International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials

Coens, Corneel; Pe, Madeline; Dueck, Amylou C; Sloan, J; Basch, Ethan; Calvert, Melanie; Campbell , A; Cleeland, Charles; Cocks, K; Collette , L; Devlin, N; Dorme, Lien; Flechtner, Hans-Henning; Gotay , C; Griebsch, I; Groenvold , M; King, M; Kluetz, Paul G; Koller , M; Malone, Daniel C

DOI:
10.1016/S1470-2045(19)30790-9

Citation for published version (Harvard):
Coens, C, Pe, M, Dueck, AC, Sloan, J, Basch, E, Calvert, M, Campbell , A, Cleeland, C, Cocks, K, Collette , L, Devlin, N, Dorme, L, Flechtner, H-H, Gotay , C, Griebsch, I, Groenvold , M, King, M, Kluetz, PG, Koller , M, Malone, DC, Martinelli, F, Mitchell, SA, Musoro, J, O'Connor, D, Oliver , K, Piault-Louis, E, Piccart, M, Quinten, C, Reijneveld, JC, Schurmann, C, Smith , AW, Soltys, KM, Taphoorn , M, Velikova, G & Bottomley , A 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*, vol. 21, no. 2, pp. e83-e96. https://doi.org/10.1016/S1470-2045(19)30790-9

To cite this article:

Coens C, Pe M, Dueck AC, Sloan J, Basch E, Calvert M, Campbell A, Cleeland C, Cocks K, Collette L, Devlin N, Dorme L, Flechtner HH, Gotay C, Griebsch I, Groenvold M, King M, Kluetz PG, Koller M, Malone DC, Martinelli F, Mitchell SA, Musoro J, O'Connor D, Oliver K, Piault-Louis E, Piccart M, Quinten C, Reijneveld JC, Schürmann C, Smith AW, Soltys KM, Taphoorn M, Velikova G, Bottomley A. (in press). International Standards for the Analysis of Quality of Life and Patient Reported Outcomes Endpoints in Cancer Randomised Controlled Trials; Recommendations based on critical reviews of the literature and international multi-expert, multi-stakeholder collaborative process. *The Lancet Oncology.* www.thelancet.com/journals/lanonc/home

International Standards for the Analysis of Quality of Life and Patient Reported Outcomes Endpoints in Cancer Randomised Controlled Trials:

Recommendations based on critical reviews of the literature and international multi-expert, multi-stakeholder collaborative process

Corneel Coens[1*], Madeline Pe[1*], Amylou C Dueck[2], Jeff Sloan[3], Ethan Basch[4], Melanie Calvert[5], Alicyn Campbell[6], Charles Cleeland[7], Kim Cocks[8], Laurence Collette[1], Nancy Devlin[9] , Lien Dorme[1], Hans-Henning Flechtner[10], Carolyn Gotay[11], Ingolf Griebsch[12], Mogens Groenvold[13], Madeleine King[14], Paul G Kluetz[15], Michael Koller[16], Daniel C Malone[17], Francesca Martinelli[1], Sandra A Mitchell[18], Jammbe Z Musoro[1], Daniel O'Connor[19], Kathy Oliver[20], Elisabeth Piault-Louis[21], Martine Piccart[22], Chantal Quinten[23], Jaap C Reijneveld[24], Christoph Schürmann[25], Ashley Wilder Smith[18], Katherine M Soltys[26], Martin J B Taphoorn[27], Galina Velikova[28, 29], and Andrew Bottomley[1] for the Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data (SISAQOL) Consortium.

*joint first authors

[1]European Organisation for Research and Treatment of Cancer (EORTC Headquarters), Brussels, Belgium (C Coens Msc, M Pe PhD, L Collette PhD, L Dorme MSc, F Martinelli MSc, J Z Musoro PhD, A Bottomley PhD)

[2]Alliance Statistics and Data Center, Mayo Clinic, Scottsdale, AZ, USA (A C Dueck PhD)

[3]Alliance Statistics and Data Center, Mayo Clinic, Rochester, MN, USA (J Sloan PhD)

[4]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA (E Basch MD)

[5]Centre for Patient Reported Outcomes Research, Institute of Applied Health Research and NIHR Birmingham Biomedical Research Centre, University of Birmingham, UK (Prof M Calvert PhD)

[6]Patient Relevant Evidence, San Francisco, CA, USA (A Campbell MPh)

[7]Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, TX, USA (Prof C Cleeland PhD)

[8]Adelphi Values, Bollington, Cheshire, UK (K Cocks PhD)

[9] Centre for Health Policy, School of Population and Global Health, University of Melbourne, Australia (N Devlin PhD)

[10]Clinic for Child and Adolescent Psychiatry and Psychotherapy, University of Magdeburg, Magdeburg, Germany (Prof H-H Flechtner MD)

[11]School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada (Prof C Gotay PhD)

[12]Boehringer Ingelheim International GmbH, Ingelheim, Germany (I Griebsch PhD)

[13]Department of Public Health,Bispebjerg Hospital and University of Copenhagen, Copenhagen, Denmark (Prof M Groenvold MD)

[14]University of Sydney, School of Psychology, Sydney, NSW, Australia (Prof M King PhD)

[15]Office of Hematology and Oncology Products, Center for Drug Evaluation and Research,US Food and Drug Administration, Silver Spring, MD, USA (P Kluetz MD)

[16]Center for Clinical Studies, University Hospital Regensburg, Regensburg, Germany (M Koller PhD)

[17]College of Pharmacy, University of Arizona, Tucson, AZ, USA (Prof D Malone PhD)

[18]Outcomes Research Branch, Healthcare Delivery Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA (S A Mitchell PhD, A W Smith PhD)

[19]Medicines and Healthcare products Regulatory Agency, London, UK (D O'Connor MBChB)

[20]International Brain Tumour Alliance, Surrey, UK (K Oliver BA)

[21]Genentech, a member of the Roche group, San Francisco, CA, USA (E Piault-Louis PharmD)

[22]Institut Jules Bordet, Université Libre de Bruxelles (ULB), Brussels, Belgium (M Piccart MD)

[23]European Centre for Disease Prevention and Control, Surveillance and Response Support Unit, Epidemiological Methods Section, Stockholm, Sweden (C Quinten MSc)

[24] VU University Medical Center, Department of Neurology & Brain Tumor Center, Amsterdam, The Netherlands (J C Reijneveld MD)

[25]Institute for Quality and Efficiency in Health Care, Cologne, Germany (C Schürmann PhD)

[26]Health Canada, Ottawa, ON, Canada (K Soltys PhD)

[27]Leiden University Medical Center/Haaglanden Medical Center, Leiden/The Hague, Netherlands (M Taphoorn MD)

[28]Leeds Institute of Cancer and Pathology, University of Leeds, St James's Hospital, Leeds, UK (Prof G Velikova PhD)

[29] International Society for Quality of Life Research, Milwaukee, Wisconsin, USA (Prof G Velikova PhD)

Acknowledgments

EORTC received an unrestricted education grant from Boehringer Ingelheim GmbH to initiate this work and from Genentech, a member of the Roche Group, for continuity funding. We thank ISOQOL and ISPOR for their support and review of the final manuscript and we thank Gina Mazza for the help with the survey research in the missing data part.

This publication reflects the views of the individual authors and should not be construed to represent official views or policies of the US Food and Drug Administration, US National Cancer Institute, Medicines and Healthcare products Regulatory Agency, Institute for Quality and Efficiency in Health Care (IQWIG), Health Canada, the NHS, the National Institute for Health Research, or the Department of Health, UK.

Corresponding Author

Madeline Pe, Ph.D., Quality of Life Department, European Organization for Research and Treatment of Cancer, 83/11 Avenue E. Mounier,1200 Brussels, Belgium; Tel: +32 (0) 2 774 16 61; madeline.pe@eortc.org

Total number of words: 3818

Search strategy and selection criteria

References for this Review were identified through searches of PubMed with the search terms *("patient reported outcome analysis") OR ("("quality of life analysis") AND "cancer" AND "clinical trials"*. No date restrictions were included. Articles were also identified through searches of the authors' own files. Only papers published in English were reviewed. The final reference list was generated on the basis of originality and relevance to the broad scope of this Review.

Abstract (150 words unstructured summary)

Patient-reported outcomes (PROs), such as symptoms, function and other health-related quality of life aspects, are increasingly evaluated in cancer randomized controlled trials (RCTs) to provide information on treatment risks, benefits, and tolerability. However, expert opinion and critical literature review demonstrated no consensus on optimal methods of PRO analysis in cancer RCTs, hindering interpretation of results. The Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data (SISAQOL) Consortium was formed to establish PRO analysis recommendations. Four issues were prioritized: developing a taxonomy of research objectives that can be matched with appropriate statistical methods, identifying appropriate statistical methods for PRO analysis, standardizing statistical terminology related to missing data, and determining appropriate ways to manage missing data. This paper presents PRO analysis recommendations developed through critical literature reviews and a structured collaborative process with diverse international stakeholders, providing a robust foundation for widespread endorsement. Further developments are also discussed.

Introduction

The use of patient-reported outcomes (PRO) in cancer clinical trials allows the patient voice to be incorporated in the evaluation of risks and benefits of cancer therapies. It can also facilitate patient, provider, payer and regulatory decision making [1–3]. Although PROs are now frequently collected in cancer clinical trials, evidence from systematic reviews shows a lack of standards and clear guidelines on how to analyze and interpret PRO data [4–6]. This shortcoming makes it difficult to evaluate conclusions drawn from PRO findings [7]. Although recommendations exist to improve reporting of PROs in protocols (Standard Protocol Items: Recommendations for Interventional Trials-PRO extension; SPIRIT-PRO[8]) and publications (Consolidated Standards of Reporting Trials Statement-PRO extension; CONSORT-PRO[9]), it is critical that reported PRO findings are obtained from good methodological practices and are analyzed consistently across studies to ensure that they can meaningfully and reliably inform patient safety, treatment choices and policy decisions, especially in an era where resources for cancer care are becoming limited and treatment costs are increasing [10]. To address this need, the Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data (SISAQOL) Consortium was formed [7]. The SISAQOL Consortium is a global multi-stakeholder Consortium, involving PRO experts, statisticians, regulators, representatives from international academic societies, industry, cancer institutes and patient organizations.This document presents a set of consensus recommendations for PRO analysis in cancer randomized controlled trials (RCTs) to address four key priorities [11]: (a) developing a taxonomy of research objectives that can be matched with appropriate statistical methods, (b) identifying appropriate statistical methods to address specific PRO research objectives, (c) standardizing statistical terminology related to missing data, and (d) determining appropriate ways of managing missing data.

Development of Recommendations

Described below are key developments that led to the SISAQOL recommendations (see also Figure 1 for an overview).

1. Selection of expert and multi-stakeholder panel

AB and CC, co-authors of this manuscript, invited experts and stakeholders experienced with PROs in cancer RCTs with the goal to form an international, multi-stakeholder consortium. Experts were consulted to recommend colleagues to ensure that SISAQOL is a broad international group representing different disciplines. The idea was described at major events and meetings such as the bi-annual EORTC Quality of Life Group meeting and at international society meetings (e.g., International Society for Quality of Life Research, International Society for Pharmacoeconomics and Outcomes Research, American Society of Clinical Oncology, European Society for Medical Oncology) to secure representatives. When requested, a memorandum of understanding was set-up between EORTC and the international societies. Expertise and profiles of the invited experts at every stage of the development of these recommendations can be found in Appendix page 1.

2. Expert views and systematic reviews

Twenty-six experts and stakeholders attended the SISAQOL kick-off meeting in 2016 to discuss challenges in PRO analysis in cancer RCTs. Agreement was reached on the lack of international standards and that this work was urgently needed [7]. Systematic reviews assessing the current state of PRO analysis in RCTs in different cancer disease sites supported this view [4–6]. Four key findings were highlighted: a lack of specific PRO hypotheses, use of various analysis methods, failure to address the clinical relevance of PRO findings, and ignoring missing data. These findings were also consistent with systematic reviews evaluating inclusion of PROs in protocols [12], and reporting of PROs in publications [13–17].

3. Strategic meeting

Twenty-nine experts and stakeholders attended a strategy meeting in 2017. Based on the evidence gathered, it was agreed that no international standards for PRO analysis in cancer RCTs exist. A core issue was identified: current PRO objectives and hypotheses tend to be broad and uninformative for PRO analysis. As such, the consortium agreed to focus on four key priorities*:*

- Developing a taxonomy of research objectives that can be matched with appropriate statistical methods
- Identifying statistical methods appropriate to address specific PRO research objectives
- Standardizing statistical terminology related to missing data
- Determining appropriate ways to manage missing data

4. Working Groups

Based on the agreed priorities, four working groups were assembled: (1) research objectives, (2) statistical methods, (3) standardization of statistical terms (with an initial focus on defining and evaluating missing data), and (4) management of missing data [11]. Described below are specific goals and methods of each working group. Final outputs from each working group were used as proposed statements for the SISAQOL recommendations. More information describing this process for each working group can be found in Appendix page 2-3.

Research objectives working group. Systematic reviews consistently showed a lack of well-defined PRO research hypotheses in cancer RCTs [5,6,12,15,17]. A well-defined PRO hypothesis should clearly align with the objectives of the study and provide a clear understanding of what needs to be estimated from the PRO data, which can then inform appropriate analysis decisions. Research objectives working group members were tasked with developing a framework for PRO research objectives that can inform the statistical method to use (taxonomy of PRO research objectives), and to provide standardized definitions for key PRO objectives. An initial framework was developed through discussions. The framework was circulated to all research objectives working group members for further refinement. A survey was conducted among the working group members to standardize definitions of key research PRO objectives: improvement, worsening and stable state (Appendix pages 4-12 for survey results).

Statistical methods working group. Findings from systematic reviews demonstrated that there is no consensus on appropriate statistical methods for PRO data analysis [4–6]. Moreover, there is no single analysis method that can address all clinical, trial design and analytical concerns. It was agreed that having set criteria to evaluate statistical methods for PRO analysis would be critical to allow the choice to be more scientifically informed [11].

A list of 19 statistical criteria was developed through literature search and expert discussions. A survey was conducted among the statistical methods working group members, in which they rated each proposed statistical criterion as “essential,” “desirable,” or “non-essential” for analysis of PROs in cancer clinical trials. An open-ended question was also included to capture additional criteria. Survey results were discussed and the set of criteria was updated until all individual concerns were addressed (Appendix pages 13-15 for survey results).

The agreed set of statistical criteria was used by the statistical methods working group to evaluate the initial list of statistical methods identified in the metastatic breast cancer systematic review [5]. A draft report on the evaluation of statistical methods was circulated and reviewed by the statistical methods working group members (see Appendix pages 16-26 for detailed results of this report). Recommended methods for each PRO objective were discussed and amended until all individual concerns from working group members were addressed.

Standardizing statistical terms working group (focus on defining and evaluating missing data). Missing PRO data is the on-going challenge in cancer clinical trials, as patients drop out of study for different reasons, including (predefined) progression of disease, death, intolerable toxicity, and patient or clinician decision [18–20]. In order to evaluate the extent of missing data, missing data rates should be reported in a standardized way since PRO estimates may be biased if a large number of patients fail to complete the PRO assessments [21]. However, the very definition of “missing data” remains opaque and elusive. For example, it is unclear whether unobserved assessments after a patient drops out of a study because of disease progression is truly missing data if administration is not expected per the protocol test schedule. Therefore, the aim of this working group was to standardize the definition of missing data and the reporting of missing data rates; and to clarify their relationship with the PRO study population (i.e., all patients who consented and were eligible to participate in the PRO data collection), and PRO analysis population (i.e., patients that will be included in the primary PRO analysis). A first set of definitions/calculations for missing data rates was extracted from a systematic review of metastatic breast cancer RCTs [5]. An exploratory literature search in additional peer-reviewed publications was conducted to identify other definitions of missing data and approaches to calculate missing data rates. Consortium members responded to a survey to standardize these definitions (Appendix pages 27-29 for survey results). Findings were discussed and iteratively refined until all individual concerns from the working group were addressed.

Missing data working group. The missing data working group was tasked with identifying whether it was possible to set a threshold for acceptable rates of missing data based on simulation studies (how much missing data is too much?); develop a standardized case report form (CRF) to identify reasons for non-completion of PROs; recommend a general strategy for managing missing data; and test a set of macros for various missing data settings for sensitivity analysis.

Monte Carlo simulations were performed to assess how increasing missing data rates impact bias and power in a typical RCT. The simulation results were planned as the basis for later recommendations on thresholds for missing data[22]

In an effort to develop a standardized CRF with possible reasons for PRO non-completion, existing CRF templates from seven different clinical trial networks were collected (e.g., the CRF from the Alliance for Clinical Trials in Oncology was previously published[23]). An initial list of 27 reasons for PRO non-completion was compiled. A survey was conducted among all consortium members, where members indicated whether the reason for non-completion (a) should be included in the standard CRF, (b) is related to the patient's health, and (c) affects data quality (Appendix pages 30-31 for survey results).

5. SISAQOL recommendations meeting

Thirty-one experts and stakeholders attended the SISAQOL recommendations meeting in 2018. The meeting aimed to ratify the statements proposed by the different working groups. The meeting was divided into four sessions, representing each working group: (1) taxonomy of research objectives; (2) recommending statistical methods; (3) standardizing terminology related to missing data; and (4) managing missing data.

For each statement, participants voted either to agree, disagree, or abstain. A proposed statement was *ratified* if at least two-thirds of the voters agreed on the statement. A statement was *rejected* if less than half of the voters agreed on the statement. A statement was *postponed* or *for discussion* if it did not meet the agreement or rejection criteria, or if it was agreed by the consortium that more discussion was needed. A statement was *cancelled* if it was conditional on the ratification of a previous statement, and the previous statement was not ratified. Participants who abstained or did not vote for a specific statement were not included in the total number of voters.

SISAQOL recommendations and their considerations are presented in Table 1. A brief overview is presented in Table 2. Statements that were not ratified, including reasons for non-ratification, can be found in Appendix pages 35-36. A brief summary of the recommendations for each section is described below.

SISAQOL recommendations

Forty-three statements were presented at the recommendations meeting, of which 32 were ratified (32/43; 74%), 8 were postponed, (8/43; 19%), 1 was rejected (1/43; 2%) and 2 were cancelled (2/43; 5%). Appendix pages 37- 40 (Table 2) shows the voting results of all proposed statements.

