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

Corneel Coens, Madeline Pe*, Amylou C Dueck, Jeff Sloan, Ethan Basch, Melanie Calvert, Alicyn Campbell, Charles Cleeland, Kim Cocks, Laurence Collette, Nancy Devlin, Lien Dorme, Hans-Henning Flechtner, Carolyn Gotay, Ingolf Griebsch, Mogens Groenvold, Madeleine King, Paul G Kluetz, Michael Koller, Daniel C Malone, Francesca Martinelli, Sandra A Mitchell, Jammbe Z Musoro, Daniel O'Connor, Kathy Oliver, Elisabeth Piault-Louis, Martine Piccart, Chantal Quinten, Jaap C Reijneveld, Christoph Schürmann, Ashley Wilder Smith, Katherine M Soltys, Martin J B Taphoorn, Galina Velikova, Andrew Bottomley; for the Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data Consortium*

Patient-reported outcomes (PROs), such as symptoms, function, and other health-related quality-of-life aspects, are increasingly evaluated in cancer randomised controlled trials (RCTs) to provide information about treatment risks, benefits, and tolerability. However, expert opinion and critical review of the literature showed 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 Consortium was formed to establish PRO analysis recommendations. Four issues were prioritised: developing a taxonomy of research objectives that can be matched with appropriate statistical methods, identifying appropriate statistical methods for PRO analysis, standardising statistical terminology related to missing data, and determining appropriate ways to manage missing data. This Policy Review presents recommendations for PRO analysis developed through critical literature reviews and a structured collaborative process with diverse international stakeholders, which provides a foundation for endorsement; ongoing developments of these recommendations are also discussed.

Introduction

The use of patient-reported outcomes (PROs) 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 showed no consensus on standards and unclear guidelines on how to analyse 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 important that PRO findings are obtained from good methodological practices and are analysed consistently across studies to ensure that they can meaningfully and reliably inform patient safety, treatment choices, and policy decisions, especially in an era in which 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, and representatives from international academic societies, industry, cancer institutes, and patient organisations. This Policy Review presents a set of consensus recommendations for PRO analysis in cancer randomised controlled trials (RCTs) to address four key priorities:[11] developing a taxonomy of research objectives that can be matched with appropriate statistical methods, identifying appropriate statistical methods to address specific PRO research objectives, standardising statistical terminology related to missing data, and determining appropriate ways of managing missing data.

Development of recommendations

Selection of the expert and multi-stakeholder panel

Figure 1 shows an overview of the key developments that led to the SISAQOL recommendations. Two authors of this manuscript (AB and CCo) invited experts and stakeholders experienced with PROs in cancer RCTs. The goal was 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 biannual European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group meeting and at international society meetings (eg, International Society for Quality of Life Research, International Society for Pharmacoeconomics and Outcomes Research, American Society of Clinical Oncology, and European Society for Medical Oncology) to recruit representatives. When requested, a memorandum of understanding was set up between EORTC

Lancet Oncol 2020; 21: e83–96

*Joint first authors

European Organisation for Research and Treatment of Cancer, Brussels, Belgium (C Coens Msc, M Pe PhD, L Collette PhD, L Dorme MSc, F Martinelli MSc, J Z Musoro PhD, A Bottomley PhD); **Alliance Statistics and Data Center, Mayo Clinic, Scottsdale, AZ, USA** (A C Dueck PhD); **Alliance Statistics and Data Center, Mayo Clinic, Rochester, MN, USA** (J Sloan PhD); **Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA** (E Basch MD); **Centre for Patient Reported Outcomes Research, Institute of Applied Health Research and National Institute for Health Research Birmingham Biomedical Research Centre, University of Birmingham, Birmingham, UK** (Prof M Calvert PhD); **Patient Relevant Evidence, San Francisco, CA, USA** (A Campbell MPh); **Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, TX, USA** (Prof C Cleeland PhD); **Adelphi Values, Bollington, Cheshire, UK** (K Cocks PhD); **Centre for Health Policy, School of Population and Global Health, University of Melbourne, Melbourne, VIC, Australia** (N Devlin PhD); **Clinic for Child and Adolescent Psychiatry and Psychotherapy, University of Magdeburg, Magdeburg, Germany** (Prof H-H Flechtner MD); **School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada** (Prof C Gotay PhD); **Boehringer Ingelheim**

International GmbH, Ingelheim, Germany (I Griebsch PhD); **Department of Public Health, Bispebjerg Hospital and University of Copenhagen, Copenhagen, Denmark** (Prof M Groenvold MD); **School of Psychology, University of Sydney, Sydney, NSW, Australia** (Prof M King PhD); **Office of Hematology and Oncology Products, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA** (P G Kluetz MD); **Center for Clinical Studies, University Hospital Regensburg, Regensburg, Germany** (M Koller PhD); **College of Pharmacy, University of Arizona, Tucson, AZ, USA** (Prof D C Malone PhD); **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); **Medicines and Healthcare products Regulatory Agency, London, UK** (D O'Connor MBChB); **International Brain Tumour Alliance, Surrey, UK** (K Oliver BA); **Genentech, Roche Group, San Francisco, CA, USA** (E Piault-Louis PharmD); **Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium** (M Piccart MD); **European Centre for Disease Prevention and Control, Surveillance and Response Support Unit, Epidemiological Methods Section, Stockholm, Sweden** (C Quinten MSc); **Department of Neurology and Brain Tumor Center, VU University Medical Center, Amsterdam, Netherlands** (J C Reijneveld MD); **Institute for Quality and Efficiency in Health Care, Cologne, Germany** (C Schürmann PhD); **Health Canada, Ottawa, ON, Canada** (K M Soltys PhD); **Department of Neurology, Leiden University Medical Center, Leiden** (M J B Taphoorn MD); **Department of Neurology, Haaglanden Medical Center, The Hague, Netherlands** (M J B Taphoorn); **Leeds Institute of Cancer and Pathology, University of Leeds, St James's Hospital, Leeds, UK** (Prof G Velikova PhD);

