PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Establishing Anchor-based Minimally Important Differences (MID) with the EORTC Quality of Life Measures: a meta-analysis protocol
AUTHORS	Musoro, Zebedee Jammbe; Hamel, Jean-Francois; Ediebah, Divine Ewane; Cocks, Kim; King, Madeleine; Groenvold, Mogens; Sprangers, Mirjam; Brandberg, Yvonne; Velikova, Galina; Maringwa, John; Flechtner, Hans-Henning; Bottomley, Andrew; Coens, Corneel

VERSION 1 – REVIEW

REVIEWER	Cheryl D. Coon Principal at Outcometrix Boston, Massachusetts, USA
REVIEW RETURNED	06-Sep-2017

GENERAL COMMENTS	It is exciting that EORTC is undertaking such a large analysis project with a robust heterogenous database to establish interpretation thresholds for the QLQ-C30. The field will greatly benefit from this work, and the comparison of results across methods, studies, and subgroups will be of particular methodological interest for psychometricians looking to inform their own research. My main concern with this protocol is the use of the term MID to apply to both group-level differences and individual-level changes. The PRO field has generally moved away from using the term MID with individual-level change. Some experts in the field (particularly in Europe) use the term MIC because the threshold is identifying the minimal amount of within-person change that would be important to patients (e.g., De Vet et al., 2006). In the US, the FDA uses the term responder definition to indicate "the individual patient PRO score change over a predetermined time period that should be interpreted as a treatment benefit" (PRO Guidance, 2009). The authors acknowledge that interpretation thresholds *may* differ when interpreting group-level data versus individual-level data, and if that is confirmed with this research, then there will be a need for them to refer to group-level MID and individual-level MID, which just perpetuates the confusion. I strongly recommend that the authors reconsider the use of the blanket term MID in this article and instead either (a) use "MID" to refer to group-level data while using "MIC" or "responder threshold" for individual-level data, or (b) refrain from picking a term and just use the generic terms "group-level threshold" and "individual-level threshold".

	One other suggestion I have to improve this protocol and get organized for data analysis and results reporting is to create a table that summarizes domains (physical function, emotional function, etc.) methods (mean change, linear regression, etc.), anchors (performance status, CTCAE, etc.), subgroups (age, gender, etc.), direction (improvement, worsening), and cancer site (melanoma, lung, etc.) to show the different permutations that will be applied to these data. This table would allow for you to compute how many tables and figures will be produced for this work. It wasn't clear to me if subgroups will each get their own set of tables and figures, so summarizing this process will help to set expectations. Additionally, I have some feedback specific to pages and lines within the manuscript: p. 4 line 19-23: This sentence quotes a definition of MID but references 3 different articles. Shouldn't the citation be the single source of the quote? p. 8 line 19-21: Have you considered using the global QOL item from the C30 as an anchor? That would be available for all studies and has been used as an anchor in similar analyses previously (e.g., Bedard G, Zeng L, Zhang L, Lauzon N, Holden L, Tsao M et al. Minimal important differences in the EORTC QLQ-C30 in patients with advanced cancer. Asia-Pacific journal of clinical oncology. 2014;10(2):109-17). p. 8 line 25-30: In the table of anchors, it would be good to include the way in which important change will be defined using the anchor (e.g., a score of "somewhat better" on the PGIC) and the correlation with the QLQ-C30 domains. p. 9 line 34; p. 11 line 52: EORTC is misspelled p. 9 line 48-p. 10 line 9: It is not clear how these cross-sectional differences will be used, as this is an assessment of construct validity. Further, it is not clear that adjacent anchor categories will be distinct and equidistant (or if they need to be). p. 10 line 23-25: Depending on the study design, treatment type, and cancer staging, you may need to consider shorter time intervals prior to the end of treatment. For example, if you're working with Phase 2 data in advanced cancer, an investigational drug may show an initial symptom improvement within weeks or months of beginning treatment, and you would want use this earlier time point as a source of improved data, as their end of treatment visit may actually reflect disease progression and worsened symptoms. p. 11 line 26-48: You should use all subjects for the ROC analysis rather than limiting to no change and small change. E.g., for defining improvement, you would create a "minimally or greater importantly changed" group using small positive CCG and greater improvements, and a "not minimally importantly changed" group using no change and any level of worsening. Your results should be more robust, and sensitivity and specificity will be more likely to reach 75% because by including subjects who exceeded a small level of change, you will be increasing the number of true positives and true negatives.

