scales.

Clinical anchors will be pre-selected based on availability (i.e. the total that can be successfully matched to existing QLQ-C30 assessments), strength of correlation with the corresponding EORTC QLQ-C30 scale and finally clinical plausibility. A clinical anchor will be matched to an EORTC QLQ-C30 form if their respective assessment dates are within a predefined window. This time window will be determined on a per trial basis to ensure that the underlying true associations in the data are preserved. First, a candidate list of relevant clinical variables will be assembled based on the availability within each disease site. The acceptable compliance rate (i.e. availability of complete information on both the anchor and the HRQOL scale) will depend on both relative and absolute available numbers. We aim for compliance rates ≥50% and an effective sample size of at least 200 patients with repeated observations after pooling data for each cancer site separately. Thereafter, we will evaluate how well the anchors correlate with the corresponding QLQ-C30 scale at various time-points of interest. Either a Spearman's rank, polyserial or polychoric correlation will be used, depending on the distribution of the pair of variables. The correlation between their change scores will also be checked. Revicki et al. [11] suggested a correlation of ≥ |0.30| as a measure of an acceptable association. Where achievable, however, anchors with much stronger correlations will be prioritized as suggested by recent simulation studies [18]. The list of retained anchors will be independently scrutinized for clinical relevance by the clinical experts, who will help to define clinical relevant cut-offs points in the anchor. Multiple anchors will be constructed for each QLQ-C30 scale where possible. If no suitable anchors can be identified for a given scale, no anchor-based MID will be estimated and reported for that scale.

Availability of anchors

When an anchor is only available for a subset of trials, only that subset will be used. A table will be constructed to summarise the availability of each anchor in the set of trials, and the QLQ-C30 scales to which each anchor is related (conceptually, clinically and empirically). For each anchor, we will present how important change will be defined (as prescribed by our panel of clinical experts), along with the estimated correlation with the corresponding QLQ-C30 scale.

2.3 Preliminary analyses

a. Descriptive tabulation of the distribution of anchors and the EORTC QLQ-C30 scales will be made by trial, and pooled across trials. If insufficient variation is present or missing data is substantial in any anchor or scale, its inclusion in further analyses will be re-evaluated.

b. As a first step to establish the validity of an anchor, correlations between the anchors and their corresponding QLQ-C30 scales will be calculated using all matched anchor/ HRQOL scale pairs, regardless of time point. Scatterplots of the correlations will be inspected to gain greater understanding of bivariate distributions. The correlations will be calculated taking potential confounding factors into account (e.g. treatment, gender, age, disease stage, country, trial etc.,), to investigate the robustness of the associations in the overall population. Anchor/HRQOL scale pairs that fail to correlate at least 0.30 in at least one subgroup will be excluded from further consideration. Subgroups with associations < 0.30 may be excluded from further analysis, after discussion with the clinical experts. Similarly, we will investigate the correlation between change scores of the anchor and HRQOL scale over time. Priority will be given to anchor/HRQOL scale pairs with correlations of at least 0.30 when MIDs for change scores are to be calculated.

c. The HRQOL score will be presented descriptively (e.g. mean, median, range, SD) at every time point of interest, within various subgroups (e.g. treatment, gender, age group, disease stage, country, trial etc.,), as well as in the overall population.

2.4 Handling of missing data

Missing HRQOL data:

We will cross-check compliance with the protocol schedule and verify the reasons for missing data. A cross-tabulation of the clinical anchors with HRQOL compliance will be made. We will evaluate the proportion of missing HRQOL forms per category of the anchor, and also check if subjects with missing HRQOL forms differ systematically from those with complete HRQOL data. If systematic differences are found, a panel of methodological experts will be consulted to suggest appropriate sensitivity analyses (e.g. imputation techniques) to check the robustness of the estimated thresholds.

Missing clinical anchor data:

Clinical anchors will be selected in such a way that missing data is minimized. For each EORTC QLQ-C30 scale the subset of anchors with the least amount of missing data will be prioritized. Similar to the handling of missing HRQOL data, we will also explore the anchor data to identify patterns as well as reasons for missingness.

