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

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

International Standards for the Analysis of Quality of Life and Patient Reported Outcomes Endpoints in Cancer Randomized 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], Laura Lee Johnson[14], Madeleine King[15], Paul G Kluetz[14], 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], Rajeshwari Sridhara[14], Martin J B Taphoorn[27], Galina Velikova[28], 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

[2]Alliance Statistics and Data Center, Mayo Clinic, Scottsdale, AZ, USA

[3]Alliance Statistics and Data Center, Mayo Clinic, Rochester, MN, USA

[4]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA

[5]Centre for Patient Reported Outcomes Research, Institute of Applied Health Research and NIHR Birmingham Biomedical Research Centre, University of Birmingham, UK

[6]Patient Relevant Evidence, San Francisco, CA, USA

[7]Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[8]Adelphi Values, Bollington, Cheshire, UK

[999] Centre for Health Policy, School of Population and Global Health, University of Melbourne, Australia

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

[11]School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada

[12]Boehringer-Ingelheim International GmbH, Ingelheim, Germany

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

[14]US Food and Drug Administration, Silver Spring, MD, USA

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

[16]Center for Clinical Studies, University Hospital Regensburg, Regensburg, Germany

[17]College of Pharmacy, University of Arizona, Tucson, AZ, USA

[18]Outcomes Research Branch, Healthcare Delivery Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA

[19]Medicines and Healthcare products Regulatory Agency, London, UK

[20]International Brain Tumour Alliance, Surrey, UK

[21]Genentech, a member of the Roche group, San Francisco, CA, USA

[22]Internal Medicine/Oncology, Institut Jules Bordet, Brussels, Belgium

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

[24]VU University Medical Center, Department of Neurology & Brain Tumor Center, Amsterdam, The Netherlands

[25]Institute for Quality and Efficiency in Health Care, Cologne, Germany

[26]Health Canada, Ottawa, ON, Canada

[27]Leiden University Medical Center/Haaglanden Medical Center, Leiden/The Hague, Netherlands

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

Acknowledgments of research support for the study

EORTC received an unrestricted education grant from Boehringer Ingelheim GmbH to initiate this work and from Genentech, a member of the Roche Group, and Fonds Cancer (FOCA) in Belgium to continue the work.

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: XXXX

Note

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) or Health Canada.

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". 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 outcome (PRO) data, assessing symptoms, functioning and other aspects of health-related quality of life, are increasingly being evaluated in cancer randomized controlled trials (RCTs) to provide information on treatment risks, benefits, and tolerability. However, expert opinion and critical review of the literature have demonstrated no consensus on the analysis of PRO data 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 set recommendations for PRO analysis in cancer RCTs. Four specific issues were prioritized: (a) developing a taxonomy of research objectives that can be matched with appropriate statistical methods, (b) identifying appropriate statistical methods to analyze PRO data, (c) standardizing statistical terminology related to missing data, and (d) determining appropriate ways of handling missing data. In this review, we present SISAQOL's first set of PRO analysis recommendations. They were developed through critical reviews of the literature and a structured collaborative process with diverse international experts and stakeholders, providing a strong foundation for widespread endorsement of these recommendations.

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]. As resources to cover costs of cancer care become scarcer and treatment costs increase [8], it is critical that PRO findings are obtained and analyzed consistently across studies to produce meaningful and reliable results that can aid treatment choices and policy decisions. 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 comprises international experts including PRO researchers and statisticians, representatives from regulatory bodies, academic societies, pharmaceutical industry, cancer institutes, and patient organizations. This document presents SISAQOL's first set of recommendations for PRO analysis in cancer randomized controlled trials (RCTs), excluding preference weighted measures. It focuses on four key priorities [9]: (a) developing a taxonomy of research objectives that can be matched with appropriate statistical methods, (b) identifying appropriate statistical methods to analyze PRO data, (c) standardizing statistical terminology related to missing data, and (d) determining appropriate ways of handling missing data.

Methods

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 were appointed by the European Organisation for Research and Treatment of Cancer (EORTC) to standardize PRO analysis in cancer RCTs. In 2016, AB and CC invited experts and different 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 discussed 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. An internal EORTC team was formed to support the consortium. Expertise and profiles of the invited experts at every stage of the development of these recommendations can be found in Appendix 1, Table 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 [10], and reporting of PROs in publications [11–15].

3. Strategic meeting

Twenty-nine experts and stakeholders attended the strategy meeting in 2018. 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 of handling missing data

4. Working Groups

Based on the agreed on 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) general handling of missing data [9]. 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 1, Table 2.