	p.12 line 7-20: When triangulating across results, you should consider what Kathy Wyrwich has called "state change" (see her PRO Consortium 2017 slides https://c-path.org/wp-content/uploads/2017/05/2017_session5_scoringfinal.pdf). If you're setting individual-level thresholds on domains that are computed based on a single item, then the minimum an individual can change on that domain is 33.3. Thus, a threshold of, say, 10 would not make sense because that would correspond to a change of less than one category on the 4-point item response scale. Thus, the state change concept is another version of the minimum detectable change because change less than that value is impossible on the scale. p. 12 line 54-p. 13 line 7: It is not clear what data will be input into the regression model - will it be the threshold estimates across each of the data sources? If so, this means that you will have to repeat all of the analyses by sex, age, disease stage, country, etc., which seems like a daunting task. I would recommend skipping this analysis, at least for sex, age, and country, as it is unlikely that there are inherent reasons for why the threshold should differ across these demographic groups. If you put them into a regression model, then there may be significant differences due to sample size, but these differences may not be meaningful or worth setting different threshold for different subgroups. It may, however, be worth considering disease stage, as patients may perceive symptom changes differently depending on their baseline symptom severity. p. 14 line 3-21: While I believe that different methods should be used for individual versus group interpretation, it is good that you are going into this with an open mind and will consider if the values should indeed differ. The literature is not aligned on the topic of method selection, and this study will offer a great opportunity to compare results across methods. If you were to match methods to level of interpretation a priori, my recommendation is to use the mean change method to define within-group change and linear regression to define between-group differences. The mean change method identifies subjects who experienced the a minimal amount of change, and their mean change score would be akin to looking at the mean change score for a treatment group in a trial (i.e., within-group change). The linear regression method compares change scores for subjects who experienced the a minimal amount of change to subjects who experienced no change, and the difference between the scores for these two groups would be akin to comparing the mean change score for a target treatment group to a comparator group (i.e., between-group differences). ROC and ECDF should then be used for within-person change because the estimates are focused on individuals rather than group-level means.

REVIEWER	Anna Dencker RN, RM, PhD, Associate professor, Senior lecturer Institute of Health and Care Sciences, Sahlgrenska Academy, University of Gothenburg Sweden
REVIEW RETURNED	12-Sep-2017

GENERAL COMMENTS	Thank you for letting me review this interesting protocol: Establishing Anchor-based Minimally Important Differences (MID) with the EORTC Quality of Life Measures: a meta-analysis protocol. I have a few comments listed below:

	Introduction: • Please add page reference for quotation in second paragraph. • In second paragraph it is said that MIDs are often described as interchangeable at group- and individual level. This statement needs references, is that really the case? Maybe better to re-formulate. • MIDs vary substantially between patient groups and different kind of anchors, as for ex. described by Nordin et al in BMC Medical research methodology, 2016. The view of the patient would surely differ from the view of the clinician or from results on a test/lab. This problem is not at all taken into consideration in the first paragraph on page 5 where “patient/physician-derived ratings of change” is mentioned. Please elaborate on probable problems to encounter. Methods and analysis: • It is said that data from all published EORTC clinical trials (II and III) will be used. Are the data already collected? From all countries? • Why exclude samples <200 patients? A meta-analysis is a good method to compile data from both small and large studies. • What is meant by “compliance rate”, is it compliance to treatment? Maybe important information is lost if non-compliant patients are left out in the analysis. Ethics and dissemination: • I would suggest an ethics application for the study because individual data are used.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Cheryl D. Coon
Institution and Country: Principal at Outcometrix, Boston, Massachusetts, USA Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
It is exciting that EORTC is undertaking such a large analysis project with a robust heterogenous database to establish interpretation thresholds for the QLQ-C30. The field will greatly benefit from this work, and the comparison of results across methods, studies, and subgroups will be of particular methodological interest for psychometricians looking to inform their own research.

1. My main concern with this protocol is the use of the term MID to apply to both group-level differences and individual-level changes. The PRO field has generally moved away from using the term MID with individual-level change. Some experts in the field (particularly in Europe) use the term MIC because the threshold is identifying the minimal amount of within-person change that would be important to patients (e.g., De Vet et al., 2006). In the US, the FDA uses the term responder definition to indicate "the individual patient PRO score change over a predetermined time period that should be interpreted as a treatment benefit" (PRO Guidance, 2009). The authors acknowledge that interpretation thresholds *may* differ when interpreting group-level data versus individual-level data, and if that is confirmed with this research, then there will be a need for them to refer to group-level MID and individual-level MID, which just perpetuates the confusion. I strongly recommend that the authors reconsider the use of the blanket term MID in this article and instead either (a) use "MID" to refer to group-level data while using "MIC" or "responder threshold" for individual-level data, or (b) refrain from picking a term and just use the generic terms "group-level threshold" and "individual-level threshold".