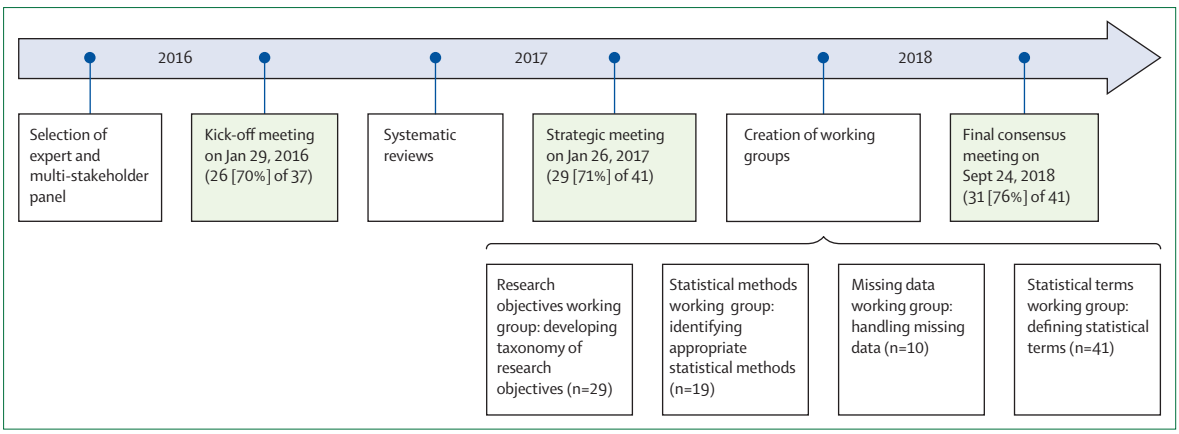

Figure 1: **Overview of the development of the SISAQOL recommendations**
The number of consortium members attending the meetings from the invited members is shown; non-attendees received the full meeting reports and could comment and add suggestions. The final version of the report was approved by the SISAQOL Consortium. n=number of working group members. SISAQOL=Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data.

and the international societies. Expertise and profiles of the invited experts at every stage of the development of these recommendations can be found in the appendix (p 1).

Expert views and systematic reviews

26 experts and stakeholders attended the SISAQOL kick-off meeting on Jan 29, 2016, to discuss challenges in PRO analysis in cancer RCTs. Agreement was reached on the absence of international standards and that this work was urgently needed.[7] Systematic reviews assessing the current state of PRO analysis in RCTs in different cancers supported this view.[4–6] Four key findings were highlighted: the absence 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]

Strategic meeting

29 experts and stakeholders attended a strategy meeting on Jan 26, 2017. On the basis of 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, standardising statistical terminology related to missing data, and determining appropriate ways to manage missing data.

Working groups

On the basis of the agreed priorities, four working groups were assembled: research objectives, statistical methods, standardisation of statistical terms (with an initial focus on defining and evaluating missing data), and management of missing data;[11] each working group had specific goals and methods (appendix pp 2–3). Final outputs from each working group were used as proposed statements for the SISAQOL recommendations.

Research objectives working group

Systematic reviews consistently showed an absence 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. Members of the research objectives working group were tasked with developing a framework for PRO research objectives that can inform the statistical method to use (ie, taxonomy of PRO research objectives), and providing standardised definitions for key PRO objectives. An initial framework was developed through discussions. The framework was circulated to all members of the research objectives working group for further refinement. A survey was done among the working group members to standardise definitions of key research PRO objectives: improvement, worsening, and stable state (appendix pp 4–12).

Statistical methods working group

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

A list of 19 statistical criteria was developed through a literature search and expert discussions. A survey was

done among the members of the statistical methods working group, 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 pp 13–15).

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 members of the statistical methods working group (appendix pp 16–26). Recommended methods for each PRO objective were discussed and amended until all individual concerns from working group members were addressed.

Standardising statistical terms working group

Missing PRO data is an ongoing challenge in cancer clinical trials, as patients drop out of a 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 this issue, the proportion of missing data in a trial should be reported in a standardised way because PRO estimates might be biased if a large number of patients do not 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 standardise the definition of missing data and the reporting of missing data, and to clarify their relationship with the PRO study population (ie, all patients who consented and were eligible to participate in the PRO data collection) and PRO analysis population (ie, patients who will be included in the primary PRO analysis). A first set of definitions and calculations for missing data was extracted from a systematic review of metastatic breast cancer RCTs.[5] An exploratory literature search in additional peer-reviewed publications was done to identify other definitions of missing data and approaches to calculate proportions of missing data. Consortium members responded to a survey to standardise these definitions (appendix pp 27–29). 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 on the basis of simulation studies (eg, how much missing data is too much?), develop a standardised case-report form 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 done to assess how increasing proportions of missing data affect 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 standardised case-report form with possible reasons for PRO non-completion, existing case-report form templates from seven different clinical trial networks were collected (eg, the case-report form 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 done among all consortium members in which they indicated whether the reason for non-completion should be included in the standard case-report form, is related to the patient's health, and affects data quality (appendix pp 30–31).

SISAQOL recommendations meeting

31 experts and stakeholders attended the SISAQOL recommendations meeting on Sept 24, 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: taxonomy of research objectives, recommending statistical methods, standardising terminology related to missing data, and managing missing data.

For each statement, participants voted 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 noted 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.

Figure 2 shows the SISAQOL recommendations and their considerations; the table shows a brief overview of these recommendations. Statements that were not ratified, including reasons for non-ratification, are shown in the appendix (pp 35–36).

