Quality of Life Research (2023) 32:1247–1253
https://doi.org/10.1007/s11136-022-03181-4

SPECIAL SECTION: METHODOLOGIES FOR MEANINGFUL CHANGE

How scoring limits the usability of minimal important differences (MIDs) as responder definition (RD): an exemplary demonstration using EORTC QLQ-C30 subscales

Kim Cocks[1,2] · Jacqueline Buchanan[3]

Accepted: 18 June 2022 / Published online: 9 July 2022

Abstract

Purpose The recommended method for establishing a meaningful threshold for individual changes in patient-reported outcome (PRO) scores over time uses an anchor-based method. The patients assess their perceived level of change and this is used to define a threshold on the PRO score which may be considered meaningful to the patient. In practice, such an anchor may not be available. In the absence of alternative information often the meaningful change threshold for assessing between-group differences, the minimally important difference, is used to define meaningful change at the individual level too. This paper will highlight the issues with this, especially where the underlying measurement scale is not continuous.

Methods Using the EORTC QLQ-C30 as an example, plausible score increments (“state changes”) are calculated for each subscale highlighting why commonly used thresholds may be misleading, including leading to sensitivity analyses that are inadvertently testing the same underlying threshold.

Results The minimal possible individual score change varies across subscales; 6.7 for Physical Functioning, 8.3 for Global Health Scale and Emotional Functioning, 11.1 for fatigue, 16.7 for role functioning, cognitive functioning, social functioning, nausea and vomiting, pain and 33.3 for single items.

Conclusions The determination of meaningful change for an individual patient requires input from the patients but being mindful of the underlying scale ensures that these thresholds are also guided by what is a plausible change for patients to achieve on the scale.

Keywords Meaningful change · Responder definition · State change · EORTC QLQ-C30

Introduction

It is common to convert a patient-reported outcome measure (PRO) into scores for a variety of concepts being measured, e.g. pain or physical function scores. In a clinical trial, these scores are used to compare between treatments, normally to see if one treatment is superior to another in relation to patients’ quality of life. In order to summarise these scores for each treatment group, mean scores are often used and then compared across the groups to check for differences. Another method is to compare the proportion of patients in each group that have experienced a meaningful change, referred to as a ‘PRO responder’ or ‘PRO non-responder’. This would be used to demonstrate a treatment benefit if a higher proportion of patients experienced a PRO response in one treatment group compared to the other. The PRO scores are dichotomised using a threshold, commonly referred to as the responder definition (RD), which represents a change in an individual’s score that would provide evidence of a treatment benefit. Although statistically sub-optimal, since an ordinal or continuous score is dichotomized for analysis resulting in loss of information, there is an increasing emphasis on defining importance of individual patient change [1]. Therefore, there is a need for careful consideration of appropriate thresholds for responders and standardisation across studies so that results are comparable [2]. Often, due to the uncertainty in defining these thresholds, sensitivity analyses are also required to test alternative thresholds. Since PROs use different response scales,

✉ Kim Cocks
kim@kcstats.co.uk

[1] KCStats Consultancy, Leeds, UK

[2] Adelphi Values, Cheshire, UK

[3] Amgen Ltd, South San Francisco, CA, USA

different numbers of items and different scoring techniques these thresholds need to be considered for each PRO instrument separately and also within the PRO instrument for the different domains or subscales. This paper focusses on how the limitations of the PRO scores can be a starting point to help to define appropriate RD thresholds.

When considering changes in individual patient scores it depends on how the scores are constructed as to whether all integer scores between the minimum and maximum are plausible for a patient to score. If patients are responding to items using discrete Likert responses (such as 'Not at all', 'Very much' etc.) then the score can be constructed by scaling these limited response options up so the minimum score is 0 and maximum score is 100. The underlying measurement scale is not actually continuous though as the scores will be limited by the response options on the Likert scale. The measurement scale will vary according to how many items are summated to obtain a subscale or domain score and also according to the number of options on the response scale. The more items included, or the more response options available to patients, the more continuous the measurement scale will appear. Therefore, a five-item subscale has a wider range of scores between 0 and 100 that are plausible than the range of scores provided by a single item. This means that for an individual patient only certain thresholds for defining a responder will be plausible.

