



UNIVERSITY OF LEEDS

This is a repository copy of *International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/161154/>

Version: Supplemental Material

Article:

Coens, C, Pe, M, Dueck, AC et al. (32 more authors) (2020) International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium. *The Lancet Oncology*, 21 (2). e83-e96. ISSN 1470-2045

[https://doi.org/10.1016/s1470-2045\(19\)30790-9](https://doi.org/10.1016/s1470-2045(19)30790-9)

© 2020, Elsevier. All rights reserved. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



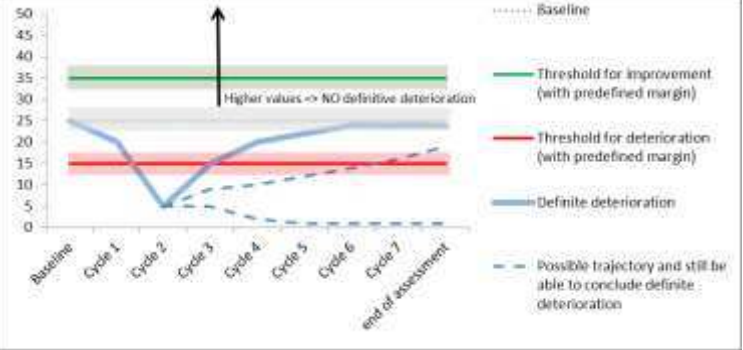
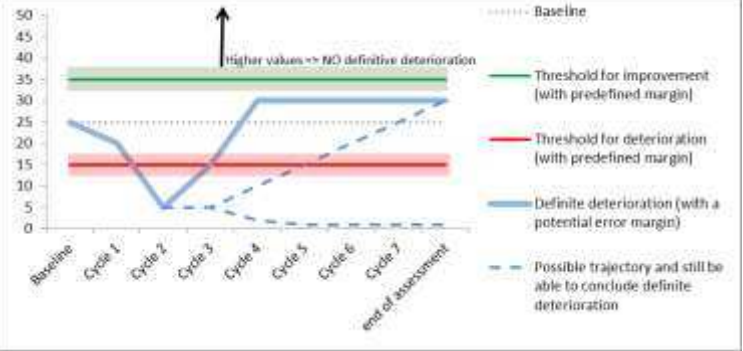
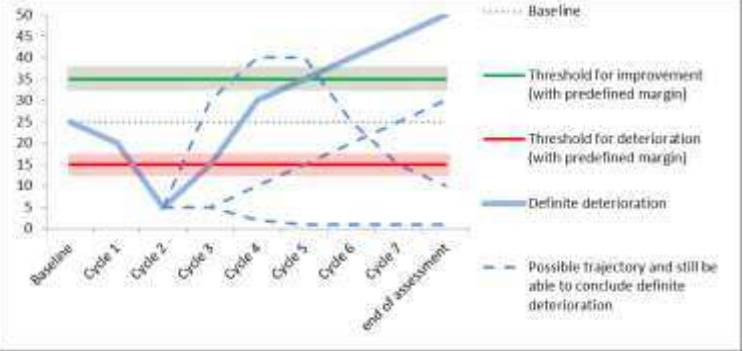
eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Appendix 2

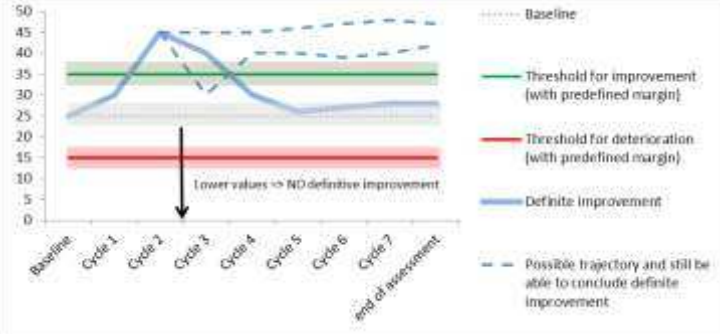
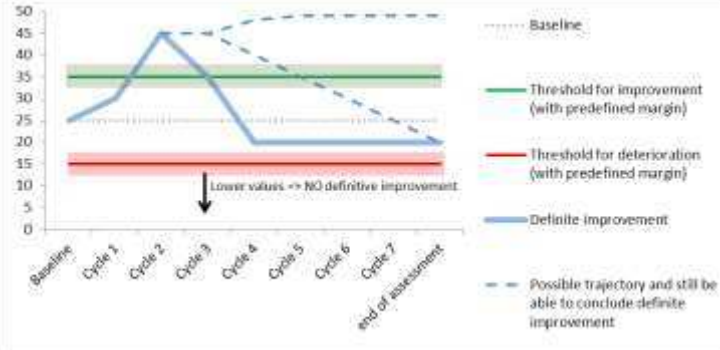
Table 1. Research objectives Working Group Survey Results on Standardizing Definitions of Improvement, Maintenance (or Stable State) and Deterioration (or Worsening)

(N = 26).

Definition	Graphic Visualization	Primary Scoring (% agree ¹)
1. Definitive deterioration		
<ul style="list-style-type: none"> • Post-baseline deterioration • After the post-baseline deterioration: <ul style="list-style-type: none"> • no follow-up scores are higher than one's own deterioration level (or its pre-defined margin); • no follow-up scores are higher than the deterioration threshold (or its pre-defined margin); • no follow-up scores are higher than one's own baseline level (or its predefined margin) • no follow-up scores are higher than the improvement threshold (or its pre-defined margin) 		22 (85%)
<ul style="list-style-type: none"> • Post-baseline deterioration • After the post-baseline deterioration: <ul style="list-style-type: none"> • follow-up scores may be higher than one's own deterioration level (or its pre-defined margin); • no follow-up scores are higher than the deterioration threshold (or its pre-defined margin); • no follow-up scores are higher than one's own baseline level (or its predefined margin) 		21 (81%)*

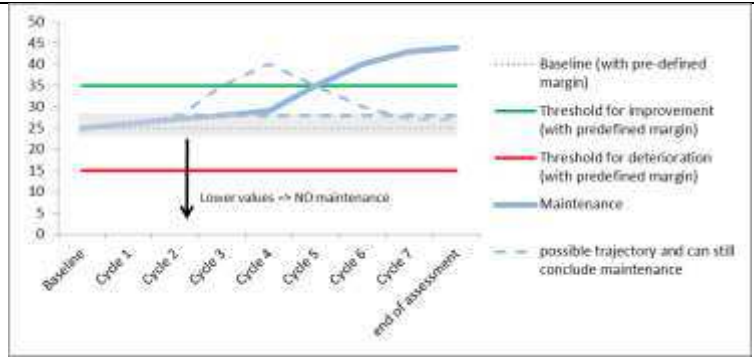
<ul style="list-style-type: none"> no follow-up scores are higher than the improvement threshold (or its pre-defined margin) 		
<ul style="list-style-type: none"> Post-baseline deterioration After the post-baseline deterioration: <ul style="list-style-type: none"> follow-up scores may be higher than one's own deterioration level (or its pre-defined margin); follow-up scores may be higher than the deterioration threshold (or its pre-defined margin); no follow-up scores are higher than one's own baseline level (or its predefined margin) no follow-up scores are higher than the improvement threshold (or its pre-defined margin) 	 <p>The graph plots a score over eight time points: Baseline, Cycle 1, Cycle 2, Cycle 3, Cycle 4, Cycle 5, Cycle 6, and Cycle 7 (end of assessment). The y-axis ranges from 0 to 50. A dotted line represents the Baseline at approximately 25. A solid green line at 35 is the 'Threshold for improvement (with predefined margin)'. A solid red line at 15 is the 'Threshold for deterioration (with predefined margin)'. A solid blue line shows the 'Definite deterioration', starting at 25, dipping to 5 at Cycle 2, and then rising to 25 by Cycle 7. A dashed blue line shows a 'Possible trajectory and still be able to conclude definite deterioration', starting at 25, dipping to 5 at Cycle 2, and then rising to 18 by Cycle 7. An arrow points to the top of the graph with the text 'Higher values => NO definitive deterioration'.</p>	4 (8%)
<ul style="list-style-type: none"> Post-baseline deterioration After the post-baseline deterioration: <ul style="list-style-type: none"> follow-up scores may be higher than one's own deterioration level (or its predefined margin); follow-up scores may be higher than the deterioration threshold (or its predefined margin); follow-up scores may be higher than one's own baseline level (or its predefined margin) no follow-up scores are higher than the improvement threshold (or its predefined margin) 	 <p>The graph plots a score over eight time points: Baseline, Cycle 1, Cycle 2, Cycle 3, Cycle 4, Cycle 5, Cycle 6, and Cycle 7 (end of assessment). The y-axis ranges from 0 to 50. A dotted line represents the Baseline at approximately 25. A solid green line at 35 is the 'Threshold for improvement (with predefined margin)'. A solid red line at 15 is the 'Threshold for deterioration (with predefined margin)'. A solid blue line shows the 'Definite deterioration (with a potential error margin)', starting at 25, dipping to 5 at Cycle 2, and then rising to 30 by Cycle 7. A dashed blue line shows a 'Possible trajectory and still be able to conclude definite deterioration', starting at 25, dipping to 5 at Cycle 2, and then rising to 25 by Cycle 7. An arrow points to the top of the graph with the text 'Higher values => NO definitive deterioration'.</p>	1 (4%)
<ul style="list-style-type: none"> Post-baseline deterioration After the post-baseline deterioration: <ul style="list-style-type: none"> follow-up scores may be higher than one's own deterioration level (or its pre-defined margin); follow-up scores may be higher than the deterioration threshold (or its pre-defined margin); follow-up scores may be higher than one's own baseline level (or its predefined margin) 	 <p>The graph plots a score over eight time points: Baseline, Cycle 1, Cycle 2, Cycle 3, Cycle 4, Cycle 5, Cycle 6, and Cycle 7 (end of assessment). The y-axis ranges from 0 to 50. A dotted line represents the Baseline at approximately 25. A solid green line at 35 is the 'Threshold for improvement (with predefined margin)'. A solid red line at 15 is the 'Threshold for deterioration (with predefined margin)'. A solid blue line shows the 'Definite deterioration', starting at 25, dipping to 5 at Cycle 2, and then rising to 50 by Cycle 7. A dashed blue line shows a 'Possible trajectory and still be able to conclude definite deterioration', starting at 25, dipping to 5 at Cycle 2, and then rising to 40 by Cycle 7. An arrow points to the top of the graph with the text 'Higher values => NO definitive deterioration'.</p>	1 (4%)