2.5 Statistical Analysis

Cross-sectional analysis of HRQOL scores

Cross-sectional differences (i.e. at the same time point) of HRQOL scores will be calculated between distinct subgroups of patients, where the grouping has been done on the clinical anchor. The categorization based on the clinical anchor are expected to yield groups that are distinct in health state, as this property is part of the clinical anchor building and evaluation process. For each HRQOL scale, the difference in mean HRQOL between each pair of adjacent group categories will be calculated at specific time-points of interest e.g. at baseline, at the end of treatment and at the end of follow-up. In addition, we will calculate effect sizes for these groups by dividing the difference of the mean HRQOL score from both groups by the standard deviation between patients in either group [8].

Anchor-based method for change scores:

The focus will be on examining both group-level and individual-level change over time. We will compute all possible pairwise time point differences in HRQOL scores and combine the data. This means that a subject can contribute multiple change scores that are calculated across different pairs of time-points, and the resulting dependency within the data will be accounted for whenever a regression model is applied. We will consider specific time intervals, namely changes in HRQOL scales in the periods between start and end of treatment, and between end of treatment and end of follow-up as these are often well defined across several studies. Furthermore, depending on the study design and setting, we will consider additional shorter time intervals prior to the end of treatment where feasible. Subjects will be assigned to distinct subgroups reflecting various levels of change (e.g. no

change, small positive changes, large positive changes, small negative changes, or large negative changes) based on the clinical anchor(s). These groups will be referred to as clinical change groups (CCG) and they are mutually exclusive. For each pair of time-points and for a given anchor, a patient can thus belong to only one CCG category.

Change in HRQOL score between two time-points is commonly expressed as a simple difference. We will explore other ways to express this change e.g. using relative differences that correct for the scores at baseline or another previous time-point. Table 1 presents a list of alternative summary scores for expressing change in HRQOL scores that will be explored. For each CCG, the summary scores will be presented descriptively (e.g. mean, median, range, SD). Differences in HRQOL summary scores between adjacent CCGs will be evaluated using primarily non-parametric techniques.

a. **Mean change method:** For a given HRQOL scale and its corresponding anchor, the MID for improvement is equal to the mean summary score of the "small positive change" CCG and the MID for deterioration is equal to "small negative change" CCG. The mean summary scores of the "small change" CCGs and that of the "no change" CCG will be compared. If the mean summary score for "no change" CCG is similar to any of the two "small change" CCGs, the estimated MID is doubtful [19].

b. **Linear regression:** The estimate of the numerical change in HRQOL summary scores (see table 1) that is associated with the transition between adjacent CCG categories will be determined using a linear regression. Separate models will be fitted for improving and deteriorating scores based on the anchor. The outcome variable is the summary score, and the covariate is a binary anchor variable; coded as "no change" = 0, and "small positive change" = 1 for model on improvement, and "no change" = 0 and "small negative change" = 1 for model on deterioration. The resulting β's (i.e. slope parameters) correspond to the MIDs for improvement and deterioration respectively. This approach can be extended to correct for other covariates that could possibly affect the MID estimates [20].

c. **Receiver operating characteristic (ROC) curves:** For each summary score, the ROC analysis will be used to estimate RTs based on an anchor. Changes in different directions will be examined separately. For example, for defining improvement, we will create an "at least

minimally important change" group using all CCGs for improvements, i.e. small positive and large positive CCGs, and a "no minimally important change" group using no change CCG and any level of worsening (i.e. small negative and large negative CCGs). Different approaches will be used to calculate threshold values, e.g. by; (i) minimizing the gap between sensitivity and specificity, (ii) minimizing the sum of 1- sensitivity and 1- specificity and (iii) minimizing the sum of squares of 1- sensitivity and 1- specificity [21]. The various estimates will be compared and triangulation considered in order to establish robust guidelines. The assurance with which an estimated threshold can be used will depend on their corresponding sensitivity and specificity values. It is commonly not recommended to apply thresholds to individual patients when sensitivity and specificity are less than 75% [22].

d. **Empirical cumulative distribution function (ECDF)**: For each possible value of a given summary score (see Table 1) expressing change in the EORTC QLQ-C30 domains over time, the percentage of patients achieving at least that amount of change will be plotted separately for each CCG, and also separately for improving and deteriorating scores. The benefit of this approach is that the separation between CCGs may be visually compared across all values of the summary scores, thus offering a range of possible RTs for clinical relevance that can be considered simultaneously [23, 24].

