# Appendix 5

## Development of statements in the scientific work packages

The work in the four scientific work packages (WPs) was, in general, performed according to the predefined steps described in the Guidebook, chapter 7 (Figure 4). Additional actions were performed within each WP as needed.

## Review and discussion of current views and practices

To achieve the aims of the different WPs, an important first step was to get an overview of the current practise. All WPs had "kick off" meetings to present the tasks, activities and work plans. Members of each WP participated with their expertise, provided their organisations' views and practices, and if available, literature (including guidance documents) considered relevant for use of PRO data in cancer clinical trials. It was important to identify knowledge gaps and develop strategies on how to address and provide recommendations to fill these gaps. The WPs agreed upon themes to be explored in the following literature reviews related to randomised controlled trials (RCTs) and single arm trials (SATs), definitions of PRO score interpretation thresholds, presentations of PRO data and templates for guiding graphical presentations of PRO data in the best possible way. Both patient representatives and the other stakeholder groups were involved in the initial discussions.

## Reviews of literature and existing guidelines

All WPs conducted literature reviews to provide an overview of the current practise and literature. The information sources and methods used to select and extract relevant data were described in detail in research protocols. The results of the literature reviews fed into the formulation of the initial set of recommendation statements for RCTs, SATs, PRO score interpretation thresholds and for presentations of PRO data.

Existing stakeholder guidelines and recommendations were reviewed (e.g., US Food and Drug Administration [FDA], European Medicines Agency [EMA], European Network for Health Technology Assessment [EUnetHTA]), in addition to key published methodological/statistical recommendations

for clinical trials (e.g., SPIRIT-PRO, CONSORT-PRO, SISAQOL, ICH guidelines including the estimand framework).

**The literature reviews on RCTs** focused on current practices of PRO analysis, existing stakeholder guidelines and key methodological recommendations for PRO analysis in RCTs (Coens, Pe et al., 2020). The literature review on RCTs included 41 protocols, 31 European public assessment reports (EPARs) , 21 methodological guidelines, six systematic reviews and 104 stakeholder guidelines. In addition, the review included public information by regulators such as assessment reports of PRO data submitted for cancer medicines and qualification guidance that was conceptually relevant. Relevant literature sources were collated and summarised, and areas of agreement and disagreement were identified. The overview of the current practices confirmed the established SISAQOL-IMI framework based on the SISAQOL taxonomy (see Chapter 6 of the Guidebook, Read more about the SISAQOL taxonomy). Time-to-event and magnitude-of-change endpoints were most frequently used while responder-type analysis was infrequently used. Best/worst/ last PRO score was not routinely used in practice, but these endpoints were listed in a minority of the guidelines. The heterogeneity in analysis populations used in practice or proposed in the guidelines were highlighted. Results pertaining to the handling of intercurrent events were mixed (Machingura et al., 2025).

**The literature reviews on SATs focused** on current practices and methodological recommendations on design, analysis, reporting and interpretation of SATs. Sixty papers published between 2018 and 2021 of SATs of cancer treatment with PRO data were examined. The review also examined the studies' handling of potential bias and how they informed decision making. Most studies analysed PROs without stating a predefined research hypothesis and only 22% (13/60) used PRO as a primary or co-primary endpoint. Definitions of PRO objectives, study population, endpoints, and missing data strategies varied widely. Some studies (23/60) compared the PRO data with external information. One study included a historical control group. Appropriateness of methods to handle missing data and intercurrent events (including death) were seldom discussed. Most studies concluded that PRO results supported treatment (Liu et al., 2023).

**The literature reviews on PRO score interpretation thresholds** was performed in two phases. First, a systematic scoping review focused on publications on clinically meaningful change thresholds between 2009 and 2021. Especially, on studies establishing **PRO score interpretation thresholds** for the most commonly used PRO measures in oncology (Smith et al., 2014, Gnanasakthy et al., 2019, Giesinger et al., 2021), and methodological articles discussing application of these thresholds. The review included non-cancer literature identified via hand-search of reference lists of included publications from the literature search, and publications on change thresholds from relevant professional organizations (e.g., ISPOR, ISOQOL) and regulatory authorities (e.g., FDA, EMA). Of the 5,394 unique abstracts and 439 full texts that were screened, 134 articles were eligible for inclusion. There were 93 original articles reporting the results from analyses establishing **PRO score interpretation** thresholds for a specific PRO measure, and 41 reviews, editorials, or other articles discussing the methodological aspects of thresholds. The most common approaches to establishing PRO score interpretation thresholds were anchor-based (82 articles, 88%) and distribution-based (64 articles, 69%). The anchor was primarily patient-reported (62 articles,