<ul style="list-style-type: none"> • follow-up scores may be higher than the improvement threshold (or its pre-defined margin) 		
<ul style="list-style-type: none"> • Post-baseline deterioration 		<p>1 (4%)</p>
<p>2. Definitive improvement</p>		
<ul style="list-style-type: none"> • Post-baseline improvement • After the post-baseline improvement: <ul style="list-style-type: none"> • no follow-up scores are lower than one's own improvement level (or its pre-defined margin); • no follow-up scores are lower than the improvement threshold (or its pre-defined margin); • no follow-up scores are lower than one's own baseline level (or its predefined margin) • no follow-up scores are lower than the deterioration threshold (or its pre-defined margin) 		<p>21 (81%)</p>
<ul style="list-style-type: none"> • Post-baseline improvement • After the post-baseline improvement: <ul style="list-style-type: none"> • follow-up scores may be lower than one's own improvement level (or its pre-defined margin); • no follow-up scores are lower than the improvement threshold (or its pre-defined margin); • no follow-up scores are lower than one's own baseline level (or its predefined margin) 		<p>22 (85%)*</p>

<ul style="list-style-type: none"> no follow-up scores are lower than the deterioration threshold (or its pre-defined margin) 		
<ul style="list-style-type: none"> Post-baseline improvement After the post-baseline improvement: <ul style="list-style-type: none"> follow-up scores may be lower than one's own improvement level (or its pre-defined margin); follow-up scores may be lower than the improvement threshold (or its pre-defined margin); no follow-up scores are lower than one's own baseline level (or its predefined margin) no follow-up scores are lower than the deterioration threshold (or its pre-defined margin) 	 <p>The graph plots a score over time from Baseline to Cycle 7, ending at 'end of assessment'. The y-axis ranges from 0 to 50. A dotted line represents the 'Baseline' at approximately 25. A solid blue line shows the 'Definite improvement' trajectory, which rises to a peak of 45 at Cycle 2, then falls to about 25 by Cycle 4 and remains there. A green horizontal band between 35 and 40 represents the 'Threshold for improvement (with predefined margin)'. A red horizontal band between 15 and 20 represents the 'Threshold for deterioration (with predefined margin)'. A dashed blue line shows a 'Possible trajectory and still be able to conclude definite improvement', which peaks at 45 at Cycle 2 and then gradually declines to about 40 by Cycle 7. An arrow points to the score at Cycle 3 with the text 'Lower values => NO definite improvement'.</p>	<p>6 (23%)</p>
<ul style="list-style-type: none"> Post-baseline improvement After the post-baseline improvement: <ul style="list-style-type: none"> follow-up scores may be lower than one's own improvement level (or its pre-defined margin); follow-up scores may be lower than the improvement threshold (or its pre-defined margin); follow-up scores may be lower than one's own baseline level (or its predefined margin) no follow-up scores are lower than the deterioration threshold (or its pre-defined margin) 	 <p>The graph plots a score over time from Baseline to Cycle 7, ending at 'end of assessment'. The y-axis ranges from 0 to 50. A dotted line represents the 'Baseline' at approximately 25. A solid blue line shows the 'Definite improvement' trajectory, which rises to a peak of 45 at Cycle 2, then falls to about 20 by Cycle 4 and remains there. A green horizontal band between 35 and 40 represents the 'Threshold for improvement (with predefined margin)'. A red horizontal band between 15 and 20 represents the 'Threshold for deterioration (with predefined margin)'. A dashed blue line shows a 'Possible trajectory and still be able to conclude definite improvement', which peaks at 45 at Cycle 2 and then gradually declines to about 40 by Cycle 7. An arrow points to the score at Cycle 3 with the text 'Lower values => NO definite improvement'.</p>	<p>2 (8%)</p>

<ul style="list-style-type: none"> • Post-baseline improvement • After the post-baseline improvement: <ul style="list-style-type: none"> • follow-up scores may be lower than one's own improvement level (or its pre-defined margin); • follow-up scores may be lower than the improvement threshold (or its pre-defined margin); • follow-up scores may be lower than one's own baseline level (or its predefined margin) • follow-up scores may be lower than the deterioration threshold (or its pre-defined margin) 	<p>The graph shows a score starting at a baseline of approximately 25. It rises to about 45 by Cycle 2, then drops to 35 at Cycle 3, 20 at Cycle 4, 15 at Cycle 5, and ends at 10 at the end of assessment. The improvement threshold is at 35, and the deterioration threshold is at 15. A dashed line shows a possible trajectory that stays above the improvement threshold.</p>	<p>1 (4%)</p>
<ul style="list-style-type: none"> • Post-baseline improvement 	<p>The graph shows a score starting at a baseline of approximately 25. It rises to about 45 by Cycle 2, then drops to 35 at Cycle 3, 15 at Cycle 4, 10 at Cycle 5, and ends at 40 at the end of assessment. The improvement threshold is at 35, and the deterioration threshold is at 15. A dashed line shows a possible trajectory that stays above the improvement threshold. An arrow labeled 'Irrelevant information' points to the end of the assessment.</p>	<p>2 (8%)</p>
<p>3. Maintenance</p>		
<ul style="list-style-type: none"> • Follow-up scores are similar to baseline score (by a pre-defined margin) <ul style="list-style-type: none"> • No follow-up scores are better than the baseline score. • No follow-up scores are worse than the baseline score. 	<p>The graph shows a score starting at a baseline of approximately 25. It remains stable around 25 throughout the assessment, ending at 25. The improvement threshold is at 35, and the deterioration threshold is at 15. A dashed line shows a possible trajectory that stays between the improvement and deterioration thresholds. Arrows indicate that higher values imply no maintenance and lower values imply no maintenance.</p>	<p>23 (88%)*</p>

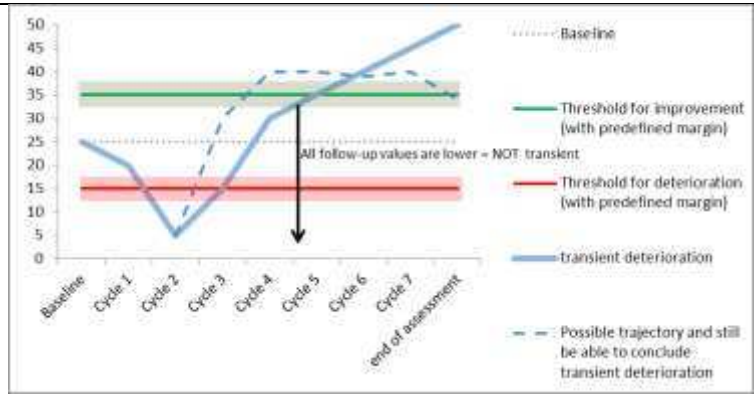
- Follow-up scores are not worse than the baseline score (by a pre-defined margin)
 - Follow-up scores may be better than baseline score.
 - No follow-scores are worse than the baseline score.



13 (50%)**

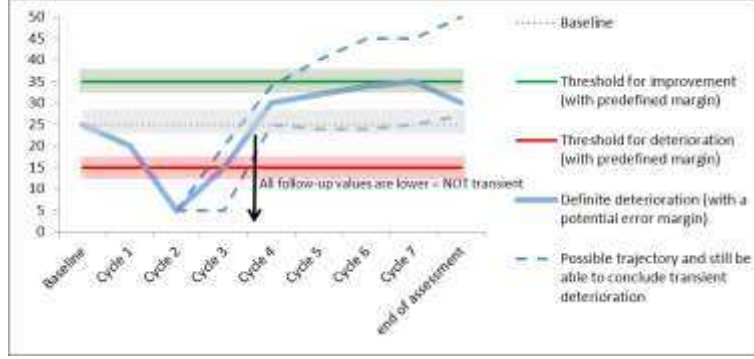
4. Transient deterioration

- Post-baseline deterioration
- After the post-baseline deterioration, there is an increase in scores:
 - At least one follow-up score should be higher than or be at the level of the improvement threshold (or its pre-defined margin).