Research objectives working group. Systematic reviews consistently showed a lack of well-defined PRO research hypotheses in cancer RCTs [5,6,10,13,15]. A well-defined PRO hypothesis is needed to 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 2, Table 1 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 [9].

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 PRO analysis. An open-ended question was also

included to capture additional criteria. Survey results were discussed and the set of criteria was updated until individual concerns were addressed (Appendix 2, Table 2 for survey results).

The agreed on 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 2, Table 3 for detailed results of this report). Recommended methods for each PRO objective were discussed and updated until all individual concerns were addressed.

Standardizing statistical terms (focus on defining and evaluating missing data). Missing PRO data is the on-going challenge in cancer clinical trials because patients drop out of study for different reasons, including predefined progression of disease, death, intolerable toxicity, and patient or clinician decision [16–18]. 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 [19]. However, the very definition of "missing data" remains opaque and elusive. For example, it is unclear whether unobserved assessments after patients drop out of study because of disease progression is missing data. 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 relationships with the PRO study population 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 published RCTs in a systematic literature review of metastatic breast cancer reports [5]. An exploratory literature search in additional peer-reviewed publications was conducted to check for other definitions of missing data and ways to calculate missing data rates. Consortium members responded to a survey to standardize these definitions (Appendix 2, Table 4 for survey results). Findings were discussed until all individual concerns were addressed.

Handling of missing data. The missing data working group was tasked with identifying whether it was possible to set a threshold for missing data based on simulation studies (how much missing data is too much?); develop a standardized case report form (CRF) to identify reasons for PRO non-completion; recommend a general strategy for handling 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 a the basis for later recommendations on thresholds for missing data[20]

In an effort to develop a standardized case report form (CRF) with possible reasons for PRO non-completion, existing CRF templates from seven different institutions were collected [21]. 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 2 Table 5 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) handling of 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.

Results

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 3, Table 1. 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 3, 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 will 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 PRO domain(s) and time frame of interest should be pre-specified [22,23]. 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).

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 could not be recommended as appropriate for PRO analysis.

Two essential statistical properties were identified: the ability to perform a comparative test (statistical significance) and to produce interpretable treatment effect estimates (clinical relevance). Highly desirable criteria were the ability to adjust for covariates, including baseline PRO score, handle missing data with the least restrictions, and handle 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 are advantages in recommending more complex models (e.g., pattern mixture models), this often comes at the cost of strong and

untestable assumptions and produces results that may not be easily interpretable for non-statisticians. A balance between feasibility, usefulness, interpretability and statistical correctness remains 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 analysis, parametric methods were recommended over non-parametric methods. However, parametric methods have limitations, most importantly, their reliance on distributional assumptions [24]. To address this limitation, it was recommended that non-parametric methods should be used for sensitivity analysis to investigate deviations from these assumptions [24].

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 ordinal data can be analyzed as a continuous measure 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%). Recommendation on the definition of missing PRO data was proposed: missing PRO data is data that would be meaningful for the analysis of a given research objective, but were not collected for any reason (Table 1, RS 16-17; [25,26]. This definition clarifies 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 [27]. Recommendations on how to specifically deal with missing data for each recommended method is the next step for 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 (a fixed denominator rate) has the number of patients on PRO assessment submitting a valid PRO assessment at the designated time point as numerator and the number of patients in the PRO study population as denominator (i.e. all patients who consented and were eligible to participate in the PRO data collection). 2) The completion rate

(a variable denominator rate) also has number of patients on PRO assessment submitting a valid PRO assessment at the designated time point as numerator and the number of patients on PRO assessments at the designated time point as denominator (i.e. all patients who are still expected to provide PRO assessments at that time point). Of note, the denominator of the completion rate depends on the chosen research question, e.g. if PRO 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: Handling of 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[20]. 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; [18]. A case report form to collect in a standardized way reasons for missing data is needed and will be further developed. General recommendations on how to handle missing data were proposed (see Table 1, RS 25 - 30).

Discussion

The aim of SISAQOL is to develop a set of recommendations to facilitate standard methods 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 test was recommended as an appropriate analysis method for time-to-event outcomes; and the linear mixed model was recommended for magnitude of event at time t, and response patterns/profiles. Recommendations on a standardized definition of missing PRO data, and reporting of completion and available data rates were proposed. Some general recommendations for handling missing PRO data were also suggested.