There are various aids to interpretation which are used to quantify the size of difference in PRO scores which would be considered meaningful or beneficial. The minimally important difference (MID) provides a measure of the smallest change in the PRO of interest that patients perceive as important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in management [3]. It is used for the purpose of comparing mean scores between groups and, aligned with this is the minimally important change (MIC) which provides guidance on the smallest level of change over time for a group of patients that is meaningful. Often the MID published for an instrument is assumed to be appropriate also for use as the MIC and RD. There are a number of reasons why these thresholds may need to be different. The underlying concepts are different, a change an individual patient needs to achieve in order for it to be meaningful to them is not the same as how different a mean score needs to be between two groups in order for a treatment benefit to be declared. The way between-group MIDs are derived also lends itself to defining meaningful differences between an average score for a group rather than for individuals over time. The accepted method for deriving MIDs is based on comparing mean scores on a different measure where the meaning is already known (an anchor). This results in a threshold that lies anywhere on the continuous scale between the minimum and maximum score for a subscale. When deriving a RD, the threshold needs to be achievable by an individual patient completing the questionnaire at two points in time. If we use a MID threshold in the place of a RD threshold we may be referring to an integer score that an individual patient cannot achieve by filling in their answers at two points in time. The smallest of these plausible scores that an individual can achieve has been previously referred to as a single 'state change'. Wyrwich et al. [4] highlighted the state change for the SF-36 PRO measure. Note that this refers to individual scores and is not equivalent to the minimum detectable change (MDC) which is defined as the smallest amount of change that is greater than measurement error, based on the confidence interval around the standard error of measurement [5]. The MDC is outside of the scope of this paper but will also be linked to the underlying distribution of the PRO scores.

Using the EORTC QLQ-C30, which is an example of a PRO measure with scores derived from single and multiple Likert scales, we aimed to calculate the state change for each subscale as one of the steps in choosing an appropriate responder threshold and provide a diagram highlighting the other considerations when choosing a RD for each subscale.

Methods

The European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 is a validated, self-rating questionnaire including 30 items (or questions) [6]. The QLQ-C30 includes 15 subscales; One scale for Global Health Status/QoL; five functional scales including Physical Functioning, Role Functioning, Emotional Functioning, Cognitive Functioning, Social Functioning and nine symptom scales comprising of Fatigue, Nausea/Vomiting, Pain, Dyspnoea, Insomnia, Appetite Loss, Constipation, Diarrhoea and Financial Difficulties.

For all Functional and Symptom subscales, items are answered on a 4-point Likert scale ranging from 'Not at all' through to 'Very much', with the exception of the Global Health Status/QoL subscale which has two items using a 7-point Likert scale. Scoring procedures can be found in the EORTC QLQ-C30 Scoring Manual, ver. 3 [7]. All scale scores range from 0 to 100. The score for each subscale is made up of a differing number of items, some are single item scales and the maximum number of items in a subscale is 5, Table 1.

For the Global Health Status/QoL scale and functional subscales, a higher score represents a better health state and for the symptom subscales a lower score represents a better health state.

MS Excel was used to generate all possible scores from all possible combinations of responses to the items within each subscale, based on the QLQ-C30 Scoring Manual, ver. 3 [7]. It was assumed all items had been answered, though

Table 1 EORTC QLQ-C30

	QLQ-C30
Total number of items	30
Subscales	5 functional scales - Physical functioning (5 items) - Role functioning (2 items) - Emotional functioning (4 items) - Cognitive functioning (2 items) - Social functioning (2 items) 9 symptom scales - Fatigue (3 items) - Pain (2 items) - Nausea and vomiting (2 items) - Dyspnoea (1 item) - Appetite loss (1 item) - Insomnia (1 item) - Constipation (1 item) - Diarrhoea (1 item) - Financial difficulties (1 item) A global health status (GHS)/Quality of life (QOL) scale (2 items)
Response scales	4- or 7-point scales (for GHS/QOL items only)
Score range	0–100 (high score = high response) - Functional scales: High score = better functioning - GHS/QOL: High score = better GHS/QOL - Symptom scales: High score = worse symptoms
Recall period	Past week

in practice a subscale score will still be calculated with up to 50% of items missing. These generated scores represent the plausible scores an individual completing the questionnaire could achieve and therefore provide the range of possible values for the individual level of change. Table 2 shows an example of some of the score generation for the global health status/QoL subscale, with two items on a 7-point Likert scale. First a raw score is created as the average of the two responses and then the subscale score is created using the following formula from the scoring manual:-

$$Global\,health\,score = \left\{ \frac{(RawScore - 1)}{6} \right\} \times 100$$

Once all the possible combinations of responses have been generated for each subscale, the range of generated scores were summarised on a graph as the 'plausible' scores for that subscale (Fig. 1).