19 (73%)

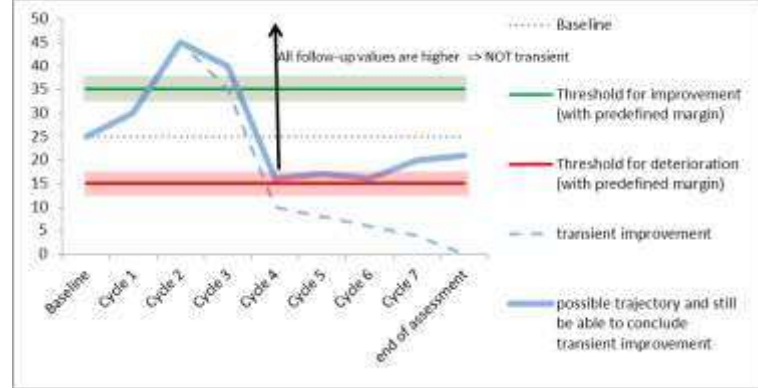
- Post-baseline deterioration
- After the post-baseline deterioration, there is an increase in scores:
 - At least one follow-up score should be higher than or at least be at the baseline level (or its pre-defined margin).



21 (81%)*

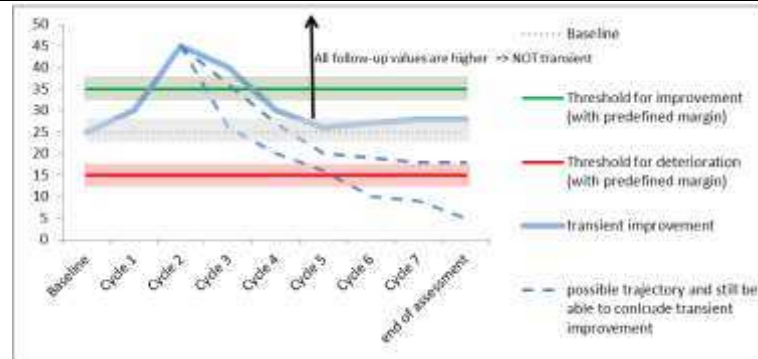
<ul style="list-style-type: none"> • Post-baseline deterioration • After the post-baseline deterioration, there is an <u>increase</u> in scores: <ul style="list-style-type: none"> ○ At least one follow-up score should be higher than or at least be at the deterioration threshold (or its pre-defined margin). 		11 (43%)
<ul style="list-style-type: none"> • Post-baseline deterioration • After the post-baseline deterioration, there is an <u>increase</u> in scores: <ul style="list-style-type: none"> ○ At least one follow-up score should be higher than or at least be at the deterioration level (or its pre-defined margin). 		3 (12%)
<ul style="list-style-type: none"> • Post-baseline deterioration 		2 (8%)
<p>5. Transient improvement</p>		

- Post-baseline improvement
- After the post-baseline improvement, there is a decrease in scores:
 - At least one follow-up score should be lower than or at least be at the deterioration threshold (or its pre-defined margin).



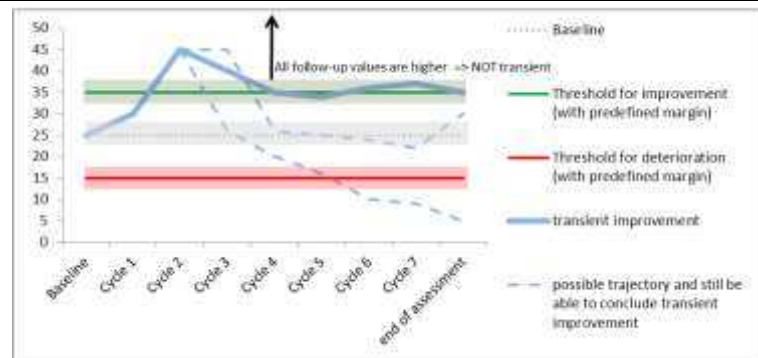
19 (73%)

- Post-baseline improvement
- After the post-baseline improvement, there is a decrease in scores:
 - At least one follow-up score should be lower than or at least be at the baseline level (or its pre-defined margin).



23 (88%)*

- Post-baseline improvement
- After the post-baseline improvement, there is a decrease in scores:
 - At least one follow-up score should be lower than or at least be at the improvement threshold (or its pre-defined margin).



12 (46%)

<ul style="list-style-type: none"> • Post-baseline improvement • After the post-baseline improvement, there is a <u>decrease</u> in scores: <ul style="list-style-type: none"> ○ At least one follow-up score should be lower than or at least be at the improvement level (or its pre-defined margin). 		<p>5 (19%)</p>
<ul style="list-style-type: none"> • Post-baseline improvement 		<p>2 (8%)</p>

Note. Maintenance was the original term used for stable state; and deterioration was the original term used for worsening.

¹Primary scoring decision rule: Accept as soon as >/70% respondents rated “(completely) agree” (rating 4 or 5) AND </ 15% votes “(completely) disagree” (rating 1 or 2). Reject as soon as >/30% votes “(completely) disagree” (rating 1 or 2). When 2 or more options received a >/70% agreement, they were discussed and a final decision was agreed upon during a WebEx meeting; the less strict definition was usually chosen. For maintenance, it was agreed during discussions that both definitions of maintenance are needed.

*Agreed definition by the research objectives working group.

**The first definition remains the primary definition of maintenance, but the second definition (i.e., the definition of maintenance is combined with improvement) can be applied in exceptional cases.

Table 2. Statistical Methods Working Group Survey Results on Essential Statistical Features for Patient Reported Outcome Analysis (N = 16).

Code	Statistical feature	Considerations	Primary Scoring ¹ (% essential)	Secondary Scoring ²	Rationale for the scoring (summarized comments from WG members)
Essential / highly desirable statistical features					
S1	Compare 2 treatment arms	The ability of the model to perform a statistical test between two samples.	16 (100%)	40	<ul style="list-style-type: none"> <input type="checkbox"/> Comparing groups is the main goal of an RCT <input type="checkbox"/> To compare groups, a statistical test is needed.
S5	Adjust for baseline score	The ability to include the baseline assessment in the model either as a covariate or as the first of repeated measures.	14 (88%)	29	<ul style="list-style-type: none"> <input type="checkbox"/> Although randomization should take care of the confounding factors, there is still a need to stratify or correct for baseline variables for the primary outcome <input type="checkbox"/> It provides a more accurate estimate of the treatment effect.
S16	Be clinically relevant	The ability of the model to produce results that guide informative clinical-decision making and influence clinical practice. This means the ability of the model to produce results on the size, certainty, and direction of the estimate and precision of the treatment effect (point estimate, confidence interval and error margin) that has a direct link with the clinical relevance classification of the PRO instrument.	13 (81%)	36	<ul style="list-style-type: none"> <input type="checkbox"/> Essential for proper interpretation of results
S3	Allow for confounding factors	The ability of the model to include baseline covariates that are believed to be associated with the outcome variable or compliance. Covariates can be: - Demographic variables: age, gender,...	12 (75%)	32	<ul style="list-style-type: none"> <input type="checkbox"/> Although randomization should take care of the confounding factors, there is still a need to stratify or correct for baseline variables for the primary outcome

		- Disease characteristics: duration, stage,... - Others: country, center, investigator,.			<input type="checkbox"/> It provides a more accurate estimate of the treatment effect.
S6	Handle missing data (Part I)	The ability of the model to deal with missing data due to non-compliance. Thereby, we mean a method that allows for incomplete data, i.e. a method that makes the least restrictive assumptions about their relationship with missing data.	11 (69%)	26	<input type="checkbox"/> Missing data is a problem in PRO analysis. <input type="checkbox"/> Model should allow for incomplete data (that makes the least restrictive assumptions about missingness).
S9	Handle clustered data (Part I – over time)	The ability of the model to allow for correlations over time (longitudinal repeated assessment within the same patient)	11 (69%)	25	<input type="checkbox"/> PRO data is often longitudinal and this should be reflected in the analysis method <input type="checkbox"/> Essential in the case of a longitudinal study objective (e.g., comparing means over time) <input type="checkbox"/> Not essential for time to event objectives
Other statistical features that did not meet the essential / highly desirable criteria					
S2	Compare more than 2 treatment arms	The ability of the model to perform a statistical test between more than two samples in an integrated test	9 (56%)	9	<input type="checkbox"/> Only needed if the trial hypothesis calls for an integrated test <input type="checkbox"/> It is more efficient but not essential. Similar to other clinical endpoints, several independent tests may be considered (with error correction)
S13	Handle unbalanced designs (Part II)	The ability of the model to handle situations where the schedule of assessment is planned to be different over patients because the assessment time is dependent on a certain event in an individual (e.g. 3-weekly vs 4-weekly assessment schedule due to treatment cycles)	9 (56%)	14	<input type="checkbox"/> This should have already been taken into account during the trial design rather than requiring the analysis to handle it.