Response:
Thank you for highlighting the need to make the distinction between group-level and individual-level threshold. We agree with this proposal.
On page 4 (line 15-17), we have added that “MID” will be used to refer to group-level threshold and “responder threshold” for individual level change.
We have also added that this project will focus more on group-level change (page 4, line 17).

2. One other suggestion I have to improve this protocol and get organized for data analysis and results reporting is to create a table that summarizes domains (physical function, emotional function, etc.) methods (mean change, linear regression, etc.), anchors (performance status, CTCAE, etc.), subgroups (age, gender, etc.), direction (improvement, worsening), and cancer site (melanoma, lung, etc.) to show the different permutations that will be applied to these data. This table would allow for you to compute how many tables and figures will be produced for this work. It wasn't clear to me if subgroups will each get their own set of tables and figures, so summarizing this process will help to set expectations.

Response:
We appreciate this suggestion to add a more structured overview of the complete envisioned MID project. However it is very difficult to create such a table beforehand. While some of the components can be pre-specified (e.g. QLQ-C30 domains), other components such as clinical anchors will be dependent on the available data. It may very well be possible that e.g. we are able to construct multiple valid anchors for a scale in one disease setting but that no such anchor may be found in another setting.

Nonetheless, we adopted this suggestion by constructing a flow diagram (see Figure 1 below) to summarize the key data component, the clinical anchor construction procedure and the main statistical methods which will be applied in this project. This flow diagram has been added on page 20 and has been referred to in the main text on page 14 (line 14-15).

3. Additionally, I have some feedback specific to pages and lines within the manuscript:

a. p. 4 line 19-23: This sentence quotes a definition of MID but references 3 different articles. Shouldn't the citation be the single source of the quote?

Response:
We have retained one reference for the quoted definition of MID.

b. p. 8 line 19-21: Have you considered using the global QOL item from the C30 as an anchor? That would be available for all studies and has been used as an anchor in similar analyses previously (e.g., Bedard G, Zeng L, Zhang L, Lauzon N, Holden L, Tsao M et al. Minimal important differences in the EORTC QLQ-C30 in patients with advanced cancer. Asia-Pacific journal of clinical oncology. 2014;10(2):109-17).

Response:
The use of the global QLQ-C30 scale as an anchor for the patient-reported scales was not considered in this project. The intention of the anchor construction was to use clinical data, originating from clinician evaluation or test results as surrogate measures of clinical relevance. The use of the QLQ-C30 global scale as a valid anchor is controversial as it is subject to response shift over time, subjective interpretation and lacks an established MID. The authors of the proposed article (Bedard et al. 2014) themselves acknowledge this use of the global scale to be “questionable”.

We have added the following text on page 7 (line 14-15)
“No patient ratings of change (e.g. subjective significance questionnaires) are available in our database. HRQOL scores will only be considered as anchors if valid MIDs are known.”

c. p. 8 line 25-30: In the table of anchors, it would be good to include the way in which important change will be defined using the anchor (e.g., a score of "somewhat better" on the PGIC) and the correlation with the QLQ-C30 domains.

Response:
We cannot say beforehand how important change groups will be defined, since it depends on the selected clinical variables. We have mentioned (on page 8 line 12) that clinically relevant cut-offs points in the selected anchor will be defined with the help of our clinical experts.

We have also added the text below on page 8 line 20-22.
“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.”

d. p. 9 line 34; p. 11 line 52: EORTC is misspelled

Response:
Thank you for noticing this error. It has been corrected.

e. p. 9 line 48-p. 10 line 9: It is not clear how these cross-sectional differences will be used, as this is an assessment of construct validity. Further, it is not clear that adjacent anchor categories will be distinct and equidistant (or if they need to be).

Response:
Cross-sectional differences will be calculated between distinct subgroups of patients where the grouping has been done on the clinical anchor. As an example, a patient population may be split according their CTCAE fatigue rating (0 vs 1 vs 2 vs 3 vs 4) at baseline. The resulting QLQ-C30 fatigue score will be presented for each of these subgroups. While differences between these subgroups are indications of construct validity, the magnitude of these differences represent the difference in the patient-reported scale between two groups with distinct clinical health states. This does inform us about the size of the MID based on clinical anchor classification.