SISAQOL recommendations

43 statements were presented at the recommendations meeting, of which 32 (74%) were ratified, eight (19%) were postponed, one (2%) was rejected, and two (5%) were cancelled. The appendix (pp 37–40) shows the voting results of all proposed statements.

and International Society for Quality of Life Research, Milwaukee, WI, USA
(Prof G Velikova)

Correspondence to:
Dr Madeline Pe, European Organisation for Research and Treatment of Cancer, 1200 Brussels, Belgium
madeline.pe@eortc.org

See **Online** for appendix

For more on the **SISAQOL Consortium** see http:// www.eortc.org/sisaqol

	Recommended statement	Considerations
Section 1: taxonomy of research objectives		
RS 1	Clearly state the broad PRO research objectives for each PRO domain or item of interest: • Treatment efficacy or clinical benefit • Exploratory or describe patient perspective	**Treatment efficacy or clinical benefit** If a PRO domain will be used to provide formal comparative conclusions between treatment groups, 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.[24] 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 RS 2–5. **Exploratory or 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 or exploratory objective are 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 or clinical benefit. Findings should be reported as either descriptive (ie, summarising estimates with or without confidence intervals but no statistical testing is involved), or exploratory (ie, choice of hypothesis might be data-driven and statistical testing might be involved), but this option should not be used as basis of evidence for clinical benefit or treatment efficacy.[24] Both PRO objectives are important and complement each other,[25] 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 or clinical benefit, describe the patient perspective, or are exploratory.
RS 2	Clearly state the between-treatment group comparison that will be used for each PRO domain or item of interest: • Superiority • Equivalence or non-inferiority	Superiority design and analysis techniques differ from equivalence or non-inferiority techniques.[24,26] 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 groups are similar (equivalent) or not worse (non-inferior). **Superiority** A superiority PRO objective aims to show that for the prespecified PRO domain, the treatment group is superior to the reference group by a clinically relevant treatment effect size. The effect size to show a clinically relevant treatment difference should be predefined 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.[24,27,28] **Equivalence or non-inferiority** An equivalence or non-inferiority PRO objective aims to show that for the prespecified PRO domain, the treatment group is similar (equivalent) or not worse (non-inferior) than the reference group by a prespecified clinically relevant margin. It is important that these margins are prespecified 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 or inferior treatment effect.[27] The choice of effect size (superiority) and margins (equivalence or non-inferiority) should be tailored to the PRO instrument and clinical context, and should be justified on both clinical and statistical grounds.[27] Trials might include any combination of these between-treatment group PRO objectives. However, the protocol should clearly specify which PRO domains or items will be tested for superiority, equivalence, or non-inferiority.
RS 3–5	Clearly state the within-patient or within-treatment group PRO objective in protocol. Valid within-individual or within-group PRO objectives include the following: • 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* • Overall effects o Overall PRO score over time o Response patterns or profiles	**Within-treatment group PRO assumption: improvement, worsening, stable state, or overall effect** 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 (ie, summarising 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 might be less robust. When deciding which within-treatment group PRO assumption will be used, patients' observed baseline levels on the specific PRO domain should be taken into account; this decision will help inform the feasibility of assessing a clinically relevant change for that PRO domain. **Within-patient or within-treatment group 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 or profiles** Various within-patient or within-treatment group 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 might yield different conclusions. Depending on the endpoint, the clinically relevant threshold for the PRO domain might be at the patient level (eg, within patient: classifying a patient as a responder or not) or at the group level (eg, within group: mean change within the group).[29] **Within-patient PRO objective** The primary interest is in identifying which patients had a clinically relevant response before doing further analysis. The clinically relevant threshold is specified at the individual level (ie, responder definition), which identifies whether 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 group 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 6–9 provide more specific definitions for these PRO objectives.

(Figure 2 continues on next page)

	Recommended statement	Considerations
RS 6	Improvement is defined as change from baseline that reaches a predefined 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; figure 3).	**Time to improvement** A clinically relevant within-patient level improvement is predefined, 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 predefined improvement threshold is ignored. **Magnitude of improvement at time *t*** A clinically relevant within-treatment group improvement is predefined, and the interest is in assessing the mean or median improvement (with corresponding CIs) at a predefined, clinically relevant timepoint. Variability in the observed scores is taken into account. **Proportion of responders with improvement at time *t*** A clinically relevant within-patient level improvement is predefined, and the interest is in evaluating the number of patients with improvement at a predefined clinically relevant timepoint. Variability in the scores above or below this predefined improvement threshold is ignored.
RS 7	Worsening is defined as change from baseline that reaches a predefined 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 (figure 3).	**Time to worsening** A clinically relevant within-patient level worsening is predefined, and the interest is in evaluating the time it takes before a clinically relevant worsening is observed. Variability in the scores above or below this predefined worsening threshold is ignored. **Magnitude of worsening at time *t*** A clinically relevant within-treatment group worsening is predefined, and the interest is in assessing the mean or median improvement (with corresponding CIs) at a predefined clinically relevant timepoint. 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 predefined, and the interest is in evaluating the number of patients with worsening at a predefined clinically relevant timepoint. Variability in the scores above or below this predefined worsening threshold is ignored.
RS 8	Stable state is defined as no change from baseline, or as change from baseline within the predefined baseline margin. This stable state is maintained if follow-up assessments remain at the baseline predefined margin. The stable state is discontinued once the follow-up assessment leaves the predefined baseline margin and reaches the improvement or worsening threshold. There might be circumstances where the relevant PRO objective would include improvement in the definition of stable state (ie, at least stable). In this case, the definition is that as long as follow-up assessments do not reach the deterioration threshold, then stable state can still be concluded (figure 3).	Disagreement among consortium members (during discussion) arose because the current definition of stable state implies distinction between three possible categories (improvement, worsening, or stable state). However, situations might occur where categories exist between improvement and stable state; or worsening and stable state (five categories). These additional two categories might be used as an error margin between stable state and improvement or worsening, or be included as meaningful categories (eg, 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 predefined, and the interest is in evaluating the time it takes before a clinically relevant stable state is observed. This endpoint might 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 or worsening is observed. **Proportion of responders with a stable state at time *t*** A clinically relevant within-patient level stable state is predefined, and the interest is in evaluating the number of patients with a stable state at a predefined clinically relevant timepoint. Variability in the scores above or below this predefined 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 who 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 as noise—ie, random fluctuations not representing any meaningful changes.
RS 9	Overall effect is defined as summarising all available scores over time for each patient on a specific PRO domain or item.	Disagreement among consortium members (during discussion) arose on whether overall effect endpoints can be used with a treatment efficacy or clinical benefit PRO objective. The recommendation is that overall effects can be used alongside a treatment efficacy or clinical benefit PRO objective. Since information is lost with this type of endpoint, relative to improvement, worsening, and stable state (eg, an overall PRO score over time will not capture the direction and timing of an effect), 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 summarise all available scores over a given period into a single datapoint per patient for a specific PRO domain. The timeframe of interest should be predefined. 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 mean or median, minimum and maximum, and area under the curve.[30,31] These summary measures might or might not include the baseline score, depending on the research objective. Clinically relevant thresholds should also be predefined to aid interpretation of these values. However, by summarising all available data into one score, information is lost and clinically relevant changes at particular timepoints might be obscured.[31] 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 predefined period (ie, 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 predefined to aid interpretation of these values. As it is not always straightforward to predefine the exact profiles within a timeframe, this within-patient or within-treatment group PRO research objective is recommended to be used alongside a descriptive or an exploratory objective rather than evidence for treatment efficacy or clinical benefit.