S15	Calculate sample size	The ability of the model to reliably calculate sample size and perform a post-hoc power calculation	8 (50%)	8	<ul style="list-style-type: none"> <input type="checkbox"/> The preference is in using an analysis model that fits the trial design rather than whether it can calculate sample size. Sample size can be based on a simpler model with fewer assumptions. <input type="checkbox"/> Simulations can help provide sample size calculations
S12	Handle unbalanced designs (Part I)	The ability of the model to handle situations where the schedule of assessment is planned to be different over the treatment arms for practical reasons (e.g. 3-weekly vs 4-weekly assessment schedule due to treatment cycles)	7 (44%)	10	<ul style="list-style-type: none"> <input type="checkbox"/> This should have already been taken into account during the trial design rather than requiring the analysis to handle it.
S17	Robustness	The ability of the statistical procedure to be not overly dependent on critical assumptions regarding: <ul style="list-style-type: none"> a) an underlying parameter distribution (e.g. normality) b) a structural relationship between variables (e.g. linear relationship) c) the joint probability distribution of the observations/errors (e.g. independent observations) 	7 (44%)	10	<ul style="list-style-type: none"> <input type="checkbox"/> This can be assessed with sensitivity analyses <input type="checkbox"/> Desirable if we have statistical models that are robust to violations of these assumptions.
S8	Ability to maintain the ITT population	The ability of the model to use the entire intent-to-treat population in the analysis, meaning that all randomized subjects are included in the analysis according to original treatment assignment, regardless of protocol adherence (i.e. regardless the treatment actually received, patients' compliance including baseline, cross-over to other treatments or withdrawal from the study)	6 (38%)	7	<ul style="list-style-type: none"> <input type="checkbox"/> ITT is the standard in most protocols. <input type="checkbox"/> ITT is needed for generalizability of findings. <input type="checkbox"/> Too restrictive if needed for all analyses. <input type="checkbox"/> The use of ITT depends on the study objectives.

S18	Handle multiplicity	The ability of the model to statistically test multiple outcomes (due to multiple scales of interest and/or repeated measures of the same outcome) in an integrated test	6 (38%)	-1	<ul style="list-style-type: none"> <input type="checkbox"/> Only needed if the trial hypothesis calls for an integrated test <input type="checkbox"/> It is more efficient but not essential. Similar to other clinical endpoints, several independent tests may be considered (with error correction)
S4	Allow for time-varying covariates	The ability of the model to include time-varying covariates that are believed to be associated with the outcome variable or compliance	5 (31%)	2	<ul style="list-style-type: none"> <input type="checkbox"/> It depends on the study. <input type="checkbox"/> It may be useful but will not be used for the primary analysis <input type="checkbox"/> It makes the findings more difficult to interpret
S10	Handle clustered data (Part II – within groups)	The ability of the model to allow for correlations within groups (between subjects within the same institution/country,..)	5 (31%)	1	<ul style="list-style-type: none"> <input type="checkbox"/> Similar to controlling or stratifying for confounding factors / covariates <input type="checkbox"/> Not often part of the primary analysis even with other endpoints such as overall survival <input type="checkbox"/> Depends on the study objectives: probably needed if comparing centers or countries
S19	Handle a bounded scale	The ability of the model to analyze an outcome variable that has a defined maximum and minimum value (e.g. 0-100)	5 (31%)	2	<ul style="list-style-type: none"> <input type="checkbox"/> In practice, having a bounded scale rarely generates problems <input type="checkbox"/> This depends on the distribution of the data
S11	Handle clustered data (Part III – between outcomes)	The ability of the model to allow for correlations between outcomes (if multiple dimensions)	4 (25%)	-2	<ul style="list-style-type: none"> <input type="checkbox"/> It is only needed when a study calls for multiple outcomes to be tested at once. Even then, this can be handled by several independent tests (with error correction) <input type="checkbox"/> Pre-specifying the PRO domains is important rather than modelling multiple PROs

					<input type="checkbox"/> This adds too much complexity and model will be difficult to interpret
S14	Handle unbalanced designs (Part III)	The ability of the model to handle situations where the schedule of assessment is planned to be equal across patients, but differs across patients due to non-adherence to the protocol (patients respond to the assessment point based on the protocol not exactly on the same time)	3 (19%)	-8	<input type="checkbox"/> This is a post-hoc issue that can be addressed with sensitivity analyses. <input type="checkbox"/> This is something that can be dealt with using time windows
S7	Handle missing data (Part II)	The ability of the model to deal with missing data due to non-compliance. Thereby, we mean a method that provides an uncertainty estimate to address the impact of the missing data/how sensitive the method is to missing data	2 (13%)	-1	<input type="checkbox"/> This is not essential as a primary analysis. The impact of missing data can be assessed via sensitivity analyses

Note. Members from the statistical methods working group were asked to rate each statistical feature from a scale of 1 – 5. 1 = not essential; 3 = desirable; 5 = essential.

¹Primary scoring decision rule: Accept as soon as >/70% respondents rated “essential” (rating 4 or 5) AND </ 15% votes “not essential” (rating 1 or 2). Reject as soon as >/30% votes “not essential” (rating 1 or 2).

²Secondary scoring (sensitivity analysis): Ranking based on weighted sums. Ratings of 5, 4, 3, 2, 1 are transformed to scores of +3, +1, 0, -1, -3 respectively. For example, if a statistical feature is given a rating of 5, the transformed score is + 3. The sum of the transformed scores for each statistical feature was used to rank the statistical features. Highest possible score: 48 (16 * 3). Lowest possible score: -48 (16 * -3).

Table 3a. Coding scheme for the evaluation of each statistical method based on agreed essential/highly desirable statistical feature for PRO analysis

Statistical Feature	Codes	Examples
<p>Clinical relevance: produce results on the size, certainty and direction of the estimation and precision of the treatment effect that have a direct link with the clinical relevance classification of the instrument</p>		
<p>1. Clinical relevance at the within-individual level*</p> <p>*Note that this is not a feature of the statistical method.</p>	<p>(Yes)</p> <p>The within-individual level outcome can be directly linked to the clinical relevance classification of the instrument AND the clinical relevance of the result is interpreted at the within-individual level</p>	<ul style="list-style-type: none"> □ Definition of event for “time to event”: change score is computed for each individual; if the change score reaches a pre-defined threshold, individual data is coded as an event.
	<p>(No)</p> <p>Clinical relevance of the result cannot be directly linked to the clinical relevance classification of the instrument OR clinical relevance of the result is not interpreted at the within-individual level</p>	<ul style="list-style-type: none"> □ Raw or change scores are used as an outcome variable, and the clinical relevance of the result is interpreted through an estimate of the mean on the group level □ Individual summary measures that cannot be directly linked to the clinical relevance classification of the instrument
<p>2. Clinical relevance of the <u>treatment effect</u>: Within-group/ Between groups*</p> <p>*Note that all evaluations are based on comparison of only two arms</p>	<p>(Yes)</p> <p>Statistical models that produce not only statistical significance estimates, but also the magnitude of the treatment effect</p> <p>Between group: Clinical relevance of the result is interpreted as a difference between groups; and this difference can be directly linked to the clinical relevance classification of the instrument</p> <p>Within-group: Clinical relevance of the result is interpreted as a change within a group; and this group change can be directly linked to the clinical relevance classification of the instrument</p>	<ul style="list-style-type: none"> □ Between-group: Mean difference between groups (with CI); Odds ratio (with CI) □ Within-group: This can be seen in longitudinal models (e.g., mixed models) which estimates the main effect of time (mean change within group with the corresponding CI).
	<p>(No)</p> <p>Statistical models that give a statistical significance estimate, but the magnitude of the treatment effect is not estimated or the treatment effect is distorted</p> <p>Between group: Clinical relevance of the result for the difference between groups cannot be directly linked to the clinical relevance classification of the instrument</p> <p>Within-group: Clinical relevance of the result for the change within groups cannot</p>	<ul style="list-style-type: none"> □ Between-group: Results are derived from a sum of squares or sum of ranks □ Within-group: Results are derived from a sum of squares

	be directly linked to the clinical relevance classification of the instrument	
3. Adjust for covariates including baseline	(Yes) Covariates and stratification can be included	
	(Limited) Can only include stratification	
	(No) Inclusion of covariates and stratification are not possible	
4. Missing data with least restrictions	(Informative missingness) Method has the ability to take into account informative missingness (The process which caused the missing data is informative and can be used to estimate the true response; MAR or MNAR) ¹	
	(Non-informative missingness) Method provides valid inference only in the case of non-informative missingness (the process which caused the missing data is not informative about the parameter that is to be estimated; MCAR) ¹	
5. Clustered data (repeated assessments)	(Yes) Repeated assessments of each individual is taken into account; the order of measurements over time is also taken into account.	<input type="checkbox"/> Covariance structure of the repeated assessments can be specified.
	(Limited) Repeated assessments of each individual is taken into account. However the order of measurements over time cannot be taken fully into account.	
	(No) Repeated assessments are not taken into account. Each assessment is treated as an independent observation.	<input type="checkbox"/> Techniques designed for independent observations (i.e.. one observation per patient, e.g. techniques for cross-sectional data) are used even though the data set contains repeated (non-independent) observations per individual