The estimated thresholds across these methods will be compared, and the percentage of patients with improved or deteriorated HRQOL scores will be reported. Recommendation for using estimated RTs for classifying individual patients will be based on whether the probability of misclassification is low [22] or whether the RT values exceed the measurement error level by comparing the thresholds to the minimum detectable change (MDC) [13, 3, 22, 25]. The $\text{MDC} = 1.96 * \sqrt{2} * SEM$ [25] represents the smallest change that can be considered to be above the measurement error. Usually if the MDC is > the RT then the measure is insufficiently precise to monitor individual patients. Furthermore, when setting RTs, especially on domains that are computed based on a single item, we will check that the RTs align with the underlying change levels of the scale scores [26].

Distribution-based methods

We will examine the distribution-based approaches based on the standard deviation criteria, e.g. 0.2 SD, 0.3 SD, 0.5 SD, and the standard error of mean, SEM [13]. The SDs and SEM will be calculated on the summary scores (see Table 1) yielding MIDs corresponding to the rules above. Since this approach requires that the data are normally distributed, those summary scores that violate this assumption (based on standard testing techniques) will not be considered.

Effect sizes (ES) [14] will be calculated by dividing the summary scores in Table 1 by the pooled standard deviation of subjects at baseline (i.e., before treatment). This will be done for any two adjacent time points, e.g. depending on whether the level of compliance is acceptable. As a variation we will also calculate the effect sizes between adjacent time points by using the standard deviations of subjects at the previous time point [27].

3. VALIDATION AND SENSITIVITY ANALYSIS

3.1 Stability of the estimated MIDs

Internal validation:

Characteristics such as age, gender, disease stage, country, etc. typically influence the absolute score outcomes of many HRQOL scales [28]. The stability of the estimated MIDs will therefore be investigated by including these factors (one at a time) and an interaction term with the anchor in a regression model. We will include as many socio-demographic and clinically relevant covariates as are available from the study database and that can be evaluated by the available sample size.

External validation:

For each cancer site in order to perform external validation, we will examine external (i.e. non-EORTC) studies having comparable data. This is subject to the availability of such data.

3.2 Handling the boundaries (floor and ceiling) effects

We will check for the proportion of patients with boundary (extreme) scores. For those patients where the later time-point was a boundary score, the change over time may be incorrectly estimated by

simple subtraction. The change in clinical anchor for these patients at the boundaries will be used to estimate the magnitude of the problem. The proportion of patients with a change in clinical anchor that is not reflected in the HRQOL change due to the boundary constraints would be an indication of a limiting boundary problem. As a sensitivity check, we will investigate how much the MIDs are affected if we include or exclude these patients.

4. ETHICS AND DISSEMINATION

All patient data originate from completed clinical trials with mandatory written informed consent, approved by local ethical committees. Our findings will be presented at scientific conferences, disseminated via peer-reviewed publications, and also compiled in a MID "blue book" which will be made available online on the EORTC Quality of Life Group (QLG) web site as a free guideline document.

5. CONCLUSION

In this project we will determine minimally important differences (MID) for HRQOL scores of the EORTC QLQ-C30, using empirical individual patient data. The main focus is on the anchor-based approach. We aim to construct multiple anchors per QLQ-C30 multiple-item or single-item scale and apply and compare results from several anchor-based methods as recommended in the literature [11, 19]. Figure 1 presents a flow diagram summarizing the key data component, the clinical anchor construction procedure and the main statistical methods which will be applied in this project. Hopefully the resulting MID estimates can triangulate to one value or a small range of values.

It is important to highlight that there are diverse opinions in the literature on whether or not it is plausible to use the same methods for interpreting individual-level change versus group-level differences/change. For instance, the mean change method and the ROC curve method have been labelled to be appropriate for comparing group-level and individual-level change respectively [20, 29]. On the other hand, both methods have been recommended to be useful for estimating MIDs that are useful for interpreting either group-level or individual-level change as long as the anchor is available at the individual level [22, 30, 31]. There is a need for a consensus on this matter. We will compare and

contrast MID estimates from the different methods to provide empirical evidence, and assess whether it is possible to apply a simplified guideline to between group differences/change and individual-level change.

A strength of our research is its integral combination of both clinical and methodological expertise. The findings will ultimately improve the interpretation of the QLQ-C30 scale scores in clinical trials by providing empirical guidelines for relevant improvements and deteriorations.