(Figure 2 continues on next page)

	Recommended statement	Considerations
Section 2: recommending statistical methods		
RS 10	Essential statistical features for analysing PRO data include the following: • Do a statistical test between two treatment groups • Produce clinically relevant results Highly desirable statistical features include the following: • Adjust for covariates, including baseline PRO score • Handle missing data with the least restrictions • Handle clustered data (repeated assessments)	The appendix (pp 13–15) provides more details on how this statement was developed, including the list of other statistical features considered. **Do a statistical test between two groups** The current scope of these recommendations is on randomised controlled trials, and testing for statistical differences between groups is the main goal of a randomised controlled trial.[32] **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 influence clinical practice. Statistically significant results do not necessarily imply that results are clinically relevant.[33] 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 methods are favoured over non-parametric methods. Since parametric methods rely on distributional assumptions, it is recommended that non-parametric methods are used for sensitivity analyses to investigate deviations from these assumptions especially when sample sizes are small.[34,35] **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.[36,37] For example, baseline PRO scores are often correlated with PRO scores at follow-up;[38] therefore, it is important to have an analytical method that can incorporate baseline covariates. Other covariates could include demographic variables (eg, age, sex), disease characteristics (eg, disease site, stage), and other relevant variables (eg, 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 to potential bias of estimates.[39] Missing data are almost always inherent when analysing PRO data in cancer clinical trials, and the most restrictive assumption that the probability of missing data is unrelated to the PRO domain or item of interest is highly unlikely.[40] **Handle clustered data (repeated assessments)** To capture changes in the PRO domain or item of interest, PROs are often assessed repeatedly over time in cancer clinical trials. Analysing these kind of data would require taking into account both the clustering of PRO assessments within each patient, and the temporal order of the measurements.[41]
RS 11	For evaluating time to event outcomes (improvement, stable state, or worsening), it is recommended to use the Cox proportional-hazards test instead of the log-rank test.	The appendix (pp 16–26) provides more details about how the statistical methods were evaluated based on the agreed set of criteria. When using the Cox proportional-hazards test, the proportional-hazards assumption should be checked.[42] If this assumption is not met, doing a sensitivity analysis with a log-rank or Cox non-proportional-hazards 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.[43]
RS 12	For evaluating magnitude of event (improvement or worsening) at time *t* (where the design is baseline plus more than one follow-up), it is recommended to use the linear mixed model (time as discrete) over the other statistical methods evaluated.	The appendix (pp 16–26) provides more details about 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 (eg, 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 small number of follow-up assessments. General assumptions of linear mixed models hold. For example, the missing at random assumption has to be satisfied—ie, the linear mixed model will provide an unbiased estimate of the treatment effect that would have been observed if missing data are dependent on known and observed factors.[44]
RS 13	For evaluating magnitude of event (improvement or worsening) at time *t* (where the design is baseline plus one follow-up only), it is recommended to use the linear regression over the ANOVA, ANCOVA, *t* test, and Wilcoxon rank-sum test.	The appendix (pp 16–26) provides more details about 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 (eg, SAS).[45] Estimates resulting from such analysis will only provide valid inference when missing data are missing completely at random.
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 combining the repeated assessments of a PRO domain per patient over a specific period into a single outcome (eg, area under the curve, overall means, and minimum and maximum). 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.[46] Moreover, information is lost when data are pooled and summarised into one value, which might then affect 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 identifying the most appropriate way to analyse these data.

(Figure 2 continues on next page)