Table 3b. Evaluation of each statistical method based on agreed essential/highly desirable statistical feature for PRO analysis

Stat Method	Clinical relevance		Descriptive	Adjust for covariates including baseline	Missing data with least restrictions ^{2,3}	Clustered data – repeated assessments	Recommended # of follow-up assessments	Comments
	Within-individual	Within-group and between group (treatment effect)						
<p>Improvement / worsening (event): time to event</p> <p>Maintenance (event): time to (end of) maintenance</p> <p>Time to event: Time to event</p>								
Cox PH (Kaplan-Meier) ⁴⁻⁶	<p>Yes</p> <p>Clinical relevance of the result is interpreted at the within-individual level (through a clinically relevant definition of a within-individual event)</p>	<p>Yes</p> <p>Between group:</p> <p>Clinical relevance of the difference between groups can be assessed using a hazard ratio (with CI)</p>	<p>- Median duration for each group</p> <p>- Survival probabilities for each group at a time point</p>	<p>Yes</p> <p>Covariates and stratification can be included</p>	<p>Can handle informative missingness</p> <p>Method provides valid inference when censored* data are MCAR or MAR.</p> <p>*Non-informative censoring: censoring is independent from the possibly unobserved time-to-event applies ⁶</p>	<p>Limited:</p> <p>Cluster of repeated assessments per patient (with event time), but the order of measurements over time is ignored (i.e., measurements before or after the specified event is ignored).</p>	<p>Baseline + Sufficient # of follow-ups</p> <p>Sufficient follow-up assessments needed to capture occurrence of event</p>	<p>Strong assumption of proportional hazards</p> <p>Results need to be checked to assess whether assumption of proportional hazards is met. If not met, consider using log-rank test + restricted mean survival time (RMST)</p> <p>Assumption of independent censoring should be met⁷</p>

<p>Log-rank test (Kaplan-Meier)⁴⁻⁶</p>	<p><u>Yes</u></p> <p>Clinical relevance of the result is interpreted at the within-individual level (through a clinically relevant definition of a within-individual event)</p>	<p><u>No</u></p> <p>Between group:</p> <p>Indicates whether survival between two groups is significantly different, but does not indicate how different they are.</p>	<p>- Median duration for each group</p> <p>- Survival probabilities for each group at a time point</p>	<p><u>Limited</u></p> <p>Can only include stratification</p>	<p>Can handle <u>informative</u> missingness</p> <p>Method provides valid inference when censored* data are MCAR or MAR.</p> <p>*Non-informative censoring: censoring is independent from the possibly unobserved time-to-event ⁶</p>	<p><u>Limited:</u></p> <p>Cluster of repeated assessments per patient (with event time), but the order of measurements over time is ignored (i.e., measurements before or after the specified event is ignored).</p>	<p>Baseline + <u>Sufficient</u> # of follow-ups</p> <p>Sufficient follow-up assessments needed to capture occurrence of event</p>	<p>Less efficient when proportional hazards assumption is not met, but does not require the assumption of proportional hazards.</p> <p>Assumption of independent censoring should be met</p>
---	--	--	--	---	---	---	--	--

Improvement / worsening (response): Proportion of patients with a response at time t

Maintenance: Proportion of patients with a maintained response at time t

Fisher's exact test ⁸⁻¹¹	<u>Yes</u> Clinical relevance of the result is interpreted at the within-individual level (through a clinically relevant definition of a within-individual event or discrete outcomes)	<u>No</u> Between group: Discrete/binary outcome: Only indicates whether there is an association between treatment and frequency of their response, but does not indicate the magnitude of this association.	-Proportion (or percentage) of responders for each group -Odds/risk ratio	<u>No</u> Inclusion of covariates and stratification are not possible	Can only handle <u>non-informative</u> missingness Method provides valid inference only for MCAR. Listwise deletion/complete case analysis: Patients with no data at baseline and/or specific timepoint are not included in the analysis.	<u>No</u> - Does not cluster repeated assessments per patient - Does not take into account longitudinal nature of data	Baseline + <u>1</u> follow-up	Ideal for smaller sample sizes Does not require the assumption of normality
(Pearson's) Chi-square test ⁸⁻¹¹	<u>Yes</u> Clinical relevance of the result is interpreted at the within-individual level (through a clinically relevant definition of a within-individual event or discrete outcomes)	<u>No</u> Between group: Discrete/binary outcome: Only indicates whether there is an association between treatment and frequency of their response, but does not indicate the magnitude of this association.	-Proportion (or percentage) of responders for each group -Odds/risk ratio	<u>No</u> Inclusion of covariates and stratification are not possible	Can only handle <u>non-informative</u> missingness Method provides valid inference only for MCAR. Listwise deletion/complete case analysis: Patients with no data at baseline and/or specific timepoint are not included in the analysis.	<u>No</u> - Does not cluster repeated assessments per patient - Does not take into account longitudinal nature of data	Baseline + <u>1</u> follow-up	Large data set is needed. Assumption of normality is required
(Cochran) Mantel-Haenszel test ¹²⁻¹⁵	<u>Yes</u> Clinical relevance of the result is interpreted at the within-individual level (through a clinically relevant definition of a within-	<u>Yes</u> Between group: Discrete/binary outcome: Clinical	-Proportion (or percentage) of responders for each group	<u>Limited</u> Can only include stratification	Can only handle <u>non-informative</u> missingness	<u>No</u> - Does not cluster repeated assessments per patient	Baseline + <u>1</u> follow-up	

	individual event or discrete outcomes)	relevance of the difference between groups can be assessed using odd/risk ratio (with CI)	-Odds/risk ratio		Method provides valid inference only for MCAR. Listwise deletion/complete case analysis: Patients with no data at baseline and/or specific timepoint are not included in the analysis.	- Does not take into account longitudinal nature of data		
<p>Improvement / worsening (response): level of response at time t</p> <p>Maintenance: not applicable (by definition of maintenance. For example, we cannot say “level of maintenance is higher/lower” in one arm vs the other)</p>								
(Generalized) linear mixed model (time as discrete - specific time point) ¹⁶	No Clinical relevance of the result is not interpreted at the within-individual level, but as a change on the group level	Yes Between group: Continuous outcome: Clinical relevance of the result can be assessed using the mean difference between the two groups at a specific time point (with CI) Within-group: Clinical relevance of the result can be assessed using an estimate assessing change within group (with CI) (i.e. main effect of time). *Clinical relevance of the estimated mean	-Mean baseline level (with CI) & mean specific time point level (with CI) for each group -Mean change between baseline and each assessed time point (with CI) for each group	Yes Covariates and stratification can be included	Can handle informative missingness Method provides valid inference when missing data are MCAR or MAR.	Yes - Cluster of repeated assessments per patient - Order of measurements can be taken into account (i.e., covariance structure can be specified to take into account that measurements that are closer in time tend to have higher correlations)	Baseline + sufficient but limited # of follow-ups As the number of follow-up assessments increases, the number of parameters to estimate also increases	Since time is treated as discrete, a parameter needs to be estimated for every assessment over time. This is not ideal if there are too many follow-up assessments. Does not require an assumption regarding the relationship between time and outcome variable (e.g., assumption of a linear relationship). The assumption under MAR is that the treatment estimate is based on the assumption that patients will continue on treatment for the full study duration. ¹⁷

		difference (between group) and change (within-group) can be interpreted by comparison with effect size, or PROM-specific MID or interpretation guidelines, if available.						Generalized linear mixed models can be used for discrete, count or binary outcome.
(Generalized) linear mixed model (time as continuous) ¹⁶	No Clinical relevance of the result is not interpreted at the within-individual level, but as a change on the group level	Yes Between group: Continuous outcome: Clinical relevance of the result can be assessed using the mean difference between the two groups at a specific time point (with CI) Within-group: Clinical relevance of the result can be assessed using an estimate assessing change within group (with CI) (i.e. main effect of time). *Clinical relevance of the estimated mean difference (between group) and change (within-group) can be interpreted by	-Mean baseline level (with CI) & mean specific time point level (with CI) for each group -Rate of change between baseline and the specific time point (with CI)	Yes Covariates and stratification can be included	Can handle informative missingness Method provides valid inference when missing data are MCAR or MAR.	Yes - Cluster of repeated assessments per patient - Order of measurements can be taken into account (i.e., covariance structure can be specified to take into account that measurements that are closer in time tend to have higher correlations)	Baseline + sufficient # of follow-ups	May be suitable if there are many follow-up assessments and the relationship between time and outcome variable is linear. Since time is treated as continuous, only one parameter needs to be estimated regardless of the number of follow-up assessments over time. This implies a strong assumption that the influence of time on the outcome variable is linear. More complex models are available to assess non-linear relationships between time and outcome. For example, time is treated as continuous; and linear, quadratic and cubic polynomial terms may be used to approximate the time curves. But this also implies more

		comparison with effect size, or PROM-specific MID or interpretation guidelines, if available.						parameters to estimate and making strong assumptions regarding the non-linear relationship between time and the outcome variable. The assumption under MAR is that the treatment estimate is based on the assumption that patients will continue on treatment for the full study duration. ¹⁷ Generalized linear mixed models can be used for discrete, count or binary outcome.
Generalized estimating equation 18-24	No Clinical relevance of the result is not interpreted at the within-individual level, but as a change on the group level	Yes Between group: Continuous outcome: Clinical relevance of the result can be assessed using the mean difference between the two	Continuous outcome: Mean baseline level (with CI) & mean specific time point level (with CI) for each group	Yes Covariates and stratification can be included	Can only handle non-informative missingness Method provides valid inference only for MCAR.*	Yes - Cluster of repeated assessments per patient - Order of measurements can be taken into account (i.e.,	Time as continuous: Baseline + sufficient # of follow-ups Time as discrete:	Parameter estimates are consistent and asymptotically normal even under misspecified correlation structure of responses. ²⁵