Section 1: Taxonomy of research objectives

All proposed statements from the research objectives working group were ratified (9/9; 100%). A taxonomy of PRO research objectives for cancer RCTs was recommended. The framework is intended to aid the development of well-defined PRO objectives that can be matched with appropriate statistical methods. An overview of this framework can be found in Table 2.

When developing a PRO objective, the Consortium concluded that the PRO domain(s) and time frame of interest should be pre-specified [24,8]. Critically, four key attributes need to be considered *a priori* for each PRO domain:

- Broad PRO research objective: treatment efficacy / clinical benefit (confirmatory), or describe patient perspective (exploratory / descriptive)
- Between-arm PRO objective: superiority or equivalence / non-inferiority
- Within-treatment group PRO assumption for the treatment or control arm: worsening, stable state, improvement or overall effect
- Within-patient/within-treatment PRO objective: time to event, magnitude of event at time *t*, proportion of responders at time *t*, overall PRO score over time or response patterns/profiles

Considerations for each attribute are found in Table 1, RS 1-5. Recommended standardized definitions of improvement, stable state, worsening, and overall effects were ratified (see Table 1, RS 6-9). Sample illustrations of the recommended definitions of improvement, stable state and worsening can also be found in Figure 2.

Section 2: Recommended statistical methods

The majority of the proposed statements for this section were ratified (6/7; 86%). A set of essential and highly desirable statistical criteria for defining appropriate statistical methods for PRO analysis was recommended. If a statistical method did not satisfy an essential criterion, then the method was not recommended as appropriate for PRO analysis.

Two essential statistical properties were identified: the ability to perform a comparative test (statistical significance) and the ability to produce interpretable treatment effect estimates (clinical relevance). Highly desirable criteria included: the ability to adjust for covariates, including baseline PRO score, handling missing data with the least restrictions, and handling clustered data (repeated assessments). More information on these criteria can be found on Table 1 (RS 10). When two or more statistical methods fit the essential and highly desirable criteria equally, the simpler method was prioritized. Although there may be advantages in recommending more complex models for specific purposes (e.g., pattern mixture models), this often comes at the cost of strong and untestable assumptions and can produce results that may not be easily interpreted by non-statisticians. A balance between feasibility, usefulness, interpretability and statistical correctness was determined to be critical for the primary PRO analysis; however, more complex models can be deployed as sensitivity analysis to test the robustness of the primary result.

Based on the agreed set of statistical criteria and selection criteria, statistical methods were recommended for each PRO objective. Two statistical methods were recommended: (a) Cox proportional hazards for time to event PRO objectives (Table 1, RS 11), and (b) linear mixed models for magnitude of event at time *t* (Table 1, RS 12) and response patterns/profiles (Table 1, RS 15). In exceptional cases where the PRO design only required baseline and one follow-up assessment, linear regression was recommended as the appropriate statistical method (Table 1, RS 13).

Notably, because clinical relevance was agreed to be an essential criterion for PRO interpretation, parametric methods were recommended over non-parametric methods. However, parametric methods have limitations, most importantly, they rely on distributional assumptions [25]. To address this limitation, it was recommended that non-parametric methods be used for sensitivity analyses to investigate deviations from these assumptions [25].

No agreement was reached on appropriate statistical methods to evaluate longitudinal data for proportion of responders, prompting further discussions. Also, no agreement was reached on recommended summary measures for PRO data over time (e.g., min/max, AUC, overall means), but it was recognized that summary measures should be part of SISAQOL's future work (Table 1, RS 14). Whether it is appropriate to analyze ordinal data as continuous needs further investigation; discussions on this issue revolved around statistical approximation, complexity of the model, and ease of interpretation.

Section 3: Standardizing Terminology related to Missing Data

The majority of the proposed statements for this section were ratified (8/11; 73%). A recommendation on the definition of missing PRO data was proposed: missing PRO data is defined as 'data that would be meaningful for the analysis of a given research objective, but were not collected (Table 1, RS 16-17) [26,27]. This definition implies that not all unobserved assessments are considered as missing data depending on the scientific question (e.g., unobserved assessments after death; unobserved assessments off-treatment if the PRO objective focuses on on-treatment patients; or unobserved assessments after the PRO objective has been reached). However, depending on the analysis method, all unobserved assessments may implicitly be treated similarly as missing data [28]. Recommendations on how to specifically deal with missing data for each recommended method is the next step for the SISAQOL work.

The current document stresses the importance of differentiating missing observations in relation to a reference set of expected data (see Table 1, RS 19-22). The discussion resulted in two definitions: 1) The 'available data rate' has a fixed denominator, the number of patients in the PRO study population (i.e. all patients who consented and were eligible to participate in the PRO data collection at baseline). 2) The 'completion rate' has a variable denominator, the number of patients on PRO assessments at the designated time point (i.e. all patients who are still expected to provide PRO assessments at that time point). The numerator of both rates are the number of patients on PRO assessment submitting a valid PRO assessment at the designated time point. Of note, the denominator of the 'completion rate' depends on the chosen research question, e.g. whether PROs should be collected only up to progression or also after progression. It was recommended that patients who died are excluded from the denominator of the 'completion' rate at assessment points after death. However, these patients are included in the denominator of the available data rate as that rate always refers to a fixed set of patients at baseline (see Table 1, RS 18).

Section 4: Missing Data

More than half of the proposed statements were ratified in this section (9/16; 56%). A simulation study was conducted to assess whether it was possible to have a threshold to define *substantial* missing data[22]. Although no agreement was reached for a threshold, the simulation study showed that impact of missing data rates on PRO findings depends on the type of missing data (i.e., informative or non-informative missing data). It was recommended that collecting reasons for missing data is key in assessing the impact of missing data on PRO findings (see Table 1, RS 24; [20]. A case report form to collect reasons for missing data in a standardized way is needed and will be further developed. General recommendations on how to handle missing data were proposed consistent with existing regulatory guidelines (see Table 1, RS 25 - 30).

Discussion

The aim of SISAQOL is to develop a set of recommendations to facilitate standard approaches for PRO analysis in cancer RCTs. Through critical literature reviews and discussions with international experts and stakeholders, SISAQOL provides a framework of well-defined PRO research objectives matched with appropriate statistical method(s) (see Table 2). The Cox proportional hazards model was recommended as an appropriate analysis method for time-to-event outcomes. The linear mixed model was recommended for the analyses of magnitude of event at time *t*, and response patterns/profiles. Recommendations on a standardized definition of missing PRO data, completion rates and available data rates were proposed, with corresponding standardized calculation and reporting. Some general recommendations for managing missing PRO data were also suggested.

Generating robust PRO conclusions from cancer clinical trials is not only about agreeing on and using standardized research objectives and analysis standards. It also entails thoughtful trial planning and design with meaningful involvement of patient representatives from the beginning of the process, high-quality data collection and transparent reporting of results. We believe this set of recommendations will support clinical researchers, trialists and statisticians to improve the conceptualization and design of PRO studies, the quality of statistical analysis and the clinical interpretation of PROs in cancer clinical trials. SISAQOL adds to a growing toolbox of methodological recommendations on best practices for PRO in cancer trials, including Standard Protocol Items: Recommendation for Interventional Trials in Patient Reported Outcomes (SPIRIT-PRO) [8], the Consolidated Standards of Reporting Trials in Patient Reported Outcomes (CONSORT-PRO) [9], and other relevant guidelines [29,30]. Whereas SPIRIT-PRO and CONSORT-PRO recommendations focus on good, high quality reporting for both the protocol and final report, allowing readers to judge the robustness of the design, analysis and interpretation of the PRO endpoint, SISAQOL recommendations focus on improving the quality of PRO design and analysis. Good quality reporting and good methodology are not interchangeable. The overarching goal is to improve both reporting and methodology in PROs in clinical trials.

Given the substantial need for safe and effective cancer therapeutics, and the cost and complexity of cancer clinical trials, it is critical that clinical and healthcare policy decisions made by regulators, payers, clinicians, and patients and their families are based on robust scientifically sound international standards and the limited research resources are not wasted[10].

Limitations and Future Work

The standards for PRO analysis have some limitations. First, we focused on cancer RCTs; while many issues may generalize to other health conditions, this warrants further scrutiny. Another limitation relates to the relevance of these standards to preference weighted measures of HRQOL, also called preference-based measures, multi-attribute utility measures. Such measures can be used for two purposes: 1) as utility scores which represent a special type of HRQOL summary score, i.e. with domains of HRQOL weighted by preferences, usually the general population’s preferences but sometimes patient preferences; 2) as quality weightings in QALYs and cost-utility analysis. Whether the standards reported in this paper apply for any of these purposes need to be further discussed.

Much work still needs to be done to further finesse these standards for cancer RCTs. First, several proposed statements were not agreed upon and will need more discussion (e.g., statistical method for proportions of patients at time *t*, summary measures and several issues on missing data; see Appendix pages 35-36 for more details). Second, the taxonomy of research objectives needs to be applied in future cancer clinical trials to evaluate whether they are fit-for-purpose when planning trials with a PRO endpoint, with further revisions made if necessary. Third, the choice of statistical methods to be evaluated for each PRO objective was largely based on commonly used statistical methods for PRO analysis found in systematic reviews. Although consortium members had opportunities to suggest other methods, there may be additional appropriate statistical methods for PRO analysis in the evaluation that were missed. Nonetheless, the set of statistical methods evaluated are time-tested and scientifically rigorous and can be applied in the majority of the cases. Fourth, best statistical practices for each of the recommended methods need to be agreed upon, including how to handle missing data. Fifth, an agreement on which summary measures are relevant to address specific PRO objectives is also needed. In addition to working on the identified limitations, future steps would include identifying the target population and intercurrent events relevant for PRO analysis. Finally, how these recommendations relate to the recently suggested estimands framework [27] is yet to be examined.

Conclusion

Patient-reported outcome (PRO) data, such as symptoms, functioning and other HRQOL endpoints are increasingly assessed in cancer RCTs to provide valuable evidence on risks, benefits, safety and tolerability of treatment. PRO findings inform patients, providers, payers and regulatory decision-makers. For these reasons, it is imperative that PRO findings are robust and derived consistently across studies to yield meaningful results. The current SISAQOL recommendations represent an important first step towards generating international consensus-based standards for PRO analysis in cancer RCTs.

References

1 Vodicka E, Kim K, Devine EB, Gnanasakthy A, Scoggins JF, Patrick DL. Inclusion of patient-reported outcome measures in registered clinical trials: Evidence from ClinicalTrials.gov (2007-2013). *Contemp Clin Trials* 2015; **43**: 1–9.

2 Basch E, Geoghegan C, Coons SJ, *et al.* Patient-Reported Outcomes in Cancer Drug Development and US Regulatory Review: Perspectives From Industry, the Food and Drug Administration, and the Patient. *JAMA Oncol* 2015; **1**: 375–9.

3 Kluetz PG, O'Connor DJ, Soltys K. Incorporating the patient experience into regulatory decision making in the USA, Europe, and Canada. Lancet Oncol. 2018; **19**: e267–74.

4 Hamel J-F, Saulnier P, Pe M, *et al.* A systematic review of the quality of statistical methods employed for analysing quality of life data in cancer randomised controlled trials. *Eur J Cancer* 2017; **83**: 166–76.

5 Pe M, Dorme L, Coens C, *et al.* Statistical analysis of patient-reported outcome data in randomised controlled trials of locally advanced and metastatic breast cancer: a systematic review. *Lancet Oncol* 2018; **19**: e459–69.

6 Fiteni F, Anota A, Westeel V, Bonnetain F. Methodology of health-related quality of life analysis in phase III advanced non-small-cell lung cancer clinical trials: a critical review. *BMC Cancer* 2016; **16**: 122.

7 Bottomley A, Pe M, Sloan J, *et al.* Analysing data from patient-reported outcome and quality of life endpoints for cancer clinical trials: a start in setting international standards. *Lancet Oncol* 2016; **17**: e510–4.

8 Calvert M, Kyte D, Mercieca-Bebber R, Slade A, Chan AW, King MT. Guidelines for inclusion of patient-reported outcomes in clinical trial protocols the spirit-pro extension. *JAMA* 2018; **319**: 483–94.

9 Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. *JAMA.* 2013; **309**: 814–22.

10 Chalmers I, Bracken MB, Djulbegovic B, *et al.* How to increase value and reduce waste when research priorities are set. *Lancet* 2014; **383**: 156–65.

11 Bottomley A, Pe M, Sloan J, *et al.* Moving forward toward standardizing analysis of quality of life data in randomized cancer clinical trials. *Clin Trials* 2018; **15**: 624–30.

12 Kyte D, Duffy H, Fletcher B, *et al.* Systematic evaluation of the patient-reported outcome (PRO) content of clinical trial protocols. *PLoS One* 2014; **9**: e110229.

13 Fielding S, Ogbuagu A, Sivasubramaniam S, MacLennan G, Ramsay CR. Reporting and dealing with missing quality of life data in RCTs: has the picture changed in the last decade? *Qual. Life Res.* 2016; **25**: 2977–83.

14 Efficace F, Fayers P, Pusic A, *et al.* Quality of patient-reported outcome reporting across cancer randomized controlled trials according to the CONSORT patient-reported outcome extension: A pooled analysis of 557 trials. *Cancer* 2015; **121**: 3335–42.

15 Brundage M, Bass B, Davidson J, *et al.* Patterns of reporting health-related quality of life outcomes in randomized clinical trials: Implications for clinicians and quality of life researchers. *Qual Life Res* 2011; **20**: 653–64.

16 Mercieca-Bebber, R., Friedlander M, Stockler M, Calvert M, *et al.* A systematic

evaluation of compliance and reporting of patient-reported outcome endpoints in ovarian cancer randomised controlled trials: implications for generalisability and clinical practice. *J Patient-Reported Outcomes* 2017; **1:5**. DOI:10.1186/s41687-017-0008-3.

17 Kyte D, Retzer A, Ahmed K, *et al.* Systematic evaluation of Patient-Reported Outcome protocol content and reporting in cancer trials. *JNCI J Natl Cancer Inst* 2019. https://doi.org/10.1093/jnci/djz038.

18 Bell M, Fairclough D. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Stat Methods Med Res* 2014; **23**: 440–59.

19 Fayers P, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. John Wiley & Sons, 2013.

20 Palmer MJ, Mercieca-Bebber R, King M, Calvert M, Richardson H, Brundage M. A systematic review and development of a classification framework for factors associated with missing patient-reported outcome data. *Clin Trials* 2018; **15**: 95–106.

21 Machin D, Weeden S. Suggestions for the presentation of quality of life data from clinical trials. *Stat Med* 2002; **17**: 711–24.

22 Mazza G, Pe M, Dorme L, *et al.* How Much Missing Data is Too Much? Monte Carlo Simulations to Develop SISAQOL Guidelines for Missing Data Handling. 25th Annual Conference of the International Society for Quality of Life Research, Dublin, Ireland October 2018. *Qual Life Res* 2018; **27**: ab208.1, p:43.

23 Atherton PJ, Burger KN, Pederson LD, Kaggal S, Sloan JA. Patient-reported outcomes questionnaire compliance in Cancer Cooperative Group Trials (Alliance N0992). *Clin Trials* 2016; **13**: 612–20.

24 Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of Patient-Reported Outcomes in Randomized Trials. *JAMA* 2013; **309**: 814–22.

25 Altman DG, Bland JM. Parametric v non-parametric methods for data analysis. *BMJ.* 2009; **339**: 170.

26 Little RJ, Ph D, Agostino RD, *et al.* The Prevention and Treatment of Missing Data in Clinical Trials. *N Engl J Med* 2012; **367**: 1355–60.

27 ICH Expert Working Group. International council for harmonisation of technical requirements for pharmaceuticals for human use: Estimands and sensitivity analysis in clinical trials E9(R1). 2017; **9**. [Available at https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9-R1EWG_Step2_Guideline_2017_0616.pdf].

28 Kurland BF, Johnson LL, Egleston BL, Diehr PH. Longitudinal Data with Follow-up Truncated by Death: Match the Analysis Method to Research Aims. *Stat Sci* 2010; **24**: 211–22.

29 European Medicines Agency. Appendix 2 to the Guideline on the Evaluation of Anticancer Medicinal Products in Man: The use of patient-reported outcome (PRO) measures in oncology studies. 2016. [Available at http://www.ema.europa.eu/docs/en_GB/document_library/Other/2016/04/WC500205159.pdf].

30 US Dep Heal Hum Serv Food Drug Adm. Guidance for Industry Patient Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. [Available at: http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf. 2009].

31 ICH Expert Working Group. ICH Harmonised Tripartite: Guideline Statistical Principles for Clinical Trials E9. 1998. [Available at https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf]

32 Tukey JW. We need both exploratory and confirmatory. *Am Stat* 1980; **34**: 23–5.

33 Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW. Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. *J Am Med Assoc* 2006; **295**: 1152–60.

34 European Medicines Agency. Guideline on the choice of non-inferiority margin. 2008. 1–9. [Available at https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-choice-non-inferiority-margin_en.pdf]

35 European Medicines Agency. Points to consider on switching between superiority and non-inferiority. 2000. [Available at https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-switching-between-superiority-non-inferiority_en.pdf]

36 King MT. A point of minimal important difference.pdf. *Expert Rev Pharmacoeconomics Outcomes Res* 2011; **11**: 171–84.

37 Fairclough DL. Summary measures and statistics for comparison of quality of life in a clinical trial of cancer therapy. *Stat Med* 1997; **16**: 1197–209.

38 Curran D, Aaronson N, Standaert B, *et al.* Summary measures and statistics in the analysis of quality of life data: An example from an EORTC-NCIC-SAKK locally advanced breast cancer study. *Eur J Cancer* 2000; **36**: 834–44.

39 National Cancer Institute. NCI Dictionary of Cancer Terms. Natl. Cancer Inst. 2013. [Available at https://www.cancer.gov/publications/dictionaries/cancer-terms].

40 Wasserstein RL, Nicole AL. The ASA Statement on P -Values: Context, Process and Purpose. *The American Statistician* 2016; ***70:** 129-133.* DOI:10.1080/00031305.2016.1154108.

41 European Medicines Agency. Points To Consider on Adjustment for Baseline Covariates. 2003. [Available at https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf]

42 Senn SJ. Covariate imbalance and random allocation in clinical trials. *Stat Med* 1989; **8**: 467–75.

43 Vickers AJ, Altman DG. Statistics Notes: Analysing controlled trials with baseline and follow up measurements. *BMJ* 2002; **323**: 1123–4.

44 Fairclough DL, Peterson HF, Cella D, Bonomi P. Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. *Stat Med* 1998; **17**: 781–96.

