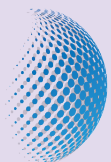




**Guidebook for designing, analysis,
interpreting and presenting patient-
reported outcomes in cancer
clinical trials: the SISAQOL-IMI
recommendations**



SISAQOL | IMI

Cover photograph © nd3000/istock

© All rights reserved.

Reproduction or use of this content requires citation or permission from the copyright holder.

The suggested citation for the SISAQOL-IMI Guidebook is:

Guidebook for designing, analysing, interpreting, and presenting patient-reported outcomes in cancer clinical trials: the SISAQOL-IMI recommendations. Amdal et al. on behalf of the SISAQOL-IMI Consortium. 2025. Available at <https://www.sisaqol-imi.org/>

We welcome feedback from our readers; please send your comments and suggestions using our [contact form](#). You will not receive a response, but the document will be updated regularly based on the feedback received.





Writing committee

CD Amdal^{1,2}, RS Falk^{1,3}, K Bjordal^{1,3}, A Alanya⁴, KLH Joseph¹, L Wintner⁵, A Regnault⁶, A Ingelgård⁷, JKrisam⁷, VPawar⁸, Sten Seldam⁹, M Calvert¹⁰, C Coens⁴, M Schlichting¹¹, S Le Cessie¹², S Roychoudhury¹³, B Holzner⁵, J Chang¹³, M Taphoorn¹⁴, P Cisko¹³, J Giesinger⁵, J Cappelleri¹³, E Papadopoulos¹⁵, M Pe⁴.

Contributing co-authors in alphabetical order

Jl Arraras¹⁶, GL Astrup², E Basch¹⁷, A Belančić¹⁸, M Brundage¹⁹, A Campbell²⁰, N Cherny²¹, K Cocks²², S Eremenco²³, M Ferrer²⁴, M Fiero²⁵, C Gerlinger²⁶, E Goetghebeur²⁷, U Grouven²⁸, A Lauer⁷, A Machingura⁴, J Mizusawa²⁹, G Molenberghs³⁰, LA Olalekan¹⁰, MA Petersen³¹, C Quinten³², KR Rantell³³, B Reeve³⁴, J Reijneveld³⁵, J Ringash³⁶, G Rumpold³⁷, C Rutherford³⁸, K Sail¹⁵, M Sasseville³⁹, W Sauerbrei⁴⁰, A Schiel⁴¹, AW Smith⁴², C Snyder⁴³, G Velikova⁴⁴, XS Wang⁴⁵.

List of affiliations

¹ Research Support Services, Oslo University Hospital, Norway

² Department of Oncology, Oslo University Hospital, Norway

³ University of Oslo, Norway

⁴ European Organisation for Research and Treatment of Cancer (EORTC) Headquarters, Belgium

⁵ University Hospital of Psychiatry II, Medical University of Innsbruck, Austria

⁶ Modus Outcomes, France

⁷ Boehringer Ingelheim, Germany

⁸ EMD Serono, USA

⁹ Myeloma patients Europe, Belgium

¹⁰ University of Birmingham, UK

¹¹ Merck, Germany

- ¹² Department of Neurology, Leiden University Medical Center, the Netherlands
- ¹³ Pfizer, USA
- ¹⁴ Department of Biomedical Data Sciences, Leiden University Medical Center, the Netherlands
- ¹⁵ AbbVie, USA
- ¹⁶ Hospital Universitario de Navarra, Spain
- ¹⁷ American Society for Clinical Oncology, USA
- ¹⁸ Clinical Hospital Center Rijeka, Croatia
- ¹⁹ Queen's University at Kingston, Canada
- ²⁰ Patient Relevant Evidence, USA
- ²¹ European Society for Medical Oncology, Switzerland
- ²² Adelphi Values, UK
- ²³ Critical Path Institute, USA
- ²⁴ Institut Hospital del Mar d'Investigacions Mèdiques, Spain
- ²⁵ US Food and Drug Administration, USA
- ²⁶ Bayer, Germany
- ²⁷ University of Ghent, Belgium
- ²⁸ Institute for Quality and Efficiency in Health Care, Germany
- ²⁹ The Japan Clinical Oncology Group, National Cancer Center Hospital, Japan
- ³⁰ Katholieke Universiteit Leuven, Belgium
- ³¹ Region Hovedstaden, Denmark
- ³² European Medicines Agency, the Netherlands
- ³³ Medicines and Healthcare products Regulations Agency, UK
- ³⁴ Duke University School of Medicine, USA
- ³⁵ Amsterdam University Medical Center, the Netherlands
- ³⁶ The University Health Network – Princess Margaret Cancer Centre/University of Toronto, Canada
- ³⁷ Evaluation Software Development, Austria
- ³⁸ University of Sydney, Australia
- ³⁹ Health Canada, Canada
- ⁴⁰ University of Freiburg, Germany
- ⁴¹ Norwegian Medical Products Agency, Norway
- ⁴² National Cancer Institute, USA
- ⁴³ John Hopkins University, USA
- ⁴⁴ University of Leeds, UK
- ⁴⁵ MD Anderson Cancer Center, University of Texas, USA

Other actively involved participants in the Consortium by institution in alphabetical order

AbbVie (C Bui, N Emechebe, C Ferguson, K Fitzgerald, R Kamalakar, R Sen, SC Turner, J Yu)

Bayer (Y Su, B Wolf)

Boehringer Ingelheim (M Ge, H Zettl, M Voorhaar, B Wong, I Griebisch [former])

Clinical Hospital Center Rijeka (GD Arbanas, K Kuljanic, D Petranovic)

Critical Path Institute (C Coon)

European Medicines Agency (F Pignatti)

European Organisation for Research and Treatment of Cancer (J Musoro, F Martinelli, A Bottomley [former])

European Society for Medical Oncology (J Barriuso, F Chiavaro, B Kiesewetter, M Galotti, B Gyawali, N Latino, S Oosting, F Roitberg)

Health Canada (J Black, M Leach)

Institut Hospital del Mar d'Investigacions Mèdiques (O Garin, G Vilagut)

Institute for Quality and Efficiency in Health Care (S Thomas, B Wieseler)

Katholieke Universiteit Leuven (K Bogaerts)

Leiden University Medical Center (D Thomassen)

Medical University of Innsbruck (I Al-naesan, F Gross, MJ Pilz, A Thurner)

Modus Outcomes (G Desplanques, F Mazerolle)

Myeloma Patients Europe (S Alexis, V Claus, E Duncan, K Joyner, M Maguri, P Matamoros, K Morgan, D Ness, H Scheurer, A Plate, A Vallejo, J Vesseur)

National Cancer Center Hospital (M Terada)

National Cancer Institute (S Mitchell)

Pfizer (S Böhme, A Russell-Smith)

Queen's University at Kingston (D Tu)

Region Hovedstaden (M Groenvold)

The Symptoms Tool Executive Committee, University of Texas MD Anderson Cancer Center (C Cleeland, L Williams)

University of Birmingham (SC Rivera)

University of Ghent (L Liu, D Reynders)

University of Leeds (F Boele, A Gilbert, R Peacock)

University of Sydney (M King)

US Food and Drug Administration (V Bhatnagar, TY Chen, E Horodniceanu, LL Johnson, P Kluetz, L Rodriguez)

Workgroup of European Cancer Patient Advocacy Networks (J Geissler: CML advocates network, B Ryll and G Spurrier-Bernard: Melanoma Patient Network Europe, S Leto di Priolo, J Taylor)

Acknowledgments



The SISAQOL-IMI project received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking under grant agreement No. 945052. This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Programme and the European Federation of Pharmaceutical Industries and Associations (EFPIA).

This publication reflects the views of the individual authors. It should not be construed to represent official views or policies of the European Medicines Agency (EMA), the US Food and Drug Administration (FDA), US National Cancer Institute (NCI), Medicines and Healthcare products Regulatory Agency (MHRA), Institute for Quality and Efficiency in Health Care, Health Canada, the Norwegian Medical Products Agency, the American Society of Clinical Oncology, the European Society for Medical Oncology or any other institution, organization, or entity. This project received no funding from the US National Institutes of Health. Neither IMI, the European Union, nor EFPIA are responsible for any use that may be made of the information contained therein.

We are grateful for the valuable contribution of previous SISAQOL-IMI members in different roles: as project leaders, work package leaders, members of the different working groups, and representatives from the individual institutions.

We extend our gratitude to former work package leaders Carla Mamolo, Jinma Ren, Jayne Galinsky, Linda Dirven as well as former SISAQOL-IMI Steering Committee members Daniel O'Connor and Kathy Oliver, for their invaluable efforts and dedication.

We thank Rosa Abruzzese for the meticulous copy-editing work, which has improved the clarity and accuracy of the project's outputs, and Kasirajan Vellaisamy for providing invaluable graphic design support.



Contents

12	1. Introduction
21	2. How to use the SISAQOL-IMI outputs
27	3. The analytical framework
33	4. SISAQOL-IMI recommendations according to the analytical framework
105	5. SISAQOL-IMI recommendations according to the matrix in the interactive table
106	6. How the recommendations were developed
124	7. How the SISAQOL-IMI was organised
130	8. Lessons learned from the consensus project
133	9. The future
135	10. List of tables, figures and appendices

Preface



When we started SISAQOL-IMI (Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints in Cancer Clinical Trials under the Innovative Medicines Initiative), we envisioned that a consensus among international experts and stakeholders (see the section on the organisation) would increase the use and quality of patient-reported outcomes (PRO) in cancer clinical trials. The recommendations provide an essential and substantive framework for analysing, interpreting and presenting PROs in cancer clinical trials. We hope that they will fulfil our expectations. The Guidebook is a living document, updated regularly based on the feedback received.

SISAQOL-IMI was not the first initiative to develop guidelines for the design and analysis of PROs in cancer clinical trials. Instead, the project sought to create a unified set of guidelines by leveraging best practices from existing frameworks. While many organisations have established valid, tailored guidelines, achieving a broader impact for PROs requires moving discussions beyond individual experts or stakeholder groups and reaching consensus on a shared set of guidelines. To ensure that resources, including patients' time and efforts, are fully optimized, PRO results must address the needs of all stakeholders and lead to meaningful, standardized conclusions. The development of SISAQOL-IMI proved to be timely and successful, with the alignment of interests across various expert and stakeholder groups, both within the European Union and globally.

This Guidebook is a comprehensive document presenting the SISAQOL-IMI recommendations (statements with accompanying explanations and examples) as well as an example protocol and statistical analysis plan (SAP) to illustrate how the recommendations can be implemented in practice. The Guidebook also provides an overview of the development process, including information about the background, methods, lessons learned, and the sustainability plan. It is intended for a broad audience, including statisticians, PhD students, clinicians, regulatory and health technology assessment (HTA) bodies, industry and academics, patient advocates and other stakeholders involved in implementing PROs in clinical research. It provides instructions on how to read and use this document and other accompanying SISAQOL-IMI outputs (the interactive table, the glossary, and the plain language checklist). The Guidebook is available as both a **HTML** and a PDF version. To facilitate interpretation, an integrated glossary (scientific and plain language) is available in the HTML version. The current guidebook is also linked to the SISAQOL-IMI interactive table, which is designed to present statements related to a specific research question.

Although the SISAQOL-IMI recommendations were developed within the framework of, and with all empiricism from cancer clinical trials, the concepts and thought process behind the development of these recommendations may apply to other diseases. The recommendations are also designed to be PRO measure (PROM) agnostic and can be applied to any validated PROM.

We believe these consensus recommendations will help us take an important step toward standardising how cancer clinical trials implement PROs, allowing for better use of this unique information by all stakeholders.

We trust you will find this document helpful, and, on behalf of the SISAQOL-IMI Consortium, we thank you for your time.

Abbreviations list



Abbreviation	Definition
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
AUC	Area Under the Curve
CRO	Contract Research Organizations
EFPIA	European Federation of Pharmaceutical Industries and Associations
EMA	European Medicines Agency
EORTC	European Organisation for Research and Treatment of Cancer
FDA	Food and Drug Administration
GEE	Generalized Estimating Equations
GEN	General
GLMM	Generalized Linear Mixed Model
HCP	Health Care Providers
HRQoL	Health-Related Quality of Life
HTA	Health Technology Assessment
ICE	Intercurrent Events
ICH	International Council for Harmonisation
IMI	Innovative Medicines Initiative
ITT	Intention-to-Treat
LMM	Linear Mixed Models
LOCF	Last Observation Carried Forward
MAR	Missing at Random
MCAR	Missing Completely at Random
MHRA	Medicines and Healthcare Products Regulatory Agency
MNAR	Missing Not at Random
MPE	Myeloma Patients Europe
NCI	National Cancer Institute
PRO	Patient-Reported Outcomes
PROMS	Patient-Reported Outcome Measures
PROTEUS	Patient-Reported Outcomes Tools: Engaging Users & Stakeholders
QoL	Quality of Life

Abbreviation	Definition
RCT	Randomised Controlled Trial
SACE	Survivor Average Causal Effect
SAP	Statistical Analysis Plan
SAT	Single Arm Trial
SISAQOL	Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints Data
STRATOS	STRengthening Analytical Thinking for Observational Studies
WECAN	Workgroup of European Cancer Patient Advocacy Networks
WP	Work Package

1. Introduction



Patient-reported outcomes (PROs), such as symptoms and side effects, functioning, and other health-related quality of life (HRQoL) issues, are recognised as important endpoints in new cancer therapies' benefit/risk assessments. The terms PRO and HRQoL are sometimes used interchangeably, but it is important to distinguish between them. HRQoL is defined as a multidimensional concept that usually includes patients' reports of their physical, emotional, social, and sometimes spiritual well-being. HRQoL, therefore, is a type of PRO, as it is best reported by patients themselves. HRQoL is limited in relation to quality of life (QoL) in that it focuses on areas in life that are affected by health or treatment. PROs are broader, covering other issues such as patient reports of a single or a group of symptoms, coping strategies, and other experiences (Marquis et al., 2006). The use of PROs in clinical trials is a scientific methodology used to measure patient experience with the disease and treatment. PRO endpoints document the effect of treatment from the patient's perspective. PROs can be used to capture the benefits of a cancer treatment (such as improvements in physical functioning or alleviation of disease symptoms).

PROs can also capture the harms of a cancer treatment (such as the level and impact of the treatment toxicity on patients' functioning) and complement clinician-reported symptomatic adverse event data (Thanarajasingam et al., 2015). The concept of "tolerability" covers this aspect of the evaluation of cancer treatment reported by the patients on how they feel and function while on treatment (Basch et al., 2020; Peipert et al., 2022). The terms outcome and endpoint are often used interchangeably but are not synonymous. The endpoints (primary or secondary) are defined in a specific manner through the attributes of the estimand framework (Lawrance et al., 2020) and the time point or interval of interest, the analysis metric, and the relevant PRO score interpretation threshold) ([SPIRIT-PRO PROtocol Reporting Template | The Proteus Consortium](#); Calvert et al., 2018). On the other hand, the term outcome refers more generally to the variable to be measured (the estimator). PRO is also included in the US Food and Drug Administration (FDA) term "Clinical Outcomes Assessment" as an important category.

SISAQOL-IMI was established to provide international consensus-based recommendations on designing, analysing, interpreting, and presenting PROs in cancer clinical research among multiple stakeholders.

Read more about the background of the project

In cancer clinical trials, patient-reported outcomes (PROs) capture treatment effects from the patient's perspective. While PRO data collection is growing, procedures for analysing and interpreting PRO data vary widely across researchers and organisations.