		<p>groups at a specific time point (with CI)</p> <p>Within-group:</p> <p>Clinical relevance of the result can be assessed using an estimate assessing change within group (with CI) (i.e. main effect of time).</p> <p>*Clinical relevance of the estimated mean difference (between group) and change (within-group) can be interpreted by comparison with effect size, or PROM-specific MID or interpretation guidelines, if available.</p>	Ordinal/binary outcome: Odds ratio (with CI)		*Weighted GEE method is available to take into account MAR.	covariance structure can be specified to take into account that measurements that are closer in time tend to have higher correlations)	Baseline + sufficient but limited # of follow-ups	Generalized estimating equations can be used for discrete, count or binary outcome.
Linear regression	No	<p>Between group:</p> <p>Continuous outcome: Clinical relevance of the result can be assessed using the mean difference between the two</p>	Wilc	Yes	<p>Can only handle non-informative missingness</p> <p>Method provides valid inference only for MCAR.</p> <p>Listwise deletion/complete case</p>	No	Baseline + 1 follow-up	<p>- Does not cluster repeated assessments per patient</p> <p>- Does not take into account</p>

		groups at a specific time point (with CI) *Clinical relevance of the estimated mean difference (between group) and change (within-group) can be interpreted by comparison with effect size, or PROM-specific MID or interpretation guidelines, if available.			analysis: Patients with no data at baseline and/or specific timepoint is not included in the analysis.	longitudinal nature of data		
ANOVA ¹⁶ or ANCOVA	No Clinical relevance of the result is not interpreted at the within-individual level, but as a change on the group level	No Between group: Continuous outcome: Indicates whether the difference between two groups is significantly different, but does not indicate how different they are.	-Mean baseline level (with CI) & mean specific time point level (with CI) for each group -Mean change between baseline and specific time point (with CI) for each group (if change score is used as outcome)	Yes Covariates and stratification can be included	Can only handle non-informative missingness Method provides valid inference only for MCAR. Listwise deletion/complete case analysis: Patients with no data at baseline and/or specific timepoint is not included in the analysis.	No - Does not cluster repeated assessments per patient - Does not take into account longitudinal nature of data	Baseline + 1 follow-up	
(Independent samples) t-test	No Clinical relevance of the result is not interpreted at the within-individual level, but	Yes Between group:	-Mean baseline level (with CI) &	No Inclusion of covariates and	Can only handle non-informative missingness	No - Does not cluster repeated	Baseline + 1 follow-up	Assumption of normal distribution is needed

	as a change on the group level	<p>Continuous outcome: Clinical relevance of the result can be assessed using the mean difference between the two groups at a specific time point (with CI)</p> <p>*Clinical relevance of the estimated mean difference (between group) and change (within-group) can be interpreted by comparison with effect size, or PROM-specific MID or interpretation guidelines, if available.</p>	<p>mean specific time point level (with CI) for each group</p> <p>-Mean change between baseline and specific time point (with CI) for each group (if change score is used as outcome)</p>	stratification are not possible	<p>Method provides valid inference only for MCAR.</p> <p>Listwise deletion/complete case analysis: Patients with no data at baseline and/or specific timepoint is not included in the analysis.</p>	<p>assessments per patient</p> <p>- Does not take into account longitudinal nature of data</p>		
Wilcoxon rank sum test	<p>No</p> <p>Clinical relevance of the result is not interpreted at the within-individual level, but as a change on the group level</p>	<p>No</p> <p>Between group:</p> <p>Continuous outcome: Indicates whether the difference between two groups is significantly different, but does not indicate how different they are.</p>	<p>- Mean baseline level (with CI) & mean specific time point level (with CI) for each group</p> <p>-Mean change between baseline and specific time point (with CI) for each group (if change</p>	<p>No</p> <p>Inclusion of covariates and stratification are not possible</p>	<p>Can only handle non-informative missingness</p> <p>Method provides valid inference only for MCAR.</p> <p>Listwise deletion/complete case analysis: Patients with no data at baseline and/or specific timepoint is not included in the analysis.</p>	<p>No</p> <p>- Does not cluster repeated assessments per patient</p> <p>- Does not take into account longitudinal nature of data</p>	<p>Baseline + 1 follow-up</p>	Does not assume normal distribution

			score is used as outcome)					
Pattern mixture model ²⁶⁻²⁸	No Clinical relevance of the result is not interpreted at the within-individual level, but as a change on the group level	Yes Between group: Time as discrete: Clinical relevance of the result can be assessed using the difference in levels between the two groups at a specific time point (with CI) Time as continuous: Clinical relevance of the result can be assessed using the mean difference in the rate of change between groups at a specific time point (with CI) Within-group: Clinical relevance of the result can be assessed using an estimate assessing change within group (with CI) (i.e. main effect of time). *Clinical relevance of the estimated mean difference (between group) and change (within-group) can be	-Mean baseline level (with CI) & mean specific time point level (with CI) for each group -Mean change between baseline and specific time point (with CI) for each group (if time is discrete) -Rate of change between baseline and specific time point (with CI) for each group (if time is continuous)	Yes Covariates and stratification can be included	Can handle informative missingness Method provides valid inference when missing data are MCAR or MAR. Method can take into account potential MNAR data -> missing values can be modeled (takes time of missingness as explanatory missing variable)	Yes - Cluster of repeated assessments per patient - Order of measurements can be taken into account (i.e., covariance structure can be specified to take into account that measurements that are closer in time tend to have higher correlations)	Time as continuous: Baseline + sufficient # of follow-ups Time as discrete: Baseline + sufficient but limited # of follow-ups As the number of follow-up assessments increases, the number of parameters to estimate also increases	Validity of the pattern mixture model depends on the choice of patterns which is often a subjective choice of the investigator and is not verifiable from the data ²⁷ . However it is often advised to use pattern mixture models as a sensitivity analysis. Investigators should have several sensitivity analyses performed over a variety of pattern choices (e.g., where each analysis has a different set of clinical assumptions regarding unobserved data) to ensure robustness of findings ²⁶⁻²⁸ Because of the many parameters to be estimated, time is often treated as continuous in this statistical model Generalized linear mixed models can be used for discrete, count or binary outcome.

		interpreted by comparison with effect size, or PROM-specific MID or interpretation guidelines, if available.						
Joint model for longitudinal and survival data ²⁹⁻³⁵	No Clinical relevance of the result is not interpreted at the within-individual level, but as a change on the group level	Yes Between group: Continuous outcome: Clinical relevance of the result can be assessed using the mean difference in the rate of change between two groups at a specific time point (with CI) Within-group: Clinical relevance of the result can be assessed using an estimate assessing the rate of change within group (with CI) (i.e. main effect of time). *Clinical relevance of the estimated mean difference (between group) and change (within-group) can be interpreted by comparison with effect size, or PROM-	-Mean baseline level (with CI) & mean specific time point level (with CI) for each group -Rate of change between baseline and the specific time point (with CI)	Yes Covariates and stratification can be included	Can handle informative missingness Method provides valid inference when missing data are MCAR or MAR. Method can take into account potential MNAR data* -> missing values can be modeled (see comments)	Yes - Cluster of repeated assessments per patient - Order of measurements can be taken into account (i.e., covariance structure can be specified to take into account that measurements that are closer in time tend to have higher correlations)	Baseline + sufficient # of follow-ups	Joint modeling of longitudinal data and survival data. Possibility to account for informative patterns of missing data by jointly modeling the longitudinal PRO outcome (longitudinal process) and time to informative PRO dropout (survival data). ³⁶ Joint models rely on the conditional independence assumption (event process and longitudinal responses are independent conditionally on a latent process expressed by a set of random effects) ³³ Many parameters (such as the association between the longitudinal and the TTE process, baseline hazard function, random effects, defining the

		specific MID or interpretation guidelines, if available.						<p>'event' for the time to informative drop-out...) are to be specified³⁴ and the model can be very computationally demanding³¹.</p> <p>Because of the many parameters to be estimated, time is often treated as continuous in this statistical model</p> <p>Generalized linear mixed models can be used for discrete, count or binary outcome.</p>
Overall effect: Describe trajectory of outcome over time								
(Generalized) linear mixed model (time as discrete - omnibus test): group*time interaction ^{16,37,38}	No Clinical relevance of the result is not interpreted at the within-individual level, but as a change on the group level	No Between group: Assesses whether the mean response profiles between the two groups are statistically significantly different (non-parallel profiles), but does not provide	-Mean baseline level (with CI) & levels at each assessed time point (with CI) for each group -Mean change between	Yes Covariates and stratification can be included	Can handle informative missingness Method provides valid inference when missing data are MCAR or MAR.	Yes - Cluster of repeated assessments per patient - Order of measurements can be taken into account (i.e., covariance	Baseline + sufficient but limited # of follow-ups As the number of follow-up assessments increases, the number of parameters to	Profiles are reported cross-sectionally and not longitudinally. That is, every assessment point has a mean and CI. If individual longitudinal profiles are of interest, more