45 Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical Analysis of Quality of Life With Missing Data in Cancer Clinical Trials. *Stat Med* 1998; **17**: 653–66.

46 Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis, 2nd edition. John Wiley & Sons, 2011. DOI:10.1198/jasa.2005.s24.

47 Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *Br J Cancer* 2003; **89**: 431–6.

48 Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. *Br J Cancer* 2003; **89**: 232–8.

49 Cnaan A, Laird NM, Slasor P. Tutorial in biostatistics: Using the General Linear

Mixed Model to Analyse Unbalanced Repeated Measures and Longitudinal Data. *Stat Med* 1997; **16**: 2349–80.

50 SAS. The REG Procedure: Missing Values. SAS/STAT(R) 9.3 User's Guid. Available at [https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_reg_sect026.htm].

51 Bell ML, King MT, Fairclough DL. Bias in Area Under the Curve for Longitudinal Clinical Trials With Missing Patient Reported Outcome Data. *SAGE Open* 2014; **4**: doi.org/10.1177/2158244014534858.

52 Fairclough DL, Peterson HF, Chang V. Why are missing quality of life data a problem in clinical trials of cancer therapy? *Stat Med* 2002; **17**: 667–77.

53 Mercieca-Bebber R, Palmer MJ, Brundage M, Calvert M, Stockler MR, King MT. Design, implementation and reporting strategies to reduce the instance and impact of missing patient-reported outcome (PRO) data: A systematic review. *BMJ Open* 2016; **6**. doi:10.1136/bmjopen-2015-010938.

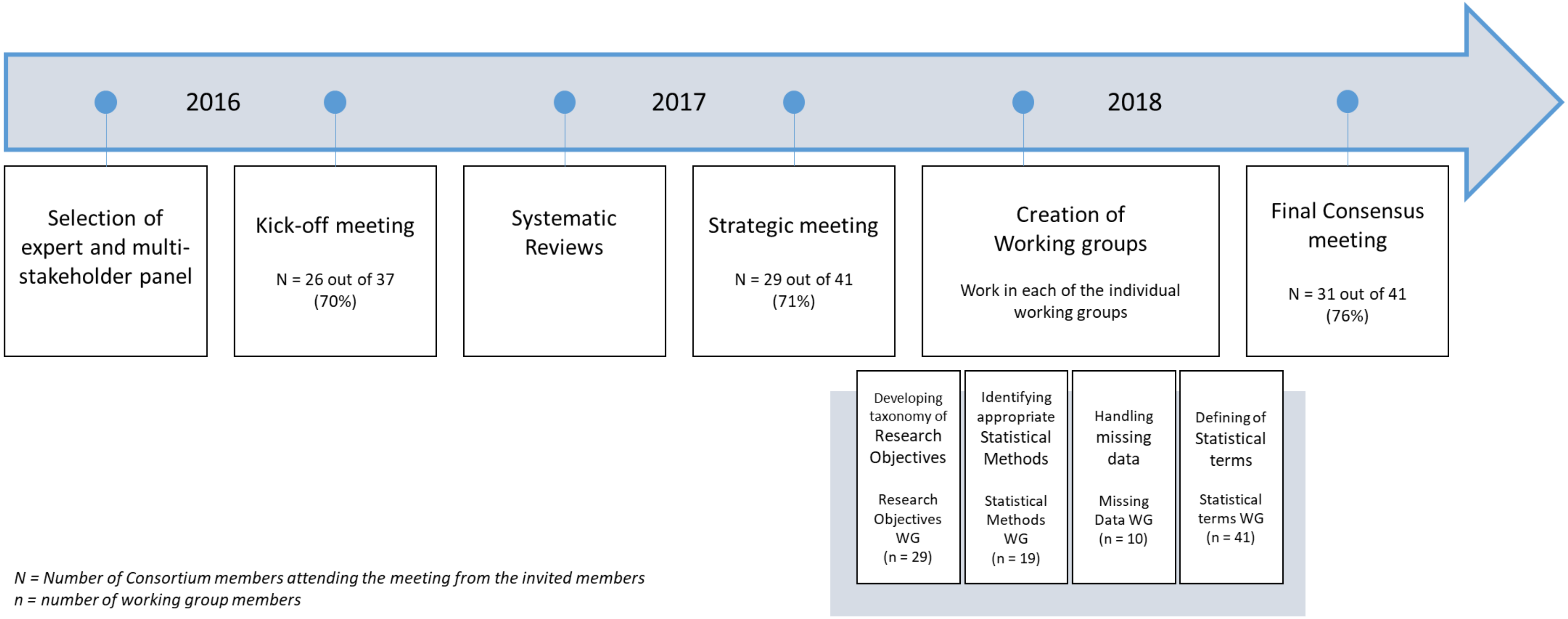
54 European Medicines Agency. Guideline on Missing Data in Confirmatory Clinical Trials Guideline on Missing Data in Confirmatory Clinical Trials. 2011. [Available at https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-missing-data-confirmatory-clinical-trials_en.pdf].

55 Fielding S, Fayers P, Ramsay CR. Analysing randomised controlled trials with missing data: Choice of approach affects conclusions. *Contemp. Clin. Trials.* 2012; **33**: 461–9.

56 Fayers PM, Curran D, Machin D. Incomplete quality of life data in randomized trials: missing forms. *Stat Med* 1998; **17**: 679–96.

57 Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: From design to analysis. *Yale J. Biol. Med.* 2013; **86**: 343–58.

58 Liu G, Gould AL. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *J Biopharm Stat* 2002; **12**: 207–26.


2016
2017
2018
Selection of expert and multi-stakeholder panel
Kick-off meeting
N = 26 out of 37 (70%)
Systematic Reviews
Strategic meeting
N = 29 out of 41 (71%)
Creation of Working groups
Work in each of the individual working groups
Final Consensus meeting
N = 31 out of 41 (76%)
Developing taxonomy of Research Objectives
Research Objectives WG (n = 29)
Identifying appropriate Statistical Methods
Statistical Methods WG (n = 19)
Handling missing data
Missing Data WG (n = 10)
Defining of Statistical terms
Statistical terms WG (n = 41)

N = Number of Consortium members attending the meeting from the invited members
n = number of working group members

685 *Figure 1.* Overview of the development of the SISAQOL Recommendations. Non-attendees received the full meeting reports and could comment and add suggestions.The final version of the report was approved by the Consortium.

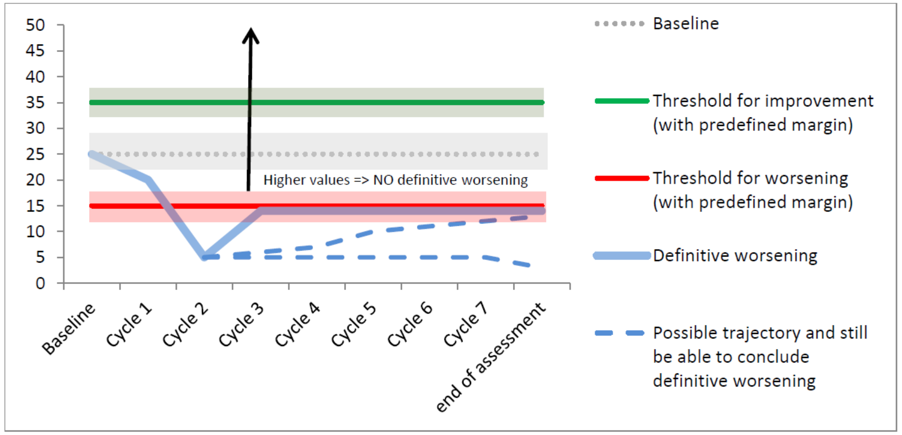


Figure 2a. Worsening is defined as change from baseline that reaches a pre-defined worsening threshold level (post-baseline worsening). Worsening is maintained if follow-up assessments remain at or are lower than the worsening threshold (definitive worsening). Worsening is discontinued once a follow-up assessment is above the worsening threshold (transient worsening). See also RS 7.

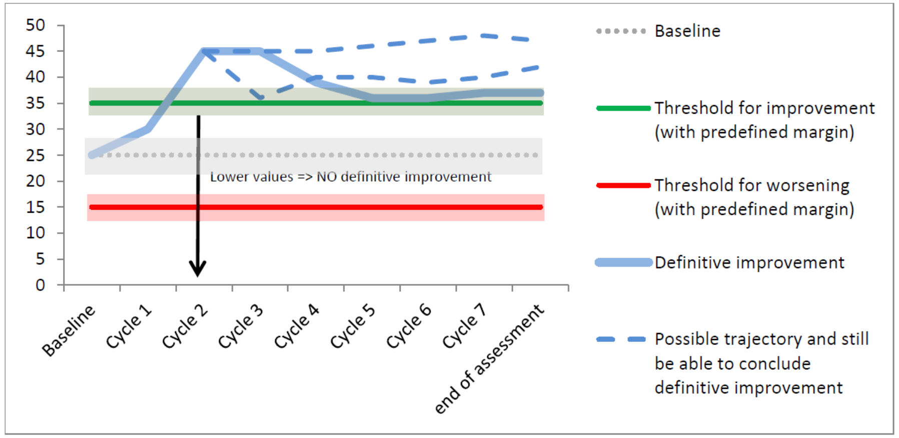


Figure 2b. Improvement is defined as change from baseline that reaches a pre-defined improvement threshold level (post-baseline improvement). Improvement is maintained if follow-up assessments remain at or are higher than the improvement threshold (definitive improvement). Improvement is discontinued once a follow-up assessment is below the improvement threshold (transient improvement). See also RS 6.

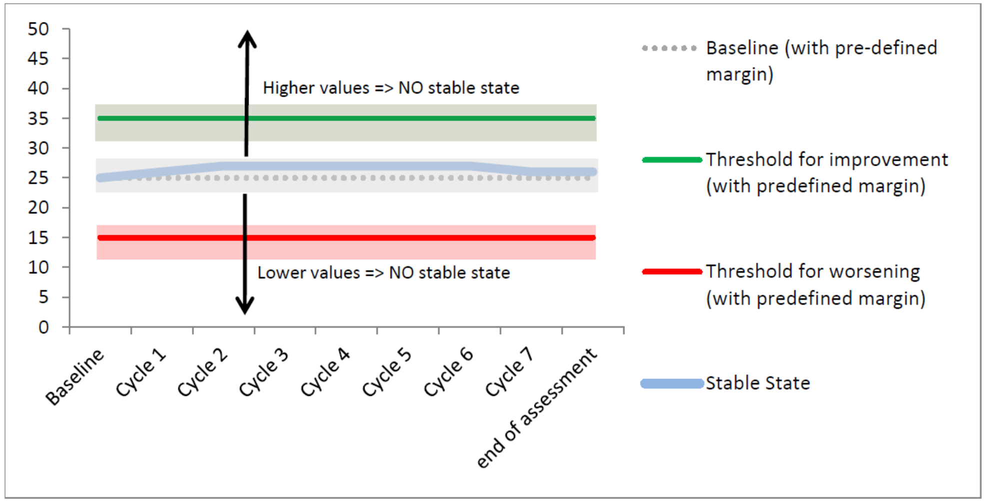


Figure 2c. Stable state is defined as no change from baseline is observed, or change from baseline is within the pre-defined baseline margin. This stable state is maintained if follow-up assessments remain at the baseline pre-defined margin. The stable state is discontinued once the follow-up assessment leaves the pre-defined baseline margin (and reaches the improvement or worsening threshold). There may be circumstances where the relevant PRO objective would include improvement in the definition of stable state (i.e., at least stable). In this case, the definition is as long as follow-up assessments do not reach the deterioration threshold, then stable state can still be concluded. See also RS 8.

Table 1. SISAQOL recommended statements and their considerations

Section 1: Taxonomy of Research Objectives		
Statement No.	Recommended statement (RS)	Considerations
RS 1	Clearly state the broad PRO research objectives for each PRO domain(s)/item(s) of interest: - Treatment efficacy / clinical benefit, - Exploratory / describe patient perspective	*Treatment efficacy / clinical benefit:* If a PRO domain will be used to provide formal comparative conclusions between treatment arms, then the rules for a confirmatory objective are followed: an *a-priori* hypothesis is needed for each PRO domain, which will then be statistically tested at the end of the trial [31]. If multiple PRO domains or multiple assessment points of a PRO domain are of interest, then correction for multiple testing is needed. Components for a well-defined *a priori* PRO hypothesis are detailed in the subsequent recommended statements (see RS 2 to 5). *Exploratory / describe patient perspective:* If a PRO domain will be used to describe the patient perspective during the trial or to explore the PRO data and use its findings to inform future studies, then the rules for descriptive/exploratory objective is followed: an a-priori hypothesis is not required for the PRO domain. However, these outcomes cannot be used to draw comparative conclusions or used as support for treatment efficacy/clinical benefit. Findings should be reported as either descriptive (i.e., summarizing estimates with or without confidence intervals but no statistical testing is involved), or exploratory (i.e., choice of hypothesis may be data-driven and statistical testing may be involved, but this should not be used a basis of evidence of clinical benefit / treatment efficacy [31]. Both PRO objectives are important and complement each other [32]; and can be included together within a trial. However, the protocol should clearly specify which PRO domains will be used to provide evidence of treatment efficacy/clinical benefit, describe the patient perspective or are exploratory.
RS 2	Clearly state the between treatment-arm	Superiority design and analysis techniques differ from equivalence /

	comparison that will be used for each PRO domain/item of interest: - Superiority, - Equivalence / non-inferiority	non-inferiority techniques [31,33]. Non-significant *p*-values from a statistical test aimed to assess treatment difference (superiority test) should not be used as evidence that the two treatment arms are "similar" (equivalent) or "not worse" (non-inferior). *Superiority:* A superiority PRO objective aims to show that for the pre-specified PRO domain, the treatment arm is superior to the reference arm by a clinically relevant treatment effect size. The effect size to demonstrate a clinically relevant treatment difference should be pre-defined in the protocol. The trial should be designed as to allow unbiased and adequately powered testing for the rejection of the hypothesis of no treatment effect. [31,34,35]. *Equivalence / non-inferiority:* An equivalence/non-inferiority PRO objective aims to show that for the pre-specified PRO domain, the treatment arm is similar (equivalent) or not worse than (non-inferior) the reference arm by a pre-specified clinically relevant margin. It is critical that these margins are pre-specified in the protocol. The trial should be designed as to allow unbiased and adequately powered testing for the rejection of the hypothesis of non-equivalence / inferior treatment effect [34]. The choice of effect size (superiority) and margins (equivalence / non-inferiority) should be tailored to the PRO instrument and clinical context; and should be justified on both clinical and statistical grounds [34]. Trials may include any combination of these between-treatment arm PRO objectives. However the protocol should clearly specify which PRO domain(s)/item(s) will be tested for superiority or equivalence / non-inferiority.
RS 3	Clearly state the within-patient/within-treatment arm PRO objective in protocol. Valid within-individual/within-group PRO	**Within-treatment arm PRO assumption: improvement, worsening, stable state or overall effect.** The choice of whether a worsening, stable state or improvement is

	objectives are: - Improvement: o time to improvement, o magnitude of improvement at time *t*, o proportion of responders with improvement at time *t*, - Worsening: o time to worsening, o magnitude of worsening at time *t*, o proportion of responders with worsening at time *t*, - Stable state: o time to [end of] stable state, o proportion of responders with stable state at time *t*,	expected within the treatment group should be based on previous literature, expert knowledge or early phase trials. It is also possible that the interest for the within-treatment group is not on a specific direction of the effect, but rather on an overall effect (i.e., summarizing all available scores over time for each patient on a specific PRO domain). However caution should be noted that for overall effects, since there is no *a priori* within-treatment group assumption, the conclusions drawn may be less robust. When deciding which within-treatment arm PRO assumption will be used, patients' observed baseline levels on the specific PRO domain should be taken into account; this will help inform the feasibility of assessing a clinically relevant change for that PRO domain. **Within-patient/within-treatment PRO objective: time to event, magnitude of event at time *t*, proportion of responders at time *t*, overall PRO score over time or response patterns/profiles** Various within-patient/within-treatment arm PRO endpoints are possible, however these are often ignored and erroneously interpreted as synonymous. For example, a PRO endpoint examining "time to first worsening while on treatment" is not equivalent to the endpoint "magnitude of worsening at 6 weeks". In fact, these PRO endpoints will use different analytical techniques and may yield different conclusions. Depending on the endpoint, the clinically relevant threshold for the PRO domain may be at the patient-level (e.g., within-patient: classifying a patient as a responder or not), or at the group level (e.g., within-group; mean change within the group) [36]. *Within-patient PRO objective:* The primary interest is in identifying which patients had a clinically relevant response before performing further analysis. The clinically relevant threshold is specified at the
RS 4	Valid within-patient/within-treatment arm PRO objectives is: - Overall effects: o overall PRO score over time	
RS 5	Valid within-patient/within-treatment arm PRO objectives is: - Overall effects: o Response patterns/profiles	

		individual level (i.e., responder definition), which identifies which patients had a clinically relevant change or not. This objective is linked to endpoints such as time to event or proportion of responders. *Within-treatment arm PRO objective:* The primary interest is in evaluating whether on average the specified group had a clinically relevant change. The clinically relevant threshold is specified at the group level which identifies whether the group had a clinically relevant change or not. This objective is linked to endpoints such as magnitude of change. RS 6 to 9 provide more specific definitions for these PRO objectives.
RS 6	*Improvement* is defined as change from baseline that reaches a pre-defined improvement threshold level (post-baseline improvement). Improvement is maintained if follow-up assessments remain at or are higher than the improvement threshold (definitive improvement). Improvement is discontinued once a follow-up assessment is below the improvement threshold (transient improvement). See Figure 2 for illustration.	*Time to improvement:* A clinically relevant within-patient level improvement is pre-defined, and the interest is in evaluating the time it takes before a clinically relevant improvement is observed. Variability in the scores above or below this pre-defined improvement threshold is ignored. *Magnitude of improvement at time t:* A clinically relevant within-treatment arm improvement is pre-defined, and the interest is in assessing the mean/median improvement (with corresponding confidence intervals) at a pre-defined, clinically relevant time point. Variability in the observed scores are taken into account. *Proportion of responders with improvement at time t:* A clinically relevant within-patient level improvement is pre-defined, and the interest is in evaluating the number of patients with improvement at a pre-defined clinically relevant time point. Variability in the scores above or below this pre-defined improvement threshold is ignored.
RS 7	*Worsening* is defined as change from baseline that reaches a pre-defined worsening threshold level (post-baseline	*Time to worsening:* A clinically relevant within-patient level worsening is pre-defined, and the interest is in evaluating the time it takes before a clinically relevant worsening is observed. Variability in the

	worsening). This worsening is maintained if follow-up assessments remain at or are lower than the worsening threshold (definitive worsening). Worsening is discontinued once a follow-up assessment is above the worsening threshold. See Figure 2 for illustration.	scores above or below this pre-defined worsening threshold is ignored. *Magnitude of worsening at time t:* A clinically relevant within-treatment arm worsening is pre-defined, and the interest is in assessing the mean/median improvement (with corresponding confidence intervals) at a pre-defined clinically relevant time point. Variability in the observed scores are taken into account. *Proportion of responders with worsening at time t:* A clinically relevant within-patient level worsening is pre-defined, and the interest is in evaluating the number of patients with worsening at a pre-defined clinically relevant time point. Variability in the scores above or below this pre-defined worsening threshold is ignored.
RS 8	*Stable state* is defined as no change from baseline is observed, or change from baseline is within the pre-defined baseline margin. This stable state is maintained if follow-up assessments remain at the baseline pre-defined margin. The stable state is discontinued once the follow-up assessment leaves the pre-defined baseline margin (and reaches the improvement or worsening threshold). There may be circumstances where the relevant PRO objective would include improvement in the definition of stable state (i.e., at least stable). In this case, the definition is as long as follow-up assessments do not reach the deterioration threshold, then stable state can still be	Disagreement arose because the current definition of stable state implies distinction among three possible categories (improvement, worsening or stable state). However, situations may occur where categories exist between improvement and stable state; and/or worsening and stable state (five categories). These additional two categories may be used as an error margin between stable state and improvement/worsening; or be included as meaningful categories (e.g., partial improvement or partial worsening). *Time to (end of) stable state:* For time to stable state, a clinically relevant within-patient stable state level is pre-defined, and the interest is in evaluating the time it takes before a clinically relevant stable state is observed. This endpoint may be useful when worsening is expected to occur before a stable state is reached. For time to (end of) stable state, the interest is in evaluating the time until the stable state ends or time until a clinically relevant improvement and/or worsening is observed.