These differences can produce conflicting results from the same trials. Although conclusions may differ based on the research question, consistent answers are ideal when specific questions are set. Therefore, robust international standards for designing, analysing, interpreting, and presenting PRO data are essential. The SISAQOL-IMI project addresses this need.

Choosing design and statistical methods requires balancing feasibility, usefulness and robustness, and is highly dependent on the aims of the study. Health technology assessment (HTA) bodies also differ in guidance on analysing, interpreting and reporting PRO data in oncology (Chassany et al., 2022).

SISAQOL-IMI builds on work from the 2016 SISAQOL Consortium convened by the European Organisation for Research and Treatment of Cancer (EORTC), which engaged international stakeholders to develop recommendations for PRO analysis in randomised controlled trials (RCTs) (Bottomley et al., 2016; Coens, Pe, et al. 2020). These recommendations included a taxonomy of PRO objectives, identification of suitable statistical methods for PRO analysis, standardised statistical terminology for missing data and determination of appropriate ways to manage missing data (Coens, Pe, et al. 2020). This work marked a first step towards establishing good practice standards for PRO endpoint analysis in cancer clinical trials, providing a solid framework for specifying PRO endpoint analysis. However, comprehensive guidance on designing, analysing interpreting and presenting PROs in cancer clinical trials was still needed.

SISAQOL-IMI sought to expand stakeholder engagement, harmonise existing recommendations, and update them based on stakeholder needs and recent methodological advancements. Using a consensus-based approach, the Initiative aimed to develop a set of recommendations for designing, analysing, interpreting and presenting PROs for cancer clinical trials, with potential applicability to other therapeutic areas. The recommendations were intended to be broadly applicable to all validated PRO measures (PROMs), rather than tailored to specific PROMs.

Rationale for the creation of the SISAQOL-IMI project

One might anticipate that the same clearly defined PRO research endpoints would yield comparable results across different researchers. However, many researchers and organisations have their own procedures and standards for the analysis and interpretation of PRO data. It is evident that disparate approaches to analysing and interpreting PRO data from clinical trials can lead to contradictory or confusing conclusions. The lack of consistency in the way PRO data are evaluated may result in erroneous and inconsistent decisions being made by different stakeholders, which could ultimately have a negative impact on patient care and outcomes (Coon & Cappelleri, 2016).

For instance, if a study aims to assess changes in HRQoL over a six-month period, a cross-sectional HRQoL analysis at the six-month mark is not comparable to an area under the curve (AUC) analysis over the same duration. These two methods may lead to different outcomes (see Pe et al., 2018 for more examples). Similarly, two trials assessing change from baseline at month six could produce divergent findings if they apply different clinically meaningful difference thresholds, handle intercurrent events differently, or use varying statistical methods. Designing PRO endpoints involves numerous decisions that can affect the conclusions drawn. SISAQOL-IMI seeks to establish a framework that guides researchers in designing more targeted PRO endpoints based on the trial's context while offering recommendations on appropriate analytical approaches to standardise and enhance the transparency of PRO data evaluation.

Several international initiatives have provided good practices for planning and conducting clinical trials with PRO endpoints (Calvert et al., 2013; Calvert et al., 2018; Snyder et al., 2022). However, these efforts to develop high-quality PRO research have not resulted in the desired consolidation of PRO reporting. Instead, these efforts have highlighted the need for standardisation and further improvement of this area of research across different groups of stakeholders.

Read more about international PRO guidelines

PROs now exist for writing protocols (SPIRIT-PRO), publishing PRO findings (CONSORT-PRO), and for graphically displaying PRO data (Snyder et al., 2022). Guidelines also exist for selecting PRO measures, and SISAQOL has developed standards for PRO data analysis in RCTs.

A checklist for interpreting PRO data in research (Wu et al., 2014) is available. The PROTEUS Consortium (Patient-Reported Outcomes Tools: Engaging Users and Stakeholders) facilitates the dissemination and implementation of the methodologic tools developed to inform the use of PRO data from clinical trials (Snyder, Crossnohere et al., 2022). For additional guidance, see the scoping review by Kaneyase et al.

Table 1. Overview of available guidance on PRO in clinical research

Steps of clinical research	Guidelines
Developing protocols with PRO	Standard Protocol Items: Recommendations for Interventional Trials-PRO Extension (SPIRIT-PRO) (Calvert et al., 2018)
Selecting PRO measures	ISOQOL Minimum Standards for PRO Measures in Patient-Centered and Comparative Effectiveness Research COSMIN guidelines https://www.cosmin.nl/
Analysing PRO data	Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints Data (SISAQOL) Consortium (Coens et al., 2020)
Interpreting PRO findings	Clinician’s Checklist for reading and using an article about patient reported outcomes (Wu et al., 2014)
Reporting PRO findings	Consolidated Standards of Reporting Trials-PRO Extension (CONSORT-PRO) (Calvert et al., 2013) Graphical display guidelines (Snyder et al., 2019)

These guidelines are essential as they provide key information for evaluating the design and analysis used to inform PROs. Initiatives to develop high-quality PRO research in clinical trials also highlight the importance of standardising this research area. Establishing evidence-based, harmonised methodological standards is crucial to ensure that PRO data from cancer clinical trials are analysed, interpreted and presented appropriately. Such standardisation ensures PRO results can be reproduced and are comparable, enabling PRO data to meaningfully inform patient tolerability, treatment choices, and policy decisions.

The SISAQOL-IMI recommendations build on this prior guidance, including the recent estimand framework, to offer more comprehensive recommendations. These cover not only what to include, justify, or report in trial protocols and publications, but also provide direction on designing, analysing, interpreting, and presenting PROs. The SISAQOL-IMI recommendations include detailed explanations and examples, such as calculating completion and available data rates, addressing intercurrent events for specific PRO objectives, and selecting meaningful PRO interpretation thresholds. Additionally, they offer practical tools, including graphical templates to visualise PRO findings in a standardised format, aimed at enhancing scientific exchange, regulatory review, evidence-based and shared decision-making.

While many of these initiatives focused on randomised controlled trials (RCTs), there is still a need to agree on a set of standards for designing, analysing and interpreting PRO data in RCTs.

Read more about randomised controlled trials

Randomised controlled trials are best suited for examining cause-effect relationships between an intervention and a primary outcome. Randomisation helps balance baseline characteristics across arms, ensuring that any observed outcome differences are attributed to the intervention.

Even though RCTs are still considered the gold standard for demonstrating treatment effect, they are not always feasible or ethical to conduct (Baumfeld Andre et al., 2020). In pragmatic research settings, single arm trials (SATs) might better reflect real-life clinical practice (Liu et al., 2023). Some important subgroups of patients may not qualify to be included in an RCT. This limits the external validity and the implementation of the findings. Although there are still ongoing discussions on the reliability of PROs in SATs, it is the reality that many such trials have included PROs. It is, therefore, important to provide some recommendations for good practices on the use of PROs in SATs.

Read more about single arm trials

In pragmatic research settings, single arm trials (SATs) might better reflect real-life clinical practice (Liu et al. 2023), particularly for patient subgroups not eligible for randomised controlled trials. Over the last 20 years, many haematological and oncological drug approvals have relied on SATs (Hatswell et al. 2016, Agrawal et al., 2023) due to prevalent durable response rates. However, the lack of a formal control group in SATs presents a major challenge in avoiding or controlling for bias, making it essential to assess the robustness and limitations of PRO conclusions drawn from these trials.

Presenting and interpreting PRO data clearly remains an often-cited challenge (Snyder et al., 2019; Hsiao et al., 2019). While stakeholder-driven, evidence-based standards for the graphical display of PRO data have previously been developed, the recommendations are limited to the presentation of average scores over time and proportions in line with a responder definition. There is a need for a broader set of standards for the presentation and visualisation of PRO data that are comprehensible and relevant to a variety of stakeholders, including patients, clinicians, regulators, health technology assessment (HTA) bodies, and policymakers.

The importance of a shared understanding of terminology is often underestimated. To interpret PROM findings, it is necessary to harmonise concepts referring to meaningful change to patients.

The literature suggests many different terms such as clinical significance, clinical meaningful change, minimal important difference, and the need to harmonise terminology and definitions to facilitate the interpretation of PRO results (King, 2011).

SISAQOL-IMI aims

The overall aim of the SISAQOL-IMI was to achieve agreement among diverse stakeholders on the optimal use of PRO data in cancer clinical trials. An extensive network of international experts and patient representatives was established to collaborate and reach a consensus on recommendations on a minimum standard for designing, analysing, interpreting, and presenting PRO data (Pe et al., 2023).

The specific aims of SISAQOL-IMI were:

- To harmonise and update available recommendations based on stakeholder needs and recent developments in the methodological literature.
- To develop recommendations for analyses of both confirmatory and descriptive PRO endpoints
- To develop recommendations that are applicable to all validated PROMs.
- To develop recommendations for both RCTs and SATs.
- To improve the standards of PRO reporting in terms of reliable interpretation and graphical presentation of PRO results.

SISAQOL-IMI outputs

The SISAQOL-IMI final outputs are the following:

- The online interactive table: a tool organised according to the study objective and PRO variable of interest to give researchers easy access to recommendations relevant for their specific study design.
- The guidebook: the current document guiding the use of the recommendations and the final tools and providing information about the background, the development of the final recommendations, lessons learned and the sustainability plan.
- The glossary: a comprehensive compilation of all terms and concepts used in scientific and plain language, interlinked to the online guidebook and table.
- The plain language materials: the recommendations adapted for an audience without scientific PROM knowledge, such as checklists for patient representatives and other user/stakeholder groups, and tutorial videos.
- The scientific publication to disseminate the SISAQOL-IMI recommendations within the international scientific community.

The SISAQOL-IMI project does not provide guidance on:

- which instrument to choose
- the choice of specific PRO domains or concepts in clinical trials
- how to select PROMs for inclusion in clinical trials
- analyses of PROs other than binary, ordinal, or continuous variables
- issues specific to different stakeholder groups
- how to run PRO analysis using statistical software.

Applicability and expected impact of the SISAQOL-IMI recommendations

The project's motivation is that these consensus recommendations will guide the development, planning, and reporting of PRO endpoints in protocols and SAPs to ensure that the presentation and interpretation of PRO results align with the PRO objective in the trial. These recommendations are applicable in industry-sponsored trials, as well as in academic and healthcare personnel settings.

Implementing these guidelines will improve the standards of PRO endpoints in cancer clinical trials and guide the generation of high-quality PRO data in RCTs and SATs to be used to evaluate cancer treatment therapies. This way, PRO endpoints can reliably inform decision-making by regulators, bodies HTA agencies, health policymakers, clinicians, and patients. The guidance aims to inform and support the inclusion of PROs in pharmaceutical-sponsored trials and clinical trials initiated by health care providers (HCPs) and academic organisations. Establishing the SISAQOL-IMI recommendations as a recognised source in the scientific community can increase the number of clinical trials that include PROs and facilitate the comparison of results across different clinical trials.

The SISAQOL-IMI recommendations are expected to be relevant for patients in different ways. Increased focus on PRO as the primary endpoint and the improved quality of the analyses, interpretation and visualisation will benefit the patients and their relatives. Since the recommendations were developed in close collaboration with patient representatives, they may improve shared decision-making. Furthermore, standardised recommendations will facilitate involvement and engagement from patient research partners.

Although the SISAQOL-IMI recommendations have been developed for cancer clinical trials, they deal with more generic concepts and methods. Examples are handling missing data and intercurrent events, interpretation and visualisation of results, and most of the statistical analyses. Therefore, the recommendations are likely to be applied to other disease groups, such as different chronic diseases, but this assumption needs to be validated.

References

- Acquadro C, Arnould B, Marquis P, Roberts WM. Patient-reported outcomes and health-related quality of life in effectiveness studies: Pros and cons. *Drug Development Research*. 2006;67(3):193–201. <https://doi.org/10.1002/ddr.20079>
- Agrawal S, Arora S, Amiri-Kordestani L, de Claro RA, Fashoyin-Aje L, Gormley N, et al. Use of single-arm trials for US Food and Drug Administration drug approval in oncology, 2002–2021. *JAMA Oncology*. 2023;9(2):266–272. <https://doi.org/10.1001/jamaoncol.2022.5985>
- Altman DG, Blazeby J, Calvert M, Moher D, Revicki DA, Brundage MD. Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. *JAMA*. 2013;309(8):814–822. <https://doi.org/10.1001/jama.2013.879>
- Basch E, Campbell A, Hudgens S, Jones L, King-Kallimanis B, Kluetz P, et al. Broadening the definition of tolerability in cancer clinical trials to better measure patient experience. *Friends of Cancer Research [Internet]*. 2020 [cited 2025 Mar 10]. Available from: <https://friendsofcancerresearch.org>
- Baumfeld Andre E, Caubel P, Azoulay L, Dreyer NA, Reynolds R. Trial designs using real-world data: The changing landscape of the regulatory approval process. *Pharmacoepidemiology and Drug Safety*. 2020;29(10):1201–1212. <https://doi.org/10.1002/pds.5079>
- Bottomley A, Pe M, Sloan J, Basch E, Bonnetain F, Calvert M, Campbell A, & SISAQOL Consortium. Analysing data from patient-reported outcome and quality of life endpoints for cancer clinical trials: A start in setting international standards. *The Lancet Oncology*. 2016;17(11):e510-e514. [https://doi.org/10.1016/S1470-2045\(16\)30510-1](https://doi.org/10.1016/S1470-2045(16)30510-1)
- Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, & CONSORT PRO Group. Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. *JAMA*. 2013;309(8):814-822. <https://doi.org/10.1001/jama.2013.879>
- Calvert M, Chan A-W, Kyte D, Mercieca-Bebber R, Slade A, King MT. Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: The SPIRIT-PRO extension. *JAMA*. 2018;319(5):483–494. <https://doi.org/10.1001/jama.2017.21903>
- Chassany O, Engen AV, Lai L, Borhade K, Ravi M, Harnett J, Chen CI, Quek RG. A call to action to harmonize patient-reported outcomes evidence requirements across key European HTA bodies in oncology. *Future Oncology*. 2022;18(29):3323-3334. <https://doi.org/10.2217/fon-2022-0374>
- Coon CD, Cappelleri JC. Interpreting change in scores on patient-reported outcome instruments. *Therapeutic Innovation & Regulatory Science*. 2016;50(1):22–29. <https://doi.org/10.1177/2168479015622667>
- Hsiao CJ, Dymek C, Kim B, Russell B. Advancing the use of patient-reported outcomes in practice: Understanding challenges, opportunities, and the potential of health information technology. *Quality of Life Research*. 2019;28(6):1575–1583. <https://doi.org/10.1007/s11136-019-02121-z>
- Kaneyasu T, Hoshino E, Naito M, Suzukamo Y, Miyazaki K, Kojima S, Yamaguchi T, Kawaguchi T, Miyaji T, Nakajima TE, Shimozuma K. How to select and understand guidelines for patient-reported outcomes: A scoping review of existing guidance. *BMC Health Services Research*. 2024;24(1):334. <https://doi.org/10.1186/s12913-024-10707-8>

- King MT. A point of minimal important difference (MID): A critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research*. 2011;11(2):171–184.
<https://doi.org/10.1586/erp.11.9>
- Lawrance R, Degtyarev E, Griffiths P, Trask P, Lau H, D'Alessio D, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? *Journal of Patient-Reported Outcomes*. 2020;4(1):68.
<https://doi.org/10.1186/s41687-020-00218-5>
- 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](https://doi.org/10.1016/S1470-2045(23)00110-9)
- Pe M, Alanya A, Falk RS, Amdal CD, Bjordal K, Chang J, et al. Setting international standards in analyzing patient-reported outcomes and quality of life endpoints in cancer clinical trials—Innovative Medicines Initiative (SISAQOL-IMI): Stakeholder views, objectives, and procedures. *The Lancet Oncology*. 2023;24(6):e270–e283. [https://doi.org/10.1016/S1470-2045\(23\)00137-6](https://doi.org/10.1016/S1470-2045(23)00137-6)
- Peipert JD, Smith ML, & Team ES. Reconsidering tolerability of cancer treatments: Opportunities to focus on the patient. *Supportive Care in Cancer*. 2022;30(5):3661–3663.
<https://doi.org/10.1007/s00520-021-06669-y>
- Snyder C, Crossnohere N, King M, Reeve BB, Bottomley A, Calvert M, et al. The PROTEUS-Trials Consortium: Optimizing the use of patient-reported outcomes in clinical trials. *Clinical Trials*. 2022;19(3):277–284. <https://doi.org/10.1177/17407745221078028>
- Snyder C, Smith K, Holzner B, Rivera YM, Bantug E, Brundage M. Making a picture worth a thousand numbers: Recommendations for graphically displaying patient-reported outcomes data. *Quality of Life Research*. 2019;28(2):345–356. <https://doi.org/10.1007/s11136-018-2020-3>
- Thanarajasingam G, Hubbard JM, Sloan JA, & Grothey A. The imperative for a new approach to toxicity analysis in oncology clinical trials. *J Natl Cancer Inst*. 2015;107(10).
<https://doi.org/10.1093/jnci/djv216>



2. How to use the SISAQOL-IMI outputs

The following is a list of proposed uses of SISAQOL-IMI outputs; the guidebook may be used to:

- learn more about the concepts, the background, and the development process of the recommendations.
- know which statements belong to each topic of the framework.
- find instructions on how to use the content of the interactive table.