	Recommended statement	Considerations
RS 15	To describe a response trajectory over time, a linear mixed model (omnibus test; time as discrete variable; time × group interaction) is recommended over the repeated measures ANOVA (time × group interaction).	The appendix (pp 16–26) provides more details about 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 descriptions of these profiles are reported cross-sectionally and not longitudinally. That is, every assessment point has a mean and CI. Therefore, interpretation is not on the (mean) longitudinal profile of the sample, but the mean outcome at each timepoint. 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 might be used to approximate the time curves. However, many of these models rely on specific assumptions and might yield results, estimates, or 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: standardising statistical terms related to missing data		
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,[40,47] non-completion of PRO assessments that were expected to be available,[21] or any missing values that would be meaningful for analysis if they were observed.[48,49] Adopting the definition of International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use 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 affect the interpretability of PRO findings 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 condition might not always be needed or feasible. Reasons to deviate from the ITT population and not to collect PROs at all from a specific subgroup should be strongly justified in the protocol. The PRO study population is a subgroup of the ITT population, which excludes those patients in which PRO outcomes could not be collected at all because of consent or eligibility, or both. Patients of the PRO study population should be identifiable at the beginning of the study irrespective of their follow-up status or observations. The PRO study population is therefore the intention-to-collect PRO population. The PRO analysis population refers to the patients who 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.[48] Depending on the analysis method, elimination of planned assessments from some patients might 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 might or might 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.[21,49] 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 19–20	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 timepoint as a proportion of the number of patients on PRO assessment at the designated timepoint. The term completion rate should be used to express the rate with the variable denominator rate.	The number of patients on PRO assessment identifies those patients who are still expected to provide PRO assessments at that timepoint. Conversely, patients who are off PRO assessments are defined as patients who are no longer expected to provide PRO assessments from that timepoint onwards. It was agreed to standardise that PRO assessments after death are considered off PRO assessment and will no longer be included in the denominator of the completion rates (ie, 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 standardise other reasons, such as off PRO protocol, patient withdrawal, and loss to follow-up in the number of patients on PRO assessment, needs further discussion (appendix pp 35–36).
RS 21–22	A fixed 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 timepoint as a proportion of the number of patients in the PRO study population (ie, all patients who consented and were eligible to participate in the PRO data collection). The term available data rate should be used to express the rate with the fixed denominator rate.	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 23	In addition to percentages, absolute numbers for both numerator and denominator should be reported at every timepoint (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 prespecified in the protocol, and reasons not prespecified in the protocol. Further work is needed to develop this idea.

(Figure 2 continues on next page)

	Recommended statement	Considerations
Section 4: general handling of missing data		
RS 24	When doing clinical trials, exploring the reasons for missing PROs is important.	Results from a simulation study showed that the effect of missing data rates on PRO findings depends on the reasons for missing data (eg, informative, non-informative, or a combination of both). Therefore, collecting reasons for missing data is key in assessing the effect of missing data rates on the robustness of PRO findings.
RS 25	Missing data should be minimised prospectively through clinical trial and PRO design strategies and by training or monitoring approaches.	No analysis method recovers the potential for robust treatment comparisons derived from complete assessments of all patients.[48] Therefore, preventing missing PRO assessments through careful design and planning should be the first-line strategy in handling missing PRO data.[49] More information is detailed in Mercieca-Bebber et al.[50]
RS 26	Capturing data that will be needed for handling missing PRO data in the statistical analysis plan is recommended (ie, reasons for missing data and auxiliary data for interpretation or imputation).	Missing data might still be unavoidable despite careful planning and collection strategies. With missing data, unverifiable assumptions would have to be made during the analysis.[51] Collecting reasons for missing data and auxiliary data would be helpful in justifying how these patients are handled in the primary and sensitivity analyses.[18,51]
RS 27	Primary statistical analysis approach: missing data approach at the item level and scale level should be specified a priori within the protocol or statistical analysis plan.	Similar to the choice of statistical analysis, different approaches to deal with missing data can lead to different results.[52] 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,[53] PRO measures are developed with a scoring algorithm to standardise how missing items should be handled. These scoring algorithms should be used in the primary analysis; and other ways to deal with missing items can be included as part of the 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 group and timepoint should be done.	Many possible reasons for missing data exist (eg, patient withdrawal, patient moving). Depending on the reason and amount of missing data, the approach to handle missing data might differ.[18,51]
RS 30	Primary statistical analysis approach: use all available data, using the specified method from the statistical methods working group.	Approaches that require ignoring missing data and only doing analysis with patients with complete data are not recommended (eg, complete case analysis).[51] Methods that allow the use of all available data is recommended as they make weaker assumptions about missing data compared with complete case analysis.[54]
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.[54] Imputing a fixed constant will result in lower variability and therefore a lower p value.[55]
RS 32	Sensitivity analysis should be specified a priori within the protocol or statistical analysis plan. At least two different approaches to handle missing data are recommended to assess the effect of missing data across various assumptions.	Handling missing data requires 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.[56] Results that are consistent with the primary analysis provide some assurance that the missing data did not have an important effect on the study conclusions. However, if sensitivity analyses produce inconsistent results, missing data implications on the conclusions of the trial must be discussed.[51] Disagreement arose because of the increase in the workload of trialists to prespecify, analyse, and report additional sensitivity analyses.

Figure 2: **SISAQOL recommended statements and their considerations**
ITT=intention to treat. PRO=patient-reported outcome. RS=recommended statement. SISAQOL=Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data.

Taxonomy of research objectives

All proposed statements from the research objectives working group were ratified (nine [100%] of nine). 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 the table.

When developing a PRO objective, the consortium concluded that the PRO domains and timeframe of interest should be prespecified.[8,34] Essentially, four key attributes need to be considered a priori for each PRO domain: broad PRO research objective comprising treatment efficacy and clinical benefit (confirmatory), or describe patient perspective (exploratory or descriptive); between-group PRO objective consisting of superiority, or equivalence or non-inferiority; within-treatment group PRO assumption for the treatment or control group, such as worsening, stable state, improvement, or overall effect; and within-patient or within-treatment PRO objective consisting of time to event, magnitude of event at time *t*, proportion of responders at time *t*, overall PRO score over time, or response patterns or profiles.

Considerations for each attribute are found in recommended statements 1–5 in figure 2. Recommended standardised definitions of improvement, stable state, worsening, and overall effects were ratified (recommended statements 6–9 in figure 2). Figure 3 illustrates the recommended definitions of improvement, stable state, and worsening.