Table 2 Example score generation for the global health/QoL subscale

Question 29 response	Question 30 response	Raw score	Global health score
1	1	1	0
1	2	1.5	8.33
1	3	2	16.67
1	4	2.5	25.00
1	5	3	33.33
1	6	3.5	41.67
1	7	4	50.00
2	1	1.5	8.33
2	2	2	16.67
2	3	2.5	25.00
2	4	3	33.33
2	5	3.5	41.67
2	6	4	50.00
2	7	4.5	58.33
…repeated for 3 up to 7			

Results

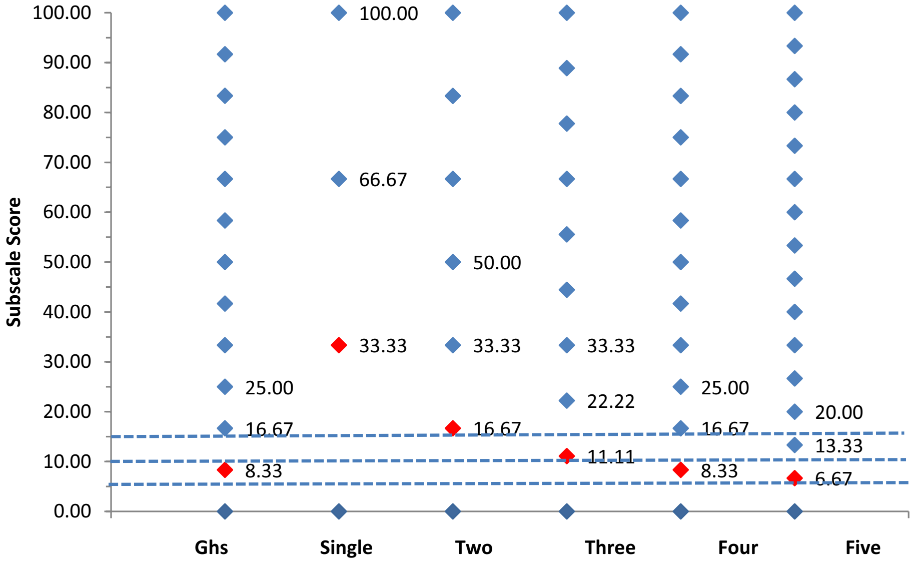
Plausible thresholds

The minimum change an individual can achieve on the QLQ-C30 subscales range from ± 6.7 (Physical Functioning scale comprising 5 items) to ± 33.3 (single item scales), Fig. 1. These minimum changes can be achieved when one item in the subscale changes by one category on the Likert-response scale assuming all items are answered and all other items in the scale have remained the same. For example, if a patient responded to the single item for Insomnia with 'Not at all' at baseline and then four weeks later responded to the same item with 'A little' then their score would have increased from 0 to 33.3. If the patient had responded similarly to the two pain items with 'Not at all' for both at baseline and then four weeks later had 'A little' pain on one item and remained at 'Not at all' for the other item then their pain score would increase from 0 to 16.67.

Key: GHS/QoL—Global health status/Quality of life subscale, Single—single-item subscales (Dyspnoea, Insomnia, Appetite loss, Constipation, Diarrhoea, Financial Difficulties), Two—two-item subscales (Role Functioning, Cognitive Functioning, Social Functioning, Nausea and Vomiting, Pain), Three—three-item scale (Fatigue), Four—four-item scale (Emotional Functioning), Five—five-item scale (Physical Functioning). Dashed lines show commonly used thresholds (5, 10, 15).