	concluded. See Figure 2 for illustration.	*Proportion of responders with a stable state at time t:* A clinically relevant within-patient level stable state is pre-defined, and the interest is in evaluating the number of patients with a stable state at a pre-defined clinically relevant time point. Variability in the scores above or below this pre-defined worsening threshold is ignored. *Magnitude of stable state at time t:* Unlike worsening or improvement, stable state will not have a PRO objective examining *magnitude of stable state at time t*. When comparing two patients that both meet the criteria for stable, one cannot rank or order them so that one patient is considered more stable than the other. By definition, differing values within the stable state threshold are considered 'noise', i.e., random fluctuations not representing any meaningful changes.
RS 9	*Overall effect* is defined as summarizing all available scores over time for each patient on a specific PRO domain/item.	Disagreement arose on whether overall effect endpoints can be used with a treatment efficacy / clinical benefit PRO objective. The recommendation is that overall effects can be used alongside a treatment efficacy / clinical benefit PRO objectives. Since information is lost with this type of endpoint (relative to improvement, worsening and stable state), caution should be taken when planning to use overall effect endpoints. For example, an overall PRO score over time will not capture the direction and timing of an effect. *Overall PRO score over time*: The goal is to summarize all available scores over a given time period into a single data point per patient for a specific PRO domain. The time frame of interest should be pre-defined. The resulting outcome can then be used to compare two groups. To capture overall PRO score over time, several summary measures exist such as the average, minimum/maximum, and area under the curve [37,38]. These summary measures may or may not include the baseline score, depending on the research objective. Clinically relevant thresholds should also be pre-defined to aid

		interpretation of these values. However, by summarizing all available data into one score, information is lost and clinically relevant changes at particular time points may be obscured [38]. Therefore, the analysis and presentation of an overall PRO score over time should always also include the presentation of the time course of the PRO over a pre-defined time period (the period included in the overall PRO measure) to support interpretation of the overall PRO score. Recommended summary measures are not included in this document, but will be part of future work. *Response patterns or profiles:* The goal is to describe response trajectories over time. Clinically relevant thresholds should also be pre-defined to aid interpretation of these values. As it is not always straightforward to pre-define the exact profiles within a time frame, this within-patient/within-treatment arm PRO research objective is recommended to be used alongside a descriptive / exploratory objective rather than evidence for treatment efficacy / clinical benefit.
Section 2: Recommending statistical methods		
Recommendation No.	Recommended statement (RS)	Considerations
RS 10	Essential statistical features for analyzing PRO data are: - perform a statistical test between two treatment groups, - produce clinically relevant results. Highly desirable statistical features are: - adjust for covariates, including baseline PRO score, - handle missing data with the least	For more details on how this statement was developed, including the list of other statistical features considered, please see Appendix pages 13-15. *Perform a statistical test between two groups:* The current scope of these recommendations is on RCTs, and testing for statistical differences between groups is the main goal of an RCT [39]. *Produce clinically relevant results:* The chosen statistical method should be able to produce results that are easily interpretable for non-statisticians, guide informative clinical-decision making and

	restrictions, - handle clustered data (repeated assessments).	influence clinical practice. Statistically significant results do not imply that results are clinically relevant [40]. Therefore, in addition to statistically testing for a difference, the method should be able to produce estimates on the magnitude, certainty and direction of the treatment effect that can be directly linked with the PRO measure. This criterion implies that for PRO analysis, parametric is favored over non-parametric methods. Since parametric methods rely on distributional assumptions, it is recommended that non-parametric methods are used for sensitivity analysis to investigate deviations from these assumptions especially when sample sizes are small [24]. *Adjust for baseline covariates, including baseline PRO score:* When baseline covariates are correlated with the outcome of interest, it is recommended to adjust for such covariates to improve the efficiency of the analysis and avoid conditional bias from the covariates [41,42]. For example, baseline PRO scores are often correlated with PRO scores at follow-up [43]; therefore it is important to have an analytical method that can incorporate baseline covariates. Other covariates could include demographic variables (e.g., age, gender), disease characteristics (e.g., disease site, stage) and other relevant variables (e.g., country). *Handle missing data with the least restrictions:* When the probability of missingness is related to the outcome of interest, this could lead not only to a loss of power but also potential bias of estimates [44]. Missing data is almost always inherent when analyzing PRO data in cancer clinical trials; and the most restrictive assumption that the probability of missing data is unrelated to the PRO domain/item of interest is highly unlikely [45]. *Handle clustered data (repeated assessments):* To capture changes in the PRO domain/item of interest, PROs are often assessed

		repeatedly over time in cancer clinical trials. Analyzing this kind of data would require taking into account both the clustering of PRO assessments within each patient, and the temporal order of the measurements [46].
RS 11	For evaluating time to event outcomes (improvement, stable state or worsening), it is recommended to use the Cox proportional hazards (PH) instead of the log-rank test.	Please refer to Appendix pages 16-26 to find more details on how the statistical methods were evaluated based on the agreed set of criteria. When using Cox PH test, the proportional hazards assumption should be checked [47]. If this assumption is not met, performing a sensitivity analysis with a log-rank and/or Cox non-PH model to assess the robustness of findings is recommended. Also, general assumptions of time-to-event analysis must hold, most notably that the censoring is independent of the event time [48].
RS 12	For evaluating magnitude of event (improvement or worsening) at time *t* (where the design is baseline + >1 follow-up), it is recommended to use the linear mixed model (time as discrete) over the other statistical methods evaluated.	Please refer to Appendix pages 16-26 to find more details on how the statistical methods were evaluated based on the agreed set of criteria. Although the linear mixed model (time as continuous), pattern mixture model, and joint longitudinal model satisfy the set criteria, the linear mixed model (time as discrete) was recommended because less assumptions were needed to be made *a priori* (e.g., regarding the relationship between time and outcome variable). The analysis strategy would be to fit a linear mixed model to the data and then obtain the test estimate for specific time *t.* This method is suitable if a study has a limited number of follow-up assessments. General assumptions of linear mixed models hold. For example, the missing at random assumption has to be satisfied; that is, the linear mixed model will provide an unbiased estimate of the treatment effect that would have been observed if missing data is dependent on known and observed factors [49].
RS 13	For evaluating magnitude of event	Please refer to Appendix pages 16-26 to find more details on how the

	(improvement or worsening) at time *t* (where the design is baseline + 1 follow-up only), it is recommended to use the linear regression over the AN(C)OVA, t-test and Wilcoxon-ranks sum test.	statistical methods were evaluated based on the agreed set of criteria. Caution is needed for this recommended analysis because many statistical programs use complete case analysis for linear regression (e.g., SAS; [50]. Estimates resulting from such analysis will only provide valid inference when missing data are missing completely at random (MCAR)
RS 14	Summary measures should be considered in SISAQOL recommendations	In the original statement, the goal was to recommend a method for evaluating an overall PRO score over time. In this context, a summary measure is defined as a combining the repeated assessments of a PRO domain per patient over a specific time period into a single outcome (e.g., AUC, overall means and min/max). The proposed recommendation is that, if a summary measure is used, a linear regression is recommended to compare outcomes between groups. Although commonly used in PRO analysis, there was a general hesitation in recommending this proposal because it might be seen as a recommendation for two-step procedures in general [51]. Moreover, information is lost when data are pooled and summarized into one value, which may then impact the interpretability of the PRO findings. It was agreed that depending on the context, summary measures can be useful in understanding PRO data and should be considered in the SISAQOL recommendations. However, future work should involve evaluating which summary measures are recommended, and to identify the most appropriate way to analyze these data.
RS 15	For describing a response trajectory over time, it is recommended to use a linear mixed model (omnibus test; time as discrete	Please refer to Appendix pages 16-26 to find more details on how the statistical methods were evaluated based on the agreed set of criteria.

	variable; time*group interaction) over the repeated measures ANOVA (time*group interaction)	The focus of this method is not to interpret the *p-value* from the time*group interaction, but to fit a model and then interpret the resulting parameters. However, post-hoc description of these profiles are reported cross-sectionally and not longitudinally. That is, every assessment point has a mean and confidence interval. Therefore, interpretation is not on the (mean) longitudinal profile of the sample, but the mean outcome at each time point. If individual longitudinal profiles are of interest, more 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. However, many of these models rely on specific assumptions and may yield results/estimates/graphs that are difficult to interpret. Deciding which time curve is most appropriate is not straightforward and should ideally be informed by historical data.
Section 3: Standardizing statistical terms related to missing data		
Recommendation No.	Recommended statement (RS)	Considerations
RS 16	Missing data are data that would be meaningful for the analysis of a given research objective or estimand, but were not collected.	Although the literature has given considerable attention to the importance of reporting and handling of missing data [13], it remains unclear what is considered as missing data. Missing data can refer to: - any PRO assessment that is missing regardless of the reasons for missingness; [45,52]; - non-completion of PRO assessments that were expected to be available [21]; - any missing value that would be meaningful for analysis (if they were observed) [26,27]. Adopting the definition of ICH E9 implies that only those data that are considered “meaningful” for analysis would contribute to the PRO

		findings. It is the missing PRO data within this framework that can impact the interpretability of PRO findings either by reducing the sample size (non-informative missing data), distorting the treatment estimate (informative missing data) or both.
RS 17	“Meaningful for analysis” refers to the PRO analysis population, which is based on the given research objective (or estimand).	A differentiation between the PRO study population from the PRO analysis population is needed. The PRO study population is defined as all patients who consented and were eligible to participate in the PRO data collection. Ideally, the PRO study population would be the same as the ITT population, but this might not always be needed or feasible. Reasons to deviate from the ITT population and not to collect PROs at all from a specific sub-group should be strongly justified in the protocol. The PRO study population is a subgroup of the ITT population which excludes those patients where PRO outcomes could not be collected at all due to consent and/or eligibility. Patients of the PRO study population should be identifiable at the beginning of the study irrespective of their follow-up status/observations. The PRO study population is therefore the ITC (intention-to collect) PRO population. The PRO analysis population refers to the patients that will be included in the primary PRO analysis; and should be as close as possible to the PRO study population. Since PROs are assessed repeatedly over time on the same patient, caution should be noted when some planned assessments are not observed [26]. Depending on the analysis method, elimination of planned assessments from some patients may imply removing those patients altogether from the intended PRO analysis population. The PRO analysis population exists only in relation to a defined PRO analysis. If there are several primary PRO analysis planned, each will correspond to its own PRO analysis population which may or may not differ from each other.
RS 18	PRO assessments are no longer expected from patients who have died (although these patients were part of the PRO study	PRO assessments after death should not be expected because a meaningful value for these observations will not exist [21,27]. These assessments are also not “meaningful for analysis” because they will

	population).	not have a relevant contribution to the PRO estimate, and are therefore not considered as missing.
RS 19	A “variable denominator rate” should be reported. This rate is defined as the ‘number of patients on PRO assessment *submitting a valid PRO assessment* at the designated time point’ as a proportion of ‘the number of patients on PRO assessment at the designated time point’.	The term ‘on PRO assessments’ identifies those patients who are still expected to provide PRO assessments at that time point. Conversely, patients that are off-PRO assessments are defined as patients who are no longer expected to provide PRO assessments from that time point onwards. It was agreed to standardize that PRO assessments after death are considered “off-PRO assessment” and will no longer be included in the denominator of the completion rates (i.e., *number of patients on PRO assessment*). This implicitly implies that unobserved assessments after death will not be considered as missing data. Whether or not to standardize other reasons such as off PRO protocol, patient withdrawal and loss to follow-up in the number of patients on PRO assessment need further discussion (see Appendix pages 35-36).
RS 20	The term ‘completion rate’ should be used to express the rate with the variable denominator rate.	
RS 21	A “fixed denominator rate” should be reported. This is defined as the ‘number of patients on PRO assessment *submitting a valid PRO assessment* at the designated time point’ as a proportion of ‘the number of patients in the PRO study population’ (i.e., all patients who consented and were eligible to participate in the PRO data collection).	The need for an available data rate (fixed denominator rate) was to help address questions on both survivorship bias (which will not be reflected in the variable denominator rate); and the number of patients contributing observed data to the PRO estimate.
RS 22	The term ‘available data rate’ should be used to express the rate with the fixed denominator rate.	
RS 23	In addition to percentages, absolute numbers for both numerator and	It was proposed that a CONSORT diagram would be helpful to report the reasons for missing data. It was suggested to have three broad

	denominator should be reported at every time point (for both rates).	categories for the reasons: death, reasons pre-specified in the protocol, and reasons not pre-specified in the protocol. Further work is needed to develop this idea.
Section 4: General handling of missing data		
Recommendation No.	Recommended statement (RS)	Considerations
RS 24	When conducting clinical trials, exploring the reasons for missing PROs is important.	Results from a simulation study showed that the impact of missing data rates on PRO findings depends on the reasons for missing data (e.g., informative, non-informative or a combination of both). Therefore, collecting reasons for missing data is key in assessing the impact of missing data rates on the robustness of PRO findings.
RS 25	Missing data should be minimized prospectively through clinical trial and PRO design strategies and by training/monitoring approaches.	No analysis method recovers the potential for robust treatment comparisons derived from complete assessments of all patients [26]. Therefore preventing missing PRO assessments through careful design and planning should be the first line strategy in handling missing PRO data [27]. For more information, refer to [53].
RS 26	Capturing data that will be needed for handling missing PRO data in the statistical analysis plan is recommended (i.e. reasons for missing data and auxiliary data for interpretation/imputation).	Missing data may still be unavoidable despite careful planning and collection strategies. With missing data, unverifiable assumptions would have to be made during the analysis [54]. Collecting reasons for missing data and auxiliary data would be helpful in justifying how these patients are handled in the primary and sensitivity analysis [18,54].
RS 27	Primary statistical analysis approach: Missing data approach at the item- and scale-level should be specified *a priori* within the protocol/statistical analysis plan.	Similar to the choice of statistical analysis, different approaches to deal with missing data can lead to different results [55]. It is therefore important to document *a priori* the missing data approach that will be used for the primary analysis [8].
RS 28	Primary statistical analysis approach: Item-level missing data within a scale should be handled according to the scoring algorithm developed during the scale's development (when available).	Although general recommendations on how to deal with missing items exist [56], PRO measures are developed with a scoring algorithm to standardize how missing items should be handled. This should be used in the primary analysis; and other ways to deal with missing items can be included as part of sensitivity analysis.

		If changes in official scoring algorithms for the PRO occur, the resulting updated guidelines from the developers should be followed.
RS 29	Primary statistical analysis approach: Critical assessment of missing data reasons and rates (by arm and time point) should be undertaken.	Many possible reasons for missing data exist (e.g., patient withdrawal, patient moving). Depending on the reason and amount of missing data, the approach to handle missing data may differ [18,54].
RS 30	Primary statistical analysis approach: Use all available data, using the specified method from Statistical Methods WG.	Approaches that require ignoring missing data and only performing analysis with patients with complete data are not recommended (e.g., complete case analysis) [54]. Methods that allow the use of all available data is recommended as they make weaker assumptions about missing data compared to complete case analysis [57].
RS 31	Primary statistical analysis approach: Explicit imputation is not recommended unless justified within the context of the clinical trial.	Explicit simple imputation methods, such as last observation carried forward, will result in underestimating the variability of the estimate because a constant is used to impute the missing value regardless of differing patient characteristics [57]. Imputing a fixed constant will result in lower variability; and therefore a lower p-value [58].
RS 32	Sensitivity analysis should be specified *a priori* within the protocol/statistical analysis plan. At least two different approaches to handle missing data are recommended to assess the impact of missing data across various assumptions.	Handling missing data require making unverifiable assumptions regarding the relationship between the missing value and the outcome. Sensitivity analyses are required to test the robustness of the conclusions using a different set of assumptions regarding missing data[30]. Results that are consistent with the primary analysis provide some assurance that the missing data did not have an important impact on the study conclusions. However, if sensitivity analyses produce inconsistent results, missing data implications on the conclusions of the trial must be discussed [54]. Disagreement arose because of the increase in the workload of trialists to pre-specify, analyze and report additional sensitivity analyses.