76%) and determined the meaningful change threshold via descriptive mean change (58 articles, 71%). In distribution-based analyses, the most common statistics used as meaningful change threshold were a measure of effect size (58 articles, 92%) or the standard error of measurement (48 articles, 76%). The review identified and extracted terminology and definitions used for PRO score interpretation thresholds. In total, there were 57 different references for definitions of threshold -related terms, of which 13 were cited more than once. Another literature review of past and current practice of using PRO score interpretation thresholds in cancer trials was performed, capitalizing on the relevant publications of RCTs and SATs identified by the dedicated WPs.

**The literature review on visualisation** was performed in two phases and focused on the evidence on the graphical representation of PRO data and more general information for the design of PRO visualisations. In the first phase, trials investigating ways to present and visualize PRO data in oncology were reviewed. The second phase identified more broadly existing methodological advice on how to present and depict PRO and/or medical data. References of extracted studies were screened for additional literature. In the two searches, 1,223 records were identified and screened. The final full-paper review included six trials on visualisation methods/ options in the field of oncology and 32 publications with different methodological approaches providing more general information for the design of PRO visualisations. Only sources with group-level data and reviews including group-level data were included in the qualitative synthesis. The descriptive results were grouped per stakeholder type, i.e. patients, health care professionals and PRO researchers or experts.

Within each stakeholder group, findings were thematically categorised:

- graph type: information on which format to use;
- data type: information on which type of data to present;
- additional information: findings on how much and which additional information to include in the graph;
- textual explanations: e.g. explanatory figure descriptions, legends, labelling statistics: information on which statistical details to include;
- interpretation aids: options facilitating the understanding of a visualization;
- complexity: amount of depicted information and details;
- directionality: findings on which direction of scores to choose, e.g. higher scores always mean better scores;
- scaling: information on scale values, interval widths or ranges;
- colour: data on which colours to use in a graphical presentation and what for;
- other suggestions: e.g., on how to design a certain graph, on the use of symbols or on general principles to consider. To complement the systematic approach, a hand search screened websites of renowned organisations (e.g., ASCO, ESMO, ISOQOL, ISPOR) for additional abstracts, information, or educational material on graphical presentation of PRO data. For feasibility reasons, the search of conference materials was limited to events held in the past three years (until 2021).

## Work package-specific surveys

To provide input for the development of the statements in the WPs, smaller WP-specific surveys were performed involving the members of each WP or other defined groups.

**RCT Statements:** prior to each consensus meeting, a survey on selected topics would be circulated among the WP2 members. These surveys focused on the topics that were not considered sufficiently covered and/or resolved by the review or on topics raised by the previous consensus meeting. Results from these surveys were first discussed within among the WP2 steering committee and then further opened for discussion with all WP2 members. During these surveys and the resulting meeting discussions, members were invited to submit methodological techniques and/or references deemed relevant for the in-depth evaluation. The resulting discussion outcomes were minuted and made available to all WP2 members for comments (to accommodate those members that could not attend the discussions). The resulting minutes formed then the basis to develop specific recommendation statements for the next consensus meeting.

**SAT statements:** a survey was sent out to WP members focusing on best practices for the analysis of PRO in SATs in oncology. Contributors were asked to indicate the appropriate study population and how they would handle terminal events (death) and other intercurrent event in the ICH E9R1 estimand framework. They were also asked to indicate what analysis methods were considered appropriate and how the absence of a randomised control group should be addressed, Furthermore, the use of minimal clinically important differences for PROs in SATs and visualisation of PRO results of a SAT were discussed.

**PRO score interpretation threshold statements:** a group of members involved in the work on terminology and definition of clinically meaningful change/PRO interpretation thresholds participated in an initial survey of preliminary statements that followed the SISAQOL-IMI matrix distinguishing analyses of time-to-event, magnitude of change, and responders. The results from this survey provided an overview of the various stakeholder perspectives and later supported the development of a list of proposed statements. Another survey was conducted to collect feedback on the use of clinically meaningful change in SATs. The survey showed a preference in both safety and benefit analyses, for applying clinically meaningful change thresholds to individual-level data (within-patient change) rather than group-level (within-group change). In the second year of the project, a third survey evaluated further statements on the use of clinically meaningful change thresholds and asked for feedback on their relevance and applicability for patient- and group-level data. This survey also explored what information should be provided to guarantee transparency when using clinically meaningful change thresholds and collected feedback on terminology and definitions identified in the literature. The online survey was a major source of information for developing additional statements including a checklist with information that should be reported when using PRO score interpretation thresholds.