Figure 1 highlights some of the commonly used thresholds, e.g. 5 points, 10 points and 15 points. These are generally based on thresholds that were originally estimated

Fig. 1 Possible scores for EORTC QLQ-C30 subscales



for group-level analyses. It is good practice to pre-specify a RD threshold when planning the study but to also include a sensitivity analysis which uses a different, normally a larger threshold to represent a more stringent hurdle. This way there is a check on how much the choice of threshold has influenced a result and whether treatment differences hold across different choices of threshold. Figure 1 shows that choosing, for example, a 10-point responder threshold with a 15-point threshold for a sensitivity analysis is not appropriate for the Global Health Status/QoL scale, single-item scales, two-item scales and four-item scales. The figure shows there are no plausible values for a patient between these lines, therefore if you set the threshold anywhere in that range you will identify the same number of responders and will not be conducting a true test of choosing a different threshold. Using the global health status/QoL as an example again, a 10-point responder threshold will not identify patients with the minimum change (a 1 response category change on one of the items) since their score would be 8.33. It would identify anyone with a change of 2 points on the response scale on one of the items and no change on the other item, or a change on both items in the same direction by one point on the Likert scale, since either of these scenarios would result in a 16.67 point change. Therefore, for this scale, valid thresholds for the main analysis and sensitivity analyses, respectively, may be 5 points and 10 points, 10 points and 20 points or 5 and 20 points depending on whether a change in only one of the items by one category is considered sufficient to indicate a meaningful change on the scale.

Similarly, if 5 points was chosen as the response definition with a 10-point threshold for the sensitivity analysis this would not be appropriate for single-, two- and three-item scales (since no values between 5 and 10 are possible).

Figure 1 highlights that for all subscales, a responder definition of 5 points would consider any patient who had the minimal change in only one item as a responder.

Table 3 shows the plausible choice of responder definitions for each subscale based on the achievable scores for an individual patient on that scale. Numbers provided represent the minimal change and the next two change increments as more stringent estimates of a response.

Since the actual scores can involve recurring decimal places, we recommend to round the choice of threshold down to the nearest 5 points from the exact score in order to capture the required patients as responders. These represent the same thresholds since no scores are possible within 5 points.

Conclusions

Defining responders based on PRO scales enables comparison between treatments with respect to the proportion of patients achieving a PRO response and the time until deterioration or improvement in PRO. The nature of these scales means that the definition of a responder requires careful consideration. Thresholds used to indicate a meaningful difference between groups may be available for an instrument but may not be directly applicable for change over time experienced by an individual patient and should not be automatically applied to define responders. We have observed that it is common to use estimates of the MID or a percentage of the scale to define responders, with 5, 10 (10%) and 15 (15%) points commonly quoted for responder analyses [8–13] with the EORTC QLQ-C30, with reliance on methods papers from 2005 to 2007 [14–16]. Further, a recent paper [17] used a threshold per subscale based on

Table 3 EORTC QLQ-C30 state changes

Subscale	Items	Response scale (Likert)	Increments in individual patient change scores[a]	Possible responder definitions (RDs)		
				A	B	C
				RD (minimal change[b])	RD (minimal change + 1 increment[c])	RD (minimal change + 2 increments)
GHS/QoL	Two	7	8.3	≥ +5	≥ +15[d]	≥ +25
Physical functioning	Five	4	6.7	≥ +5	≥ +10	≥ +15
Emotional functioning	Four	4	8.3	≥ +5	≥ +15[d]	≥ +25
Fatigue	Three	4	11.1	≤ − 10	≤ − 20	≤ − 30
Role functioning, cognitive functioning, social functioning	Two	4	16.7	≥ +15	≥30	≥50
Nausea and vomiting, pain	Two	4	16.7	≤ − 15	≤ − 30	≤ − 50
Dyspnoea, insomnia, appetite loss, constipation, diarrhoea, financial difficulties	Single	4	33.3	≤ − 30	≤ − 65	≤ − 100

GHS/QoL Global health status/quality of life

[a] Increments refer to the change in score when a patient moves a single response category on the Likert scale on one item in the scale. For example, moving from 'A little' to 'Quite a bit' on one item in the physical functioning scale would mean a change in the patient's score of 6.7 points. Increments have been rounded to 1 decimal place

[b] Minimal change is the suggested threshold to capture any change in score, i.e. the smallest increment. Thresholds are rounded down to the nearest 5 points