Table 2: Overview of taxonomy of research objectives matched with recommended primary statistical methods

	Treatment efficacy / Clinical benefit (Confirmatory objective)		Describe patient perspective (Exploratory / Descriptive objective)
Within-treatment PRO assumption Within-patient/within-treatment PRO objective	Between-treatment arms objective		
	Superiority	Equivalence / Non-inferiority	
1. Improvement			
a. Time to improvement	- Cox proportional hazards (with pre-defined effect size for the between treatment arm difference)	Equivalence - Cox proportional hazards (with pre-defined equivalence margin for the between treatment arm difference) Non-inferiority - Cox proportional hazards (with a pre-defined non-inferiority margin for the between treatment arm difference)	Exploratory - Cox proportional hazards Descriptive - Median time to improvement; - Probability of improvement at a specific time point - Hazards ratio (with CI);
b. Proportion of patients with improvement at time t	*Further discussion needed on whether logistic mixed model, (Cochrane) Mantel-Haenszel test, or the simple logistic model would be recommended*	*Further discussion needed on whether logistic mixed model, (Cochrane) Mantel-Haenszel test, or the simple logistic model would be recommended*	Exploratory - *Further discussion needed on whether logistic mixed model, (Cochrane) Mantel-Haenszel test, or the simple logistic model would be recommended* Descriptive - Proportion of responders at time *t*; - Odds/risk ratio (with CI)

c. Magnitude of improvement at time t	- Linear mixed model; Time as discrete (with pre-defined effect size for the between treatment arm difference)	Equivalence - Linear mixed model; Time as discrete (with pre-defined equivalence margin for the between treatment arm difference) Non-inferiority - Linear mixed model; Time as discrete (with a pre-defined non-inferiority margin for the between treatment arm difference)	Exploratory - Linear mixed model; time as discrete Descriptive - Mean magnitude at baseline & at time *t* (with CI); - Mean magnitude of improvement at time *t* (with CI)
2. Stable state			
a. Time to (end of) stable state	- Cox proportional hazards (with pre-defined effect size for the between treatment arm difference)	Equivalence - Cox proportional hazards (with pre-defined equivalence margin for the between treatment arm difference) Non-inferiority - Cox proportional hazards (with a pre-defined non-inferiority margin for the between treatment arm difference)	Exploratory - Cox Proportional Hazards Descriptive - Median time to (end of) stable state; - Probability of (end of) stable state at a specific time point - Hazards ratio (with CI)
b. Proportion of patients with stable state at time t	*Further discussion needed on whether logistic mixed model, (Cochrane) Mantel-Haenszel test, or the simple logistic model would be recommended*	*Further discussion needed on whether logistic mixed model, (Cochrane) Mantel-Haenszel test, or the simple logistic model would be recommended*	Exploratory - *Further discussion needed on whether logistic mixed model, (Cochrane) Mantel-Haenszel test, or the simple logistic model would be recommended*

			Descriptive - Proportion of responders at time *t*; - Odds/risk ratio (with CI)
c. Magnitude of stable state at time t	*Not applicable* (When comparing two patients that both meet the criteria for stable, one cannot rank or order them so that one patient is considered more stable than the other. By definition, differing values within the stable state threshold are considered 'noise', i.e., random fluctuations not representing any meaningful changes)	*Not applicable* (When comparing two patients that both meet the criteria for stable, one cannot rank or order them so that one patient is considered more stable than the other. By definition, differing values within the stable state threshold are considered 'noise', i.e., random fluctuations not representing any meaningful changes)	*Not applicable* (When comparing two patients that both meet the criteria for stable, one cannot rank or order them so that one patient is considered more stable than the other. By definition, differing values within the stable state threshold are considered 'noise', i.e., random fluctuations not representing any meaningful changes)
3. Worsening			
a. Time to worsening	- Cox proportional hazards (with pre-defined effect size for the between treatment arm difference)	Equivalence - Cox proportional hazards (with pre-defined equivalence margin for the between treatment arm difference) Non-inferiority - Cox proportional hazards (with a pre-defined non-inferiority margin for the between treatment arm difference)	Exploratory - Cox Proportional Hazards Descriptive - Median time to worsening; - Probability of worsening at a specific time point - Hazards ratio (with CI)
b. Proportion of patients with worsening at time t	*Further discussion needed on whether logistic mixed model, (Cochrane) Mantel-Haenszel test, or the simple logistic model would*	*Further discussion needed on whether logistic mixed model, (Cochrane) Mantel-Haenszel test, or the simple logistic model would*	Exploratory - *Further discussion needed on whether logistic mixed model, (Cochrane) Mantel-Haenszel*

	be recommended	*be recommended*	*test, or the simple logistic model would be recommended* Descriptive - Proportion of responders at time *t*; - Odds/risk ratio (with CI)
c. Magnitude of worsening at time t	Linear mixed model; Time as discrete (with pre-defined effect size for the between treatment arm difference)	Equivalence - Linear mixed model; Time as discrete (with pre-defined equivalence margin for the between treatment arm difference) Non-inferiority - Linear mixed model; Time as discrete (with a pre-defined non-inferiority margin for the between treatment arm difference)	Exploratory - Linear mixed model; time as discrete Descriptive - Mean magnitude at baseline & at time *t* (with CI); - Mean magnitude of worsening at time *t* (with CI)
4. Overall effects			
a. Overall PRO score over time	*Further discussion needed*	*Further discussion needed*	*Further discussion needed*
b. Response patterns / profiles	*Not applicable* (As it is not always straightforward to pre-define the exact profiles within a time frame, response patterns/profiles are recommended to be used alongside a descriptive / exploratory objective rather than evidence for treatment efficacy /	*Not applicable* (As it is not always straightforward to pre-define the exact profiles within a time frame, response patterns/profiles are recommended to be used alongside a descriptive / exploratory objective rather than evidence for treatment efficacy /	*Exploratory* - Linear mixed model (time as discrete / continuous) *Descriptive* - Mean magnitude at baseline & at every time point within a time frame (with CI); - Mean change at every time

	clinical benefit)	clinical benefit)	point within a time frame (with CI); - Mean profile over time (with CI)

Note: Recommended statistical methods were initially conceptualized for a superiority between-treatment arms objective. However, these methods may be extrapolated to (a) a non-inferiority / equivalence objective, but appropriate margins should be pre-specified (see Table 1, RS 2); and (b) exploratory but findings should not be used as a basis of evidence of clinical benefit / treatment efficacy (see Table 1, RS 1). Descriptive statistics are based on the work from the Statistical Methods Working Group on evaluating appropriate statistical methods with research objectives (see Appendix pages 18-26).

Contributors

The manuscript was conceptualized with the attendees of the SISAQOL kick-off meeting in Brussels on January 26 2016. All authors contributed to the work of the individual working groups. All authors discussed and finalized this work during the SISAQOL consensus meeting on September 24 2018. All authors reviewed and contributed to the revisions of the article. All authors approved the final draft of the manuscript.

Conflict of Interest Statement

EORTC received an unrestricted education grant from Boehringer Ingelheim GmbH to initiate this work and from Genentech, a member of the Roche Group, for continuity funding.

All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. AC reports having been an employee of Genentech until January 31, 2019. EPL is an employee of Genentech. GV reports personal fees from Roche, Eisai, Novartis, Pfizer, grants from NIHR UK Government, Breast Cancer NOW and EORTC, outside the submitted work. IG is an employee of Boehringer Ingelheim International GmBH that provided an unrestricted education grant to the European Organization for Research and Treatment of Cancer (EORTC). KC reports grants from EORTC, personal fees from BMS, Endomag Ltd, Celgene and Amgen, outside the submitted work. KO reports grants from Bristol-Myers Squibb, Roche, Novocure, Lilly, Pfizer, MagForce, Novartis, Medac, Photonamic, Northwest Biotherapeutics, VBL Therapeutics, AbbVie, Elekta, Apogenix and Bayer, outside the submitted work. MPi reports being a member of the Radius advisory board, grants to Institut Jules Bordet from Radius, Synthon and Servier, grants and personal fees to Institut Jules Bordet from AstraZeneca, Lilly, MSD, Novartis, Pfizer and Roche-Genentech, and personal fees from Odonate, Camel-IDS, Crescendo Biologics, Periphagen, Huya, Debiopharm, PharmaMar, G1 Therapeutics, Menarini, Seattle Genetics, Immunomedics and Oncolytics, outside the submitted work. MC reports personal fees from Astellas, Takeda, Glaukos, Merck, Daiichi Sankyo, andPatient-Centered Outcomes Research Institute (PCORI), and grants from Health Data Research UK, Innovate UK, National Institute for Health Research (NIHR), Macmillan and UCB Pharma, outside the submitted work. MKo reports grants from EORTC, Biofrontera and Komitee Forschung Naturmedizin e.V. (KFN), and personal fees from Janssen-Cilag, Lily and Verband Forschender Arzneimittelhersteller e.V (vfa)., outside the submitted work. AB reports grants to the EORTC from Boehringer Ingelheim International GmBH, Genentech, and the EORTC research fund during the conduct of the study, grants from Merck outside the submitted work and reports being a member of the EORTC Quality of Life Group executive committee. All other authors declare no competing interests.

This study received no National Institutes of Health (NIH) funding. AWS and SM are employed by NIH. No other authors were fully or partly NIH funded, employed by NIH, or are in receipt of an NIH grant.

SUPPLEMENTARY APPENDIX

Supplement to: International Standards for the Analysis of Quality of Life and Patient Reported Outcomes Endpoints in Cancer Randomised Controlled Trials: Recommendations based on critical reviews of the literature and international multi-expert, multi-stakeholder collaborative process

CONTENTS

Appendix 1 - Methods

Table 1. Profile of the Consortium Members at Each Stage of the Development of SISAQOL Recommendations

	Kick-off meeting *Jan 2016* (*N* = 26)	Strategic meeting *Jan 2017* (*N* = 29)	Working groups May 2017 – August 2018			Recommendations meeting *Sept 2018* (*N* = 31)	All[1] (*N* = 41)
			Research objectives (*N* = 26)	Statistical methods (*N* = 18)	Missing data (*N* = 10)		
Background							
Academia	7	10	9	4	1	7	14
Non-Profit	8	6	6	6	7	10	10
Government	3	5	7	5	1	8	8
Industry	4	4	2	2	1	3	4
Health Care	3	3	2	1	0	2	3
Other	1	1	0	0	0	1	2
Role[2]							
Researcher/health related academic	15	16	17	10	5	16	24
Expert advisor on PROs	8	11	9	4	3	8	14
Statistician	9	9	7	12	4	10	12
Clinician/clinical professor	4	6	6	1	1	5	10
Trials Methodologist	7	8	6	7	3	6	9
Policy maker/regulator	2	3	4	3	1	5	5
Industry representative	3	3	2	1	1	2	3
(Health) psychologist	2	3	3	2	1	2	3
Health Economist	1	1	0	0	0	2	2
Reviewer	1	2	1	1	0	2	2
Patient representative	0	1	1	0	0	1	1
Journal Editor	1	1	0	0	1	1	1
Country							
Australia	0	0	1	1	0	0	1
Austria	0	0	1	0	0	0	1
Belgium	7	4	5	5	5	7	8
Canada	0	1	2	0	1	2	2
Denmark	1	1	1	0	0	1	1
France	2	1	1	1	0	0	1
Germany	3	4	1	2	0	4	4
Portugal	0	0	0	0	0	0	1
Sweden	1	1	0	1	0	1	1
The Netherlands	1	2	1	0	0	1	2
UK	5	6	4	1	1	7	7
USA	6	9	9	7	3	8	12

[1] Membership list as on September 24th, 2018 (Recommendations meeting)

[2] Consortium members could have up to three roles

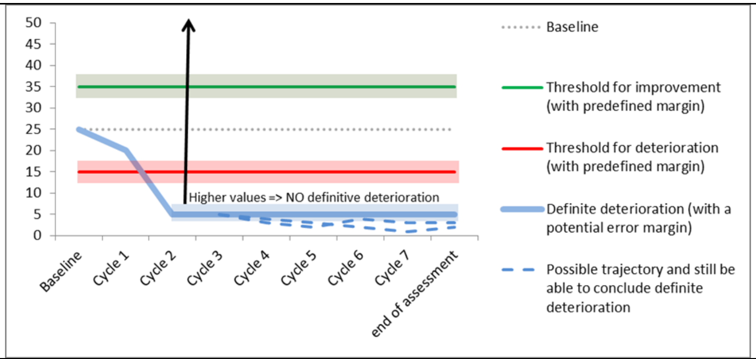
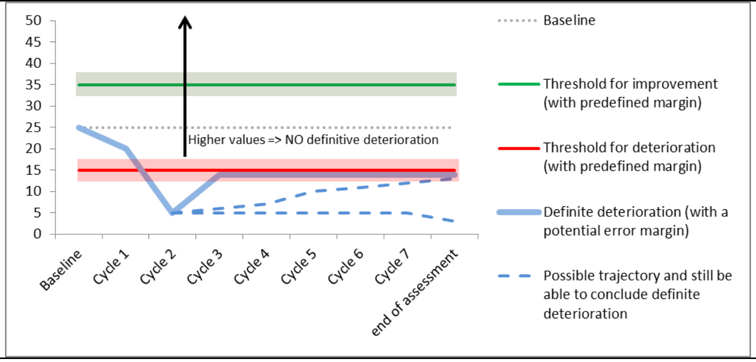
Table 2: Each Working Group's Process in Developing Proposed Statements for the SISAQOL Meeting

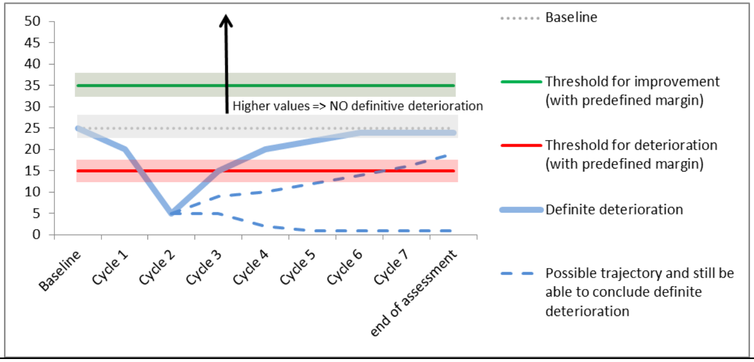
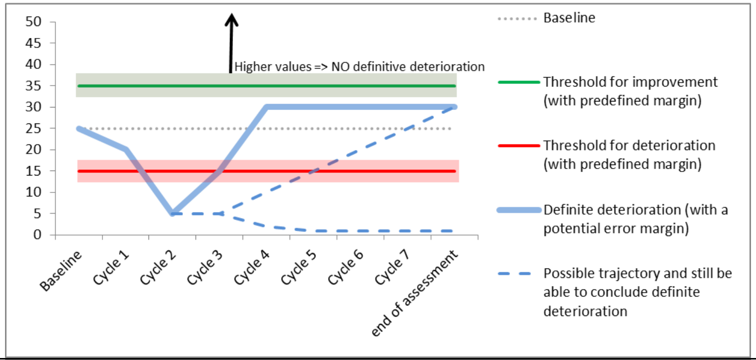
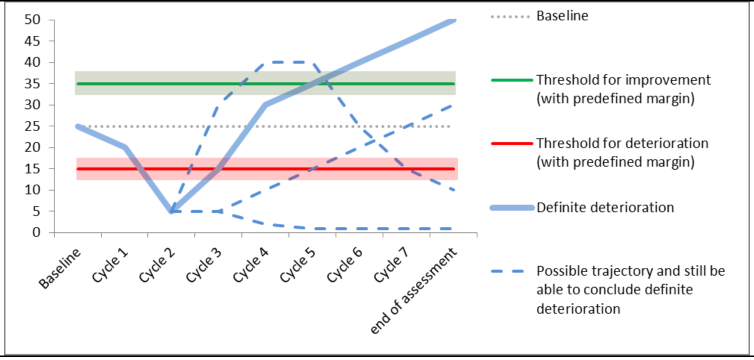
	Working Groups (WG)			
Dates	Research Objectives	Statistical Methods	Standardizing Terms Related to Missing Data	General Handling of Missing Data
2017				
May	WebEx discussions - Strategy kick-off WebEx meeting for all working group (WG) members - Presentation and discussion of content, problem description and next steps for each WG			
June	- Preparation of the initial draft of taxonomy of research objectives - Preparation of survey to standardize objectives: improvement, stable state and worsening	- Preparation of survey to list recommended statistical features for PRO/HRQOL analysis	- Preparation of survey on various definitions related to missing data: intent-to-treat (ITT), modified intent-to-treat (mITT), completion and compliance rates	- Preparation for Monte-Carlo simulations to answer the question "how much missing data is too much"
July				
August				
September	- WG members provided comments and feedback on the draft taxonomy of research objectives - WG members responded to survey on standardizing objectives			
October				
November		- WG members responded to survey on list of recommended statistical features for PRO analysis	- Consortium members responded to survey on definitions for these terms related to missing data	
December				
2018				
January	WebEx discussions (taxonomy of research objectives) - Presentation of comments on taxonomy of research objectives - Presentation of survey responses on standardizing research objectives: improvement, stable state and worsening - Agreement on updated taxonomy of research objectives			