By construction, the categorization based on the clinical anchor should yield groups that are distinct in health state. These properties are part of the clinical anchor building and evaluation process.

We have modified the text on page 10 (line 4-7) as shown below.

“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.

f. p. 10 line 23-25: Depending on the study design, treatment type, and cancer staging, you may need to consider shorter time intervals prior to the end of treatment. For example, if you're working with Phase 2 data in advanced cancer, an investigational drug may show an initial symptom improvement within weeks or months of beginning treatment, and you would want use this earlier time point as a source of improved data, as their end of treatment visit may actually reflect disease progression and worsened symptoms.

Response:
Thank you for this remark. The have edited the text in this section (page 10, line 19-23). It now reads as follows:
“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.”

g. p. 11 line 26-48: You should use all subjects for the ROC analysis rather than limiting to no change and small change. E.g., for defining improvement, you would create a "minimally or greater importantly changed" group using small positive CCG and greater improvements, and a "not minimally importantly changed" group using no change and any level of worsening. Your results should be more robust, and sensitivity and specificity will be more likely to reach 75% because by including subjects who exceeded a small level of change, you will be increasing the number of true positives and true negatives.

Response:
Our original reasoning was to limit to minimal change categories in order to reflect the “minimal” important threshold. Indeed we agree that using all data should yield more robust results.
On page 11 (line 24-27) the binary outcome indicator has been re-defined to include all subjects as suggested by the reviewer. The modified text is presented below.
“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).”

h. p.12 line 7-20: When triangulating across results, you should consider what Kathy Wyrwich has called "state change" (see her PRO Consortium 2017 slides https://c-path.org/wp-content/uploads/2017/05/2017_session5_scoringfinal.pdf). If you're setting individual-level thresholds on domains that are computed based on a single item, then the minimum an individual can change on that domain is 33.3. Thus, a threshold of, say, 10 would not make sense because that would correspond to a change of less than one category on the 4-point item response scale. Thus, the state change concept is another version of the minimum detectable change because change less than that value is impossible on the scale.

Response:
Thank you for this very important observation. We have added the text below on page P. 12 (line 20-22) and have also referred to the Wyrwich’s slides.

“. 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].”

i. p. 12 line 54-p. 13 line 7: It is not clear what data will be input into the regression model - will it be the threshold estimates across each of the data sources? If so, this means that you will have to repeat all of the analyses by sex, age, disease stage, country, etc., which seems like a daunting task. I would recommend skipping this analysis, at least for sex, age, and country, as it is unlikely that there are inherent reasons for why the threshold should differ across these demographic groups. If you put them into a regression model, then there may be significant differences due to sample size, but these differences may not be meaningful or worth setting different threshold for different subgroups. It may, however, be worth considering disease stage, as patients may perceive symptom changes differently depending on their baseline symptom severity.

Response:
We agree that the impact of these covariates is unlikely but we think it still needs to be verified. Hence these analyses are part of the section on sensitivity (and not the primary analysis). Sensitivity for gender, age, country etc. needs to be confirmed as these are characteristics that do typically influence the absolute score outcomes for many PRO scales. All analyses are subject to feasibility based on available data and distribution.

We have modified the text on page 13 (line 12-14) to read as follows:
"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."

j. p. 14 line 3-21: While I believe that different methods should be used for individual versus group interpretation, it is good that you are going into this with an open mind and will consider if the values should indeed differ. The literature is not aligned on the topic of method selection, and this study will offer a great opportunity to compare results across methods. If you were to match methods to level of interpretation a priori, my recommendation is to use the mean change method to define within-group change and linear regression to define between-group differences. The mean change method identifies subjects who experienced the a minimal amount of change, and their mean change score would be akin to looking at the mean change score for a treatment group in a trial (i.e., within-group change). The linear regression method compares change scores for subjects who experienced the a minimal amount of change to subjects who experienced no change, and the difference between the scores for these two groups would be akin to comparing the mean change score for a target treatment group to a comparator group (i.e., between-group differences). ROC and ECDF should then be used for within-person change because the estimates are focused on individuals rather than group-level means.

Response:
Thank you for this comment. We agree that the literature is not clear on these distinctions. We have now made the distinction between MID and RT (responder threshold) throughout the manuscript to distinguish between group-level versus individual-level thresholds. In the statistical analysis section, we have altered the text using this MID vs RT terminology to attribute the mean change and linear regression methods to group-level estimation, and the ROC and the ECDF methods to individual-level estimation as per your suggestion.
We will also take great care whenever publishing results from this initiative to distinguish between them.