[c] Minimal change plus one increment is the suggested threshold to capture a change in score of two increments. This may be achieved by a patient moving two response categories on the Likert scale for one item, e.g. 'Not at all' to 'Quite a bit', or moving to an adjacent response category on two of the items in a scale. Thresholds are rounded down to the nearest 5 points

[d] For these subscales a 10-point threshold is the same as a 15-point threshold. We encourage rounding down to the nearest 5 for transparency and consistency

mean differences from Cocks et al. 2011 [18]. Use of the MID in this way for an individual threshold may be misleading and may not be meaningful on an individual patient basis. Consider an analysis of the GHS/Qol scale using 10 points as a threshold in the main analysis and 15 points for a sensitivity analysis. This is common practice given the uncertainty around estimation of thresholds, to check robustness of results when using a higher threshold in a sensitivity analysis. However, these analyses would give identical results, not because the initial analysis is robust to the higher threshold but because both thresholds would define the same number of responders since patients can only change by 8.3-point increments (8.3, 16.6 and so on). The 10- and 15-point thresholds are exactly the same thresholds for this scale, they would only define a responder if the 16.6 score change or higher occurred.

Consideration of the discrete nature of the PRO subscale scores and minimum state changes must therefore be accounted for in order to appropriately define the level of change that is meaningful at an individual level. Moreover, care should be taken to use RDs for sensitivity analyses that are not essentially equivalent. It is common to conduct initial analyses with a pre-defined responder definition and then conduct a sensitivity analysis using a larger threshold to check for robustness. Analyses that have considered these thresholds as equivalent will result in misinterpretation of responder analysis being confirmed as robust when in fact the main analysis and sensitivity analysis for the subscale has simply been duplicated.

We have used the EORTC QLQ-C30 instrument as an example but the same will apply to any instrument where the underlying scale structure is discrete and responder analyses are being contemplated, for example, the SF-36 [19]. Figure 2 highlights the consideration of this 'state change' or 'plausible scores' as one of the steps to defining appropriate thresholds. Alongside the state change, the top row highlights using existing qualitative and quantitative evidence to guide choice of responder threshold, provided these were derived specifically for individual patient change. It is important to start to align on responder thresholds across studies where possible so that results can be comparable across different studies. This is with the caveat thought that previous studies have used an appropriate estimate for an individual patient change. The bottom row in the figure highlights other aspects of the study that are relevant for the choice of responder threshold including the disease, population, treatments, toxicity and the study design and hypotheses. For example, responder thresholds for a physical functioning endpoint may be different in a study of a fit and healthy population where the goal is full rehabilitation