Recommended statistical methods

Most of the proposed statements for this section were ratified (six [86%] of seven). 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

	Treatment efficacy or clinical benefit (confirmatory objective)		Describe patient perspective (exploratory or descriptive objective)
	Superiority (between-treatment groups objective)	Equivalence or non-inferiority (between-treatment groups objective)	
Improvement			
Time to improvement	Cox proportional hazards, with predefined effect size for the between treatment group difference	Cox proportional hazards, with predefined equivalence margin or predefined non-inferiority margin for the between-treatment group difference	Exploratory: Cox proportional hazards; descriptive: median time to improvement, probability of improvement at a specific timepoint, or HR with CI
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*, or OR or RR with CI
Magnitude of improvement at time *t*	Linear mixed model, and time as discrete, with predefined effect size for the between-treatment group difference	Linear mixed model, and time as discrete, with predefined equivalence margin or predefined non-inferiority margin for the between-treatment group difference	Exploratory: linear mixed model, and time as discrete; descriptive: mean magnitude at baseline and at time *t* with CI, or mean magnitude of improvement at time *t* with CI
Stable state			
Time to (end of) stable state	Cox proportional hazards, with predefined effect size for the between-treatment group difference	Cox proportional hazards, with predefined equivalence margin or predefined non-inferiority margin for the between-treatment group difference	Exploratory: Cox proportional hazards; and descriptive: median time to (end of) stable state, probability of (end of) stable state at a specific timepoint, or HR with CI
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*, or OR or RR with CI
Magnitude of stable state at time *t*	NA*	NA*	NA*
Worsening			
Time to worsening	Cox proportional hazards, with predefined effect size for the between-treatment group difference	Cox proportional hazards, with predefined equivalence margin or predefined non-inferiority margin for the between-treatment group difference	Exploratory: Cox proportional hazards; descriptive: median time to worsening, probability of worsening at a specific timepoint; or HR with CI
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*, or OR or RR with CI
Magnitude of worsening at time *t*	Linear mixed model, and time as discrete, with predefined effect size for the between-treatment group difference	Linear mixed model, and time as discrete, with predefined equivalence margin or predefined non-inferiority margin for the between-treatment group difference)	Exploratory: linear mixed model, and time as discrete; descriptive: mean magnitude at baseline and at time *t* with CI, or mean magnitude of worsening at time *t* with CI
Overall effects			
Overall PRO score over time	Further discussion needed	Further discussion needed	Further discussion needed
Response patterns or profiles	NA†	NA†	Exploratory: linear mixed model (time as discrete or continuous); descriptive: mean magnitude at baseline and at every timepoint within a timeframe with CI, mean change at every timepoint within a timeframe with CI, or mean profile over time with CI

Recommended statistical methods were initially conceptualised for a superiority between-treatment groups objective. However, these methods might be extrapolated to a non-inferiority or equivalence objective (but appropriate margins should be prespecified), and might be extrapolated to exploratory objectives (but such findings should not be used as a basis of evidence of clinical benefit or treatment efficacy [figure 2]). Descriptive statistics are based on the work from the statistical methods working group on evaluating appropriate statistical methods with research objectives (appendix pp 19–27). HR=hazard ratio. NA=not applicable. OR=odds ratio. PRO=patient-reported outcome. RR=risk ratio. *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 as noise—ie, random fluctuations not representing any meaningful changes. †As it is not always straightforward to predefine the exact profiles within a timeframe, response patterns or profiles are recommended to be used alongside a descriptive or exploratory objective rather than evidence for treatment efficacy or clinical benefit.

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

criterion, then the method was not recommended as appropriate for PRO analysis.

Two essential statistical properties were identified: the ability to do 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 about these criteria can be found in recommended statement 10 in figure 2. When two or more statistical methods fit the essential and highly desirable criteria equally, the simpler method was prioritised. Although there might be advantages in

recommending more complex models for specific purposes (eg, pattern mixture models), this advantage often comes at the cost of strong and untestable assumptions, and can produce results that might not be easily interpreted by non-statisticians. A balance between feasibility, usefulness, interpretability, and statistical correctness was determined to be essential for the primary PRO analysis; however, more complex models can be used for sensitivity analyses to test the robustness of the primary result.

On the basis of the agreed set of statistical criteria and selection criteria, statistical methods were recommended for each PRO objective. Two statistical methods were recommended: Cox proportional hazards for time-to-event PRO objectives, and linear mixed models for magnitude of event at time *t* and response patterns or profiles (recommended statements 11, 12, and 15 in figure 2). In exceptional cases in which the PRO design only required baseline and one follow-up assessment, linear regression was recommended as the appropriate statistical method (recommended statement 13 in figure 2).

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.[35] To address this limitation, it was recommended that non-parametric methods be used for sensitivity analyses to investigate deviations from these assumptions.[35]

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 (eg, minimum and maximum, area under the curve, overall means), but it was recognised that summary measures should be part of SISAQOL's future work (recommended statement 14 in figure 2). Further investigation is needed for whether it is appropriate to analyse ordinal data as continuous; discussions on this issue revolved around statistical approximation, complexity of the model, and ease of interpretation.