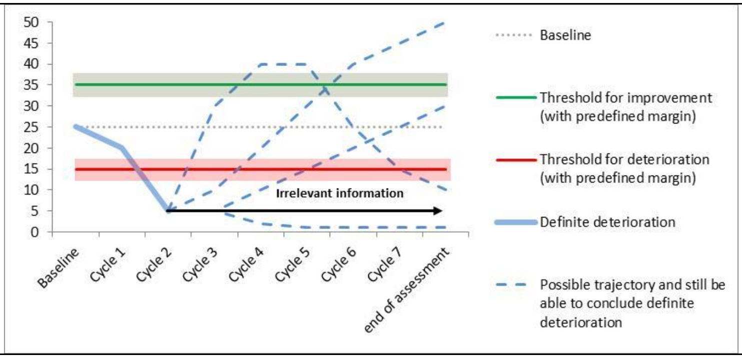
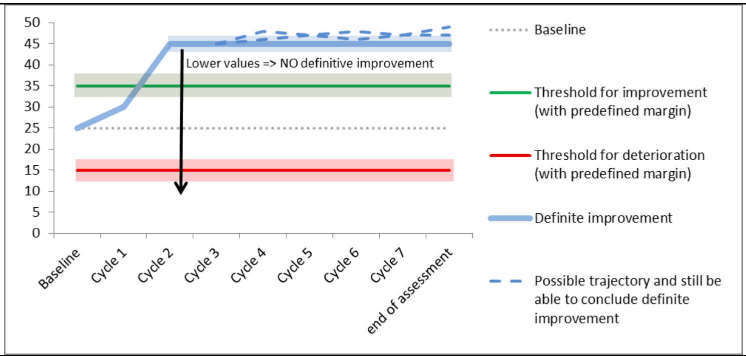
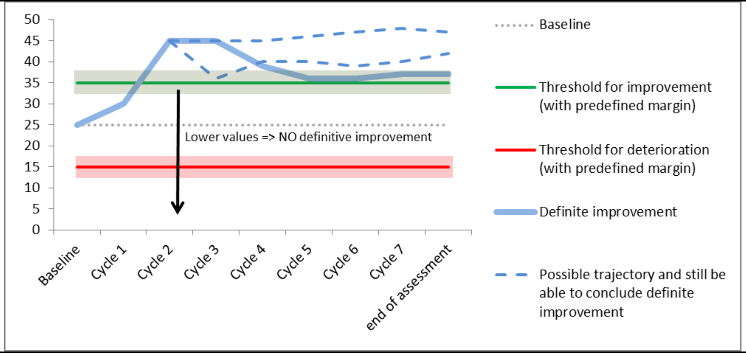
February	- Preparation of findings for recommendations meeting	WebEx discussions - Presentation of survey results on essential and highly desirable statistical properties for PRO analysis - Agreement on essential and highly desirable statistical properties for PRO analysis - Presentation of survey results on standardizing statistical terms to statistics methods WG		
March		WebEx discussions - Next steps: Work method for recommending appropriate statistical methods for each research objective based on the essential/highly desirable statistical properties - Presentation of Monte Carlo simulations for missing data thresholds - Standardized case report forms for reasons for missing data		
April		- Literature review to evaluate statistical methods for each objective based on the properties list - WG members feedback on evaluation of statistical methods	- Literature review on definitions of missing data	- Collection of case report forms for missing data reasons from SISAQOL members - Running of final Monte Carlo Simulations
May	WebEx discussions: status update from each working group			
June	- Preparation of findings for recommendations meeting	WebEx Discussions - Proposal of recommended appropriate statistical methods for each research objective - Presentation of varying definitions of missing data		- Consortium members responded to survey on reasons for non-completion based on collected case report forms - Development of missing data recommendations - Preparation of findings for recommendations meeting
July		- Preparation of findings for recommendations meeting	- Preparation of findings for recommendations meeting	
August				
September	SISAQOL recommendations meeting: Ratify proposed statements from each working group			

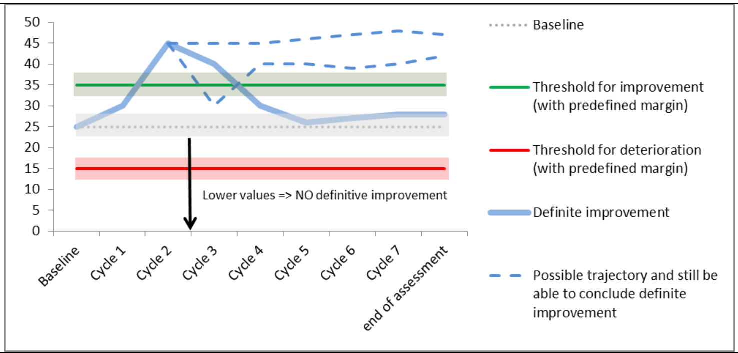
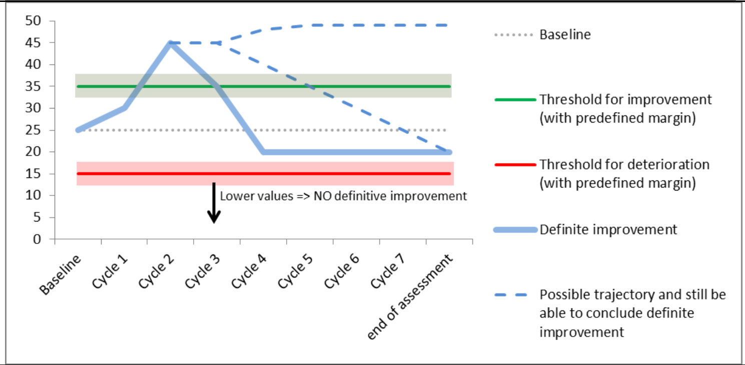
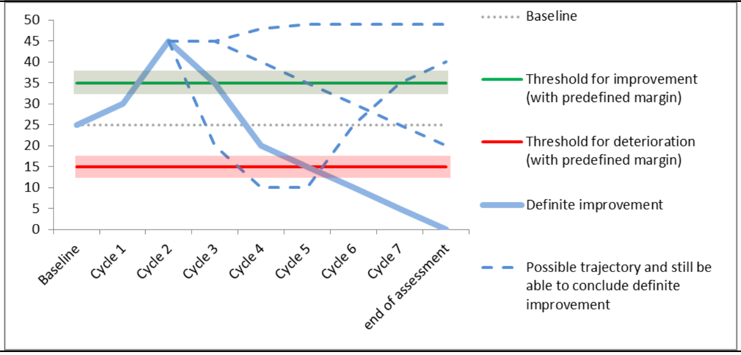
Appendix 2 - Intermediate results from each working group

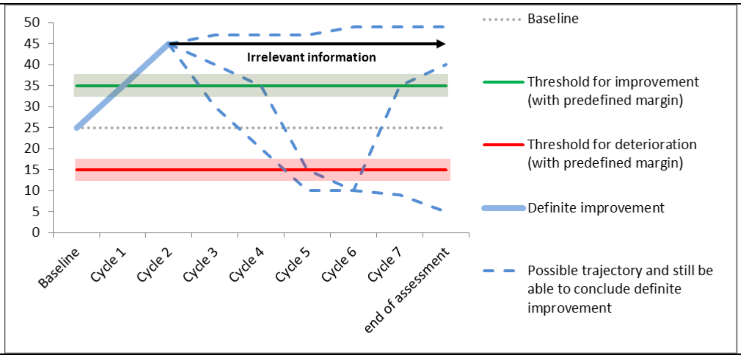
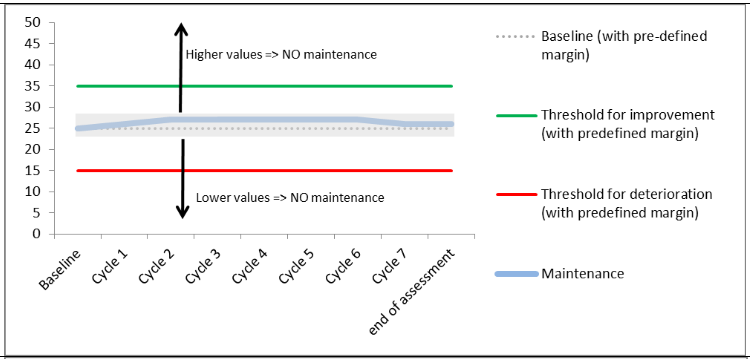
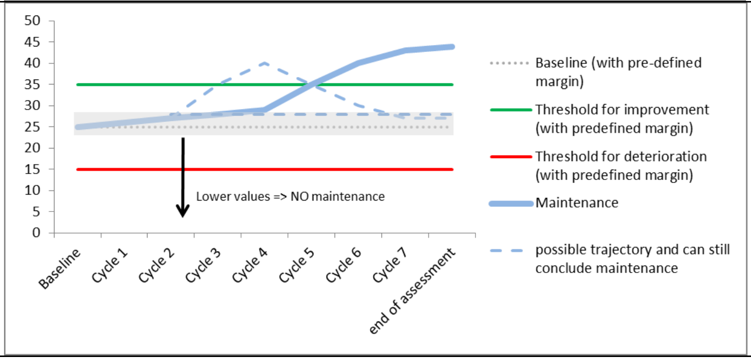
Table 1. Research objectives Working Group Survey Results on Standardizing Definitions of Improvement, Maintenance (or Stable State) and Deterioration (or Worsening) *(N = 26)*.

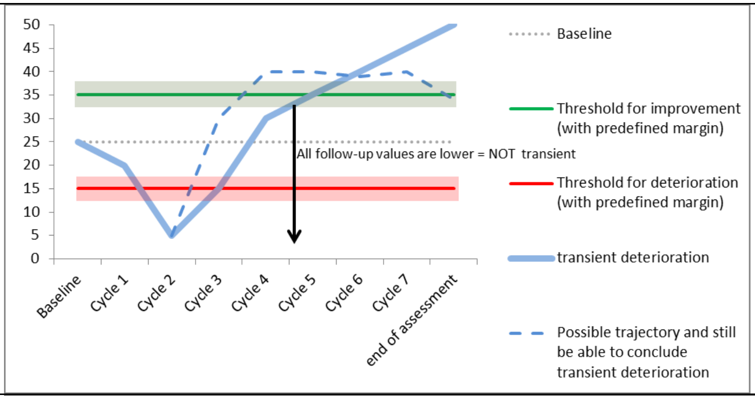
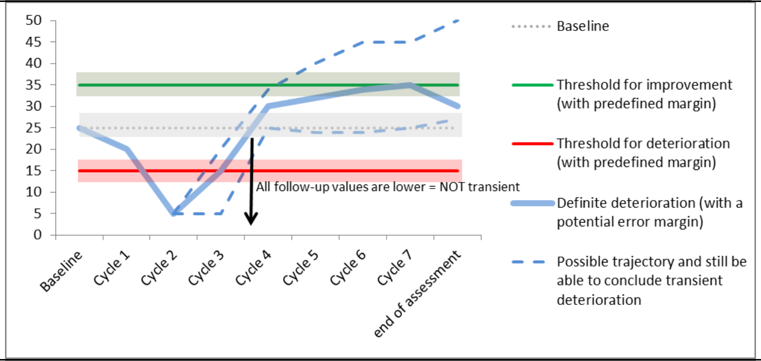
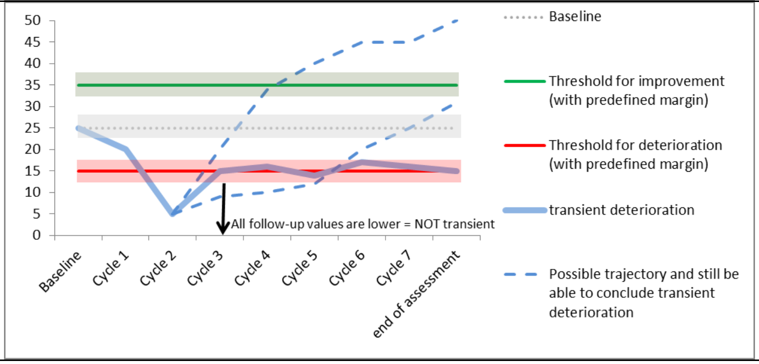
Definition	Graphic Visualization	Primary Scoring (% agree[1])
1. Definitive deterioration		
• Post-baseline deterioration • After the post-baseline deterioration: • no follow-up scores are higher than one's own deterioration level (or its pre-defined margin); • no follow-up scores are higher than the 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%)
• Post-baseline deterioration • After the post-baseline deterioration: • 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) • no follow-up scores are higher than the improvement threshold (or its pre-defined margin)		21 (81%)*

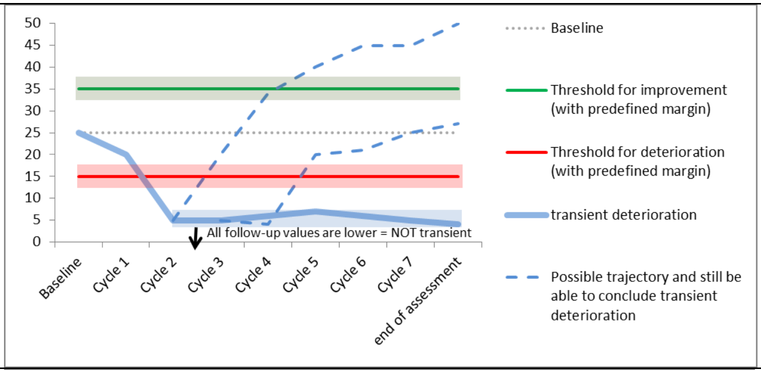
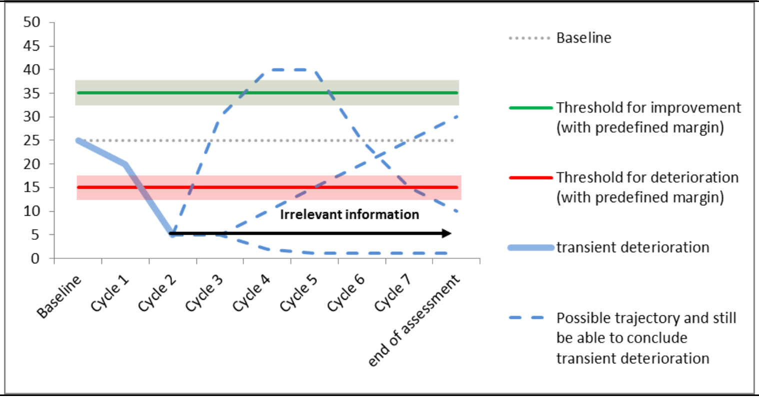
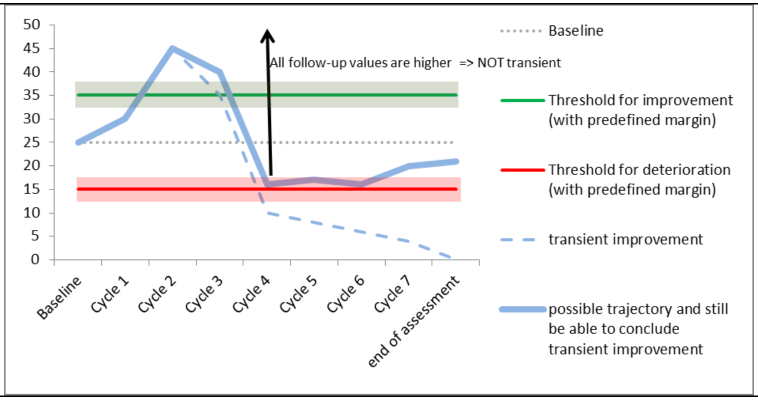
• Post-baseline deterioration • After the post-baseline deterioration: • follow-up scores may be higher than one’s own deterioration level (or its pre-defined margin); • follow-up scores may be higher than the deterioration threshold (or its pre-defined margin); • no follow-up scores are higher than one’s own baseline level (or its predefined margin) • no follow-up scores are higher than the improvement threshold (or its pre-defined margin)		4 (8%)
• Post-baseline deterioration • After the post-baseline deterioration: • 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)		1 (4%)
• Post-baseline deterioration • After the post-baseline deterioration: • 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) • follow-up scores may be higher than the improvement threshold (or its pre-defined margin)		1 (4%)

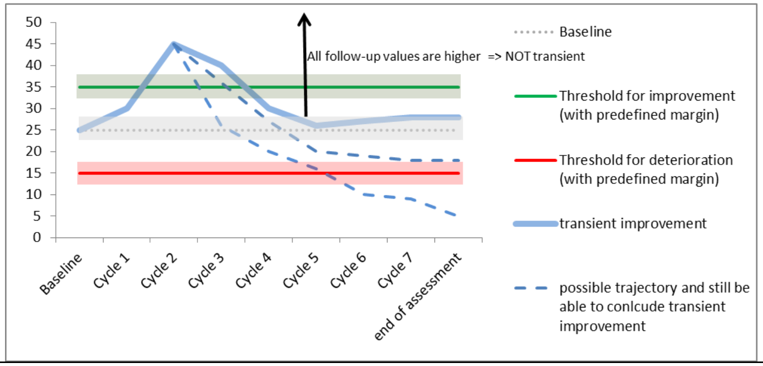
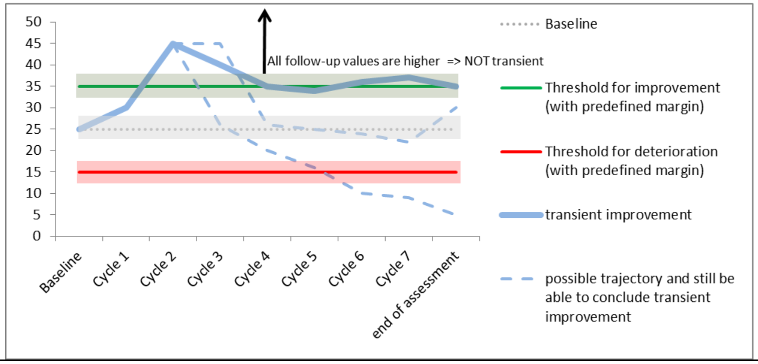
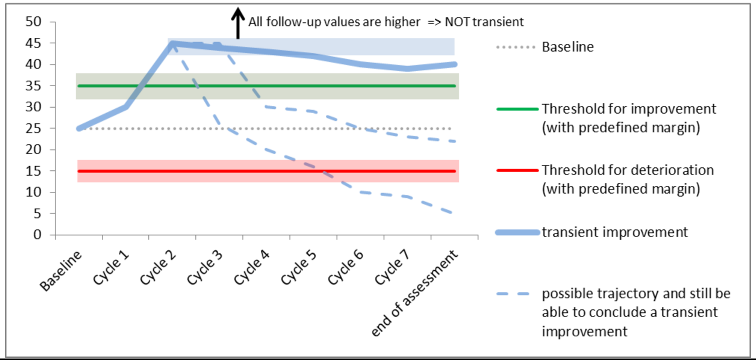
- Post-baseline deterioration		1 (4%)
2. *Definitive improvement*		
- Post-baseline improvement - After the post-baseline improvement: - no follow-up scores are lower than one's own improvement level (or its pre-defined margin); - no follow-up scores are lower than the improvement threshold (or its pre-defined 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)		21 (81%)
- Post-baseline improvement - After the post-baseline improvement: - 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) - no follow-up scores are lower than the deterioration threshold (or its pre-defined margin)		22 (85%)*

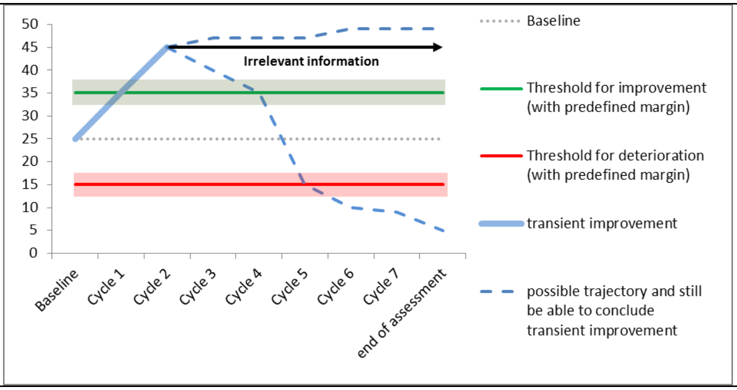
• Post-baseline improvement • After the post-baseline improvement: • follow-up scores may be lower than one's own improvement level (or its pre-defined margin); • follow-up scores may be lower than the improvement threshold (or its pre-defined margin); • no follow-up scores are lower than one's own baseline level (or its predefined margin) • no follow-up scores are lower than the deterioration threshold (or its pre-defined margin)		6 (23%)
• Post-baseline improvement • After the post-baseline improvement: • 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)		2 (8%)
• Post-baseline improvement • After the post-baseline improvement: • 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)		1 (4%)