Generating robust PRO conclusions from cancer clinical trials is not only about improving defining 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 hope this set of recommendations will support clinical researchers and improve the quality of statistical analysis and 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, such as Standard Protocol Items: Recommendation for Interventional Trials in Patient Reported Outcomes [23], the Consolidated Standards of Reporting Trials in Patient Reported Outcomes [28], and other relevant guidelines [29,30]. Given the significant unmet 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.

Limitations and Future Work

Although this document presents the first set of standards for PRO analysis in cancer RCTs, much work still needs to be done. First, several proposed statements need more discussion (e.g., statistical method for proportions of patients at time t, summary measures and several issues on missing data; see Appendix 3 Table 1 for more details). Second, the taxonomy of research objectives needs to be applied to cancer clinical trials so they can be updated and validated, ensuring that they are fit-for-use when planning trials with a PRO endpoint. 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 were allowed to suggest other methods to include in the evaluation, other potentially appropriate statistical methods for PRO analysis may not have been included. Nonetheless, the set of statistical methods evaluated are time-tested and scientifically rigorous and can be applied in the majority of the cases. Best statistical practices for each of the recommended methods need to be agreed upon, including how to handle missing data. Fourth, an agreement on which summary measures are relevant to address specific PRO objectives is also needed. Finally, it should be examined how these recommendations relate to the recently suggested estimands framework [26]. As a critical first step, this document has already defined (a) variables that are useful for PRO analysis, and (b) population level summaries for the identified PRO variables. Future steps would include identifying

the target population and intercurrent events relevant for PRO analysis. Feasibility of applying these recommendations to other clinical contexts will also be explored.

Conclusion

Patient-reported outcome (PRO) data, assessing symptoms, functioning and other aspects of health-related quality of life are increasingly assessed in cancer RCTs to provide valuable evidence on risks, benefits, safety and tolerability of treatment. PRO findings inform patient, provider, payer and regulatory decision-making. 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 a first step towards generating 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 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.

9 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.

10 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.

11 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.

12 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.

13 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.

14 Mercieca-Bebber R, Friedlander 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.

15 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; **111**: djz038.

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

59.
17 Fayers P, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes., 2nd Editio. John Wiley & Sons, 2013.
18 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.
19 Machin D, Weeden S. Suggestions for the presentation of quality of life data from clinical trials. Stat Med 2002; **17**: 711–24.
20 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.
21 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.
22 Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of Patient-Reported Outcomes in Randomized Trials. J Am Med Assoc 2013; **309**: 814–22.
23 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 - J Am Med Assoc 2018; **319**: 483–94.
24 Altman DG, Bland JM. Parametric v non-parametric methods for data analysis. BMJ 2009; **338**: a3167.
25 Little RJ, Agostino RD, Cohen ML, et al. The Prevention and Treatment of Missing Data in Clinical Trials. N Engl J Med 2012; **367**: 1355–60.
26 ICH Expert Working Group. Estimands and Sensitivity Analysis in Clinical Trials E9(R1) Statistical Principles for Clinical Trials: Addendum. 2017 DOI:10.2527/jas.2014-8546.
27 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.
28 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.
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.
30 U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health. Guidance for Industry. Patient Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. December. 2009.
31 ICH Expert Working Group. ICH harmonized Tripartite Guideline: Statistical Principles for Clinical Trials E9. 1998.
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