Uses of other outputs:

- The interactive table might be used for focused reading, to find statements applicable to your research question of interest according to design and objective.
- The protocol template may be used as an example when including PRO in your protocol.
- The statistical analysis plan (SAP) template may be used as an example when planning your PRO statistical analysis.
- The glossary may be used to explain definitions of complex terms and concepts in the scientific or the plain version.
- The plain language checklist may be used if you are a user representative and are keen to ensure that the study team has considered PRO issues relevant to patients.
- The tutorial videos may be helpful to the general audience to learn more about PRO and the SISAQOL-IMI work.
- The visualisation advice may be helpful when presenting the PRO results to scientific and plain language audiences.

For a summary of the final outputs of the project, please see publication (Amdal et al., submitted for publication 2025)

Prerequisites for applying SISAQOL-IMI recommendations to a trial

For the application of the SISAQOL IMI recommendations, it is necessary that the following elements are already available:

- The rationale for assessing PROs in the clinical trial is clearly stated.
- The specific PRO variable of interest is identified, such as magnitude of change, time-to-event, or proportion of patients.
- Relevant PRO concepts and appropriate time points or timeframes for assessment to address the PRO rationale are identified
- The chosen PRO tool is validated and suitable for measuring the concept of interest, including a scoring algorithm and an interpretation guide.
- If the PRO concept is used for confirmatory PRO objectives, it is critical to ensure that the trial has an adequate sample size (Coens, Pe et al., 2020).

How to use the interactive table

The SISAQOL-IMI interactive table organises and presents statements related to a specific research question. Use the interactive table to find statements applicable to your study. The table includes 30 cells, each representing a different study design, objective, and PRO variable of interest.

Read more about how to use the interactive table

Instructions for using the interactive table.

Before proceeding, please ensure you have:

- Defined a research objective based on the rationale for assessing patient-reported outcomes (PROs) in cancer clinical trials.
- Identified a relevant PRO concept and time point or time frame for the assessment.
- Selected a validated tool suitable for measuring the chosen PRO concept of interest, including a scoring algorithm and guidance on interpreting PRO scores.
- Confirmed that if the PRO concept will be used for confirmatory PRO objectives, the trial has a sufficient sample size.

Please open the interactive table that is provided for randomised clinical trials (RCT) and single arm trials (SAT) separately.

PRO* variable of interest		Randomised controlled trials (RCTs)			Single arm trials (SATs)	
		Evaluate clinical benefit		Describe patient perspective	Evaluate clinical benefit	Describe patient perspective
		Confirmatory objective		Descriptive objective	Confirmatory objective	Descriptive objective
		Superiority	Equivalence / non-inferiority		Superiority	
1	Magnitude of PRO (change) score at time <i>t</i>	A1	B1	C1	D1	E1
2	Responder with PRO improvement at time <i>t</i>	A2	B2	C2	D2	E2
3	Responder with PRO worsening at time <i>t</i>	A3	B3	C3	D3	E3
4	Time to PRO improvement	A4	B4	C4	D4	E4
5	Time to PRO worsening	A5	B5	C5	D5	E5
6	Overall mean or median PRO scores over a specified time frame	A6	B6	C6	D6	E6

*The term 'PRO' is used here as a placeholder; researchers should specify the domain they intend to measure.

Figure 1. The structure of the interactive table

Source: Authors' own elaboration

1. Select the trial design

Choose either a RCT or a SAT.

2. Define the overall PRO objective

In the interactive table columns, select your PRO objective:

Confirmatory objective: focus on treatment efficacy or clinical benefit

Descriptive objective: focus on understanding the patient perspective

3. Select the PRO variable of interest

From the rows in the interactive table, choose the PRO variable. Define the objective for within-patient or within-treatment group PROs in your protocol. When setting the PRO endpoint or variable of interest, consider if a worsening, stability, or improvement is expected. This consideration should be supported by previous literature, expert knowledge, or early-phase trials.

Read more about PRO objectives

Confirmatory objective: treatment efficacy or clinical benefit

For formal comparative conclusions between treatment groups using a patient-reported outcome (PRO) domain, apply the confirmatory objective guidelines:

1. *a priori* hypothesis: define a specific hypothesis for each PRO domain at the outset, as this will form the basis of the statistical testing at the trial's conclusion.

2. Multiple testing correction: if the trial involves multiple PRO domains or assessment points for a domain, apply corrections for multiple testing to ensure valid statistical results.

These measures ensure that conclusions about treatment efficacy or clinical benefit are statistically sound and reliable.

Superiority objective

A superiority PRO objective aims to demonstrate that for the pre-specified PRO domain, the treatment group is superior to the reference group by a clinically meaningful treatment effect size, pre-specified in the protocol. The trial design should ensure unbiased, adequately powered testing to allow a robust assessment of whether the hypothesis of no treatment effect can be rejected.

Equivalence or non-inferiority

An equivalence or non-inferiority of PRO objective seeks to demonstrate that, for the prespecified PRO domain, the treatment group is either similar (equivalent) or not worse (non-inferior) than the reference group, within a clinically relevant margin defined in advance in the protocol. The trial design should ensure unbiased, scientifically supported testing to allow the rejection of the hypothesis of non-equivalence or inferiority of treatment effect. It should be noted that equivalence or non-inferiority PRO objectives for SATs are currently not applicable and will not be discussed further.

Distinctions in superiority, equivalence, and non-inferiority analyses

Superiority design and analysis techniques differ from those used for equivalence or non-inferiority. Non significant p values from a statistical test aimed at assessing treatment differences (superiority test) should not be used as evidence that two treatment groups are similar (equivalent) or that one is not worse than the other (non-inferior).

The choice of effect size (superiority) and margins (equivalence or non-inferiority) should be adapted to the specific PRO instrument and clinical context, justified on both clinical and statistical grounds. Trials may include various combinations of these objectives between treatment groups. However, protocols should clearly specify the primary and secondary PRO study objectives for key domains of interest.

Descriptive objective: describe patient perspectives

When a PRO domain is used to describe patient perspectives during the trial (e.g. tolerability) or to explore PRO data and use its findings to inform future studies, descriptive or exploratory objective rules apply: an *a priori* hypothesis is not required for the PRO domain. However, these outcomes cannot support comparative conclusions or claims of treatment efficacy or clinical benefit.

Findings should be reported as either descriptive (i.e., summarising estimates with or without confidence intervals but without statistical testing), or exploratory.

Confirmatory and descriptive PRO objectives complement each other, and can coexist 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 patient perspectives, or serve as exploratory objectives.

Read more about PRO variables of interest

When defining the PRO endpoint or variable of interest, it is important to identify whether a worsening, stability, or improvement is expected within the treatment group. This assumption should be informed by previous literature, expert knowledge, or early-phase trials. For SISAQOL-IMI, statements were developed based on a prioritised list of PRO variables of interest agreed upon by Consortium members.

The following six PRO variables of interest were considered:

1. **Magnitude of PRO (change) score at time t :** the actual value or change from baseline value for a PRO domain at predefined time points.
2. **Proportion of responders with improvement at time t :** whether the value (or change from baseline value) from a PRO domain at a specific time point reaches a predefined improvement threshold or not.
3. **Proportion of responders with worsening at time t :** whether the value (or change from baseline value) from a PRO domain at a specific time point reaches a pre-specified worsening threshold or not.
4. **Time to PRO improvement:** the time taken for a clinically relevant improvement from a PRO domain to be observed within a pre-specified time frame.
5. **Time to PRO worsening:** the time taken for a clinically relevant worsening from a PRO domain to be observed within a pre-specified time frame.
6. **Overall mean or median scores over a specified time frame:** the starting point for the PRO variable of interest is the mean or median score of all available scores from a PRO domain for an individual patient over a pre-specified timeframe.

Visualisation of PRO data

Visualisation of PRO data is a crucial element when reporting PRO results to the audience. SISAQOL-IMI has developed visualisation advice for presenting the PRO results to scientific and plain language audiences. These recommendations address how to present the data from RCTs and SATs with templates and graphical examples. They are listed in Chapter 4, Section 6 and are included in all relevant cells in the table. General good advice (not consensus-based) on how to develop “good figures” is also included.

Read more about the visualisation advice

The visualisation of PRO data is essential for effectively reporting PRO results to diverse audiences.

The following recommendations for visualising PRO data from clinical trials for a general audience can be applied:

- 1) Adapt scientific figures for a general audience: first, create figures according to scientific recommendation statements, then modify them to enhance accessibility. When plain versions do not have corresponding graphic types, such as Kaplan-Meier curves and forest plots, choose an alternative visualisation, such as a bar chart.
- 2) Create a plain figure version only: if there are corresponding scientific figures, carefully consider the intended message for the plain figure audience, using guidance on plain figures serving as a framework for setup and presentation. Simplified figures should not contradict the statements of the scientific figures (e.g. VizSci9_GEN graphs should not include different directionalities; this should also be adhered to for plain graphs).

In addition to the statements on communication and visualisation of results (see Chapter 4, Subsection 6), general guidance for creating quality illustrations has been developed and included at the end of Chapter 4.

References

- Amdal CD, Falk RS, Alanya A, Schlichting M, Roychoudhury S, Bhatnagar V, SISAQOL-IMI Consortium. SISAQOL-IMI consensus-based guidelines to design, analyse, interpret and present patient-reported outcomes in cancer clinical trials. Submitted for publication
- 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](https://doi.org/10.1016/S1470-2045(19)30790-9)



3. The analytical framework

This chapter outlines the attributes used to organise the recommendations, which can be helpful to plan a study. Addressing all parts of the analytical framework is expected to increase the quality of a study, thereby providing valid and reliable PRO results. It includes the estimand framework's five attributes (Fiero et al., 2020; Lawrance et al., 2020) and was also expanded to include consideration of PRO scores interpretation thresholds, study design, analyses, description of external comparison (for SATs only), and communication and visualisation of results.

The SISAQOL-IMI introduced the term “PRO score interpretation thresholds” as the umbrella term to discuss meaningful change or differences in the PRO at both patient and group-level in a study.

Each concept and underlying issues are detailed below (Table 2).

Read more about each of the attributes of the analytical framework

Estimands framework

The estimands framework is a structured approach to align clinical trial objectives with study design, including endpoints, and analysis. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9 guidance document, “Statistical Principles for Clinical Trials” [ICH E9 (R1)], describes the estimands framework, which is essential for harmonising clinical trial endpoint analyses (ICH 2019).

An estimand precisely describes the treatment effect, reflecting the research question posed by the trial objective. It summarises, at population level, the outcomes in patients under the different treatment conditions being compared.

The estimands framework ensures alignment among objectives, design, execution, and interpretation of statistical analyses in clinical trials. Clearly defined trial objectives can be translated into key clinical questions of interest by defining an appropriate estimand.

1. Attributes of an estimand

The definition of an estimand requires specification of five attributes:

- a. Population: the patient group targeted by the clinical question.
- b. Treatment: the treatment condition of interest and, where applicable, the alternative treatment condition for comparison.
- c. Variable (or endpoint): to be obtained for each patient in order to address the clinical question.
- d. Intercurrent events: the intercurrent event strategy considering relevant events occurring after treatment initiation that affect either the interpretation or the existence of the measurements related to the clinical question of interest.
- e. Population-level summary for the variable: provides a basis for comparing treatment conditions.

The estimands framework applies whenever estimating a treatment effect or testing a related hypothesis and is applicable to randomised controlled trials (RCTs), single arm trials (SATs), and any endpoint type.

a. Population

The **target** population is defined as the study population targeted by the clinical question. This will be presented by the question's focus, typically outlined by study inclusion and exclusion criteria or by sub-populations characterised by baseline measures or principal strata affected by specific intercurrent events.

The **analysis** population includes subjects to be included for estimating the treatment effect. Different analysis populations may be used for patient-reported outcome (PRO) analyses. Ideally all enrolled patients (intent-to-treat, [ITT] population) should be considered. However, in subsequent refinements emphasis may be placed on patients with at least a valid baseline PRO assessment, patients with a valid baseline PRO assessment and a follow-up assessment, and patients with any PRO assessment, according to the clinical question of interest.

The **ITT** population is generally accepted as the relevant analysis population when drawing conclusions on treatment efficacy. Using the full analysis set will give the most reliable estimate, accounting also for individuals that do not take the intervention.

When evaluating the safety/tolerability of a treatment the SISAQOL-IMI concluded that the most relevant analysis population would be patients who actually started treatment (e.g., received at least one dose or treatment cycle).

b. Treatment

SISAQOL-IMI does not provide specific PRO recommendations for the treatment attribute within the estimands framework, as the treatment follows the intervention outlined in the clinical trial protocol.

c. PRO variable of interest

The PRO variable of interest specifies the patient outcome of interest, including the timeframe for evaluation. A generic reference to a "PRO score" is not specific; for example, a more precise variable could be "pain severity (as measured by [questionnaire X]) at six months post-baseline."