		<p>an estimate of how different they are.</p> <p>Within-group:</p> <p>Assesses whether responses over time are statistically significantly different, but does not provide an estimate of how different they are..</p>	<p>baseline and each assessed time point (with CI) for each group</p>			<p>structure can be specified to take into account that measurements that are closer in time tend to have higher correlations)</p>	<p>estimate also increases</p>	<p>complex models are available. For example, time is treated as continuous; and linear, quadratic and cubic polynomial terms may be used to approximate the time curves.</p> <p>Generalized linear mixed models can be used for discrete, count or binary outcome.</p>
<p>Repeated measures ANOVA: group*time interaction ^{16,37,38}</p>	<p>No</p> <p>Clinical relevance of the result is not interpreted at the within-individual level, but as a change on the group level</p>	<p>No</p> <p>Between group:</p> <p>Assesses whether the mean response profiles between the two groups are statistically significantly different (non-parallel profiles), but does not provide an estimate of how different they are.</p> <p>Within-group:</p> <p>Assesses whether responses over time</p>	<p>-Mean baseline level (with CI) & levels at each assessed time point (with CI) for each group</p> <p>-Mean change between baseline and each assessed time point (with CI) for each group</p>	<p>Yes</p> <p>Covariates and stratification can be included</p>	<p>Can only handle non-informative missingness</p> <p>Method provides valid inference when data are MCAR.</p> <p>Listwise deletion/complete case analysis: Patients with no data at baseline and/or any specific timepoint is not included in the analysis.</p>	<p>Limited</p> <p>- Cluster of repeated assessments per patient</p> <p>- Order of measurements cannot be taken into account (i.e., assumes compound symmetry for covariance structure, meaning covariance between pairs of assessments are equal regardless of the distance</p>	<p>Baseline + sufficient but limited # of follow-ups</p> <p>As the number of follow-up assessments increases, the number of parameters to estimate also increases</p>	<p>Profiles are reported cross-sectionally and not longitudinally. That is, every assessment point has a mean and CI.</p>

		are statistically significantly different, but does not provide an estimate of how different they are.				between occasions)		
--	--	--	--	--	--	--------------------	--	--

Table 4.a Survey Results on standardizing definitions for analysis population (intent-to-treat population and modified intent-to-treat population) (N=38)

Statement	Voting results
Intent-to-treat population (ITT): The ITT population includes all the patients that were randomized to the study. According to the strict ITT principle, all randomized subjects should be analyzed according to the allocated treatment, regardless of the treatment actually received, protocol adherence, crossover to other treatments or withdrawal from the study.	
<input type="checkbox"/> Agree	37/38 (97%)
<input type="checkbox"/> Don't know	1/38 (3%)
Modified intent-to-treat population (mITT): Acceptable modifications to the Intent-To-Treat (ITT) population for the analysis of PRO data in randomized controlled trials (multiple answers possible)	
<input type="checkbox"/> Analysis population could be limited to patients with baseline PRO assessment	12/38 (32%)
<input type="checkbox"/> Analysis population could be limited to patients with at least one post-baseline PRO assessment	6/38 (16%)
<input type="checkbox"/> Analysis population could be limited to patients with baseline + at least one post-baseline PRO assessment	17/38 (45%)
<input type="checkbox"/> Analysis population could be limited to eligible patients	9/38 (24%)
<input type="checkbox"/> No modification to the ITT population is appropriate (the analysis population should be all randomized patients, analyzed according to the allocated treatment)	6/38 (16%)
<input type="checkbox"/> Analysis population could be limited to the safety population (patients exposed to their intended treatment only)	4/38 (11%)
<input type="checkbox"/> Analysis population could be limited to patients exposed to any protocol treatment	4/38 (11%)
<input type="checkbox"/> Other (To specify)	4/38 (11%)
<input type="checkbox"/> Patients who consent to PRO substudy	<input type="checkbox"/> 1/38 (3%)
<input type="checkbox"/> Depends on the study objective	<input type="checkbox"/> 3/38 (8%)
<input type="checkbox"/> No answer/don't know	5/38 (13%)

Table 4.b. Survey results on standardizing calculation and definition of completion (variable denominator) and available data (fixed denominator) rates.

Statement	Voting results
Fixed and variable denominator rate:	
a) Fixed denominator rate – a rate with a denominator that stays the same over time (e.g. total number of enrolled patients)	
b) Variable denominator rate – a rate with a variable denominator at every time point (e.g. number of expected patients at time t)	
<input type="checkbox"/> Both the fixed denominator rate and the variable denominator rate are needed	26/38 (68%)
<input type="checkbox"/> Only the variable denominator rate is needed	6/38 (16%)
<input type="checkbox"/> Only the fixed denominator rate is needed	2/38 (5%)
<input type="checkbox"/> Other (To specify)	4/38 (11%)
<input type="checkbox"/> Both + cohort plots	<input type="checkbox"/> 1/38 (3%)
<input type="checkbox"/> Both + additional information related to the attrition	<input type="checkbox"/> 1/38 (3%)
<input type="checkbox"/> <i>Both can, but is not a 'must'</i>	<input type="checkbox"/> 1/38 (3%)
<input type="checkbox"/> Variable denominator rate + death rate	<input type="checkbox"/> 1/38 (3%)
Fixed denominator rate: Numerator	
<input type="checkbox"/> On-study patients submitting the PRO assessment at the designated time point	32/38 (84%)
<input type="checkbox"/> On-study patients submitting the PRO assessment at baseline AND at the designated time point	4/38 (11%)
<input type="checkbox"/> Other: Patients submitting any part of the PRO assessment at the designated time point	1/38 (3%)
<input type="checkbox"/> Don't know	1/38 (3%)
Fixed denominator rate: Denominator	
<input type="checkbox"/> Randomized patients (ITT population)	21/38 (55%)
<input type="checkbox"/> Patients with a PRO baseline assessment	6/38 (16%)
<input type="checkbox"/> Enrolled patients	2/38 (5%)
<input type="checkbox"/> Eligible patients ¹	2/38 (5%)

¹It was not specified in the survey whether this is patients (in)eligible for the PRO (sub)study or patients (in)eligible for the full study

<input type="checkbox"/> Safety population (patients who received intended treatment)	1/38 (3%)
<input type="checkbox"/> Other <ul style="list-style-type: none"> <input type="checkbox"/> Depends on analysis population: ITT or mITT <input type="checkbox"/> Depends on study objective <input type="checkbox"/> ITT minus patients not eligible for PRO assessment 	4/38 (11%) <ul style="list-style-type: none"> <input type="checkbox"/> 2 (5%) <input type="checkbox"/> 1 (3%) <input type="checkbox"/> 1 (3%)
<input type="checkbox"/> Don't know	2/38 (5%)
Fixed denominator rate: Terminology	
<input type="checkbox"/> Completion rate	20/38 (53%)
<input type="checkbox"/> Compliance rate	8/38 (21%)
<input type="checkbox"/> Other	6/38 (16%)
<input type="checkbox"/> Don't know/N.A.	4/38 (11%)
Variable denominator rate: Numerator	
<input type="checkbox"/> On-study patients submitting the PRO assessment at the designated time point	30/38 (79%)
<input type="checkbox"/> On-study patients submitting the PRO assessment at baseline AND at the designated time point	6/38 (16%)
<input type="checkbox"/> Don't know	2/38 (5%)
Variable denominator rate: Denominator (defining who the "available patients at time t" are)	
<input type="checkbox"/> Patients who have died prior to assessment time t to be excluded from the denominator	34/38 (89%)
<input type="checkbox"/> Patients not on study anymore to be excluded from the denominator	27/38 (71%)
<input type="checkbox"/> Patients no longer part of the PRO assessment schedule (according to protocol) to be excluded from the denominator	24/38 (63%)
<input type="checkbox"/> Ineligible patients ^{Error! Bookmark not defined.} to be excluded from the denominator	19/38 (50%)
<input type="checkbox"/> Patients not on treatment anymore to be excluded from the denominator	10/38 (26%)
<input type="checkbox"/> Patients illiterate in the language of the PRO tool to be excluded from the denominator	10/38 (26%)
<input type="checkbox"/> Patients without a valid PRO baseline assessment to be excluded from the denominator	7/38 (18%)