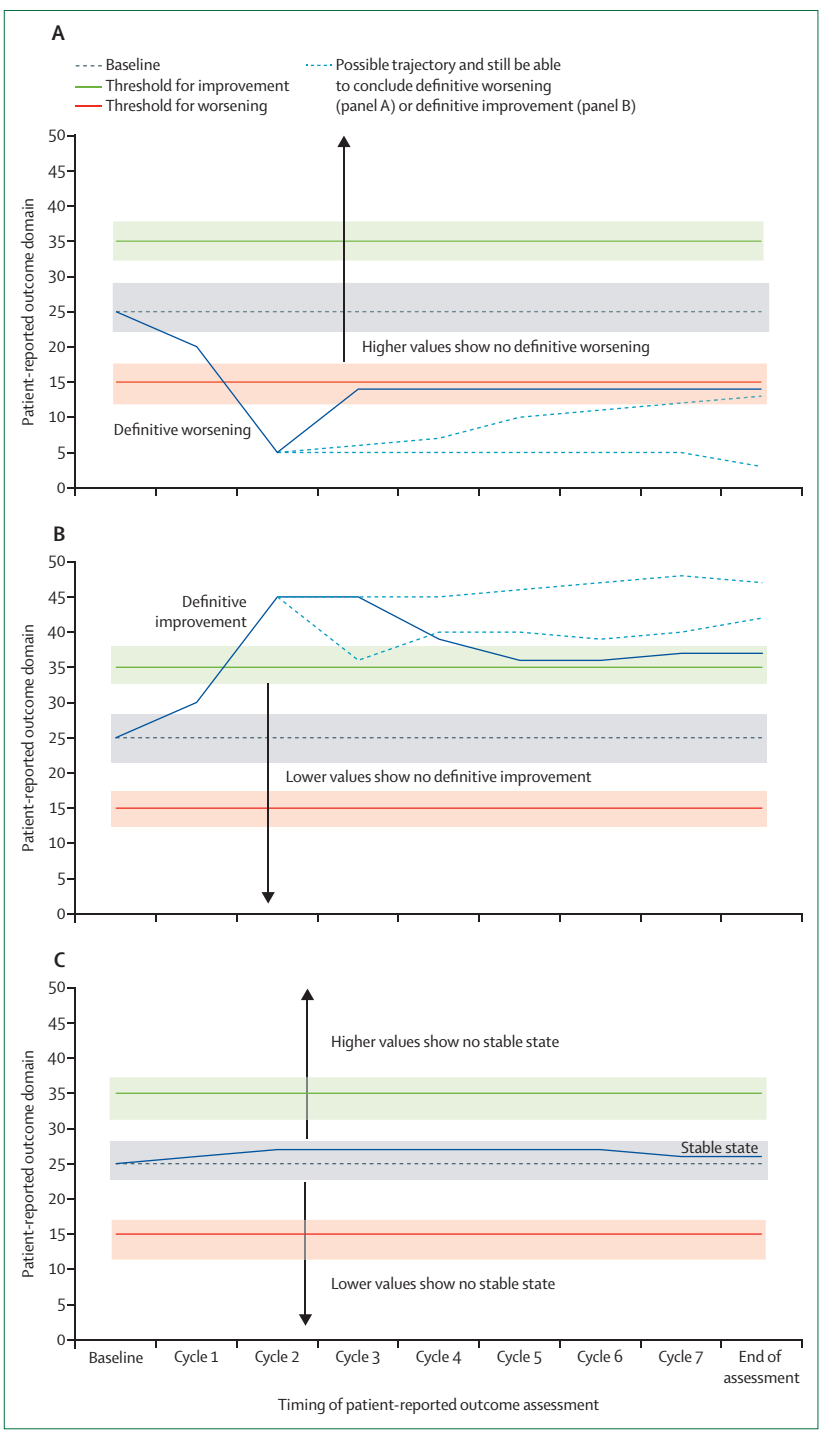

Figure 3: **Sample illustrations of the recommended definitions of improvement, stable state, and worsening**
The shaded areas are predefined margins of the threshold levels. (A) Worsening is defined as change from baseline that reaches a predefined worsening threshold level (post-baseline worsening). 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). (B) Improvement is defined as change from baseline that reaches a predefined 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). (C) Stable state is defined as no observed changes from baseline, or the change from baseline is within the predefined baseline margin. This stable state is maintained if follow-up assessments remain at the baseline predefined margin. The stable state is discontinued once the follow-up assessment leaves the predefined baseline margin (and reaches the improvement or worsening threshold). There might be circumstances in which the relevant patient-reported outcome objective would include improvement in the definition of stable state (ie, 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.

Standardising terminology related to missing data

Eight (73%) of 11 proposed statements for this section were ratified. A recommendation on the definition of missing PRO data was proposed: missing PRO data are defined as data that would be meaningful for the analysis of a given research objective, but were not collected (recommended statements 16 and 17 in figure 2).[48,49] This definition implies that not all unobserved assessments are considered as missing data depending on the scientific question (eg, 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 might implicitly be treated similarly as missing data.[57] Recommendations on how to specifically deal with missing data for each recommended method is the next step for the SISAQOL Consortium's work.

This Policy Review stresses the importance of differentiating missing observations in relation to a reference set of expected data (recommended statements 19–22 in figure 2). The discussion resulted in two definitions: the so-called available data rate has a fixed denominator, defined as the number of patients in the PRO study population (ie, all patients who consented and were eligible to participate in the PRO data collection at baseline); and the completion rate has a variable denominator, defined as the number of patients on PRO assessments at the designated timepoint (ie, all patients who are still expected to provide PRO assessments at that timepoint). The numerator of both rates are the number of patients submitting a valid PRO assessment at the designated timepoint. Of note, the denominator of the completion rate depends on the chosen research question—eg, 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 this rate always refers to a fixed set of patients at baseline (recommended statement 18 in figure 2).

Missing data

More than half of the proposed statements were ratified in this section (nine [56%] of 16). A simulation study was done 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 the effect of missing data rates on PRO findings depends on the type of missing data (ie, informative or non-informative missing data). It was recommended that collecting reasons for missing data is key in assessing the effect of missing data for PRO findings (recommended statement 24 in figure 2).[20] A case-report form to collect reasons for missing data in a standardised way is needed and will be further developed by the consortium. General recommendations on how to handle missing data were proposed consistent with existing regulatory guidelines (recommended statements 25–30 in figure 2).