Six generic PRO variables of interest include, where “PRO” is the placeholder for the specific PRO domain that is measured:

1. **Magnitude of PRO (change) score at time t:** the actual value or change from baseline value for a PRO domain at pre-specified time points.
2. **Proportion of responders with PRO improvement at time t:** whether the value (or change from baseline value) from a PRO domain at a specific timepoint reaches a pre-specified improvement threshold or not.
3. **Proportion of responders with PRO worsening at time t:** whether the value (or change from baseline value) from a PRO domain at a specific timepoint reaches a pre-specified worsening threshold or not.
4. **Time to PRO improvement:** the time taken for a clinically relevant improvement from a PRO domain to be observed within a pre-specified time frame.
5. **Time to PRO worsening:** the time taken for a clinically relevant worsening from a PRO domain to be observed within a pre-specified time frame.
6. **Overall mean or median PRO scores over a specified time frame:** the starting point for the PRO variable of interest is the mean or median score of all available scores from a PRO domain for an individual patient over a pre-specified time frame.

d. Handling of intercurrent events

Intercurrent events (ICEs) are events that can occur after treatment initiation that impact the presence and/or interpretability of the measurement associated with the clinical question of interest. The strategy for addressing intercurrent events needs to be specified when describing the clinical question, since how ICEs are handled is crucial for defining the treatment effect to be estimated. ICEs may be accounted for as part of the treatment attribute (treatment policy strategy, hypothetical strategy), in the population attribute (principal stratification strategy), or as part of the variable of interest (composite strategy, while-on-treatment strategy). If not yet addressed in the previous attributes, other ICEs and the strategies to address them should be specified in this attribute. ICEs differ from missing values. Missing data are data that would be meaningful for the analysis but were not collected (for example, due to lost-to-follow-up, unanswered questionnaires). For events like death, data are absent, but these are not classified as missing data.

e. Population-level summary

The population-level summary indicates how patient-specific endpoints are estimated to enable comparison between treatment conditions. Population-level summaries may include effect measures, such as differences in mean change for magnitude of change endpoints, hazard ratios for time-to-event endpoints, and risk difference or odds ratios for response endpoints.

2. Patient-reported outcome score interpretation thresholds

SISAQOL-IMI introduced the term “PRO score interpretation thresholds” as an umbrella term covering meaningful changes or differences in PRO scores at both patient and group levels.

a. Application of PRO score interpretation thresholds

Key considerations include the applicability at the patient (e.g., change within individual patients) or group level (e.g., mean group changes over time), the relevance to the trial population at hand, and the need for specific sensitivity analysis.

- b. Selection of PRO score interpretation thresholds
Selecting a PRO score interpretation threshold requires understanding the threshold types, methods for establishment, and compatibility between the patient population used to establish the threshold and the trial population to which it is applied.
- c. Reporting of PRO score interpretation thresholds
The methodology for deriving the PRO score interpretation threshold should be provided, specifying whether it is anchor- or distribution-based and the population in which it was established.

3. Study design considerations

SISAQOL-IMI recommendations are meant to be interpreted within the context of the overall clinical trial design, and are assumed to be appropriate for the intended trial PRO objective.

Key assumptions are as follows:

- Knowledge of treatment characteristics (start, duration, assessment schedule, discontinuation conditions)
- Validated measures with known measurement properties, including, but not limited to, PRO score interpretation thresholds
- Identification of major covariates impacting PROs or trial completion (e.g., age). Such covariates are expected to be trial-specific and part of the clinical trial design aspect
- Comparable PRO assessment schedules across trial arms
- Compilation of reasons for missing data.

4. External comparison

In single arm oncology studies, comparisons often involve baseline changes within SATs. External data could be used to contextualise or compare the SAT data to an external control group (e.g., historical data or reference data). However, when used, external data should be carefully assessed for quality, comprehensiveness, and comparability in terms of patient population, endpoints, summary measures, temporality and key data collection including but not limited to the PRO assessment schedule (Mishra-Kalyani et al., 2022).

5. Analysis considerations

SISAQOL-IMI have provided recommendations to guide the choice of statistical methods to analyse PRO data, and to match appropriate statistical methods to valid PRO objectives. Generally, four different aspects need to be considered when considering the appropriate statistical method for PRO analyses.

(a) Assumptions

The choice of method requires the assumptions of the method to be fulfilled. Parametric methods have limitations, most importantly, their reliance on distributional assumptions, primarily with smaller sample sizes. Model assumptions about missing data and the influence of outlying observations should be checked.

(b) Presentation of data

The descriptive presentation of data is crucial for understanding the population of interest and the sample included in the analysis.

(c) Main statistical analysis

The choice of statistical method should match the research question of interest according to the estimand framework.

(d) Sensitivity and supplementary analysis

It is important to conduct sensitivity and supplementary analysis for the chosen statistical analysis method addressing key PRO objectives. Sensitivity analyses test the robustness of inferences against model assumptions and data limitations, while supplementary analyses provide additional insights into the treatment effect (e.g. in a different analysis set). Both analyses help address limitations in the results.

6. Communication and visualisation of results

Clear PRO data presentation improves interpretability across stakeholder groups. In addition to scientific guidelines, SISAQOL-IMI developed guidelines for plain graphs to communicate PRO data effectively to patients and the general population.

Table 2. The SISAQOL-IMI analytical framework

1. Estimands framework
a. Population
b. Treatment
c. PRO variable of interest
d. Handling of intercurrent events
i. Disease progression
ii. Deviations from protocol-defined treatment
iii. Concomitant therapies allowed by the protocol
iv. Death
v. Treatment discontinuation or start of subsequent therapy
e. Population-level summary
2. PRO score interpretation thresholds
a. Application of PRO score interpretation thresholds
b. Selection of PRO score interpretation thresholds
c. Reporting of PRO score interpretation thresholds
3. Study design considerations
4. External comparison
5. Analyses considerations
a. Assumptions
b. Main analysis
c. Sensitivity/supplementary analysis
d. Results presentation and interpretation
6. Results communication and visualisation
a. Scientific figure types
b. Considerations applicable to all scientific figures
c. Plain figure types
d. Considerations applicable to all plain figures

References

- Fiero MH, Pe M, Weinstock C, King-Kallimanis BL, Komo S, Klepin HD, et al. Demystifying the estimand framework: a case study using patient-reported outcomes in oncology. *Lancet Oncol*. 2020;21(10):e488-e494
- International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials [Internet]. 2019 [cited 2025 Mar 10]. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf
- Lawrance R, Degtyarev E, Griffiths P, Trask P, Lau H, D'Alessio D, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? *J Patient Rep Outcomes*. 2020;4(1):68.
<https://doi.org/10.1186/s41687-020-00218-5>
- Mishra-Kalyani PS, Amiri Kordestani L, Rivera DR, Singh H, Ibrahim A, DeClaro RA, et al. External control arms in oncology: Current use and future directions. *Ann Oncol*. 2022;33(4):376-383.
<https://doi.org/10.1016/j.annonc.2021.12.015>

4. SISQOL-IMI recommendations according to the analytical framework



This chapter lists the recommendations following the structure of the analytical framework explained above. There are 146 SISAQOL-IMI recommendations in total. For each recommendation, there is a statement followed by an explanation section, itself illustrated with examples where available and relevant.

For each of the analytical framework's attributes, general statements (applicable to both RCT and SATs) are listed first, followed by statements applicable to RCTs and SATs only. The statements are marked with GEN (general), RCT, and SAT respectively, and each statement has a unique name and number, e.g. **EstFrame1_GEN**. For the general statements applicable to both RCT and SAT settings, the text of some statements/explanations was adapted to match the specific setting. Those statements/explanations are indicated with an asterisk (*). An overview of the language differences between the statements applicable to both research settings is given in Appendix 1.

Estimands framework

EstFrame1_GEN*

Statement: the choice of each estimand should depend on the PRO objective of the clinical trial and on the research question.

Explanation*: in RCTs, research objectives are often unclear and reported sporadically. The best approach to design, conduct and analyse the PRO part of an RCT depends on the

trial context. Several factors may influence the approach chosen. These include the type of treatment and the intent of the study (curative or not) and the type of PRO (symptoms, functional impacts, general HRQoL).

For both RCT and SAT setting, the PRO objective will determine how to address ICEs such as treatment discontinuation. If the PRO objective is related to patients' experience while on treatment, then PRO measurements while on treatment will be more relevant, whereas efficacy objectives measuring long-term benefits on patients would take into account PRO observations even after the treatment is discontinued. Other defining elements may include:

- Nature of the disease (disease of interest, stage of disease)
- Characteristics of the target population
- Whether PROs are measured for benefits and/or tolerability
- Treatment regime (e.g., intervention, duration and frequency)
- Treatment intent (e.g., palliative or curative)
- Treatment blinding
- What type of PRO measurements can be used and the availability
- PRO time points of interest
- The PRO-based population summaries of interest (means or medians, changes from baseline, proportion of responders, time until deterioration/improvement, etc.)
- Relation with other outcomes of interest in the study.

Different stakeholders may prefer different estimands.

Prior to the analysis, an adequate objective should be developed covering all necessary estimand attributes. Clear trial objectives should be translated into key clinical questions of interest by defining suitable estimands.

For example, if the objective is to estimate what proportion of patients on treatment improved their physical functioning at month six, the analysis is limited to patients still on treatment at month six. However, if the objective is to compare the proportion of patients who improved their physical functioning at month six, the analysis set is not limited by treatment discontinuation.

EstFrame2_SAT

Statement: the aim(s) of collecting PROs in a single arm study should be clearly defined.

Explanation: the aim(s) and related research questions on PROMs in single arm studies are often poorly defined or not mentioned at all. A clearly defined aim is needed to define the research question and the corresponding estimand.

EstFrame3_SAT

Statement: a PRO objective for single arm studies may be to describe PRO scores over time.

Explanation: descriptive PRO objectives might include assessing the means or medians of PRO scores at several time points to inform tolerability outcomes, or describing the time-to-improvement of symptoms or time-to-deterioration of symptoms, or describing change from the baseline. It is important to consider that objectives of the PRO analysis may differ from the objectives for other outcomes in the single arm study.

EstFrame4_SAT

Statement: a PRO objective for single arm studies may be to make comparisons, either to baseline or to external controls. This is on condition that appropriate care is taken in the design and the conduct of studies to reduce bias, to avoid misleading interpretations due to the absence of randomisation and treatment blinding.

Explanation: the lack of a reference arm means that single arm trials have limitations in assessing the efficacy of PRO objectives. However, comparisons can still be made either within the single arm study or using external information. It is possible, for example, to evaluate a change from the baseline, or to compare a change from the baseline against a pre-specified meaningful within-group change threshold, or to make a comparison with control PRO data from external sources (for instance, from a historical control dataset).

It is important to exercise caution to avoid undue bias. For example, in comparisons with a baseline within a single arm study, other causes for the observed change over time should be considered. These may include the use of pain killers if pain is the outcome, the natural course of the disease, or the response shift.

When using external controls for comparison, it is important to select controls that closely align with the patient group in the study, such as same centres, age distribution and cultural backgrounds. Moreover, the PRO collection should be conducted in a similar fashion in both groups, such as using the same questionnaires, that the time and frequency of measurements is the same, and the strategies chosen to handle ICE be similar. Any deviations should be justified.

EstFrame5_SAT

Statement: when formulating the research questions and translating them into estimands, one must consider several defining elements such as:

- Nature of the disease (disease of interest, stage of disease)
- Target population
- Whether PROs are collected for benefit and/or tolerability
- Treatment regime (e.g., intervention, duration, and frequency)
- Treatment intent (e.g., palliative or curative)

- What types of PROMs can be used, and their availability
- PRO time points of interest
- PRO-based summary measures (means or medians, change from baseline, proportion of responders, time-until-deterioration/improvement, etc.)
- Relation with other outcomes of interest in the study
- How to handle post-baseline events, such as treatment discontinuation or death, that is not part of the PRO measure

Explanation: in single arm trials, research objectives are often unclear and reported sporadically. The best approach to set up and analyse a single arm study depends on the trial context. Factors like the type of treatment, the aim of the study (curative yes/no) and type of PRO measure (symptoms, other functional impacts, HRQoL) may influence the approach.

EstFrame6_SAT

Statement: PRO objectives in a single arm study should be specified using the estimand framework. For clarity, it is necessary to detail the following attributes:

- The treatment condition of interest; and other treatment conditions when the aim is to make comparisons between different treatments
- The target population for the PRO research objectives
- The PRO endpoints
- Strategies to handle ICEs, such as:
 - ▣ death
 - ▣ treatment discontinuation
 - ▣ disease progression
 - ▣ use of concomitant therapy or violation of protocols
- The targeted population-level summary.

Explanation: a well-defined research question for PRO could be “can analgesic medication (treatment x) reduce pain score at time t in patients with advanced breast cancer and pain, still alive at time t and irrespective of disease treatment discontinuation?” In this example, the estimand framework can be applied as follows:

1. The treatment condition is administering pain medication using a certain dosing scheme.
2. The population targeted is patients with advanced breast cancer and pain.
3. The PRO endpoint is the mean change in pain from baseline.
4. Strategies to handle ICEs:
 - Patients who died before time t will not be included in the analysis (while-alive-strategy).
 - Patients’ PRO data at time t will be used in the analysis even if they discontinued treatment before time t (treatment policy strategy).
5. Results are summarised by mean change in pain score from baseline and a 95 % confidence interval among those still alive at time t .

1a. Population

Pop1_RCT

Statement: when the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective) by demonstrating that the treatment arm is superior to the reference arm (superiority objective) for the PRO endpoint, all randomised patients should be used as the PRO analysis set for the main PRO analysis (according to the intention-to-treat [ITT] principle). Any deviations should be justified.

Explanation: according to the ICH guidelines, superiority objectives should use the ITT population in the primary analysis because it tends to avoid over-optimistic estimates of efficacy.

The ITT principle also preserves randomisation, providing a secure foundation for statistical tests.

Example: the main goal of a trial is to demonstrate that the score of the PRO physical functioning in the treatment arm is better than in the control arm at month six. In this case, the analysis should account for all randomised patients and not be limited to those patients who are still alive or completed questionnaires at month six.

Pop2_RCT

Statement: the safety analysis set (all randomised patients who started protocol treatments) should be used as the starting point for the main PRO analysis population to answer PRO objectives related to tolerability.

Explanation: assessing tolerability should be limited to those patients who were sufficiently exposed to the interventions. If the PRO dataset for analysis is selected based on high or low PRO scores and the same PRO domain is used as an outcome, the effects of regression to the mean should be considered in the PRO analysis or should be mentioned when discussing the findings. Such selection may occur as part of the RCT entry criteria or by selecting a specific PRO dataset for analysis. The effect of regression to the mean can be estimated by a quantitative bias analysis. This requires an estimate of the intra- and inter-individual variance, which can be estimated from repeated measurements before treatment or from external data. If this information is not available or reliable enough, then the potential bias due to regression to the mean should be discussed when reporting the results.

Example: if the intention is to measure the impact of a particular side effect of the treatment/intervention such as a rash, then only patients who were exposed to the treatment/intervention should be included in the analysis. Patients who did not receive the actual treatment/intervention are not representative of the exposed population and will dilute the reporting of the side effect.

Pop3_SAT

Statement: the analysis population of a single arm trial depends on the research objectives and may differ when measuring PROs for benefit or tolerability.

Explanation: when PROs are measured for efficacy, a relevant analysis population could be all patients included in the study. Analysing data obtained from all enrolled patients will provide the most reliable estimate of what might be expected in real-world scenarios where not everyone follows the intervention according to protocol.

When the PRO measurements aim to assess safety or tolerability, the most relevant analysis population is often patients who actually started treatment.

1b. Treatment

No specific PRO statements were developed since the treatment will follow the intervention as described in the protocol.

1c. PRO variable of interest

PROvar1_GEN*

Statement: continuous or ordinal PROs should be analysed as continuous or ordinal outcomes. A motivation should be given if a different approach is used, such as dichotomising PRO values in a responder analysis or a time-until-deterioration/improvement analysis.

Explanation RCT*: for clinicians and patients, interpreting results from a responder or time-to-deterioration/improvement analysis is often easier. However, from a methodological viewpoint, these analyses have limitations. Categorising PRO scores may lead to misclassification and a reduction in statistical power. Additionally, the choice of threshold for defining responders or events may be hard to justify. Sensitivity analyses are necessary to evaluate the effects of threshold determination and misclassification rates. Additionally, analysing time-to-deterioration/improvement is complicated by missing data and the fact that the exact moment of “deterioration/improvement” is unknown, since PROs are only measured at specific time points. To address this issue, interval censoring methods can be used. Furthermore, the time-to-deterioration/improvement analysis does not consider the reversibility of the deterioration/improvement status, such as in situations where low PRO values may later improve.