<input type="checkbox"/> Patients who cannot be reached at the time of the visit to be excluded from the denominator	4/38 (11%)
<input type="checkbox"/> Patients refusing to respond the PRO assessment to be excluded from the denominator	3/38 (8%)
<input type="checkbox"/> Other to be excluded from the denominator <ul style="list-style-type: none"> <input type="checkbox"/> Patients not meeting the clinically significant change criterion <input type="checkbox"/> Patients without valid PRO baseline assessment or not, depending on the situation 	2/38 (5%) <ul style="list-style-type: none"> <input type="checkbox"/> 1/38 (3%) <input type="checkbox"/> 1/38 (3%)
Variable denominator rate: Terminology	
<input type="checkbox"/> Completion rate	9/38 (24%)
<input type="checkbox"/> Compliance rate	17/38 (45%)
<input type="checkbox"/> Other	7/38 (18%)
<input type="checkbox"/> Don't know/N.A.	5/38 (13%)

References

1. Shih W. Problems in dealing with missing data and informative censoring in clinical trials. *Curr Control Trials Cardiovasc Med.* 2002;3(1):4. doi:10.1186/1468-6708-3-4
2. Fielding S, Fayers PM, Ramsay CR. Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health Qual Life Outcomes.* 2009;7:57. doi:10.1186/1477-7525-7-57
3. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. *Test (Madr).* 2009;18(1):1-43. doi:10.1007/s11749-009-0138-x
4. Bland JM, Altman DG. The logrank test. *BMJ.* 2004;328(7447):1073. doi:10.1136/bmj.328.7447.1073
5. Bewick V, Cheek L, Ball J. Statistics review 12: survival analysis. *Crit Care.* 2004;8(5):389-394. doi:10.1186/cc2955
6. Zhao Y, Herring AH, Zhou H, Ali MW, Koch GG. A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring. *J Biopharm Stat.* 2014;24(2):229-253. doi:10.1080/10543406.2013.860769
7. Leung K-M, Elashoff RM, Afifi AA. Censoring Issues in Survival Analysis. *Annu Rev Public Health.* 1997;18(1):83-104. doi:10.1146/annurev.publhealth.18.1.83
8. Ruxton GD, Neuhäuser M. Review of alternative approaches to calculation of a confidence interval for the odds ratio of a 2×2 contingency table. Freckleton R, ed. *Methods Ecol Evol.* 2013;4(1):9-13. doi:10.1111/j.2041-210x.2012.00250.x
9. Cook JA, Bunce C, Doré CJ, Freemantle N, Ophthalmic Statistics Group on behalf of the OS. Ophthalmic statistics note 6: effect sizes matter. *Br J Ophthalmol.* 2015;99(5):580-581. doi:10.1136/bjophthalmol-2014-306303
10. Olivier J, Bell ML. Effect sizes for 2×2 contingency tables. *PLoS One.* 2013;8(3):e58777. doi:10.1371/journal.pone.0058777
11. Allison PD. Missing Data. *Quant Appl Soc Sci.* 2001:104. doi:10.1136/bmj.38977.682025.2C
12. Wittes J, Wallenstein S. The Power of the Mantel—Haenszel Test. *J Am Stat Assoc.* 1987;82(400):1104-1109. doi:10.1080/01621459.1987.10478546
13. Kuritz SJ, Landis R, Koch GG. A GENERAL OVERVIEW OF MANTEL-HAENSZEL METHODS: Applications and Recent Developments. *Ann Rev Public Heal.* 1988;9:123-160. <https://www.annualreviews.org/doi/pdf/10.1146/annurev.pu.09.050188.001011>. Accessed April 17, 2018.
14. McDonald JH. Handbook of Biological Statistics. *Handbook of Biological Statistics.* <http://udel.edu/~mcdonald/>. Published 2014. Accessed April 17, 2018.
15. SAS support. SAS/STAT(R) 9.2 User's Guide, Second Edition: Cochran-Mantel-Haenszel Statistics. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_freq_sect031.htm. Accessed April 17, 2018.
16. Liu S, Rovine MJ, Molenaar PCM. Selecting a linear mixed model for longitudinal data: Repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychol Methods.* 2012;17(1):15-30. doi:10.1037/a0026971
17. European Medicines Agency. Guideline on missing data in confirmatory clinical trials. London: European Medicines Agency. doi:10.2307/2290157
18. Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: What are the differences? *Stat Med.* 2009;28(2):221-239. doi:10.1002/sim.3478
19. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient reported outcomes. *Stat Methods Med Res.* 2014;23(5):440-459. doi:10.1177/0962280213476378
20. Lindsey JK, Lambert P. On the appropriateness of marginal models for repeated measurements in clinical trials. *Stat Med.* 1998;17(4):447-469. doi:10.1002/(SICI)1097-0258(19980228)17:4<447::AID-SIM752>3.0.CO;2-G
21. Verbeke G, Molenberghs G, Rizopoulos D. Random effects models for longitudinal data Link Random Effects Models for Longitudinal Data. doi:10.1016/S0167-7152(02)00397-8
22. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 1986;42(1):121-130. <http://www.ncbi.nlm.nih.gov/pubmed/3719049>. Accessed April 17, 2018.
23. Stephens AJ, Tchetgen Tchetgen EJ, De Gruttola V. Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level

- and individual-level covariates. *Stat Med.* 2012;31(10):915-930. doi:10.1002/sim.4471
24. SAS Institute Inc. *SAS/STAT® 14.3 User's Guide: The GEE Procedure.*; 2017. <https://support.sas.com/documentation/onlinedoc/stat/143/gee.pdf>. Accessed April 17, 2018.
 25. Wang M, Kong L, Li Z, Zhang L. Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Stat Med.* 2016;35(10):1706-1721. doi:10.1002/sim.6817
 26. Ratitch B. Implementation of Pattern-Mixture Models Using Standard SAS/STAT Procedures. 2011. <https://pharmasug.org/proceedings/2011/SP/PharmaSUG-2011-SP04.pdf>. Accessed April 17, 2018.
 27. Pauler DK, McCoy S, Moinpour C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Stat Med.* 2003;22(5):795-809. doi:10.1002/sim.1397
 28. Post WJ, Buijs C, Stolk RP, de Vries EGE, le Cessie S. The analysis of longitudinal quality of life measures with informative drop-out: a pattern mixture approach. *Qual Life Res.* 2010;19(1):137-148. doi:10.1007/s11136-009-9564-1
 29. Huang X, Li G, Elashoff RM, Pan J. A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime Data Anal.* 2011;17(1):80-100. doi:10.1007/s10985-010-9169-6
 30. Barrett J, Su L. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Stat Med.* 2017;36(9):1447-1460. doi:10.1002/sim.7209
 31. Tsiatis AA, Davidian M. Joint Modeling of Longitudinal and Time-To-Event Data: An Overview. *Stat Sin.* 2004;14:809-834. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.542&rep=rep1&type=pdf>. Accessed April 13, 2018.
 32. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. *J Clin Oncol.* 2010;28(16):2796-2801. doi:10.1200/JCO.2009.25.0654
 33. Rizopoulos D, Verbeke G, Lesaffre E, Vanrenterghem Y. A Two-Part Joint Model for the Analysis of Survival and Longitudinal Binary Data with Excess Zeros. *Biometrics.* 2008;64(2):611-619. doi:10.1111/j.1541-0420.2007.00894.x
 34. Rizopoulos D. An Introduction to the Joint Modeling of Longitudinal and Survival Data, with Applications in R. http://www.drizopoulos.com/courses/Int/JMwithR_CEN-ISBS_2017.pdf. Published 2017. Accessed April 16, 2018.
 35. Dupuy J, Mesbah M. Joint Modeling of Event Time and Nonignorable Missing Longitudinal Data. *Lifetime Data Anal.* 2002;8(2):99-115. doi:10.1023/A:1014871806118
 36. Cella D, Wang M, Wagner L, Miller K. Survival-adjusted health-related quality of life (HRQL) among patients with metastatic breast cancer receiving paclitaxel plus bevacizumab versus paclitaxel alone: results from Eastern Cooperative Oncology Group Study 2100 (E2100). *Breast Cancer Res Treat.* 2011;130(3):855-861. doi:10.1007/s10549-011-1725-6
 37. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis.* Wiley; 2011. <https://www.wiley.com/en-be/Applied+Longitudinal+Analysis%2C+2nd+Edition-p-9780470380277>. Accessed April 17, 2018.
 38. Hee Jo C, Gossett J, Simpson P. Regression Splines with Longitudinal Data. <http://www2.sas.com/proceedings/forum2007/143-2007.pdf>. Published 2007. Accessed April 17, 2018.