Each year there are over 5000 newly registered downloads of the EORTC quality of life measures. The information from our research will be of added value to all its users (e.g. pharma and academic) since a frequent issue raised by regulators and trial sponsors is an understanding of MID.

The main limitations of this project are that anchor-based MIDs can only be estimated for QLQ-C30 scales for which a suitable anchor are available in the database. Also, the available anchors rely exclusively on clinical observations or interpretations. Unfortunately, patient ratings of change (e.g. subjective significance questionnaires) are not available in the study database. We will consider using other HRQOL scores as a way to include the patient's perspective if valid MIDs are known for the given HRQOL scores.

Overall, this project will supplement previously published research by using individual patient data to estimate MIDs for different cancer sites separately, hence further providing evidence to robust and practical MID guidelines for the EORTC QLQ-C30.

6. 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. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcome Res. 2011 Apr;11(2):171-84.
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. Lydick F, Epstein RS. Interpretation of quality of life changes. Qual Life Res. 1993; 2: 221-226.
6. 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.
7. 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.
8. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res. 1996; 5: 555-567.
9. Cella D, et al. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol. 2002; 55: 285-295.
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. 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
12. Nordin A, Taft C, Lundgren-Nilsson A, Dencker A. Minimal important differences for fatigue patient reported outcome measures-a systematic review. BMC Med Res Methodol, 16 (2016), p. 62.
13. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J. Clin. Epidemiol. 1999; 52(9), 861–873.
14. Cohen J. Statistical Power Analysis for the Behavioural Sciences (2nd Edition). Lawrence Erlbaum Associates, NJ, USA (1988).

15. 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.
16. 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.
17. 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
18. 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.
19. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. COPD 2(1), 63–67 (2005).
20. Angst F., Aeschlimanna A. and Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol. 2016; 82: 128-136.
21. Froud R, Abel G. Using ROC Curves to Choose Minimally Important Change Thresholds when Sensitivity and Specificity Are Valued Equally: The Forgotten Lesson of Pythagoras. Theoretical Considerations and an Example Application of Change in Health Status. Caylà JA, ed. PLoS ONE. 2014;9(12):e114468. doi:10.1371/journal.pone.0114468
22. de Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW, Hendriks EJ, Bouter LM, Terwee CB. Three ways to quantify uncertainty in individually applied "minimally important change" values. J Clin Epidemiol. 2010 Jan;63(1):37-45. doi: 10.1016/j.jclinepi.2009.03.011.
23. Farrar JT, Dworkin RH, Max MB. Use of the cumulative proportion of responders analysis graph to present pain data over a range of cutoff points: making clinical trial data more understandable. J. Pain Symptom Manage. 2006; 31(4):369–377.
24. McLeod LD, Coon CD, Martin SA, Fehnel SE, and Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharmacoecon Outcomes Res. 2011; 11(2): 163–169. doi:10.1586/erp.11.12.
25. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual. Life Outcomes. 2006; 4, 54.
26. Wyrwich K. What's the Score? Moving from Items to Scores -Methods, Considerations, and Case Examples. Eighth Annual Patient-Reported Outcome Consortium Workshop. Bethesda, MD, April 2017; 13-28 (https://c-path.org/wp-content/uploads/2017/05/2017_session5_scoringfinal.pdf).

27. Yost, Kathleen J. et al. Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) instrument. Value in Health , 2005; 8, 117 – 127
28. Wan GJ, Counte MA, Cella DF, Hernande L, Deasy S, Shiomoto G: An analysis of the impact of demographic, clinical and social factors on health-related quality of life. Value in Health 1999, 2: 308–318. 10.1046/j.1524-4733.1999.24006.x
29. Wells GA, Beaton D, Shea B, Boers M, Simon L, Strand V, et al. Minimal clinically important differences: review of methods. J Rheumatol. 2001; 28:406-12.
30. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, the Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002; 77:371-83.
31. 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.

Acknowledgements: We thank the various EORTC disease group members and their clinical investigators, and all the patients who participated in the trials that we shall be analyzing.

Contributors: AB, CC, DEE, MTK, MG, MAGS, GV and HF contributed to the conception and design of the study. MG, MAGS, YB, GV, HF revised the proposed methodology for clinical plausibility. JZM, JH, CC, CK, DEE, JM and MTK provided critical input on the proposed statistical analysis. JZM and CC drafted the protocol. All the authors read and corrected the drafts and approved the final version.