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 methods. 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 or profiles. Recommendations on a standardised definition of missing PRO data, completion rates, and available data rates were proposed, with corresponding standardised 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 standardised 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 conceptualisation 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 PROs in cancer trials, including SPIRIT-PRO,[8] CONSORT-PRO,[9] and other relevant guidelines.[56,58] 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 important that clinical and health-care 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; although many issues

Search strategy and selection criteria

We identified references for this Policy Review through searches of PubMed using 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 Policy Review.

might generalise to other health conditions, this generalisation warrants further scrutiny. Another limitation relates to the relevance of these standards to preference-weighted measures of health-related quality of life (HRQOL), also called preference-based measures or multi-attribute utility measures. Such measures can be used for two purposes: as utility scores that represent a special type of HRQOL summary score (ie, with domains of HRQOL weighted by preferences, usually the general population's preferences, but sometimes patient preferences) and as quality weightings in quality-adjusted life-years and cost–utility analysis. Whether the standards reported in this paper apply for any of these purposes needs 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 (eg, statistical method for proportions of patients at time *t*, summary measures, and several issues on missing data). 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 might 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 they can be applied in most 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[24] is yet to be examined.

Conclusion

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.

Contributors

This manuscript was conceptualised with the attendees of the SISAQOL kick-off meeting in Brussels, Belgium, on Jan 29, 2016. All authors contributed to the work of the individual working groups. All authors discussed and finalised this work during the SISAQOL consensus meeting on Sept 24, 2018. All authors reviewed and contributed to the revisions of the article. All authors approved the final draft of the manuscript.

Declaration of interests

MC reports personal fees from Astellas, Takeda, Glaukos, Merck, Daiichi Sankyo, and the Patient-Centered Outcomes Research Institute; and grants from Health Data Research UK, Innovate UK, National Institute for Health Research (NIHR), Macmillan, and UCB Pharma, outside the submitted work. AC reports having been an employee of Genentech until Jan 31, 2019. KC reports grants from European Organization for Research and Treatment of Cancer (EORTC); and personal fees from Bristol-Myers Squibb, Endomag, Celgene, and Amgen, outside the submitted work. IG is an employee of Boehringer Ingelheim International, which provided an unrestricted education grant to EORTC. MKo reports grants from EORTC, Biofrontera, and Komitee Forschung Naturmedizin; and personal fees from Janssen-Cilag, Lily, and Verband Forschender Arzneimittelhersteller, 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. EP-L is an employee of Genentech. 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, Merck Sharp & Dohme, 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. GV reports personal fees from Roche, Eisai, Novartis, and Pfizer; and grants from the NIHR UK Government, Breast Cancer NOW, and EORTC, outside the submitted work. AB reports grants to EORTC from Boehringer Ingelheim International, 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.

Acknowledgments

The European Organisation for Research and Treatment of Cancer received an unrestricted education grant from Boehringer Ingelheim to initiate this work and from Genentech, a member of the Roche Group, for continuity funding. We thank the International Society for Quality of Life Research and the International Society for Pharmacoeconomics and Outcomes Research 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, Health Canada, the UK National Health Service, the National Institute for Health Research, or the UK Department of Health. This study received no funding from

the US National Institutes of Health (NIH). AWS and SAM are employed by NIH. No other authors were fully or partly NIH funded, employed by NIH, or are in receipt of an NIH grant.

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 (2007–13). *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:** 375–79.

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:** e510–14.

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, 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 Rep Outcomes* 2017; **1:** 5.

17 Kyte D, Retzer A, Ahmed K, et al. Systematic evaluation of patient-reported outcome protocol content and reporting in cancer trials. *J Natl Cancer Inst* 2019. **111:** 1170–78.

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. Chichester: 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.

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 ICH Expert Working Group. ICH Harmonised Tripartite: guideline statistical principles for clinical trials E9. 1998. https://database.ich.org/sites/default/files/E9_Guideline.pdf (accessed Jan 7, 2020).

25 Tukey JW. We need both exploratory and confirmatory. *Am Stat* 1980; **34:** 23–25.

26 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.

27 European Medicines Agency. Guideline on the choice of non-inferiority margin. 2008. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-choice-non-inferiority-margin_en.pdf (accessed Jan 7, 2020).

28 European Medicines Agency. Points to consider on switching between superiority and non-inferiority. 2000. https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-switching-between-superiority-non-inferiority_en.pdf (accessed Jan 7, 2020).

29 King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 2011; **11:** 171–84.

30 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.

31 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.

32 National Cancer Institute. NCI dictionary of cancer terms. 2013. https://www.cancer.gov/publications/dictionaries/cancer-terms (accessed Jan 7, 2020).

33 Wasserstein RL, Nicole AL. The ASA statement on p-values: context, process and purpose. *Am Stat* 2016; **70:** 129–33.

34 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.

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

36 European Medicines Agency. Guideline on adjustment for baseline covariates in clinical trials. 2003. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf (accessed Jan 7, 2020).

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

38 Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. *BMJ* 2002; **323:** 1123–24.

39 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.

40 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.

41 Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis, 2nd edn. Hoboken, NJ: John Wiley & Sons, 2011.

42 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:** 431–36.

43 Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. *Br J Cancer* 2003; **89:** 232–38.

44 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.

45 SAS. The REG procedure: missing values. https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_reg_sect026.htm (accessed Jan 7, 2020).

46 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:** 2.

47 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.

48 Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med* 2012; **367:** 1355–60.

49 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. https://database.ich.org/sites/default/files/E9-R1_EWG_Draft_Guideline.pdf (accessed Jan 7, 2020).

50 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:** e010938.

51 European Medicines Agency. Guideline on missing data in confirmatory clinical trials. 2011. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-missing-data-confirmatory-clinical-trials_en.pdf (accessed Jan 7, 2020).

52 Fielding S, Fayers P, Ramsay CR. Analysing randomised controlled trials with missing data: choice of approach affects conclusions. *Contemp Clin Trials* 2012; **33:** 461–69.

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

54 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.

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

56 US Food and Drug Administration. Patient-reported outcome measures: use in medical product development to support labeling claims. 2009. https://www.fda.gov/media/77832/download (accessed Jan 7, 2020).

57 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.

58 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. https://www.ema.europa.eu/documents/other/appendix-2-guideline-evaluation-anticancer-medicinal-products-man_en.pdf (accessed Jan 7, 2020).