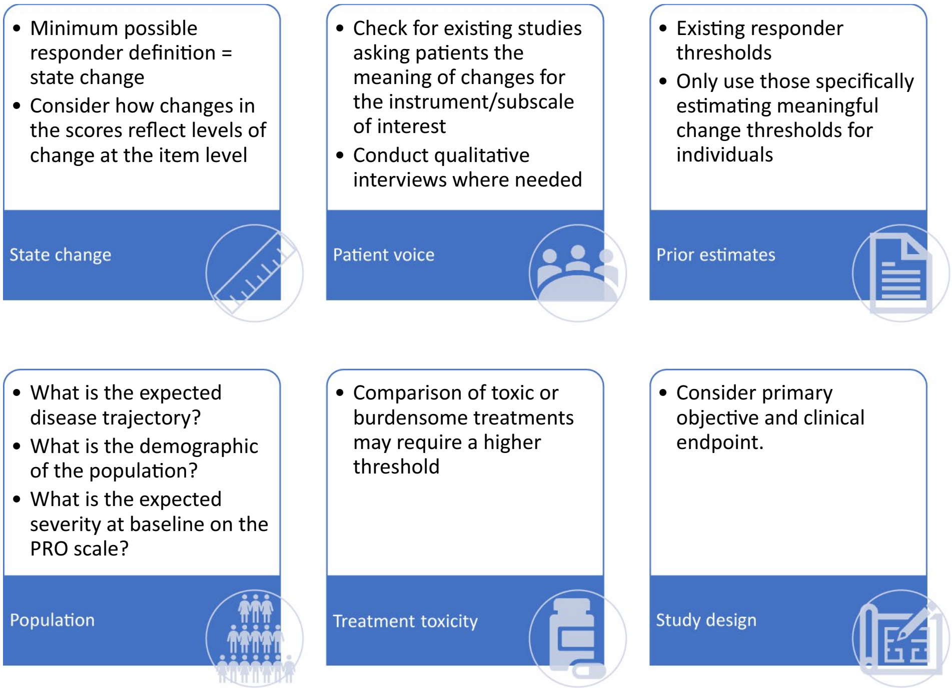

• Minimum possible responder definition = state change
• Consider how changes in the scores reflect levels of change at the item level
State change
• Check for existing studies asking patients the meaning of changes for the instrument/subscale of interest
• Conduct qualitative interviews where needed
Patient voice
• Existing responder thresholds
• Only use those specifically estimating meaningful change thresholds for individuals
Prior estimates
• What is the expected disease trajectory?
• What is the demographic of the population?
• What is the expected severity at baseline on the PRO scale?
Population
• Comparison of toxic or burdensome treatments may require a higher threshold
Treatment toxicity
• Consider primary objective and clinical endpoint.
Study design

Fig. 2 Considerations for choosing responder thresholds

following an injury compared to a study treating patients with a chronic progressive disease. The starting point on the PRO scale for a population in a study is important to consider, and this may depend on the demographic as well as expected disease trajectory. The treatment goal is also important context to consider. A study with a palliative goal may use the same PRO scale as a study with a curative aim but the changes patients view as meaningful are likely to vary in these two settings.

For multi-item scales in the QLQ-C30 we recommend, in the absence of any other information, using a responder definition that represents a larger change than simply one item changing by one response category on the Likert scale (column B or C in Table 1). The minimum change could be used as a sensitivity analysis (column A) or a larger change as deemed appropriate based on the considerations highlighted in Fig. 2. The smallest possible change may be justifiable but the threshold should not be smaller than this minimal change. The purpose here is to highlight the issues with assuming a between-group MID is appropriate for use as a responder definition. Moreover, choosing a global responder definition across subscales is not recommended given individual change scores have very different meanings across the subscales.

One limitation in our score generation is the assumption that all items are present. In practice, a subscale score can still be generated as long as at least 50% of items are answered. Typically less than 2% of patients [7] are expected to have missing items and with more PROs being administered electronically this may decrease further, so this should have minimal impact on the generated thresholds.

Further work is required to establish the sizes of changes that are meaningful to patients on the EORTC QLQ-C30 and for other similar instruments. This work could utilise patient's global ratings of change scores which are often used to determine between-group MIDs but capture individual patient views about their own change in PRO score so are directly aligned with the threshold required for responder analyses. Techniques for qualitative patient interviews are also being developed to aid definition of individual-patient thresholds and have been trialled for EORTC instruments recently too [20]. Alongside consideration of what individual score changes can be achieved these additional methods seek to identify the meaning of any changes directly from the patient, which will improve the credibility and impact of these responder analyses.

Acknowledgements None.

Author contributions Both authors contributed equally to this work and were involved in drafting the manuscript. Both authors read and approved the final manuscript.

Funding Not applicable.

Data availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Code availability The EORTC scoring manual is available from the EORTC website.

Declarations

Competing interests JB was an employee of Amgen, Inc. and owns Amgen stock. Amgen is a commercial organisation that sponsors clinical trials. KC is a statistical consultant and received funding from Amgen, Inc. for this work.

Ethical approval Not Applicable.

Informed Consent Not applicable.

Consent for publication Not applicable.