Funding: This study was funded by an unrestricted academic grant from the EORTC Quality of Life Group.

Competing interests: The authors declare no conflicts of interest.

Data Sharing Statement: Please visit http://www.eortc.org/data-sharing/ for details about the data sharing policy of the EORTC.

Table 1: Description of various summary scores for expressing changes in HRQOL scores over time

Name	Formula	Description
Absolute difference (AD)	$Q_2 - Q_1$	Simple difference in the HRQOL between two time-points. The mean of the AD values from all subjects corresponds to the Osoba's MID (Osoba et al., 1998).
Piecewise absolute difference (pAD)	$Q_{2c} - Q_{1c}$	Applying AD per subgroup c = 1 to 4, where c is based on baseline QOL values grouped as; *0-25, 26-50, 51-75* and *76-100.*
Relative difference (RD_1) – ordered	$\frac{(Q_2 - Q_1)}{Q_1} \times 100$	% change from previous value. Two variations based on direction Note: Q_1 and Q_2 are based on the Raw Scores (RS). RS is the mean of the component items for a particular scale that takes values from 1to 4 (instead of 0 to100).
Relative difference (RD_2) – sum	$\frac{(Q_2 - Q_1)}{(Q_2 + Q_1)}$	% change compared to sum. Convention: if $Q_2 = Q_1 = 0$ then RD2 = 0.
Relative difference (RD_3) – baseline	$\frac{(Q_2 - Q_1)}{Q_B}$	Q_B = baseline value Note: Q_1, Q_2 and Q_B are based on RS.
Fixed value	Q_2	MID is dependent on the observed value. This will result in a fixed threshold, not dependent on change in HRQOL score from previous value.
Slope	$\frac{(Q_2 - Q_1)}{(T_2 - T_1)}$	The absolute difference between two time points is weighted by the corresponding time period

Q_x = HRQOL outcome at time point T_x

Figure 1: A flow diagram summarizing the data (e.g. the cancer sites, QLQ-C30 scales and types of clinical variables that will be used for anchor construction), the clinical anchor construction step and the main statistical methods which will be applied in this project.

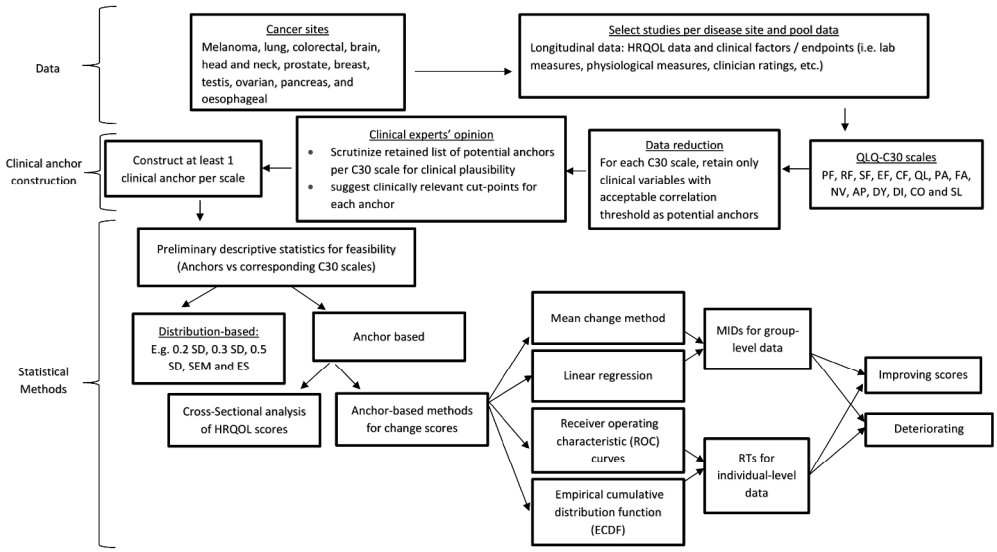

Figure 1: A flow diagram summarizing the data (e.g. the cancer sites, QLQ-C30 scales and types of clinical variables that will be used for anchor construction), the clinical anchor construction step and the main statistical methods which will be applied in this project.

275x153mm (300 x 300 DPI)