549 CONSORT statement. J Am Med Assoc 2006; **295**: 1152–60.
550 34 European Medicines Agency. Guideline on the choice of non-inferiority margin.
551 2008; : 1–9.
552 35 Committe for proprietary medicinal products. Points to consider on switching
553 between superiority and non-inferiority. 2000.
554 36 King MT. A point of minimal important difference.pdf. Expert Rev
555 Pharmacoeconomics Outcomes Res 2011; **11**: 171–84.
556 37 Fairclough DL. Summary measures and statistics for comparison of quality of
557 life in a clinical trial of cancer therapy. Stat Med 1997; **16**: 1197–209.
558 38 Curran D, Aaronson N, Standaert B, et al. Summary measures and statistics in
559 the analysis of quality of life data: An example from an EORTC-NCIC-SAKK
560 locally advanced breast cancer study. Eur J Cancer 2000; **36**: 834–44.
561 39 National Cancer Institute. NCI Dictionairy of cancer terms.
562 https://www.cancer.gov/publications/dictionaries/cancer-terms (accessed May
563 29, 2019).
564 40 Wasserstein RL, Lazar AN. The ASA's Statement on p-Values: Context,
565 Process, and Purpose. Am Stat 2016; **70**: 129–33.
566 41 European Medicines Agency. Points To Consider on Adjustment for Baseline
567 Covariates. Comm. Med. Prod. Hum. Use. 2003; : 1–10.
568 42 Senn SJ. Covariate imbalance and random allocation in clinical trials. Stat Med
569 1989; **8**: 467–75.
570 43 Vickers AJ, Altman DG. Statistics Notes: Analysing controlled trials with
571 baseline and follow up measurements. BMJ 2002; **323**: 1123–4.
572 44 Fairclough DL, Peterson HF, Cella D, Bonomi P. Comparison of several
573 model-based methods for analysing incomplete quality of life data in cancer
574 clinical trials. Stat Med 1998; **17**: 781–96.
575 45 Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical Analysis of Quality of
576 Life With Missing Data in Cancer Clinical Trials. Stat Med 1998; **17**: 653–66.
577 46 Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis, 2nd
578 edition. 2011 DOI:10.1198/jasa.2005.s24.
579 47 Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II:
580 Multivariate data analysis – an introduction to concepts and methods. Br J
581 Cancer 2003; **89**: 431–6.
582 48 Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic
583 concepts and first analyses. Br J Cancer 2003; **89**: 232–8.
584 49 Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse
585 unbalanced repeated measures and longitudinal data. Stat Med 1998; **16**:
586 2349–80.
587 50 SAS. The REG Procedure: Missing Values. SAS/STAT(R) 9.3 User's Guid.
588 https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/view
589 er.htm#statug_reg_sect026.htm (accessed May 29, 2019).
590 51 Bell ML, King MT, Fairclough DL. Bias in Area Under the Curve for
591 Longitudinal Clinical Trials With Missing Patient Reported Outcome Data.
592 SAGE Open 2014; **4**. DOI:10.1177/2158244014534858.
593 52 Fairclough DL, Peterson HF, Chang V. Why are missing quality of life data a
594 problem in clinical trials of cancer therapy? Stat Med 2002; **17**: 667–77.
595 53 Mercieca-Bebber R, Palmer MJ, Brundage M, Calvert M, Stockler MR, King
596 MT. Design, implementation and reporting strategies to reduce the instance
597 and impact of missing patient-reported outcome (PRO) data: A systematic
598 review. BMJ Open 2016; **6**. DOI:10.1136/bmjopen-2015-010938.

54 European Medicines Agency. Guideline on Missing Data in Confirmatory Clinical Trials. 2011; : 1–12.
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.