Rescaling of PROs, such as transforming an original 4-point response scale into a continuous 0–100 scale is permissible because a rescaled PRO retains the same statistical properties. However, ordinal scores that have been rescaled should still be analysed as ordinal outcomes.

Explanation SAT*: for clinicians and patients, interpreting results from a responder or time-to-deterioration/improvement analysis is often easier. However, from a methodological viewpoint, these analyses have limitations. Categorising PRO scores may lead to misclassification and a reduction in statistical power. Additionally, the choice of threshold for defining responders or events may be hard to justify. Sensitivity analyses are necessary to evaluate the effects of threshold determination and misclassification rates. Additionally, analysing time-to-deterioration/improvement is complicated by missing data and the fact that the exact moment of “deterioration/improvement” is unknown, since PROs are only measured at specific time points. To address this issue, interval censoring methods can be used. Furthermore, the time-to-deterioration/improvement analysis does not consider the reversibility of the deterioration/improvement status, such as in situations where low PRO values may later improve. This makes time-to-event endpoints difficult to interpret, especially in single arm trials without a reference arm.

Patient summary measures such as maximum PRO value and area under the curve (AUC) are dependent on the chosen time frame. These measures lack a causal interpretation, and interpretation is especially difficult without a direct comparison group. Different patterns of assessment, each reflecting different clinical scenarios, may result in comparable AUC or maximum PRO value. The maximum PRO value depends on the number of measurements. Furthermore, handling ICEs and missing data is challenging.

Rescaling of PROs, such as transforming an original 4-point response scale into a continuous 0–100 scale is permissible because a rescaled PRO retains the same statistical properties. However, ordinal scores that have been rescaled should still be analysed as ordinal outcomes.

PROvar2_RCT

Statement: when the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective), within-patient level summary measures (such as the maximum PRO value, area under the curve, etc.) are not recommended as an endpoint but may be useful to aid interpretation in a descriptive setting.

Explanation: it is possible to reduce the multiple measures on each patient to a single number, such as the average rate of change, maximum value, area under the curve, etc. The treatment effect can then be tested using a simple test like the two-sample t-test on these summary measures. Although this approach is straightforward, it is not recommended.

First, there is a considerable loss of information by using the summary measure instead of modelling the full longitudinal profiles, which results in a loss of statistical power. Second, it is challenging to properly take into account missing data. Interpolation is often used for intermediate missingness, whilst single imputation techniques, like LOCF, are often used for dropout. These simple techniques can lead to bias and misrepresentation of information.

Finally, simple tests like the two-sample t-test assume that the scores are identically and independently distributed. This assumption is not necessarily met, as, for example, the variance of the summary measure can be much smaller in the group of patients who drop out before the end of the study.

An assessment of the summary measures can be useful as an exploratory analysis, yet one should be careful not to draw conclusions from the results. An overview of the relevant assumptions should be reported and discussed to assess the extent to which these assumptions could impact the PRO results.

It should be noted that some summary measures can be meaningful depending on the context. Such a summary measure could be part of the scoring algorithm inherent to a specific PRO measurement (e.g. weekly average of daily diaries). Or it may be well-understood and accepted as a meaningful reduction of the longitudinal PRO scores pertaining to a patient. In this case, the summary score can act as an endpoint itself within the estimand framework. The PRO scores are then interpreted at the level of the summary score. In the example of the weekly average of daily diaries, the weekly average is seen as the endpoint itself with the handling of missing or incomplete daily values part of a defined scoring algorithm.

1d. Handling of intercurrent events

Independent of type of intercurrent event

ICE1_GEN

Statement: when a certain ICE can be interpreted as a treatment failure, and a plausible relationship to the PRO domain and time-to-worsening of the PRO is considered, the ICE could be incorporated into the endpoint. This can be achieved through a composite strategy, by using a composite outcome of “time-to-worsening of PRO” and “occurrence of ICE”. It is important to provide the rationale for combining the PRO outcome with the ICE, as well as information on the relative frequency of the ICE.

Explanation: if the PRO objective is tolerability, patients may discontinue treatment because it is no longer tolerable. In those cases, non-response can be defined as having a PRO score reflecting harm, or treatment discontinuation. In a confirmatory study, where improvement in PRO is considered, disease progression may be considered as treatment failure. In those cases, using a composite strategy may be an option if PROs after disease progression are not available. Consequently, non-response at a certain time point may be defined as a PRO score that indicates harm or disease progression.

However, the occurrence of disease progression and the worsening of PRO measure can only be linked if there is a plausible association between time-to-worsening of the PRO and the patient-reported symptoms. This approach should, therefore, be used with caution because, in real-life scenarios, interpreting results may be complex when it involves combining different outcomes. For example, a patient may experience growth in lung metastases whilst still reporting a good HRQoL. It is therefore, in general, recommended to measure PROs after the ICE instead of using a composite outcome.

When many patients experience the ICE during follow-up, this will dominate the composite outcome. Therefore, information on the relative frequency of ICE should be provided.

ICE2_GEN

Statement: when a certain ICE can be interpreted as a treatment failure with a plausible relationship to the PRO domain and no further PRO data are available afterwards, the ICE could be considered as part of the responder definition (composite outcome). It is important to provide the rationale for combining the PRO outcome with the ICE, as well as information on the relative frequency of the ICE.

Explanation: if the PRO objective is tolerability, patients may discontinue treatment because it is no longer tolerable. In those cases, non-response can be defined as having a PRO score reflecting harm, or treatment discontinuation. In a confirmatory study, where improvement in PRO is considered, disease progression may be considered as treatment failure. In those cases, using a composite strategy may be an option if PROs after disease progression are not available. Consequently, non-response at a certain time point may be defined as a PRO score that indicates harm or disease progression.

However, the occurrence of disease progression and the worsening of PRO measures can only be linked if there is a plausible association between the symptoms related to disease progression and the patient-reported symptoms. This approach should, therefore, be used with caution because, in real-life scenarios, interpreting results may be complex when it involves combining different outcomes. For example, a patient may experience growth in lung metastases while reporting a good HRQoL. It is therefore, in general, recommended to measure PROs after the ICE instead of using a composite outcome.

When many patients experience the ICE during follow-up, this will dominate the combined outcome. Therefore, information on the relative frequency of ICE should be provided.

ICE3_RCT

Statement: when the goal of the objective is to draw conclusions about non-inferiority or equivalence of PROs between treatment arms, treatment policy and hypothetical intercurrent event strategies should be considered to address adherence to the treatment and/or protocol. However, for evidence to be robust, results from both approaches should lead to similar conclusions.

Explanation: non-inferiority/equivalence research questions could become a study objective, in particular when the aim is to maintain a certain level of HRQoL. For instance, when a new treatment is combined with standard therapy such as radiotherapy, and the research question is if the new treatment will not worsen certain symptoms further in an advanced setting. The underlying decision to be made for non-inferiority trials is based on a one-sided nature of the hypothesis, i.e., that the new treatment will be preferred if it is either better or no worse than the existing treatment.

Unlike superiority trials, in non-inferiority and equivalence trials low adherence to the allocated treatment or deviations from protocol could contribute to the false rejection of the null hypothesis (i.e., incorrectly concluding non-inferiority/equivalence). This can occur when treatment effects are estimated under a real-world adherence pattern that compares groups

defined by treatment allocation instead of received treatment. Reduced observed differences between treatment arms could arise when, for example, study participants switch to the opposite treatment arm or take an alternative treatment with similar efficacy as the opposite treatment arm.

An “ideal treatment effect”, which estimates the difference between two randomised groups in the absence of non-adherence or protocol deviations, can represent a more suitable alternative. Estimating this hypothetical construct rather than the observed treatment effect requires accounting for deviations that may have occurred before or after the randomisation of the patient. Deviations that occurred before randomisation are not considered ICE and are to be addressed by defining the population(s) of interest in the estimand.

Deviations that occurred after randomisation can no longer be assumed to be independent of treatment allocation/exposure. Excluding all patients with a post-randomisation serious protocol deviation could introduce selection bias. Patients who do not adhere to treatment or deviate from protocol might systematically differ between treatment arms, with respect to observed or unobserved confounding factors, resulting in biased estimated treatment effects. A hypothetical estimand that addresses what would have been the treatment effect if patients did not have serious deviations and were expected to behave similar to other patients who continued on treatment, is therefore preferred.

The resulting recommendation advises minimising pre-randomisation deviations and assessing the extent to which these could impact the PRO results. Subsequent post-randomisation (severe) deviations impacting treatment/protocol adherence are to be handled via the hypothetical policy. Under this strategy, the observations after patients no longer adhere to their allocated treatment are replaced by hypothetical values. These values are based on observations from patients who followed their allocated treatments, in the same group. However, this hypothetical estimand may assume unrealistic performance (of patients and/or treating staff). A non-inferiority or equivalence trial should therefore consider both the treatment effect under this recommendation (i.e., as if no serious deviations occurred) and the effect under the ITT principle (i.e., treatment policy strategy on all randomised patients). Similar conclusions from both estimands ensure a robust interpretation.

It is advised to pre-specify what qualifies as a serious deviation that may affect the treatment/protocol adherence to be able to identify and collect the occurrence of these ICE. In addition, potential confounders should be pre-specified to collect relevant and complete data from both adherent and non-adherent participants during the trial. These data are needed to estimate the hypothetical treatment effect.

It should also be noted that PRO data collection beyond such deviations is still important to assess the ITT strategy (all randomised patients and treatment policy). In addition, when designing non-inferiority and/or equivalence trials, great care should be placed on ensuring adequate design and data quality control to prevent deviations as much as possible.

There are further points to consider in a non-inferiority setting and general existing guidelines on this topic are available.

ICE4_SAT

Statement: if the PROMs collected after an ICE are not relevant to the research question of interest (such as switching of treatment whilst assessing benefits), and furthermore, if the research question would consider a scenario in which the ICE would not occur (for example, evaluating patients who did not switch treatment), a hypothetical strategy may be followed, where the PRO values after the ICE are disregarded. To understand the impact of the model assumptions and to ensure the validity of analysis results, it is recommended to conduct sensitivity and supplementary analyses.

Explanation: for example, if a patient switches to another cancer treatment protocol the interest might not lie in the PRO scores after treatment discontinuation, but in what the PRO scores would have been had the patient continued with the initial protocol. In this situation the hypothetical scenario is: “what would have the treatment effect been had the patient not switched to another protocol and reacted like patients who remained on the initial protocol?” PRO data collected after the ICE may still contain information useful for supplementary analyses.

If PROMs taken after ICEs, such as discontinuing or switching of treatments, are not considered clinically interesting, a justification should be provided.

Furthermore, using a hypothetical strategy relies on modelling assumptions, which also need to be justified. The robustness of the results needs to be demonstrated through sensitivity analyses.

ICE5_SAT

Statement: if the PRO objective is treatment benefit and patients discontinue the treatment due to therapy-related events (for example to adverse events) or due to the disease itself (e.g., disease progression), and the data collection continues to a pre-specified defined time point or death (whichever comes first), it is recommended to include all available data from PRO assessments after the events into the analysis (treatment policy strategy).

Explanation: if the purpose of the intervention is to alleviate disease-related pain, and several patients discontinued the treatment because of its side effects, the PRO data for these patients should still be included in the analysis, even after they have terminated the treatment.

However, in some instances, the impact of treatment discontinuation or subsequent therapy on the PRO is not clinically relevant or collecting PROs after discontinuation might not be possible.

In these situations, a while-on-treatment strategy, or a composite strategy, or a hypothetical strategy can be considered. Not collecting PRO data after treatment discontinuation and using strategies other than the treatment policy strategy should be justified.

In the next sections on intercurrent events (i-iii) a condensed summary statement is listed for all intercurrent events except death to avoid repetitions (the original statements for each PRO variable of interest is listed in Appendix 1).

1d. i. Disease progression

ICEdisprog1_RCT- ICEdisprog5_RCT

Statement: when the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective) using [magnitude of PRO (change) score at time t // time-to-improvement/deterioration within a time frame // responder improvement/deterioration at time t] for a specific PRO domain: if a patient's disease progresses before time t , the main PRO analysis technique would be to use the PRO scores collected after disease progression [in the analysis at time t // to determine whether or not a PRO improvement/deterioration occurred]. Any deviations should be justified.

An overview of relevant ICEs should be reported and discussed to assess to what extent those ICEs could have impacted the PRO results. This may be supported by supplementary/sensitivity analyses.

Alternative ICE strategies might be considered as supplementary analyses to explore the robustness of outcomes and inform about the potential of bias.

Explanation: if the PRO objective is confirmatory, it is important to preserve the ITT and safety population by collecting PROs after disease progression (as long as the patient remains on study) and to use this data in the analysis. Disease progression can be an unavoidable event that may occur whilst treatments are compared. Therefore, PRO scores collected after the occurrence of such ICEs should be used in the main analysis. When defining the estimand, the feasibility and usefulness (in light of the research objectives) of collecting post-progression data should be taken into account. When such data is not collected, this should be justified. Sensitivity analyses may also be employed, where appropriate, to mitigate the effects of (partially) missing post-progression data.

An overview and reporting of relevant ICEs should help clarify to what extent those ICEs could have impacted the PRO results. Hence, reporting of ICEs is not requested for ICEs occurring after time t nor for ICEs that are disregarded for analyses. If the PRO objective is limited to estimating while-on-treatment scores, then data collection can be stopped when protocol treatment discontinuation coincides with progression.

Example: [specific to the PRO variable of interest].

1d. ii. Deviation from protocol-defined treatment

ICEprodev1_RCT- ICEprodev5_RCT

Statement: when the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective) using [magnitude of PRO (change) score at time t // time-to-improvement/deterioration within a time frame // responder improvement/deterioration at time t] for a specific PRO domain: if a patient deviated from protocol-defined treatment before [time t //a PRO improvement/deterioration occurred] (not causing treatment discontinuation), the main PRO analysis technique would be to use the PRO scores collected after deviation from protocol-defined treatment in the analysis [at time t // to determine whether or not a PRO improvement/deterioration occurred]. Any deviations should be justified.

An overview of relevant ICEs should be reported and discussed to assess to what extent those ICEs could have impacted the PRO results. This may be supported by supplementary/sensitivity analyses.

Explanation: if the PRO objective is confirmatory, it is important to preserve the ITT and safety population by collecting PROs after protocol deviation (as long as the patient remains on study) and to use this data in the analysis. Protocol deviations can be unavoidable events that may occur whilst treatments are compared. Therefore, PRO scores collected after the occurrence of such ICEs should be used in the main analysis, provided those deviations are not considered to impact the integrity of the clinical trial. When defining the estimand, the feasibility and usefulness (in light of the research objective) of collecting data after protocol deviations should be taken into account. When such data is not collected, this should be justified. Sensitivity analyses may also be employed, where appropriate, to mitigate the effects of (partially) missing post-protocol deviation data.

An overview and reporting of relevant ICEs should help clarify to what extent those ICEs could have impacted the PRO results. Hence, reporting of ICEs is not requested for ICEs occurring after time t nor for ICEs that are not expected to change the observation or interpretation of the PRO endpoint of interest. If the PRO objective is limited to estimating while-on-treatment scores, then data collection can be stopped when protocol treatment discontinuation coincides with protocol deviation.

Example: [specific to the PRO variable of interest].