References

1. Food and Durg Administration. (2009). Guidance for industry: Patient-reported outcome measures: use in medical product development to support labeling claims. *Federal Register, 74*(235), 65132–65133.
2. Anota, A., Hamidou, Z., Paget-Bailly, S., et al. (2015). Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: Do we need RECIST for quality of life to achieve standardization? *Quality of Life Research, 24*(1), 5–18.
3. Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials, 10*(4), 407–415.
4. Wyrwich, K. W., Tierney, W. M., Babu, A. N., Kroenke, K., & Wolinsky, F. D. (2005). A comparison of clinically important differences in health-related quality of life for patients with chronic lung disease, asthma, or heart disease. *Health Services Research, 40*(2), 577–592.
5. Beckerman, H., Roebroeck, M. E., Lankhorst, G. J., Becher, J. G., Bezemer, P. D., & Verbeek, A. L. (2001). Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research., 10*(7), 571–578.
6. Aaronson, N. K., Ahmedzai, S., Bergman, B., et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute., 85*(5), 365–376.
7. Fayers, P., Aaronson, N., Bjordal, K., Groenvold, M., Curran, D., & Bottomley, A. (2001). *The EORTC QLQ-C30 scoring manual* (3rd ed.). European Organization for Research and Treatment of Cancer.
8. Lenz, H.-J., Argiles, G., Yoshino, T., et al. (2019). Health-related quality of life in the Phase III LUME-Colon 1 study: Comparison and interpretation of results from EORTC QLQ-C30 analyses. *Clinical Colorectal Cancer, 18*, 269–279.
9. Anota, A., Mouillet, G., Trouilloud, I., et al. (2015). Sequential FOLFIRI 3+ Gemcitabine improves health-related quality of life deterioration-free survival of patients with metastatic pancreatic adenocarcinoma: a randomized phase II trial. *PLoS ONE, 10*(5), e0125350.
10. Stockler, M. R., Hilpert, F., Friedlander, M., et al. (2014). Patient-reported outcome results from the open-label phase III AURELIA trial evaluating bevacizumab-containing therapy for platinum-resistant ovarian cancer. *Journal of Clinical Oncology., 32*(13), 1309.
11. Harrison, C. N., Mesa, R. A., Kiladjian, J. J., et al. (2013). Health-related quality of life and symptoms in patients with myelofibrosis treated with ruxolitinib versus best available therapy. *British Journal of Haematology, 162*(2), 229–239.
12. Bonnetain, F., Dahan, L., Maillard, E., et al. (2010). Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. *European Journal of Cancer., 46*(15), 2753–2762.
13. Eisenhardt, A., Schneider, T., Scheithe, K., Colling, C., & Heidenreich, A. (2015). Health-related quality of life in patients with advanced prostate cancer undergoing treatment with TRIPTOrelin Pamoate SIX month formulation: Results of the non-interventional TRIPTOSIX study. *Journal of Clinical Oncology, 33*, 287.
14. Osoba, D., Bezjak, A., Brundage, M., et al. (2005). Analysis and interpretation of health-related quality-of-life data from clinical trials: Basic approach of The National Cancer Institute of Canada Clinical Trials Group. *European Journal of Cancer., 41*(2), 280–287.
15. Osoba, D., Bezjak, A., Brundage, M., Pater, J., National Cancer Institute of Canada Clinical Trials Group. (2007). Evaluating health-related quality of life in cancer clinical trials: The National Cancer Institute of Canada Clinical Trials Group experience. *Value Health., 10*, S138–S145.
16. Brundage, M., Osoba, D., Bezjak, A., Tu, D., Palmer, M., & Pater, J. (2007). Lessons learned in the assessment of health-related quality of life: Selected examples from the National Cancer Institute of Canada Clinical Trials Group. *Journal of Clinical Oncology., 25*(32), 5078–5081.
17. Kawahara, T., Shimozuma, K., Shiroiwa, T., et al. (2018). Patient-reported outcome results from the open-label randomized phase III select bc trial evaluating first-line s-1 therapy for metastatic breast cancer. *Oncology, 94*(2), 107–115.
18. Cocks, K., King, M. T., Velikova, G., Martyn St-James, M., Fayers, P. M., & Brown, J. M. (2011). 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. *Journal of Clinical Oncology., 29*(1), 89–96.
19. SF-36 scoring. Retrieved from https://www.rand.org/health-care/surveys_tools/mos/36-item-short-form/scoring.html.
20. Sully, K., Trigg, A., Bonner, N., Moreno-Koehler, A., Trennery, C., Shah, N., Yucel, E., Panjabi, S., & Cocks, K. (2019). Estimation of minimally important differences and responder definitions for EORTC QLQ-MY20 scores in multiple myeloma patients. *European Journal of Haematology, 103*(5), 500–509. https://doi.org/10.1111/ejh.13316