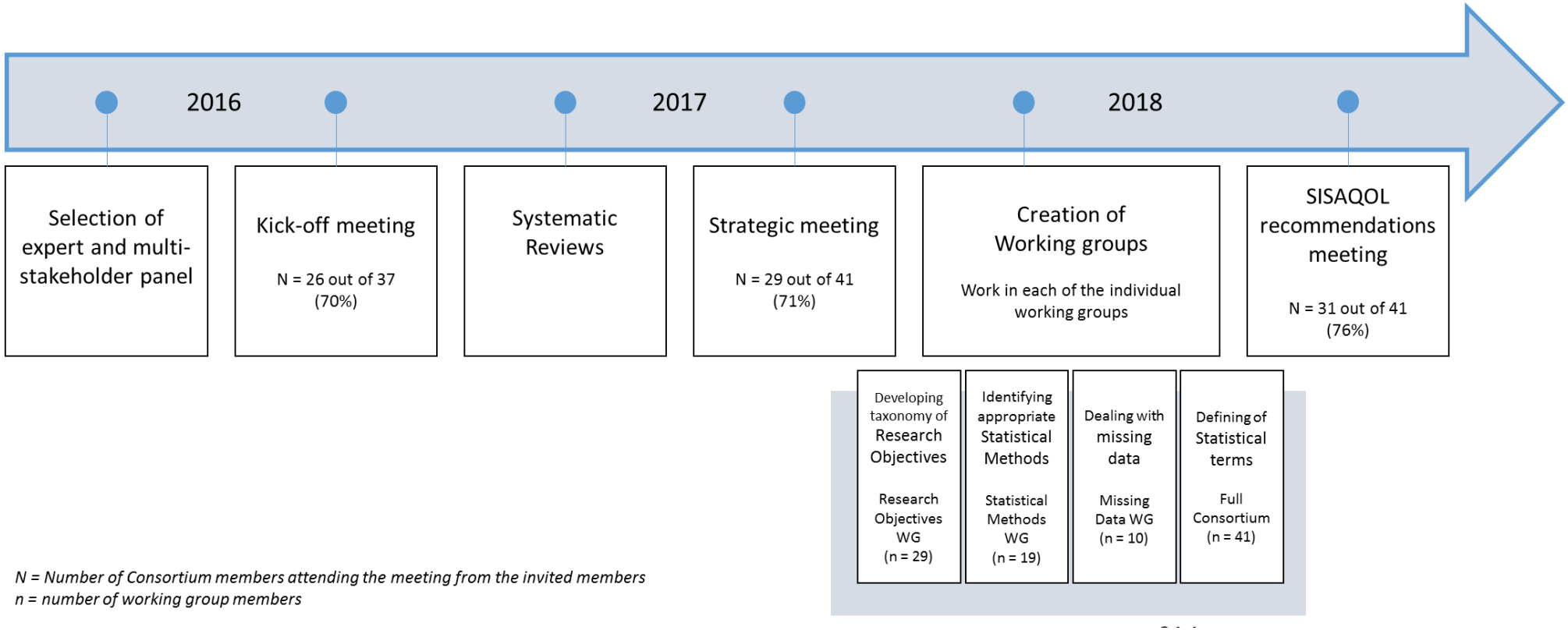

Figure 1. Overview of the process towards the development of the SISAQOL Recommendations.

615 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 comparison that will be used for each PRO domain/item of interest: - Superiority, - Equivalence / non-inferiority	Superiority design and analysis techniques differ from equivalence / 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.	**Within-treatment arm PRO assumption: improvement, worsening, stable state or overall effect.**

	Valid within-individual/within-group PRO 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,
		The choice of whether a worsening, stable state or improvement is 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

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	

		further analysis. The clinically relevant threshold is specified at the 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 Appendix 2, Table 1 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	Time to worsening: A clinically relevant within-patient level worsening is pre-defined, and the interest is in evaluating the time it takes

	worsening threshold level (post-baseline worsening). This worsening is maintained if follow-up assessments remain at or are lower than the worsening threshold. Worsening is discontinued once a follow-up assessment is above the worsening threshold. See Appendix 2, Table 1 for illustration.	before a clinically relevant worsening is observed. Variability in the 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	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.

	threshold, then stable state can still be concluded. See Appendix 2, Table 1 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 restrictions,	For more details on how this statement was developed, including the list of other statistical features considered, please see Appendix 2, Table 2. 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

	- handle clustered data (repeated assessments).	non-statisticians, guide informative clinical-decision making and 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 2, Table 3 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 2, Table 3 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 (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.	Please refer to Appendix 2, Table 3 to find more details on how the 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 variable; time*group interaction) over the repeated measures ANOVA (time*group interaction)	Please refer to Appendix 2, Table 3 to find more details on how the statistical methods were evaluated based on the agreed set of criteria. 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 [11], 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 [19]; - any missing value that would be meaningful for analysis (if they were observed) [25,26].

		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 population).	PRO assessments after death should not be expected because a meaningful value for these observations will not exist [19,26]. These assessments are also not "meaningful for analysis" because they will 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 3, Table 1).
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 denominator should be reported at every time point (for both rates).	It was proposed that a CONSORT diagram would be helpful to report the reasons for missing data. It was suggested to have three broad 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 [25]. Therefore preventing missing PRO assessments through careful design and planning should be the first line strategy in handling missing PRO data [26]. 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 [16,54].
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 [23].

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 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 [16,54].
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 analysis are required to test the robustness of the conclusions using a different set of assumptions regarding missing data. If the results are consistent with the primary analysis, this provides some assurance that the missing data did not have an important effect on the study conclusions. However, if they produce inconsistent results, their implications for 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 analysis.

616
617

618 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 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 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	Not applicable (As it is not always straightforward to pre-define the exact profiles within a time	Exploratory - Linear mixed model (time as discrete / continuous) Descriptive

	frame, response patterns/profiles are recommended to be used alongside a descriptive / exploratory objective rather than evidence for treatment efficacy / clinical benefit)	frame, response patterns/profiles are recommended to be used alongside a descriptive / exploratory objective rather than evidence for treatment efficacy / clinical benefit)	- Mean magnitude at baseline & at every time point within a time frame (with CI); - Mean change at every time 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 2, Table 3b).

624