1d. iii. Concomitant therapies allowed by the protocol

ICEconc1_RCT- ICEconc5_RCT

Statement: when the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective) using [magnitude of PRO (change) score at time t // time-to-improvement/deterioration within a time frame // responder improvement/deterioration at

time t] for a specific PRO domain: if a patient used concomitant therapies allowed by the protocol that could affect the interpretation of the PRO before [time t // a PRO improvement/deterioration], the main PRO analysis technique would be to use the PRO scores collected after patient started concomitant therapies in the analysis [at time t // to determine whether or not a PRO improvement/deterioration occurred]. Any deviations should be justified.

An overview of relevant ICEs should be reported and discussed to assess to what extent those ICEs could have impacted the PRO results. This may be supported by supplementary/sensitivity analyses.

Explanation: if the PRO objective is confirmatory, it is important to preserve the ITT and safety population by collecting PROs after the start of concomitant therapies (as long as the patient remains on study) and to use this data in the analysis. The start of concomitant therapies can be an unavoidable event that may occur whilst treatments are compared. Therefore, PRO scores collected after the occurrence of such ICEs should be used in the main analysis. When defining the estimand, the feasibility and usefulness (in light of the research objectives) of collecting data after the start of concomitant therapies should be taken into account. When such data is not collected, this should be justified. Sensitivity analyses may also be employed, where appropriate, to mitigate the effects of (partially) missing post-concomitant-therapy data.

An overview and reporting of relevant ICEs should help clarify to what extent those ICEs could have impacted the PRO results. Hence, reporting of ICEs is not requested for ICEs occurring after time t nor ICEs that are not expected to change the observation or interpretation of the PRO endpoint of interest. If the PRO objective is limited to estimating while-on-treatment scores, then data collection can be stopped when protocol treatment discontinuation coincides with the start of concomitant therapy.

Example: [specific to the PRO variable of interest].

1d. iv. Death

ICEdeath1_RCT

Statement: there are different strategies to address death as an ICE in RCTs, and the choice of strategy will have an impact on the treatment effect estimate and its interpretation. Whilst there is no uniform way of addressing death as an ICE, protocols should define and justify a clear strategy. This should be in line with the assumptions based on the pre-defined PRO objective of the clinical trial, and discussed with relevant stakeholder groups.

Explanation: the following four strategies for addressing death as an ICE in the analysis of PRO data can be considered. For each strategy, the underlying assumptions are provided in order to facilitate the selection of the most appropriate strategy or exclude strategies at odds with the context of the PRO objectives, disease setting and study constraints.

For instance, in a palliative care setting where the treatment aim is symptom relief and not intended to prolong life, a composite strategy for a patient-reported symptom relief endpoint could be excluded as it implies that death equals failure of symptom relief. Employing at least one alternative strategy as a sensitivity analysis is highly recommended in order to assess the impact of choosing a different strategy on conclusions regarding the treatment effect. The four strategies are:

1. A **hypothetical strategy** reflects the objective of assessing what the treatment effect would have been if no deaths had occurred during the time period of interest. A hypothetical strategy is most suitable when it is expected that death will not be related to the PRO variable of interest (e.g., PRO variable of interest is a specific non-lethal symptom), and the number of deaths will be limited throughout the observation period. The assumption that death is unrelated to the PRO value requires thorough consideration, as they are frequently related, and this assumption cannot be verified from the data itself. Under this strategy, it is assumed that there is no significant difference between patients who died and patients who are alive during the time period of interest. Hence, hypothetical PRO scores following death can be based on the PRO values of patients who are still alive. When the patients who died systematically differ from patients who are alive with respect to the PRO variable of interest, the hypothetical strategy will provide biased results. When implicitly imputing values for observations after the occurrence of death, the underlying imputation mechanism needs to be considered. In an analysis of the magnitude of PRO (change) score, conventional longitudinal models implicitly impute missing PRO values by modelling the intra-subject correlation. In a time-to-event analysis, a right-censoring approach can be adopted under the assumption that death is non-informative, i.e., the distribution of time to death does not depend on the parameters used to model the distribution of the event times. Alternatively, unobserved time points after death could be imputed based on observed data patterns.
2. A **composite strategy** integrates death with the PRO variable of interest by redefining the endpoint as a composite between the PRO variable and the event of death. A composite strategy assumes that a plausible value from the PRO domain of interest can be attributed to death and is, therefore, most suitable when the PRO variable of interest is assumed to be related to survival status (e.g. physical functioning). A composite PRO endpoint should not be dominated by death. If death dominates the composite PRO score, the estimated treatment effect and its interpretability would be reduced to a modified survival outcome. In the case of a time-to-deterioration analysis, patients who died are similarly categorised as those who evidenced a PRO deterioration. In this case, the composite endpoint should be named appropriately by including the term “survival”, such as “PRO deterioration-free survival”, to make it clear that death is included as part of the event definition. This composite endpoint can be interpreted as the amount of time a patient lives without experiencing a deterioration, including (but not restricted to) death. In an analysis of the magnitude of PRO (change) score, patients who died can be assigned a pre-specified score from the PRO measure, which should be aligned with the interpretation of the PRO domain of interest. In case of doubt about the pre-specified score, sensitivity analyses exploring alternative values are recommended. When the objective is to assess the proportion of responders with PRO score worsening at time t , a patient who died before or at time t will

be considered in the same category as a responder with a PRO score worsening at time t . When the objective is to assess the proportion of responders with PRO score improvement at time t , a patient who died before or at time t will be considered in the same category as a non-responder, with patients having a PRO score improvement at time t as responders. For time-to-improvement, it is not possible to use a composite strategy since the use of death as an event indicator to denote improvement cannot be considered a meaningful outcome.

3. A **while-alive strategy** is when the PRO objective refers to the treatment effect whilst the patient is alive. Under this strategy, the PRO variable is only considered relevant up to the time of death, rather than considering the same fixed endpoint for all patients at all time points. In an analysis of the magnitude of PRO (change) score at time t the while-alive strategy summarises the PRO scores at time t for patients who are still alive at time t . Consideration should be given to missing data and ICEs to determine whether the analyses are valid. The composition of alive patients in each treatment group might differ at each assessment time, and treatment arms might no longer be reflective of the randomised allocation. In such cases, only descriptive analyses are valid under this strategy. Results should be accompanied by observed mortality figures. However, if a randomised comparison is intended with all randomised patients, PRO scores might be utilised until time t to derive the magnitude of change i.e., either at time t or prior to death, whichever occurs first. This approach is feasible when there are sufficient PRO assessments prior to death, ensuring that a PRO change could be observed. Death events between treatment groups prior to time t need to be reported to assess the potential for bias. Of note, the endpoint definition and terminology need to be adjusted accordingly. For a time-to-event analysis, since death prevents the observation of the PRO variable of interest, a model for competing risks should be considered.
4. The **principal-stratum strategy** depicts a scenario where a subgroup of patients would be alive regardless of whether they were assigned to the treatment or control arm. The treatment effect is based on the stratum (i.e. a group of patients) who would have survived until a specific time point regardless of which treatment they were assigned to. It should be noted that it is not possible to correctly identify the strata with observed data since patients cannot be assigned to both the treatment and control arms. The treatment effect is referred to as the Survivor Average Causal Effect (SACE). The SACE can be estimated using a multiple imputation approach for the stratum membership of each patient. As the model to determine stratum membership is based on unverifiable assumptions, a sensitivity analysis should be done to assess the robustness of the conclusions to these assumptions.

Summary: in principle, the study protocol and analysis plan should pre-specify the ICE strategies employed to address death. At least one primary strategy should be determined *a priori* if the PRO objective is confirmatory. Supplementary analyses can consider alternative strategies according to the underlying study objectives, for instance, while-alive strategy as the primary analysis and composite strategy as supplementary analysis or vice versa. Such analyses are advised when the occurrence of death as an ICE is high or when several of the presented approaches are viable to implement the PRO objective.

ICEdeath2_SAT

Statement: using a composite strategy to address death for continuous or ordinal outcomes is generally not recommended, unless it is in a very specific situation where death is the natural ‘most severe’ state of the PRO. If this strategy is used, the value describing the state of death should be clinically justified, and it should be accompanied by the percentage of patients who died.

Explanation: using a pre-assigned value for death is appropriate only in very specific situations. The PRO value describing death should be pre-specified, clinically justified and easily interpretable. Certain PROMs already include death in their definition. An example is the EQ-5D scale, where the value 0 represents death. It is important to consider that if a significant number of patients die during treatment, death will dominate the combined outcome. Therefore, descriptive statistics related to death, such as survival probabilities, should be provided.

In scenarios such as palliative care study, where the goal is to alleviate pain until death, using a combined outcome which includes death is not suitable. This is because, in this case, death does not mean worsening of pain.

ICEdeath3_SAT

Statement: when death represents the natural ‘most severe’ state of the PRO and time-to-worsening of the PRO is considered, incorporating death into the endpoint (i.e., composite strategy) is a possibility. This can be done by using a combined outcome of “time-to-worsening or death”.

Explanation: one could use as the PRO endpoint ‘HRQoL deterioration-free survival’, defined as “time until worsening of HRQoL or death”. However, in palliative studies aiming to provide patients with effective pain relief until death, with the PRO measured by a self-reported pain score, using this combined outcome may not be suitable. This is because death does not necessarily indicate a worsening of pain if the PRO is measured by self-reported pain scores.

It should be noted that when many patients die during follow-up, death will dominate the combined outcome.

ICEdeath4_SAT

Statement: in a responder/non-responder analysis, when death can be seen as a treatment failure for the PRO endpoint, death could be defined as “non-responder” and incorporated into the responder definition (composite strategy).

Explanation: one could define non-responders at time t as patients with a certain relevant decrease in HRQoL at time t or who have died before time t .

1d. v. Treatment discontinuation or start of subsequent therapy

ICEdisc1_GEN*

Statement: when the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective), the treatment policy strategy (i.e. collect and use PRO data after treatment discontinuation) is the preferred strategy for incorporating treatment discontinuation (for reasons other than treatment completion) or the start of subsequent therapy as ICEs in the analysis. If justified, a hypothetical strategy, a composite strategy, or a while-on-treatment strategy can be considered.

Explanation RCT*: in cancer trials, some patients may discontinue treatment during the trial. There are different strategies on how to incorporate study treatment discontinuation (for reasons other than treatment completion) or the start of subsequent therapy in the analysis, and each strategy corresponds to a different interpretation of the treatment effect. Different stakeholders may be interested in different estimands and this should be discussed with the relevant stakeholder. Note that in a randomised cancer clinical trial, the study treatment refers to treatment that defines both the experimental and control arm(s).

In the **treatment policy strategy**, the ICE is considered to be part of the compared treatment effects under the ITT principle. It foresees the collection of PRO data after discontinuation. The schedule of data collection in all treatment groups should be consistent prior to and after discontinuation.

In cases where the impact of study treatment discontinuation and/or subsequent anti-cancer therapy is not of clinical interest (e.g., cross-over to the experimental therapy) or gathering PRO data after discontinuation is deemed unfeasible (e.g., in the case of advanced disease), the hypothetical strategy, composite strategy, or while-on-treatment strategy can be considered. The strategy selection should be based on the trial's objectives, with consideration given to possible scenarios after treatment discontinuation or the start of subsequent anti-cancer therapy.

If understanding the treatment effect is important, irrespective of whether the treatment is discontinued or followed by subsequent anti-cancer therapy, a treatment policy strategy is advised.

When a **hypothetical strategy** is applied in the context of study treatment discontinuation, the treatment effect is estimated in the hypothetical scenario where no treatment discontinuation and/or starting of a subsequent anti-cancer therapy occurs. This strategy could be applied if the impact of the treatment discontinuation and/or subsequent anti-cancer therapy is not of clinical interest. Such a scenario may be of interest to stakeholders, for instance, to patients who wish to understand what would have happened had they kept adhering to treatment. According to this strategy, it is assumed that there is no significant difference between patients who discontinued treatment and patients who remained on treatment.

By using a **composite strategy**, study treatment discontinuation and/or start of a subsequent anti-cancer therapy is included as a component of the PRO variable of interest. This approach

should only be used when combining PRO values with study treatment discontinuation is deemed feasible. Furthermore, an endpoint that is driven by study treatment discontinuation is not recommended. This is because a large proportion of patients who discontinued study treatment will severely impact the accuracy of the estimated treatment effect and make the results harder to interpret.

With a **while-on-treatment strategy**, the approach involves descriptively summarising the data by reporting the magnitude of PRO (change) score or proportion of responders in both arms in the group of patients who are receiving the study treatment at each assessment time.

When assessing trends over time, consideration should be given to missing data and ICEs. The group of patients included in the summary measure may vary at each assessment time. These summary measures should be supplemented with numbers on treatment discontinuation. Therefore, consideration should be given to missing data and ICEs to assess the validity of the analysis. Treatment groups might no longer be comparable if only reporting data on patients who are receiving treatment. Therefore, consideration should be given to missing data and ICE to assess the validity of the comparative analysis.

In principle, the study protocol and analysis plan should outline in advance the strategy to handle ICEs. One primary strategy should be determined at least if the PRO objective is confirmatory. Supplementary analyses can consider alternative strategies based on the objectives of the study, for instance, adopting a treatment policy strategy as the primary approach and using a hypothetical strategy for supplementary analysis.

An overview of the relevant reasons for treatment discontinuation should be reported and discussed to assess the extent to which the ICE could have impacted the PRO results.

For descriptive purposes, for instance, to explore patient experience on tolerability other strategies than treatment policy could be relevant, such as while-on-treatment strategy.

Explanation SAT*: in cancer trials, some patients may discontinue treatment during the trial. There are different strategies on how to incorporate study treatment discontinuation (for reasons other than treatment completion) or the start of subsequent therapy in the analysis, and each strategy corresponds to a different interpretation of the treatment effect. Different stakeholders may be interested in different estimands and this should be discussed with the relevant stakeholder.

In the **treatment policy strategy**, the ICE is considered to be part of the compared treatment effects under the ITT principle. It foresees the collection of PRO data after discontinuation. The schedule of data collection in all treatment groups should be consistent prior to and after discontinuation.

In cases where the impact of study treatment discontinuation and/or subsequent anti-cancer therapy is not of clinical interest (e.g., cross-over to the experimental therapy) or gathering PRO data after discontinuation is deemed unfeasible (e.g., in the case of advanced disease), the hypothetical strategy, composite strategy, or while-on-treatment strategy can be considered. The strategy selection should be based on the trial's objectives, with consideration

given to possible scenarios after treatment discontinuation or the start of subsequent anti-cancer therapy.

If understanding the treatment effect is important, irrespective of whether the treatment is discontinued or followed by subsequent anti-cancer therapy, a treatment policy strategy is advised.

When a **hypothetical strategy** is applied in the context of study treatment discontinuation, the treatment effect is estimated in the hypothetical scenario where no treatment discontinuation and/or starting of a subsequent anti-cancer therapy occurs. This strategy could be applied if the impact of the treatment discontinuation and/or subsequent anti-cancer therapy is not of clinical interest. Such a scenario may be of interest to stakeholders, for instance, to patients who wish to understand what would have happened had they kept adhering to treatment. According to this strategy, it is assumed that there is no significant difference between patients who discontinued treatment and patients who remained on treatment.

By using a **composite strategy**, study treatment discontinuation and/or start of a subsequent anti-cancer therapy is included as a component of the PRO variable of interest. This approach should only be used when combining PRO values with study treatment discontinuation is deemed feasible. Furthermore, an endpoint that is driven by study treatment discontinuation is not recommended. This is because a large proportion of patients who discontinued study treatment will severely impact the accuracy of the estimated treatment effect and make the results harder to interpret.