- Post-baseline improvement		2 (8%)
3. Maintenance		
- Follow-up scores are similar to baseline score (by a pre-defined margin) - No follow-up scores are better than the baseline score. - No follow-up scores are worse than the baseline score.		23 (88%)*
- 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 <u>increase</u> 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 <u>increase</u> 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%)*
• Post-baseline deterioration • After the post-baseline deterioration, there is an <u>increase</u> in scores: ○ 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%)

• Post-baseline deterioration • After the post-baseline deterioration, there is an <u>increase</u> in scores: ○ 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%)
• Post-baseline deterioration		2 (8%)
5. Transient improvement		
• Post-baseline improvement • After the post-baseline improvement, there is a <u>decrease</u> 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 <u>decrease</u> 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 <u>decrease</u> 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%)
- Post-baseline improvement - After the post-baseline improvement, there is a <u>decrease</u> in scores: ○ At least one follow-up score should be lower than or at least be at the improvement level (or its pre-defined margin).		5 (19%)

- Post-baseline improvement		2 (8%)

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

[1] Primary 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	□ Comparing groups is the main goal of an RCT □ 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	□ 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 □ 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	□ 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,… - Disease characteristics: duration, stage,… - Others: country, center, investigator,.	12 (75%)	32	□ 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 □ 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	□ Missing data is a problem in PRO analysis. □ 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	□ PRO data is often longitudinal and this should be reflected in the analysis method □ Essential in the case of a longitudinal study objective (e.g., comparing means over time) □ 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	□ Only needed if the trial hypothesis calls for an integrated test □ 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	□ 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	□ 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. □ 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	□ 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: 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	□ This can be assessed with sensitivity analyses □ 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	□ ITT is the standard in most protocols. □ ITT is needed for generalizability of findings. □ Too restrictive if needed for all analyses. □ 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	□ Only needed if the trial hypothesis calls for an integrated test □ 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	□ It depends on the study. □ It may be useful but will not be used for the primary analysis □ 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	□ Similar to controlling or stratifying for confounding factors / covariates □ Not often part of the primary analysis even with other endpoints such as overall survival □ 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	□ In practice, having a bounded scale rarely generates problems □ 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	□ 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) □ Pre-specifying the PRO domains is important rather than modelling multiple PROs □ 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	□ This is a post-hoc issue that can be addressed with sensitivity analyses. □ 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	□ 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.

[1] Primary 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).

[2] Secondary 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
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**		
1. Clinical relevance at the within-individual level* _*Note that this is not a feature of the statistical method._	**(Yes)** 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	□ 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.
	(No) 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	□ 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
2. Clinical relevance of the treatment effect: Within-group/ Between groups* _*Note that all evaluations are based on comparison of only two arms_	**(Yes)** Statistical models that produce not only statistical significance estimates, but also the magnitude of the treatment effect **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 **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	□ 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).
	(No) Statistical models that give a statistical significance estimate, but the magnitude of the treatment effect is not estimated or the treatment effect is distorted **Between group:** Clinical relevance of the result for the difference between groups cannot be directly linked to the clinical relevance classification of the instrument **Within-group**: Clinical relevance of the result for the change within groups cannot be directly linked to the clinical relevance classification of the instrument	□ Between-group: Results are derived from a sum of squares or sum of ranks □ Within-group: Results are derived from a sum of squares
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.	□ 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.	□ 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)						
Improvement / worsening (event): time to event **Maintenance (event):** time to (end of) maintenance **Time to event:** Time to event								
Cox PH (Kaplan-Meier)[4–6]	**Yes** Clinical relevance of the result is interpreted at the within-individual level (through a clinically relevant definition of a within-individual event)	**Yes** Between group: Clinical relevance of the difference between groups can be assessed using a hazard ratio (with CI)	- Median duration for each group - Survival probabilities for each group at a time point	**Yes** Covariates and stratification can be included	Can handle **informative** missingness Method provides valid inference when censored* data are MCAR or MAR. *Non-informative censoring: censoring is independent from the possibly unobserved time-to-event applies [6]	**Limited:** 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).	Baseline + **Sufficient** # of follow-ups Sufficient follow-up assessments needed to capture occurrence of event	Strong assumption of proportional hazards 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) Assumption of independent censoring should be met[7]
Log-rank test (Kaplan-Meier)[4–6]	**Yes** Clinical relevance of the result is interpreted at the within-individual level (through a clinically relevant definition of a within-individual event)	**No** Between group: Indicates whether survival between two groups is significantly different, but does not indicate how different they are.	- Median duration for each group - Survival probabilities for each group at a time point	**Limited** Can only include stratification	Can handle **informative** missingness Method provides valid inference when censored* data are MCAR or MAR. *Non-informative censoring: censoring is independent from the possibly unobserved time-to-event [6]	**Limited:** 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).	Baseline + **Sufficient** # of follow-ups Sufficient follow-up assessments needed to capture occurrence of event	Less efficient when proportional hazards assumption is not met, but does not require the assumption of proportional hazards. Assumption of independent censoring should be met

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]	**Yes** 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)	**No** 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	**No** Inclusion of covariates and stratification are not possible	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 are 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**	Ideal for smaller sample sizes Does not require the assumption of normality
(Pearson's) Chi-square test [8–11]	**Yes** 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)	**No** 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	**No** Inclusion of covariates and stratification are not possible	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 are 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**	Large data set is needed. Assumption of normality is required
(Cochran) Mantel-Haenszel test [12–15]	**Yes** 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)	**Yes** Between group: *Discrete/binary outcome:* Clinical relevance of the difference between groups can be assessed using odd/risk ratio (with CI)	-Proportion (or percentage) of responders for each group -Odds/risk ratio	**Limited** Can only include stratification	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 are 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**	
Improvement / worsening (response): level of response at time *t* **Maintenance:** not applicable (by definition of maintenance. For example, we cannot say "level of maintenance is higher/lower" in one arm vs the other)								
(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	**Yes** Between group: *Continuous outcome:*	-Mean baseline level (with CI) & mean specific	**Yes** Covariates and stratification	Can handle **informative** missingness Method provides valid inference when missing	**Yes** - Cluster of repeated assessments per patient	Baseline + **sufficient but limited** # of follow-ups	Since time is treated as discrete, a parameter needs to be estimated for every

	on the group level	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 comparison with effect size, or PROM-specific MID or interpretation guidelines, if available.*	time point level (with CI) for each group -Mean change between baseline and each assessed time point (with CI) for each group	can be included	data are MCAR or MAR.	- 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)	As the number of follow-up assessments increases, the number of parameters to estimate also increases	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] 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*	-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

		(within-group) can be interpreted by comparison with effect size, or PROM-specific MID or interpretation guidelines, if available.						linear, quadratic and cubic polynomial terms may be used to approximate the time curves. But this also implies more 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 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_	*Continuous outcome:* Mean baseline level (with CI) & mean specific time point level (with CI) for each group *Ordinal/binary outcome:* Odds ratio (with CI)	**Yes** Covariates and stratification can be included	Can only handle **non-informative** missingness Method provides valid inference only for MCAR.* *Weighted GEE method is available to take into account 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)	*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	Parameter estimates are consistent and asymptotically normal even under mis-specified correleation structure of responses.[25] Generalized estimating equations can be used for discrete, count or binary outcome.

		comparison with effect size, or PROM-specific MID or interpretation guidelines, if available.						
Linear regression	**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) _*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._	Wilc	**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**	
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 as a change	**Yes** Between group: *Continuous outcome:*	-Mean baseline level (with CI) & mean specific	**No** Inclusion of covariates and	Can only handle **non-informative** missingness Method provides valid	**No** - Does not cluster repeated assessments per patient	Baseline + **1 follow-up**	Assumption of normal distribution is needed

	on the group level	Clinical relevance of the result can be assessed using the mean difference between the two 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._	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)*	stratification are not possible	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.	- Does not take into account longitudinal nature of data		
Wilcoxon rank sum test	**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)*	**No** Inclusion of covariates and stratification are not possible	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**	Does not assume normal distribution
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	-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	**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	*Time as continuous:* Baseline + **sufficient** # of follow-ups *Time as discrete:* Baseline + **sufficient but limited** # of follow-ups As the number of	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

		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 interpreted by comparison with effect size, or PROM-specific MID or interpretation guidelines, if available._	*(if time is discrete)* -Rate of change between baseline and specific time point (with CI) for each group *(if time is continuous)*			tend to have higher correlations)	follow-up assessments increases, the number of parameters to estimate also increases	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.

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-specific MID or interpretation guidelines, if available._	-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 'event' for the time to informative drop-out,..) are to be specified [34] and the model can be very computationally demanding [31]. 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.
Overall effect: Describe trajectory of outcome over time								
(Generalized) linear mixed model (time as discrete - omnibus test): group*time	**No** Clinical relevance of the result is not interpreted at the within-	**No** Between group:	-Mean baseline level (with CI)	**Yes** Covariates and	Can handle **informative** missingness Method provides valid	**Yes** - Cluster of repeated assessments per	Baseline + **sufficient but limited** # of follow-ups	Profiles are reported cross-sectionally and not

interaction [16,37,38]	individual level, but as a change on the group level	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. Within-group: Assesses whether responses over time are statistically significantly different, but does not provide an estimate of how different they are..	& levels at each assessed time point (with CI) for each group -Mean change between baseline and each assessed time point (with CI) for each group	stratification can be included	inference when missing data are MCAR or MAR.	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)	As the number of follow-up assessments increases, the number of parameters to estimate also increases	longitudinally. That is, every assessment point has a mean and CI. If individual longitudinal profiles are of interest, more 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. Generalized linear mixed models can be used for discrete, count or binary outcome.
Repeated measures ANOVA: 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 an estimate of how different they are. Within-group: Assesses whether responses over time are statistically significantly different, but does not provide an estimate of how different they are.	-Mean baseline level (with CI) & levels at each assessed time point (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 only handle **non-informative** missingness Method provides valid inference when data are MCAR. Listwise deletion/complete case analysis: Patients with no data at baseline and/or any specific timepoint is not included in the analysis.	**Limited** - Cluster of repeated assessments per patient - 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 between occasions)	Baseline + **sufficient but limited #** of follow-ups As the number of follow-up assessments increases, the number of parameters to estimate also increases	Profiles are reported cross-sectionally and not longitudinally. That is, every assessment point has a mean and CI.

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.	
□ Agree	37/38 (97%)
□ 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)*	
□ Analysis population could be limited to patients with baseline PRO assessment	12/38 (32%)
□ Analysis population could be limited to patients with at least one post-baseline PRO assessment	6/38 (16%)
□ Analysis population could be limited to patients with baseline + at least one post-baseline PRO assessment	17/38 (45%)
□ Analysis population could be limited to eligible patients	9/38 (24%)
□ 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%)
□ Analysis population could be limited to the safety population (patients exposed to their intended treatment only)	4/38 (11%)
□ Analysis population could be limited to patients exposed to any protocol treatment	4/38 (11%)
□ Other (To specify) ○ Patients who consent to PRO substudy ○ Depends on the study objective	4/38 (11%) □ 1/38 (3%) □ 3/38 (8%)
□ 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*)	
□ *Both the fixed denominator rate and the variable denominator rate are needed*	26/38 (68%)
□ *Only the variable denominator rate is needed*	6/38 (16%)
□ *Only the fixed denominator rate is needed*	2/38 (5%)
□ *Other (To specify)* ○ *Both + cohort plots* ○ *Both + additional information related to the attrition* ○ *Both can, but is not a 'must'* ○ *Variable denominator rate + death rate*	4/38 (11%) □ 1/38 (3%) □ 1/38 (3%) □ 1/38 (3%) □ 1/38 (3%)
Fixed denominator rate: Numerator	
□ On-study patients submitting the PRO assessment at the designated time point	32/38 (84%)
□ On-study patients submitting the PRO assessment at baseline AND at the designated time point	4/38 (11%)
□ Other: Patients submitting any part of the PRO assessment at the designated time point	1/38 (3%)
□ Don't know	1/38 (3%)
Fixed denominator rate: Denominator	
□ Randomized patients (ITT population)	21/38 (55%)
□ Patients with a PRO baseline assessment	6/38 (16%)
□ Enrolled patients	2/38 (5%)
□ Eligible patients[3]	2/38 (5%)
□ Safety population (patients who received intended treatment)	1/38 (3%)
□ Other	4/38 (11%)

[3] It 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

○ Depends on analysis population: ITT or mITT ○ Depends on study objective ○ ITT minus patients not eligible for PRO assessment	□ 2 (5%) □ 1 (3%) □ 1 (3%)
□ Don’t know	2/38 (5%)
Fixed denominator rate: Terminology	
□ Completion rate	20/38 (53%)
□ Compliance rate	8/38 (21%)
□ Other	6/38 (16%)
□ Don’t know/N.A.	4/38 (11%)
Variable denominator rate: Numerator	
□ On-study patients submitting the PRO assessment at the designated time point	30/38 (79%)
□ On-study patients submitting the PRO assessment at baseline AND at the designated time point	6/38 (16%)
□ Don’t know	2/38 (5%)
Variable denominator rate: Denominator (defining who the “available patients at time *t*” are)	
□ Patients who have died prior to assessment time t to be excluded from the denominator	34/38 (89%)
□ Patients not on study anymore to be excluded from the denominator	27/38 (71%)
□ Patients no longer part of the PRO assessment schedule (according to protocol) to be excluded from the denominator	24/38 (63%)
□ Ineligible patients**Error! Bookmark not defined.** to be excluded from the denominator	19/38 (50%)
□ Patients not on treatment anymore to be excluded from the denominator	10/38 (26%)
□ Patients illiterate in the language of the PRO tool to be excluded from the denominator	10/38 (26%)
□ Patients without a valid PRO baseline assessment to be excluded from the denominator	7/38 (18%)
□ Patients who cannot be reached at the time of the visit to be excluded from the denominator	4/38 (11%)
□ Patients refusing to respond the PRO assessment to be excluded from the denominator	3/38 (8%)
□ Other to be excluded from the denominator ○ Patients not meeting the clinically significant change criterion ○ Patients without valid PRO baseline assessment or not, depending on the situation	2/38 (5%) □ 1/38 (3%) □ 1/38 (3%)
Variable denominator rate: Terminology	
□ Completion rate	9/38 (24%)
□ Compliance rate	17/38 (45%)
□ Other	7/38 (18%)
□ Don’t know/N.A.	5/38 (13%)

Table 5. Missing Data Working Group survey results assessing reasons for non-completion towards development of a standardized case report form (N=19 respondents; survey distributed to 41).

Reason for non-completion of the PRO assessment	Include this reason on CRF[1]	Reason is related to the patient's health[2]	Missing data due to this reason would adversely affect your evaluation of data quality[2]
□ Patient died	19/19 (100%)	16/16 (100%)	4/16 (25%)
□ Patient withdrew from study	19/19 (100%)	4/16 (25%)	6/16 (38%)
□ Not required per protocol because patient ended protocol treatment	16/19 (84%)	4/16 (25%)	3/16 (19%)
□ Unable to accommodate disability or language needs, specify: ________	18/19 (95%)	2/16 (13%)	5/16 (31%)
No clinic visit			
□ Patient missed/canceled the clinic visit	19/19 (100%)	3/16 (19%)	8/16 (50%)
□ No clinic visit due to treatment hold or delay	16/18 (89%)	6/16 (38%)	7/16 (44%)
□ No clinic visit was scheduled by mistake	16/17 (94%)	1/16 (6%)	8/16 (50%)
□ Other reason, specify: ________	17/17 (100%)	NA	NA
Not administered			
□ Staff considered patient too ill	18/18 (100%)	14/16 (88%)	11/15 (73%)
□ Staff misinterpreted protocol	14/18 (78%)	0/16 (0%)	10/16 (63%)
□ Staff unavailable	14/18 (78%)	0/16 (0%)	9/16 (56%)
□ Staff forgot to administer	14/18 (78%)	0/16 (0%)	10/16 (63%)
□ Staff gave patient incorrect questionnaire	12/18 (67%)	0/15 (0%)	10/16 (63%)
□ Paper questionnaire unavailable	14/18 (78%)	0/15 (0%)	9/16 (56%)
□ Electronic questionnaire unavailable (e.g., malfunction or technological issue)	13/18 (72%)	1/16 (6%)	9/16 (56%)
□ Other reason, specify: ___ _____	16/18 (89%)	NA	NA
Administered but patient refused or at home questionnaire not returned			
□ Patient reported being too ill	17/18 (94%)	14/16 (88%)	10/16 (63%)
□ Patient did not like content of questionnaire	14/16 (88%)	0/16 (0%)	10/16 (63%)
□ Patient felt it was inconvenient	14/17 (82%)	1/16 (6%)	10/16 (63%)
□ Patient forgot	14/17 (82%)	0/16 (0%)	8/16 (50%)
□ Patient indicated questionnaire was returned, but it was not received by site	13/17 (76%)	0/16 (0%)	8/16 (50%)
□ Patient lost paper questionnaire	13/17 (76%)	0/16 (0%)	7/16 (44%)
□ Patient reported electronic questionnaire malfunction or technological issue	15/17 (88%)	0/16 (0%)	9/16 (56%)
□ Patient did not give a reason	14/17 (82%)	NA	NA
□ Other reason, specify: ________	15/17 (88%)	NA	NA
□ Unable to contact patient	18/18 (100%)	1/16 (6%)	8/16 (50%)
□ Other reason, specify: ________	17/17 (100%)	NA	NA
Do you believe that:			
…these reasons for non-completion are easy for research personnel to understand?[3]	12/16 (75%)		
…research personnel can successfully complete this case report form?[3]	12/16 (75%)		

…including the following question on the case report form is helpful: "Is the reason for non-completion related to the patient's health?"[1]	9/15 (60%)

1. Number and percentage responding as “Yes” versus “No”.
2. Number and percentage responding as “Yes” versus “No” or “Unsure” combined into a single group.
3. Number and percentage responding as “Strongly agree” or “Agree” combined into a single group versus “Neither agree nor disagree”, “Disagree”, or “Strongly disagree” combined into a single group.