**Visualization statements:** based on the previous work of SISAQOL, a survey focusing on visualization of PRO data was conducted. The questions referred to the estimand framework and asked about preferences for visualization of each type of data (binary/nominal data, e.g., frequency of responders; metric data, e.g., mean change over time). For the specific endpoints, a main

visualisation type was indicated (e.g., the Kaplan-Meier curve can be used for a superiority time-to-event endpoint), along with whether participants wanted to include additional information and, if so, what information. In addition, the survey presented the search terms collated for the literature reviews and asked participants for terms they felt were missing.

## Interviews on communication and presentation of PRO results

Following the literature reviews and surveys, face-to-face and online interviews on communication and presentation of PRO results were performed. Using framework analysis, the qualitative content was assigned to pre-determined codes reflecting the structure of the interview guide. In a multi-stage approach, the content was abstracted from verbatim quotes to aspects that are more general. In total, 58 participants (18 patients, 17 health care professionals and 23 researchers and statisticians) shared their experience with PRO information from clinical trials and gave their preference and suggestions for improvement on seven examples of graphical displays that matched the SISAQOL framework. Furthermore, face-to-face interviews on PRO score interpretation threshold methodology were conducted to evaluate the implementation of such thresholds in the graphical presentation of PRO results for analyses of responder percentages, of time-to-event, and magnitude of change/difference. The results of the interviews on presentation of PRO findings were discussed at the Interim General Assembly 2. It was pointed out that there was a considerable difference regarding the level of knowledge between and within the stakeholder groups and that the work should focus on the level of complexity of scientific reporting with additional information targeting other audiences. Some suggestions for improvement of shown graphical examples were highlighted (such as alignment of order of colours within the graph, colour coding, indicate directionality (whether e.g., rising scores indicate an improvement or deterioration),remove unnecessary information from the caption and explain statistical terms and concepts).

## Further work within the WPs

Over the years, the WPs held regular meetings within their groups to review the relevant literature, to set up proposals for new statements and other documents and to prepare for, perform and react upon the various surveys. There was an active collaboration and formalised meetings between WP2 and WP3 to discuss issues relevant to both of them, e.g. how to handle intercurrent events versus missing data in the context of PRO data for RCTs and SATs. When statements from one WP were applicable for the other, preliminary statements with an underlying justification were drafted. After feedback and endorsement by the members in the other WP, statements were submitted to all consortium members for feedback. Separate meetings were held between the WP2 lead team and Online Focus Groups representing the patient view on the request of the patient representatives to better understand some of the more advanced issues in the surveys. Specific survey questions were discussed, and more background and examples were given by the WP2 lead team to make sure the questions were understood and intended. The patient representatives would then provide their answers and any additional comments that may have arisen from the discussion. These views from the patient representatives were added to the overall survey results received.

For the work on SATs, the initial findings on the target population, handling of terminal and intercurrent events, and handling the absence of a randomized control group were further discussed. External advice was provided from the STRATOS initiative (**https://stratos-initiative.org/**), and from the Independent Scientific Advisory Board. Following this, updated and extended statements were included in the following survey.

WP4 held several meetings with other WPs for harmonisation of the communication tool statements with the accompanying visualisation of results. Revised statements and templates were sent out to all WP members in preparation for an online meeting to discuss the statements. Subsequently, the statements adopted by the entire WP were again harmonised with the other scientific WPs and again jointly revised by WP4. A few general advice statements were not included in the voting process but were discussed directly at Consensus Meeting #4 (on topics such as the use of colours, the use of highlighting, consistency within and between graphics, the structure of the figure caption, clutter and readability).

The results from the literature review and initial discussions demonstrated that a wide variety of terms were used to describe PRO score interpretation concepts, with little consistency across. Emphasis was put on the importance of clearly distinguishing the various types of thresholds (e.g., individual versus group-level, difference versus change over time, anchor-based versus distribution-based) and suggested that rather than introducing an individual term for each type it may be more practical to have a limited number of umbrella terms accompanied by a checklist that details what needs to be defined when using a certain threshold. An in-person workshop was held to allow for further discussion of possible recommendation statements related to PRO score interpretation thresholds. The workshop focused on the applicability of thresholds at the individual- or group-level, on how to distinguish to PRO score interpretation thresholds applicable to either of the two settings. This workshop together with a third WP member survey were used to develop additional statements on PRO score interpretation thresholds. There was a refining of the statements before the re-voting of the survey statements. The survey feedback on terminology was discussed, as well as suggestions for harmonised terminology developed by the WP6 core team. For SATs, the WP core team agreed on the necessary distinctions regarding the terminology and definition for PRO score interpretation thresholds for patient-level (e.g., for use in responder analysis and analysis of time-to-improvement/deterioration) and for group-level thresholds (e.g., to interpret magnitude of differences/changes at group-level) in SATs.