With a **while-on-treatment strategy**, the approach involves descriptively summarising the data by reporting the magnitude of PRO (change) score or proportion of responders in the group of patients who are receiving the study treatment at each assessment time.

When assessing trends over time, consideration should be given to missing data and ICEs. The group of patients included in the summary measure may vary at each assessment time. These summary measures should be supplemented with numbers on treatment discontinuation. Therefore, consideration should be given to missing data and ICEs to assess the validity of the analysis.

In principle, the study protocol and analysis plan should outline in advance the strategy to handle ICEs. One primary strategy should be determined at least if the PRO objective is confirmatory. Supplementary analyses can consider alternative strategies based on the objectives of the study, for instance, adopting a treatment policy strategy as the primary approach and using a hypothetical strategy for supplementary analysis.

An overview of the relevant reasons for treatment discontinuation should be reported and discussed to assess the extent to which the ICE could have impacted the PRO results.

For descriptive purposes, for instance, to explore patient experience on tolerability other strategies than treatment policy could be relevant, such as while-on-treatment strategy.

1e. Population-level summary

Psum1_GEN*

Statement RCT*: when performing an analysis of the descriptive magnitude of (PRO) change score, the outcomes can be reported as the magnitude of change for each arm at a predefined assessment point(s) along with a measure of variability. When evaluating trends over time, consideration should be given to underlying assumptions about missing data and ICEs strategies.

Explanation RCT*: the goal of a descriptive analysis is to summarise the observed data. The data can be summarised by means, medians or the magnitude of PRO (change) score in both arms at each assessment time point. Measures of variation such as standard deviations, percentiles, interquartile ranges and interval estimates such as confidence intervals should also be reported.

The summary measure reported for each assessment time can cover a different subset of patients; for example, it might only include data from patients who are alive. Therefore apparent time trends in repeated cross-sectional outcomes must be interpreted with care. The pattern of differences over time, based on these summary statistics, is subject to selection bias due to attrition. This bias increases over time as patients drop out of the analysis due to lost-to-follow-up. Furthermore, ICEs or varying missing data over time (for instance, lower completion rates with increased follow-up) may further complicate the interpretation.

Furthermore, because treatment arms might no longer be comparable, missing data and ICEs must be taken into account before comparing treatment arms (for example, by reporting the difference in magnitude of change between arms).

Statement SAT*: when performing an analysis of the descriptive magnitude of (PRO) change score, the outcomes can be reported as the magnitude of change at a predefined assessment point(s) along with a measure of variability. When evaluating trends over time, consideration should be given to underlying assumptions about missing data and ICEs strategies.

Explanation SAT*: the goal of a descriptive analysis is to summarise the observed data. The data can be summarised by means, medians or the magnitude of PRO (change) score at each assessment time point. Measures of variation such as standard deviations, percentiles, interquartile ranges and interval estimates such as confidence intervals should also be reported.

The summary measure reported for each assessment time can cover a different subset of patients; for example, it might only include data from patients who are alive. Therefore apparent time trends in repeated cross-sectional outcomes must be interpreted with care. The pattern of differences over time, based on these summary statistics, is subject to selection bias due to attrition. This bias increases over time as patients drop out of the analysis due to lost-to-follow-up. Furthermore, ICEs or varying missing data over time (for instance, lower completion rates with increased follow-up) may further complicate the interpretation.

Psum2_GEN*

Statement RCT*: when performing a descriptive responder analysis, the outcome can be reported as the proportion of responders for each arm at pre-specified assessment point(s), along with a measure of variability.

When evaluating trends over time, consideration should be given to underlying assumptions about missing data and ICEs strategies.

Explanation RCT*: the goal of a descriptive analysis is to summarise the observed data. When performing a responder analysis, the data can be summarised by the proportion of responders in both arms. Measures of variation such as standard deviations, percentiles, interquartile ranges and interval estimates such as confidence intervals should be reported.

The proportion of responders reported for each assessment time can cover a different subset of patients– for example, it might only include data from patients who are alive. Therefore, apparent time trends in repeated cross-sectional outcomes must be interpreted with care.

The pattern of differences over time, based on these summary statistics, is subject to selection bias due to attrition. This bias increases over time as patients drop out of the analysis due to lost-to-follow-up. Furthermore, ICEs or varying missing data over time (for instance, lower completion rates with increased follow-up) may further complicate the interpretation.

Statement SAT*: when performing a descriptive responder analysis, the outcome can be reported as the proportion of responders at pre-specified assessment point(s), along with a measure of variability.

When evaluating trends over time, consideration should be given to underlying assumptions about missing data and ICEs strategies.

Explanation SAT*: the goal of a descriptive analysis is to summarise the observed data. When performing a responder analysis, the data can be summarised by the proportion of responders. Measures of variation such as standard deviations, percentiles, interquartile ranges and interval estimates such as confidence intervals should be reported.

The proportion of responders reported for each assessment time can cover a different subset of patients– for example, it might only include data from patients who are alive. Therefore, apparent time trends in repeated cross-sectional outcomes must be interpreted with care.

The pattern of differences over time, based on these summary statistics, is subject to selection bias due to attrition. This bias increases over time as patients drop out of the analysis due to lost-to-follow-up. Furthermore, ICEs or varying missing data over time (for instance, lower completion rates with increased follow-up) may further complicate the interpretation.

Psum3_GEN

Statement: when performing a descriptive time-to-event analysis, the data can be summarised by the median or by another relevant time-to-event percentile, and by the probability of experiencing an event at a specific time point. Adding a measure of variability is recommended.

Explanation: the goal of a descriptive analysis is to summarise the observed data. When a time-to-event analysis is performed, the data can be summarised by the median (or other relevant percentile) or the event probability at a specific time point. Measures of variation such as standard deviations, percentiles, interquartile ranges and interval estimates such as confidence intervals should be reported.

For descriptive purposes, other population-level summaries might be of interest such as restricted mean survival times.

It is advised to calculate the summary measures using a method that takes into account the interval-censored nature of the data. However, the median event time is not always directly available in cases of interval-censored data. To estimate the median event time, one can either take the upper bound of the interval or make a reasonable assumption regarding the distribution of the mass in the region of support (e.g., linear or exponential).

If fewer than 50% of the patients had an event during the trial, it is not possible to calculate the median time-to-event based on the data. In this case, another percentile or the event-free rate at a specific time point can be used to summarise the data. Providing the Kaplan-Meier curve and interval estimates can make the results easier to interpret.

Psum4_RCT

Statement: when performing an analysis of the magnitude of PRO (change) score and the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective), the scale of the endpoint measure should be considered. For PRO scales that are considered as continuous, the comparative treatment effect should be estimated by the difference in the mean magnitude of change from baseline between treatment arms at the time point of interest.

Explanation: the comparison between the two treatment arms is based on the population-level summary, which should take into account the PRO measurement scale. When performing an analysis of the magnitude of PRO (change) score if the PRO objective is confirmatory, the population-level summary should be the difference in the mean magnitude of change from baseline between treatment arms at the time point of interest. Underlying models should adjust for the baseline PRO score.

Non-parametric analysis (e.g. median magnitude of change from baseline) can be considered an alternative option to summarise the difference, especially if there are outliers or extreme values in the data or if the distribution of the data is heavily skewed.

Psum5_RCT

Statement: when performing a responder analysis and the goal of the PRO objective is to draw conclusions about the clinical benefit (confirmatory objective), the comparative treatment effect should be estimated by the difference in the proportion of responders between treatment arms, the odds ratio or the risk ratio at the time point of interest.

Explanation: the comparison between the two treatment arms is based on the population-level summary. When performing a responder analysis, if the PRO objective is confirmatory, the population-level summary should be the difference in the proportion of responders between treatment arms at the time point of interest, the odds ratio (the ratio of the odds of being a responder in both groups) or the risk ratio (the ratio of the proportion of responders in both groups) at the time point of interest.

As a supplementary analysis, the absolute risk difference (the difference in the proportion of responders between both groups) at the time point of interest can be derived. This may be the preferred metric in case there is a reasonable expectation of a 0% response rate in one of the treatment groups.

Psum6_RCT

Statement: when performing a time-to-event analysis and the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective), the comparative treatment effect should be estimated by the hazard ratio in cases when proportional hazards can be assumed, or the difference in the probability of deterioration or improvement at a specific time point between treatment arms.

Explanation: the comparison between the two treatment arms is based on the population-level summary. When performing a time-to-event analysis and the PRO objective is confirmatory, the population-level summary should reflect the whole time period on which events such as PRO deterioration or improvement could be observed, rather than a single point estimate such as the median. Both an absolute risk measure (e.g., the difference in risk probability) and a relative measure (e.g., hazard ratio, ratio of survival probabilities) can be used. In the case of proportional hazards (assumption should be checked) the hazard ratio should be used as a relevant population-level summary, similar to other endpoints such as progression-free survival or overall survival. Otherwise, the difference in the probability of deterioration or improvement at a specific time point between both treatment arms should be used. Under certain conditions i.e., that the empirical survival curve follows an exponential distribution with a sufficient number of events and a sufficient number of assessments prior to the expected median, the difference in median time-to-event can be considered as a relevant population summary measure for analysis.

2. PRO score interpretation thresholds

2a. Application of PRO score interpretation thresholds

CMCimp1_GEN

Statement: it should be clearly specified whether the PRO score interpretation thresholds are applied to scores at the patient-level (i.e. within-patient change) or at the group-level (i.e. within-group change, between-group difference, or between-group difference in change).

Explanation: PRO score interpretation thresholds have been defined for different types of data and analyses. Most importantly such thresholds have been established for interpreting patient-level data, such as score changes of an individual patient, and for group-level data, such as mean score change in a treatment group. Similar approaches have been USED to analyse mean differences between two treatment groups at a specific time point. Two examples are provided below.

- Example 1: threshold for meaningful within-patient change: patient X has a pain score of 12 points at the start of treatment and of 5 points at the end of treatment. A patient-level threshold helps to decide whether this change of 7 points between the start and end of treatment is meaningful or not.
- Example 2: threshold for meaningful within-group change: patients in treatment group A have a mean pain score of 5.6 points at the start of the treatment and of 7.1 points six months later. A group-level threshold helps to decide whether this mean change of 1.5 points between the two time points is meaningful or not.

Different PRO score interpretation thresholds are required for patient-level data and for group-level data. To avoid confusion, it is important to clearly specify if a threshold is used for evaluating data from individual patients, or from groups of patients. The applicability of thresholds at a patient- or group-level depends on the methodology used for defining these thresholds.

CMCimp2_GEN

Statement: to interpret the results of an analysis of magnitude of change, PRO score interpretation thresholds used should be applicable at the group-level (e.g., to interpret the between-group difference in change).

Explanation: this statement refers to group-level analyses of magnitude of change/difference and emphasises the need for group-level thresholds in this context, i.e., threshold values that can be used to interpret cross-sectional between-group differences, within-group changes over time, or between-group differences in change over time.

Example: if in a single arm trial the mean score of a PRO scale improves by 2.5 points between two time points, the meaningfulness of this improvement should be interpreted based on

thresholds for interpretation of group-level changes/differences, and not on thresholds developed for interpreting change in individual patients (i.e., patient-level thresholds).

If in this trial population a 5-point change represents a meaningful within-patient change, an observed mean change of 2.5 points may reflect a sample in which 50% of patients experienced no improvement (0 point change), while 50% of patients experienced a meaningful improvement (5 point change). While 50% of patients with a meaningful improvement may suggest a meaningful clinical benefit, this may not be considered meaningful if a patient-level threshold was applied for interpretation of the group-level mean change. Thus, a threshold for within-group changes must be established separately and applied to interpreting the meaningfulness of the 2.5 change experienced by the group on average.

A treatment group is usually composed of patients who experience a meaningful improvement over time, patients who experience a change over time that is not meaningful, no change over time, or even patients who experience a deterioration. Mean change is therefore a combination of these different changes in individual patients. Even if there is a substantial percentage of patients who experiences a meaningful improvement, the overall mean change in the group may not reach the level of change that would be meaningful to an individual patient. Applying a threshold for meaningful within-patient change when interpreting mean change at the group-level may result in an under appreciation of treatment effects.

CMCimp3_GEN

Statement: for responder analysis (e.g., percentage of patients showing meaningful improvement), thresholds for meaningful within-patient change should be used to define the threshold for improvement or deterioration.

Explanation: to calculate the percentage of treatment responders in a responder analysis, a threshold for meaningful within-patient change is required. This threshold helps to determine whether each individual patient responded to treatment.

If a decrease of 10 points or more on a pain scale of an individual patient (lower score means lower levels of pain) reflects a meaningful improvement, a responder analysis counts the number of patients experiencing such a change, i.e., number of patients who responded to treatment.

At the group-level, the number of patients reporting an improvement of 10 or more points in each treatment group can be compared between different treatment groups.

CMCimp4_GEN

Statement: for time-to-event analysis, such as time-to-deterioration of PRO scores, thresholds for meaningful within-patient change should be used to define the event of interest, whether that is an improvement or a deterioration of a PRO domain.

Explanation: the analysis of time-to-event focuses on the time until the event of interest (meaningful improvement or deterioration) occurs. To define this event, a threshold for

meaningful within-patient change is required in order to determine when the event of interest occurred for each individual patient.

If an increase of 10 points or more on a pain scale of an individual patient reflects a meaningful deterioration, this deterioration is used to define the event in a time-to-event analysis. For each patient the duration until such a deterioration occurs is recorded.

At the group-level, the average time to deterioration for each treatment group can be calculated to compare different treatment groups.

CMCimp5_GEN

Statement: if the statistical analysis of trial endpoints is based on magnitude of change (and p -values for hypothesis testing are derived from such an analysis), descriptive statistics such as percentages of treatment responders—defined by thresholds for meaningful within-patient change—may provide useful, accompanying information. Such information may be interpreted more easily by patients and health care professionals.

Explanation: the results from statistical methods used to analyse PRO data from clinical trials can be presented in different ways. Analyses focused on magnitude of change, such as mean change, have, for example, the advantage of providing information about differences between treatment groups and trajectories of symptoms and other health domains over time. However, these results may be harder to interpret compared to analyses presenting the percentage of patients responding to a treatment, such as the percentage of patients who improved, deteriorated or remained stable while on treatment.

Patients and health care professionals may want to use PRO data when guiding treatment decisions. Responder analyses, which present the percentage of patients responding to a treatment, may support interpretation of PRO results and thus may assist with decision-making.

CMCimp6_GEN

Statement: the use of PRO score interpretation thresholds should account for possible differences between improvement and deterioration of PRO scores. If no such distinction is made, the reasons should be clearly outlined.

Explanation: when evaluating change in PRO scores over time, whether at patient- or group-level, PRO score interpretation thresholds may differ. This difference depends on whether a PRO domain, like a symptom, is improving or deteriorating. For example, a 2 point deterioration on a pain scale may be regarded as a meaningful deterioration by patients, while a 2 point improvement may not be large enough to be regarded as meaningful. Such differences at the patient-level may impact on PRO score interpretation thresholds at the group-level.

Example: in some cases the threshold for patient-level deterioration is larger than for improvement. In such cases, depending on how the thresholds are defined, this may also

affect improvement and deterioration thresholds for group-level PRO score interpretation thresholds.

Since there may be differences between thresholds for meaningful improvement and deterioration, different thresholds should be used, unless there is evidence for a specific symptom or other health domain indicating that the same threshold is applicable.