Reviewer: 2
Reviewer Name: Anna Dencker
Institution and Country: RN, RM, PhD, Associate professor, Senior lecturer, Institute of Health and Care Sciences, Sahlgrenska Academy, University of Gothenburg, Sweden Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
Thank you for letting me review this interesting protocol: Establishing Anchor-based Minimally Important Differences (MID) with the EORTC Quality of Life Measures: a meta-analysis protocol. I have a few comments listed below:

1. Introduction:

a. Please add page reference for quotation in second paragraph.

Response:
A page reference has been added to the quoted definition of MID in the second paragraph.

b. In second paragraph it is said that MIDs are often described as interchangeable at group- and individual level. This statement needs references, is that really the case? Maybe better to re-formulate.

Response:
This statement has been re-formulated to address the confusion in the literature on selecting methods for estimating group-level vs individual-level thresholds.
Hence, on page 5 (line 7-10), the following text has been added in the manuscript.
"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 [13]."

We further discussed this point in the "conclusion section" (page 14 line 17-25) based on some cited examples from the literature.

c. MIDs vary substantially between patient groups and different kind of anchors, as for ex. described by Nordin et al in BMC Medical research methodology, 2016. The view of the patient would surely differ from the view of the clinician or from results on a test/lab. This problem is not at all taken into consideration in the first paragraph on page 5 where "patient/physician-derived ratings of change" is mentioned. Please elaborate on probable problems to encounter.
Response:
Thank you for pointing out this problem. We have addressed this point by adding the text below in the manuscript (page 5 line 3-7) and have also referred to the review by Nordin et al 2016.

"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. 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, 12]."

2. Methods and analysis:

a. It is said that data from all published EORTC clinical trials (II and III) will be used. Are the data already collected? From all countries?

Response:
Data that will be used in this project have already been collected and stored in the EORTC database. This comprise published Phase II and III EORTC clinical trials. Data were collected in several countries. Also, while most of the individual patient data has already been collected and is available through the various study-specific databases, these data have not yet been pooled in to a meta-database to allow our analyses.

The text on page 6 (line 15-17), has been modified and now reads;
"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."

b. Why exclude samples <200 patients? A meta-analysis is a good method to compile data from both small and large studies.

Response:
We will be pooling data for each cancer site separately (since our study focuses on estimating MIDs for the different cancer sites). We shall not exclude individual studies with samples of <200 patients. Instead, we are targeting a sample size of at least 200 patients when data from all studies for a particular cancer sites have been pooled. Two hundred patients was deemed to be the minimum feasible sample size by disease site to allow the listed analysis with sufficient reliability.
We have edited the text on page 8 (line 5-7). This now reads:
"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."

c. What is meant by "compliance rate", is it compliance to treatment? Maybe important information is lost if non-compliant patients are left out in the analysis.

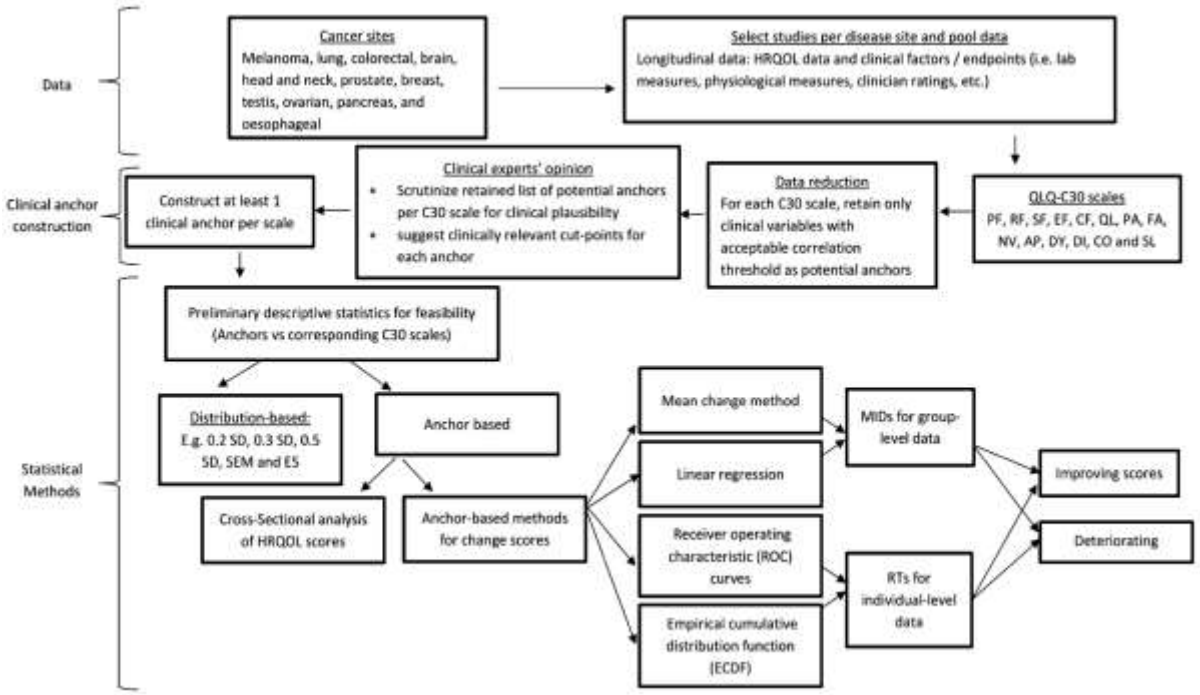
Response:
"Compliance rate" refers to the availability of complete information on both the anchor and the HRQOL scale.
Non-compliance may arise from missing data. E.g. Laboratory data may be missing due to unreadable lab results or skipped visits. Quality of life data is often missing due to patients unwilling to complete the questionnaires or hospital staff failing to administer these within the study-specific time frames.
The modified text below has been added on page 8 (line 4-5).
"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."