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.

Appendix 3 - Results from the consensus meeting: non-ratified statements and voting results

Table 1. SISAQOL non-ratified statements and their considerations

No.	Non-ratified statement (NRS)	Status	Considerations
NRS 1	For evaluating a proportion of patients (with an improvement, stable state or worsening) at time t, we recommend the Cochran Mantel-Haenszel test, logistic mixed model, simple logistic regression model.	POSTPONED	Please refer to Appendix 2 (Table 3.b) to find more details on how the statistical methods were evaluated based on the agreed set of criteria. The logistic mixed model, an extension of the linear mixed model, was proposed as alternative because of the less favorable evaluation of the (Cochran) Mantel-Haenszel test on the set criteria. The mixed model will provide an unbiased estimate of the treatment effect if missing data is dependent on known and observed factors [1], whereas the (Cochran) Mantel-Haenszel test is based on observed cases data [2] and thus only provides valid inference when missing data are missing completely at random. There were reservations for recommending the logistic mixed model due to practical considerations that limit the use of these models [3], including convergence issues. To address this potential limitation, the simpler logistic model was also proposed. The decision whether a logistic mixed model, a (Cochrane)-Mantel Haenszel test or a simple logistic model will be recommended was postponed until these methods are further explored.
NRS 2	PRO assessments are no longer expected from patients who are off the PRO protocol.	POSTPONED	There was variation in calculating the variable denominator rate. To standardize the denominator of this rate, it was agreed to standardize reasons for patients going off PRO assessment (*i.e.* patients from whom we do not expect PRO assessments anymore). The implication is that these reasons are not seen as missing data, because PRO assessments are not expected from these patients anymore. *Off PRO protocol:* The protocol describes details on timing and planning of PRO assessments. Under the assumption that the PRO assessment schedule reflects the PRO trial objectives [4] (and thus reflecting what is meaningful for PRO analysis), it was proposed to consider assessments from patients off the PRO protocol as no longer expected because these assessments are not "meaningful for analysis". This means that assessments from patients off PRO protocol do not have a relevant contribution to the PRO estimate. *Withdrawing consent*: The distinction was made between (a) a patient refusing to complete one or more PRO assessments (e.g., due to patient being too sick, questionnaire too long, ...) [5] and (b) a patient refusing (to continue) participation in the PRO study, referred to as PRO withdrawal. In the case of PRO refusal (a), the patient refuses one or more PRO assessments, but is still on PRO study. In the latter (b), the patient explicitly and voluntarily terminates informed consent to participate in the PRO study (or the broader clinical trial), for whatever reason [6], entailing that the patient is (no longer) on PRO study. It was proposed to consider assessments from patients withdrawing consent from the PRO study as off PRO study. Assessments from patients off PRO study are no longer to be collected and thus no longer to be expected. *Loss to follow-up*: Being lost to follow-up was proposed as a possible reason that can lead a patient into being off PRO study and thus off PRO assessment. The definition of loss to follow-up is vaguely defined as the loss of participants during the course of a study [7]. As a consequence, great variability exists concerning the definition of loss to follow-up in the literature [8]. It was decided to postpone the voting on this proposed statement until agreement is reached on a definition for being lost to follow-up. It was difficult to agree whether the above reasons should be considered as missing data or not, depending on the different trial settings. Further discussion on the consequences of categorizing these reasons as being off PRO assessment are needed.
NRS 3	PRO assessments are no longer expected from patients who explicitly withdraw consent from the PRO study.	POSTPONED	
NRS 4	PRO assessments are no longer expected from patients who are lost to follow-up.	POSTPONED	

No.	Non-ratified statement (NRS)	Status	Considerations
NRS 5	We should establish percentage boundaries for missing data.	REJECTED	There is currently no standard rule of how much missing data is too much [9]. To address this question, the possibility of having percentage boundaries for missing data was proposed (e.g. statistical inference is not recommended with missing data rates above 50% and caution is required with missing data rates are between 10 and 50%). Monte Carlo simulations showed mixed results on bias and power in a typical superiority RCT, depending on a number of factors such as missing data mechanism, choice of analysis method and sample size[10]. Based on these results, it was discussed that it is not possible to have one overall cut-off value (e.g. the impact of 40% missing data in a trial with 10 patients is higher than in a trial with 25000 patients or the acceptance threshold might depend on whether the disease stage is early, advanced or chronic). It was therefore agreed NOT to establish percentage boundaries for how much missing data is too much when evaluating PRO outcomes. Sensitivity analyses were suggested as an alternative way to assess the impact of missing data on PRO findings (see CS 32 on the specification of sensitivity analyses in the protocol/statistical analysis plan).
NRS 6	The lower boundary of the missing data rate should be 10% (or alternatively 15%), meaning that a missing data rate of 10% (or alternatively 15%) or less is unlikely to substantially bias a between-arm PRO analysis.	CANCELLED	Based on the outcome of NRS 5, the voting on a proposal of actual missing data thresholds was cancelled.
NRS 7	The upper boundary of the missing data rate should be 50%, meaning that we would question the data quality in a between-arm PRO analysis with a missing data rate above 50%.	CANCELLED	Based on the outcome of NRS 5, the voting on a proposal of actual missing data thresholds was cancelled.
NRS 8	Agreement with modifications to the proposed case report form (CRF)?	POSTPONED	Results from a simulation study showed that the impact of missing data rates on PRO findings depends on the reasons for missing data (e.g., informative, non-informative or a combination of both[10]). Therefore collecting reasons for missing data is key in assessing the impact of missing data rates on the robustness of PRO findings. Ideally the reason for missing data should be identified to provide more information on the possible impact of missing data and how it should be handled. This way, the level to which results may be biased can be assessed [4] and the most appropriate analysis method can be identified [11]. It was decided to develop a template for capturing these reasons of missingness, to be used in PRO reports. A standard case report form (CRF) to be administered by clinical staff during PRO collection with reasons of missingness was proposed. After expression of concern for staff burden, it was decided that further fine-tuning of the proposed template is needed. Ratification of a final template for collecting reasons of PRO non-completion was postponed.
NRS 9	Agreement with collecting the question 'Is the reason for non-completion related to the patient's health?'	POSTPONED	To assess whether the collected reason for non-completion of the PRO assessment is related to the outcome variable - and thus to determine the underlying missing data mechanism -, the inclusion of the question '*is the reason for non-completion related to the patient's health*' was proposed. The utility of this item was however questioned, as it was unclear whether we could ultimately rely on this data. To avoid redundancy and capture of unreliable data [12], it was decided to further assess the utility of this item before inclusion in the standard template for capturing reasons for PRO non-completion. It was decided to postpone the voting on this proposed statement.
NRS 10	Do you agree that the reasons in the proposed CRF for non-completion are easy for research personnel to understand?	POSTPONED	The design of the case report form is key for ensuring the quality of the data collected by the CRF. Guidelines for CRF design state that CRF design should address the needs of all users and the language used should be simple and easy to understand [12]. Based on the outcome of NRS 8, it was decided to await a more developed template before evaluating

No.	Non-ratified statement (NRS)	Status	Considerations
			whether the reasons in the CRF are easy for research personnel to understand.
NRS 11	Do you agree that research personnel can successfully complete this CRF?	POSTPONED	Based on the outcome of NRS 8, it was decided to await a more developed template before evaluating whether the reasons in the CRF are easy for research personnel to complete.

Table 2. Summary of proposed statements and voting results.

Outcome[1]	Proposed statement	Absolute number of votes					Agreement[2] (in %)
		Agree	Dis-agree	Abstain/no vote	Total incl. abstain	Total excl. abstain	
	Taxonomy of Research Objectives						
RATIFIED	1.Two broad PRO research objectives: (1) treatment efficacy/clinical benefit (2) describe patient perspective	30	0	1	31	30	100 %
RATIFIED	2. Clearly state that the PRO domain/item of interest will be used to provide evidence for pre-specifying superiority, equivalence and non-inferiority	30	0	1	31	30	100 %
RATIFIED	3. Taxonomy of PRO objectives: Valid PRO objectives for treatment efficacy/clinical benefit at the within-individual / within-treatment level (for each pre-specified domain) are: - Improvement *(time to improvement, proportion of patients with improvement at time t, magnitude of improvement at time t)* - Worsening *(time to worsening, proportion of patients with worsening at time t, magnitude of worsening at time t)* - (End of) stable state *(time to end of stable state, proportion of patients with stable state at time t)*	30	0	1	31	30	100 %
RATIFIED	4. Taxonomy of PRO objectives: A valid PRO objective for treatment efficacy/clinical benefit at the within-individual/within-treatment level (for each pre-specified domain) is the overall effect: *overall PRO score over time.*	28	1	2	31	29	97 %
RATIFIED	5. Taxonomy of PRO objectives: A valid PRO objective for treatment efficacy/clinical benefit at the within-individual/within-treatment level (for each pre-specified domain) is the overall effect: *describing response trajectory over time (response patterns/profiles)*	30	0	1	31	30	100 %
RATIFIED	6. Definition of Improvement: change from baseline that reaches a pre-defined improvement threshold level (post-baseline improvement). This improvement is maintained if follow-up assessments remain at or are higher than the improvement threshold (definitive improvement). Improvement is discontinued once a follow-up assessment is below the improvement threshold (transient improvement)	30	0	1	31	30	100 %
RATIFIED	7. Definition of Worsening: change from baseline that reaches a pre-defined worsening threshold level (post-baseline worsening). This worsening is maintained if follow-up assessments remain at or are lower than the worsening threshold (definitive worsening). Worsening is discontinued once a follow-up assessment is above the worsening threshold (transient worsening)	30	0	1	31	30	100 %
RATIFIED	8. Definition of Stable State: no change from baseline is observed, or change from baseline is within the pre-defined baseline margin. This stable state is maintained if follow-up assessments remain at the baseline pre-defined margin. The stable state is discontinued once the follow-up assessment leaves the pre-defined baseline margin (and reaches the improvement or worsening threshold)	27	3	1	31	30	90 %

RATIFIED	9. Definition of the broad ‘overall effects’: summarize all available scores over time for each patient on a specific PRO domain/item	25	2	4	31	27	93 %
Recommending Statistical Methods							
RATIFIED	10. Essential statistical features for analyzing PRO data are: • ability to perform a statistical test between two samples • ability to produce clinically relevant results Highly desirable statistical features are: • ability to adjust for covariates, including baseline PRO score • ability to handle missing data with the least restrictions • ability to handle clustered data (repeated assessments)	30	0	1	31	30	100 %
RATIFIED	11: For evaluating time to event (*improvement, stable state or worsening)* outcomes, the Cox proportional hazards instead of the log rank test is recommended.	23	0	8	31	23	100 %
RATIFIED	12: For evaluating the magnitude of event (*improvement, stable state or worsening)* at time t, the linear mixed model (time as discrete variable) is recommended	26	1	4	31	27	96 %
RATIFIED	13: For evaluating the magnitude of event at time t (simplified case where only 1 FU assessment available by design), linear regression is recommended	28	0	3	31	28	100 %
POSTPONED	14: For evaluating a proportion of patients *(with an improvement, stable state or worsening)* at time t, we recommend the Cochran Mantel-Haenszel test/logistic mixed model?	/	/	/	/	/	/
RATIFIED	15: Summary measures should be part of SISAQOL (as a way to assess overall effects)	16	4	11	31	20	80 %
RATIFIED	16: For describing a response trajectory over time (as a way to assess overall effects), it is recommended to use a linear mixed model (omnibus test; time as discrete variable; time*group interaction) over the repeated measures ANOVA (time*group interaction)	27	0	4	31	27	100 %
Standardizing Statistical Terminology							
RATIFIED	17: Definition of missing data: Missing data are data that would be meaningful for the analysis of a given research objective or estimand, but were not collected	30	0	1	31	30	100 %
RATIFIED	18: ”Meaningful for analysis” refers to the PRO analysis population, which is based on the given research objective or estimand	30	0	1	31	30	100 %
RATIFIED	19: We are not expecting data anymore from patients who have died (although these patients were part of the PRO study population)	29	0	2	31	29	100 %
POSTPONED	20: We are not expecting data anymore from patients who are off the PRO protocol	/	/	/	/	/	/
POSTPONED	21: We are not expecting data anymore from patients who explicitly withdraw consent from the PRO study	/	/	/	/	/	/

POSTPONED	22: We are no longer expecting data from patients who are lost to follow-up	/	/	/	/	/	/
RATIFIED	23: Calculation of the 'variable' denominator rate: Numerator as 'number of patients on PRO assessment submitting the PRO assessment at the designated time point' and denominator as 'Number of patients on PRO assessment at the designated time point'.	30	0	1	31	30	100 %
RATIFIED	24: Calculation of the 'fixed' denominator rate: Numerator as 'number of patients on PRO assessment submitting the PRO assessment at the designated time point' and denominator as 'number of patients in the PRO study population (all patients who consented and were eligible to participate in the PRO data collection)'.	28	0	3	31	28	100 %
RATIFIED	25: Reporting of completion/compliance rates: In addition to percentages, absolute numbers for both numerator and denominator should be reported at every time point (for both rates)	30	0	1	31	30	100 %
RATIFIED	26: The term 'completion rate' should be used to express the rate with the variable denominator rate.	30	0	1	31	30	100 %
RATIFIED	27: The term 'available data rate' should be used to express the rate with the fixed denominator rate.	25	1	5	31	26	96 %
Missing Data							
RATIFIED	28: When conducting clinical trials, exploring the reasons for missing PROs is important.	30	0	1	31	30	100 %
REJECTED	29: We should establish percentage boundaries for missing data.	5	17	9	31	22	23 %
CANCELLED	30: The lower boundary of the missing data rate should be 10/15%, meaning that a missing data rate of 10/15% or less is unlikely to substantially bias a between-arm PRO analysis.	/	/	/	/	/	/
CANCELLED	31: The upper boundary of the missing data rate should be 50%, meaning that we would question the data quality in a between-arm PRO analysis with a missing data rate above 50%.	/	/	/	/	/	/
POSTPONED	32: Agreement with modifications to the proposed CRF?	/	/	/	/	/	/
POSTPONED	33: Agreement with collecting the question 'Is the reason for non-completion related to the patient's health?'	/	/	/	/	/	/
POSTPONED	34: Do you agree that the reasons in the proposed CRF for non-completion are easy for research personnel to understand?	/	/	/	/	/	/
POSTPONED	35: Do you agree that research personnel can successfully complete this CRF?	/	/	/	/	/	/
RATIFIED	36: Minimize missing data prospectively through clinical trial and PRO design strategies and by training/monitoring approaches.	29	0	2	31	29	100 %
RATIFIED	37: We recommend capturing data that will be needed for handling missing PRO data prospectively in the statistical analysis plan (i.e., reasons for missing data and auxiliary data for interpretation/imputation).	29	0	2	31	29	100 %
RATIFIED	38: Primary statistical analysis approach: Missing data approach at the item- and scale-level should be specified *a priori* within the protocol/statistical analysis plan.	29	0	2	31	29	100 %

RATIFIED	39: Primary statistical analysis approach: Critical assessment of missing data reasons and rates (by arm and time point) should be undertaken.	29	0	2	31	29	100 %
RATIFIED	40: Primary statistical analysis approach: Item-level missing data within a scale should be handled according to the scoring algorithm developed during the scale's development (when available).	28	0	3	31	28	100 %
RATIFIED	41: Primary statistical analysis approach: Use all available data, using the specified method from Statistical Methods WG Recommendations.	29	0	2	31	29	100 %
RATIFIED	42: Primary statistical analysis approach: Explicit imputation is not recommended unless justified within the context of the clinical trial.	29	0	2	31	29	100 %
RATIFIED	43: Sensitivity analyses should be specified *a priori* within the protocol/statistical analysis plan. Use of at least two different approaches to handle missing data is recommended to assess impact of missing data across various assumptions.	26	1	4	31	27	96 %

[1] Four possible outcomes for the proposed statements: *ratified, rejected, cancelled or postponed.*
RATIFIED*:* At least two third agreed with the proposed statement.
REJECTED: More than half disagreed with the proposed statement.
CANCELLED: Voting for the proposed statement was cancelled because the statement was made obsolete due to the preceding votes or discussions.
POSTPONED: Voting for the proposed statement was postponed because the statement has to be further explored /discussed first.
[2] Agreement (in %) is calculated as the number of green votes divided by the total number of green and red votes (abstain excluded).

References

1 Xu S, Blozis SA. Sensitivity Analysis of Mixed Models for Incomplete Longitudinal Data. *J Educ Behav Stat* 2011; **36**: 237–56.

2 Ali MW, Talukder E. Analysis of Longitudinal Binary Data with Missing Data Due to Dropouts. *J Biopharm Stat* 2005; **15**: 993–1007.

3 Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J R Stat Soc Ser B Stat Methodol* 1999; **61**: 265–85.

4 Calvert MJ, Freemantle N. Use of health-related quality of life in prescribing research. Part 2: methodological considerations for the assessment of health-related quality of life in clinical trials. *J Clin Pharm Ther* 2004; **29**: 85–94.

5 Atherton PJ, Burger KN, Pederson LD, Kaggal S, Sloan JA. Patient-reported outcomes questionnaire compliance in Cancer Cooperative Group Trials (Alliance N0992). *Clin Trials* 2016; **13**: 612–20.

6 Gabriel AP, Mercado CP. Data retention after a patient withdraws consent in clinical trials. *Open Access J Clin Trials* 2011; **3**: 15.

7 Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. 2011.

8 Vervölgyi E, Kromp M, Skipka G, Bender R, Kaiser T. Reporting of loss to follow-up information in randomised controlled trials with time-to-event outcomes: a literature survey. *BMC Med Res Methodol* 2011; **11**: 130.

9 Papageorgiou G, Grant SW, Takkenberg JJM, Mokhles MM. Statistical primer: how to deal with missing data in scientific research?†. *Interact Cardiovasc Thorac Surg* 2018; **27**: 153–8.

10 Mazza G, Coens C, Pe M, *et al.* How Much Missing Data is Too Much? Monte Carlo Simulations to Develop SISAQOL Guidelines for Missing Data Handling. 25th Annual Conference of the International Society for Quality of Life Research, Dublin, Ireland October 2018. *Qual Life Res* 2018; **27**: ab208.1, p:43.

11 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.

12 Bellary S, Krishnankutty B, Latha MS. Basics of case report form designing in clinical research. *Perspect Clin Res* 2014; **5**: 159–66.