## Initial recommendations with pilot studies

A case study on SATs involving 876 participants, explored the implications of the ICH E9 (R1) estimand framework for the imputation, analysis and interpretation of quality of life in a single arm trial setting.

A range of possible estimands suitable for different research aims were presented for outcome of longitudinal quality of life. Different variables of interest and strategies to deal with intercurrent events (death, treatment discontinuation and disease progression) were explored. Statistical methods were described for handling of missing data. The results supported that decisions made

in the estimand framework are important to ensure that the analysis methods correspond to the intended estimand. Adherence to this framework can improve the quality of single arm trials (Thomassen et al., 2024). Another case study explored methods to deal with missing PRO data in SATs. This study demonstrated that the different imputation methods assumptions about the relation between PRO and the intercurrent events impacted the estimates (D Thomassen, BMC Medical Research Methodology [accepted for publication 2025]).

For RCTs, a similar case study exercise was performed in collaboration with WP5. The resulting feedback was then used to modify or expand the explanation text for those statements flagged as difficult to implement by this validation process.

## References


Coens C, Pe M, Dueck AC, Sloan J, Basch E, Calvert M, SISAQOL Consortium. International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomized controlled trials: Recommendations of the SISAQOL Consortium. *Lancet Oncol*. 2020;21(2):e83–e96. **https://doi.org/10.1016/S1470-2045(19)30790-9**

Giesinger JM, Efficace F, Aaronson N, Calvert M, Kyte D, Cottone F, Cella D, Gamper EM. Past and current practice of patient-reported outcome measurement in randomized cancer clinical trials: a systematic review. *Value Health*. 2021 Apr;24(4):585-591. **https://doi.org/10.1016/j.jval.2020.11.004**. Epub 2021 Feb 10. PMID: 33840437.

Gnanasakthy A, Barrett A, Evans E, D'Alessio D, Romano CD. A Review of Patient-Reported Outcomes Labeling for Oncology Drugs Approved by the FDA and the EMA (2012-2016). *Value Health*. 2019 Feb;22(2):203-209. **https://doi.org/10.1016/j.jval.2018.09.2842**. PMID: 30711065.

Liu L, Choi J, Musoro JZ, Sauerbrei W, Amdal CD, Alanya A, Barbachano Y, Cappelleri JC, Falk RS, Fiero MH, Regnault A, Reijneveld JC, Sandin R, Thomassen D, Roychoudhury S, Goetghebeur E, le Cessie S, Aiyegbusi OL, Van Lancker K. Single-arm studies involving patient-reported outcome data in oncology: A literature review on current practice. *The Lancet Oncology*.2023;24(5):e197–e206. **https://doi.org/10.1016/S1470-2045(23)00110-9**

Machingura A, Coens C, Pe M, Alanya A, Regnault A, Mazerolla F, et al. Methodological work and use of patient-reported outcomes data in randomized controlled trials (RCTs) in cancer: literature reviews on current practices and guidelines. *Zenodo*. 2025. **https://doi.org/10.5281/zenodo.14749434**.

Smith AB, Cocks K, Parry D, Taylor M. Reporting of health-related quality of life (HRQOL) data in oncology trials: a comparison of the European Organization for Research and Treatment of Cancer Quality of Life (EORTC QLQ-C30) and the Functional Assessment of Cancer Therapy-General (FACT-G). *Qual Life Res*. 2014 Apr;23(3):971-976. **https://doi.org/10.1007/s11136-013-0534-2**. Epub 2013 Oct 5. PMID: 24097080.

Thomassen D, Roychoudhury S, Amdal CD, Reynders D, Musoro JZ, Sauerbrei W, Goetghebeur E, le Cessie S; SISAQOL-IMI Work Package 3. Handling missing values in patient-reported outcome data in the presence of intercurrent events. *BMC Med Res Methodol*. 2025 Mar 1;25(1):56. **https://doi.org/10.1186/s12874-025-02510-8**. PMID: 40025441.