3. Ethics and dissemination:
a. I would suggest an ethics application for the study because individual data are used.

Response:
All patient data comes from existing clinical trials, approved by local ethical committees, where patients had to provide written informed consent. This use of patient data has been reviewed and approved by the EORTC Board and the EORTC Quality of Life Group. Use of the data for this project was found to be covered under the original informed consent statements.
The following text has been added in the "Ethics and dissemination" section on page 14 (line 3-4).

"All patient data originate from completed clinical trials with mandatory written informed consent, approved by local ethical committees."

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.



VERSION 2 – REVIEW

REVIEWER	Cheryl D. Coon Principal at Outcometrix, Boston, Massachusetts, USA
REVIEW RETURNED	30-Oct-2017

GENERAL COMMENTS	I appreciate the authors' careful responses to the reviewers' comments. I find the protocol to be very clear and exceptionally thorough. I understand the authors' hesitation to use the global scale as an anchor, as it lacks its own thresholds for interpretation, though I would argue that it allows for a direct connection of the scores back to the patient perspective that is not available from clinical anchors.

	I would encourage the authors to critically evaluate the appropriateness of the selected anchors and the inherent meaning of the changes they reflect. For example, ECOG performance status may be correlated > 0.30 with constipation, but should it be expected that a meaningful change in ECOG corresponds to a meaningful change in constipation? As another example, constipation as measured by the CTCAE is likely well-correlated with patient-reported constipation, but if the CTCAE only measures clinician-reported constipation as an adverse event related to treatment, then isn't it possible that there would be changes on the QLQ-C30 constipation item that wouldn't be registered on the CTCAE (e.g., in colorectal cancer)? The authors will have to be careful not to let correlations alone drive the choice and use of anchors, and the triangulation process across anchors and data sources will serve as an important and valuable check on the proposed interpretation thresholds.

REVIEWER	Anna Dencker Institute of health and care sciences, Sahlgrenska Academy, Sweden
REVIEW RETURNED	26-Oct-2017

GENERAL COMMENTS	I am pleased with the authors' answers to my questions and I have no further comments.

VERSION 2 – AUTHOR RESPONSE

We agree that it is important to take the patient's perspective into account as well. Unfortunately No patient ratings of change (e.g. subjective significance questionnaires) are available in the database of this project. We have mentioned this as one of the limitations in the conclusion section.

The text below has been added on page 15 (line 12-17).

"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."

With regards to the anchor selection, we agree with the reviewer on the need to critically evaluate the appropriateness of the selected anchors and the inherent meaning of the changes they reflect.
In this project, we will select anchors based on: (i) the strength of the correlation with the corresponding EORTC QLQ-C30 scale, and (ii) the clinical plausibility of the association between the given anchor and the QOL scale.
To judge clinical appropriateness of selected anchors, we will depend on a panel of clinical experts (per disease site) who are familiar with the specific trials, as well as with the structure of the EORTC QLQ-C30.

We have included the following text on page 7 (line 25-27)

“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.”