CMCimp7_GEN

Statement: for confirmatory analyses of PRO endpoints, the PRO score interpretation thresholds for patient- or group-level data should be stated *a priori* in the study protocol (or in a separate analysis plan). As a less desirable option, in the absence of a well-established threshold, the PRO score interpretation threshold may be established after data collection begins. This may be done by a person unaware of treatment allocation, following a predefined methodological procedure outlined in the study protocol or in a separate analysis plan.

Explanation: in the context of a confirmatory analysis, the PRO score interpretation threshold should be defined prior to database lock to avoid the use of a data-driven threshold that best shows the desired treatment effect (an approach akin to “cherry picking”). For specific cancer populations and PRO instruments, PRO score interpretation thresholds may not be available at the study’s outset. Therefore, a less desirable alternative is to define the PRO score interpretation thresholds based on trial data. This analysis may be conducted by an independent, blinded person (i.e., a person unaware of treatment allocation) and rely on data from the pooled sample, to not reveal treatment allocation. The methodological approach (such as selecting an anchor definition) and the statistical method (like regression analysis) should be well-justified and predefined in an analysis plan.

For a magnitude of change PRO endpoint, group-level PRO score interpretation thresholds need to be defined in the study protocol prior to start of the study, if required for sample size calculations.

CMCimp8_GEN

Statement: to investigate the robustness of trial results from a responder analysis or time-to-event analysis, sensitivity analyses should be conducted to evaluate the results using different patient-level thresholds for meaningful within-patient change.

Explanation: to define the “response” in a responder analysis or of “deterioration/improvement” in a time-to-event analysis, a threshold value for meaningful within-patient change is required. However, a single, well-established and “correct” threshold value for defining meaningful within-patient change is usually not available. It is therefore informative to conduct the statistical analysis of the trial data not only using a single *a priori* defined threshold value, but to investigate additional threshold values. This helps assess the robustness of results under different values of meaningful within-patient change thresholds. Depending on the available evidence supporting the validity of a given threshold value, the sensitivity analysis for the trial results may be more or less extensive.

In a confirmatory setting, threshold values for meaningful within-patient change used in the main analysis should be stated *a priori* in the study protocol (or in a separate analysis plan), while thresholds used in the sensitivity analysis may be defined either *a priori* or thereafter.

Sensitivity analyses provide information on whether or not trial results for a PRO domain depend on the choice of a specific threshold value. Our guidance aligns with the estimand framework which recommends evaluating assumptions underlying a particular estimand through a sensitivity analysis. The threshold value for meaningful within-patient change used in a responder analysis or time-to-event analysis represents one such assumption.

For example, in a trial comparing the percentage of patients with pain reduction in treatment arm A and B, the main analysis may consider patients as responders if the pain reduction (on a 0–100 scale) is at least 10 points on a pain scale. The percentage of responders is then calculated based on this threshold value.

To investigate the robustness of the results, sensitivity analyses using different thresholds should be carried out to investigate if differences between treatment arms are (also) observed when defining response (pain reduction) as a smaller or larger score changes, such as 5, 15, or 20 points on the pain scale.

CMCimp10_GEN

Statement: the specific PRO scales for which the patient- or group-level PRO score interpretation thresholds are being applied should be clearly stated.

Explanation: to allow for a better understanding of how a PRO score interpretation threshold is applied in the context of a clinical trial and to reduce ambiguity, it should be clearly stated to which PRO scale a certain threshold value is applied.

Example: a trial protocol may state that a within-patient change of at least 10 points on the fatigue scale of questionnaire X between baseline and a subsequent assessment time point is considered a meaningful within-patient change (response) in the context of a responder analysis.

CMCimp11_GEN

Statement: it should be clearly stated if the PRO score interpretation thresholds are applied to patient- or group-level data.

Explanation: to facilitate understanding of how an interpretation threshold value for PRO scores is applied in the context of a clinical trial and to reduce ambiguity, it should be clearly stated if a threshold is applied at the patient-level (to define the response in a responder analysis, or the event in an analysis of time to deterioration/improvement), or at the group-level (for instance to interpret, the mean difference between two treatment arms).

Example: a trial protocol may state that a *within-patient change* of at least 10 points on the fatigue scale of questionnaire X between baseline and a subsequent assessment time point

is considered as a response (meaningful within-patient change) in the context of a responder analysis.

CMCimp12_GEN

Statement: it should be clearly stated if the PRO score interpretation thresholds are used to interpret meaningful within-patient change, meaningful within-group change, meaningful between-group difference, or meaningful between-group difference in change.

Explanation: to better understand how a PRO score interpretation threshold is applied in the context of a clinical trial and to reduce ambiguity, it should be clearly stated if a threshold is applied to change over time, cross-sectional differences, or group differences in change over time.

Example: a trial protocol may state that a mean difference between treatment arms of at least 5 points in mean change from baseline to 3-month follow-up is considered to be meaningful. This indicates that, if treatment arm A has a mean improvement of +2 points over time and treatment arm B has an improvement of +7 points, the difference in improvement may be considered a meaningful between-group difference in change.

CMCimp13_GEN

Statement: a rationale for choosing a certain interpretation threshold for patient- and group-level PRO score should be provided.

Explanation: The selection of the interpretation threshold for the PRO score can significantly impact the trial design, the results and their interpretation. Therefore, it is important to explain why a specific threshold value has been chosen. This explanation will allow for a better understanding and appraisal of how appropriate the threshold value is to the trial population.

Example: a trial protocol may state that “a within-patient change of at least 10 points on the fatigue scale of questionnaire X between baseline and a subsequent assessment time point is considered a meaningful within-patient change (response) in the context of a responder analysis”. The threshold of 10 points has been selected to maintain consistency of the responder’s definition for fatigue with previous trials evaluating the same intervention in the same patient population. Another reason to select this threshold is that a change of 10 points has been shown to correspond to a magnitude of change that can be perceived by a patient [*add reference to study demonstrating this association*].

CMCimp14_GEN

Statement: a justification should be provided for the applicability of a PRO score interpretation threshold selected for use in the trial population.

Explanation: the trial population in which a threshold is used for analysis and interpretation may not be identical with the population in which the threshold has been established. Additionally, there may be variations in interpretation thresholds of PRO scores across

patient populations, such as terms of diagnosis, treatment types, or treatment phases. When applying a PRO score interpretation threshold (e.g., for the definition of responders in a responder analysis, or for interpretation of mean differences between treatment arms), it is important to provide a justification explaining why the selected threshold is applicable in the trial population.

Example: a PRO score interpretation threshold for a fatigue scale may be applicable in a trial population, if this threshold has been established in a patient population with similar clinical characteristics receiving the same treatment.

Alternatively, a threshold may be used in a trial population with differing characteristics, if there is evidence of limited variability of threshold values for the fatigue scale across different patient populations.

CMCimp9_RCT

Statement: for non-inferiority/equivalence endpoints with an analysis of magnitude of change of group-level PRO scores (like mean change): group-level PRO score interpretation thresholds support the specification of non-inferiority/equivalence margins. However, these thresholds may not necessarily equal such margins.

Explanation: currently, there is general information available from the literature and guidance documents on how to define non-inferiority/equivalence margins for trial endpoints (FDA, 2016; Piaggio et al., 2012; EMA, 2005). This general information also applies to PRO endpoints and may help to define such margins. The margin is the difference between two treatments that is considered acceptable and justifies conclusions such as a treatment not being worse (non-inferior) than a comparator treatment or that the treatments are similar (equivalent).

Specification of such a margin may depend, for instance, on the perspective of a stakeholder, the type of endpoint (clinical benefit versus safety), and the clinical benefit of co-endpoints. PRO score interpretation thresholds that are applicable to group-level data (e.g., for interpretation of mean differences between treatment arms) may provide helpful information for specifying the size of the non-inferiority/equivalence margin in a specific study. It is important to note that, for endpoints with a responder analysis or an analysis of time-to-improvement/deterioration, non-inferiority/equivalence margins are specified for the percentage of responders or the time variable respectively, rather than for the PRO scores directly.

For example, the non-inferiority margin for the mean difference between the experimental and the control arm for a fatigue scale may be 10 points (on a 0–100 scale). That is to say, if a mean fatigue score is up to 10 points worse in the experimental arm, it is considered to indicate non-inferiority of the experimental treatment regarding the fatigue score.

This margin may also be chosen if the group-level PRO score interpretation threshold falls below that value, provided that appropriate justification is provided. Such justification may be related to clinical benefit shown for co-endpoints, the type of comparator (e.g., placebo, or active control), and the variability of clinical benefit or safety among established treatments.

2b. Selection of PRO score interpretation thresholds

CMCselc1_GEN

Statement: the use of patient- and group-level PRO score interpretation thresholds established using anchor-based methods is recommended over the use of distribution-based methods.

Explanation: unlike distribution-based methods that rely, for example, on the standard deviation of a PRO scale for determining PRO score interpretation thresholds, anchor-based methods link PRO scales to external anchors. Most commonly, the change in an external anchor is linked statistically to the change on the PRO scale to determine an anchor-based thresholds. External anchors (e.g., clinician-reported performance status) are conceptually closer to meaningfulness than the standard deviation of a PRO scale or effect size statistics. Distribution-based thresholds may provide supportive information.

Example: an anchor-based PRO score interpretation threshold for a physical function scale may reflect change in a patients' performance status or change reported by the patient themselves. This is considered to be more meaningful than, for instance, a change equivalent to 0.5 standard deviations of a reference population.

CMCselc2_GEN

Statement: anchor-based PRO score interpretation thresholds should be based on anchors that are meaningful to patients. Any deviations should be justified.

Explanation: the most common method to define anchor-based PRO score interpretation thresholds is based on the association between changes in PRO scores and changes in another clinical external criterion measure (i.e., the external anchor). Two examples are outlined below.

- Example A: to define the threshold for meaningful within-patient change, compare the difference in physical function as reported by a patient at two time points with the amount of perceived change in physical function (improved, worsened, or stable), as reported by the patient at a later time point using a single-item patient global rating of change.
- Example B: to define the threshold for meaningful within-patient change, patient-reported changes in overall health status is compared with changes in the amount of time a patient is bed-bound during the day since the start of the trial.

When selecting the external criterion for defining the thresholds for meaningful within-patient change, it is important to select a criterion that patients themselves consider meaningful. This is to ensure that the resulting threshold is also meaningful to the patients. PRO measures are intended to reflect the patients' perspective on their health status. As a result, the definition of thresholds for meaningful within-patient change should also follow a patient-centred approach and rely on criteria that are meaningful to patients, unless there are specific reasons not to use this approach. Such reasons need to be clearly outlined.

CMCselc3_GEN

Statement: external anchors for establishing patient- and group-level interpretation thresholds for PRO scores should be conceptually associated to and statistically correlated with the PRO domain of interest.

Explanation: when selecting and PRO score interpretation threshold from the literature, this selection should be guided by existing quality criteria for such thresholds. One commonly adopted option is to provide evidence that the external anchor used to establish the threshold, shows a sufficient correlation and a conceptual link to the PRO scale.

Example: clinician-reported performance status has been shown to be highly correlated with PRO scales for physical function. In addition, the measured constructs are overlapping or similar, as both assess a patient's capability to perform a variety of physical activities. Therefore, performance status may fulfil the criteria of statistical and conceptual association.

CMCselc4_GEN

Statement: if establishing an interpretation threshold for a patient- or group-level PRO score for changes in PRO scores over time using an external patient-reported anchor, the period of time captured by the PRO-based endpoint and the anchor measure should be aligned.

Explanation: anchor-based methods for establishing interpretation thresholds for PRO scores frequently rely on anchors that assess a patient's perceived change in a symptom between two time points. Using statistical methods, the change in PRO score between these two time points is then linked to the patient's perceived change rating to establish the PRO score interpretation thresholds. This is typically done to determine a threshold that distinguishes between changes in PRO scores that a patient can perceive and changes they cannot perceive. When selecting such a threshold from the literature, the selection should be guided by specific quality criteria. One such criterion is the alignment of the recall period and the assessment schedule of the PRO scale and the anchor.

Example: a fatigue scale is administered to patients at two time points with a one month interval in between. At the second assessment, patients can also answer an anchor question on their perceived change in fatigue since the first assessment (with response categories such as "worse", "unchanged", and "better"). This anchor question should specifically refer to change in fatigue during the last month to allow to meaningfully link the change on the fatigue scale and the patients' change rating.

When the computation of the PRO score is based on multiple time points, it is important to align the recall period of the anchor with the entire window of time represented by the score. A common circumstance that would warrant this alignment would be in the case of a daily diary PRO measure that is analysed over a period of several days or weeks. For example, if the PRO measure is administered via daily diary but the resulting score is computed over a one week period, the corresponding anchor, if administered at the end of the week, should have a recall period of one week.

CMCselc5_GEN

Statement: distribution-based estimates of patient- and group-level PRO score interpretation thresholds do not necessarily reflect meaningful differences or changes.

Explanation: distribution-based estimates of PRO score interpretation thresholds are commonly determined by the standard deviation of the PRO scale in a reference population or by other statistics for effect size. Such statistics do not necessarily reflect changes or differences of a meaningful magnitude.

Example: if the standard deviation of a fatigue scale is 20 points in a reference population, a common method for determining a distribution-based estimate is to calculate 0.5 standard deviations (10 points in this example). Such methods for calculating distribution-based estimates have no inherent link to the concept of meaningfulness.

Distribution-based estimates of PRO score interpretation thresholds are not linked to meaningfulness but reflect the distribution characteristics of a PRO scale. While standard deviation and effect size statistics may provide supportive information for PRO score interpretation thresholds, it is important to note that the widely used standard error of measurement of a PRO score is conceptually different, as it relates to the uncertainty of the estimate of the PRO score. The latter indicates random variation of PRO scores from repeated assessments of a person with the same instrument around the person's "true" PRO score.

CMCselc6_GEN

Statement: when a PRO score interpretation threshold is considered the lower bound for meaningful score changes or differences the term "minimum" can only be used if justified.

Explanation: Variation in the methodology on which PRO score interpretation thresholds are based results in thresholds of different magnitudes.

The choice of which threshold to use may depend on the stakeholder, characteristics of the patient population, and context of the interpretation (e.g., regulatory decision, reimbursement decision, doctor-patient treatment continuation decision).

The choice of the threshold can have a major impact on (the interpretation of) trial results. For example, larger thresholds result in lower percentages of patients considered to experience improvement/deterioration, a longer period until improvement/deterioration, or a different appraisal of mean differences between trial arms.

Depending on the context, it may be appropriate to use thresholds of different magnitudes. However, if the interpretation of trial data intends to rely on a "minimum" threshold, this should be clearly stated, and justification should be provided for why this is a "minimum" threshold.

Example: a threshold for meaningful within-patient change can be established by determining the magnitude of change that is just large enough to be noticed by patients. This may be considered a “minimum” meaningful within-patient change in certain settings.

CMCselc7_GEN

Statement: if thresholds for meaningful within-patient change are used, the PRO scale should provide sufficient granularity to measure patient-level changes of such magnitude.

Explanation: PRO scales frequently present scores on seemingly continuous metrics that range e.g., from 0 to 100 points, even if the number of values an individual patient can actually obtain is limited. The number of different scores on a PRO scale that a patient can obtain, i.e., the granularity, directly determines the smallest possible change that can be measured by a scale. Generally, a larger number of items in a PRO scale provides a larger number of (different) obtainable scores and thus decreases the distance between obtainable values, if all scores are given on the same metric (e.g., 0 to 100). The smallest measurable change for an individual patient is determined by the distance between obtainable scores. A PRO scale should provide sufficient granularity to allow the measurement of changes within a patient that are of a meaningful magnitude. In essence, PRO scales should allow measurement of a meaningful within-patient change.

Example: patient X answers a fatigue question with four possible response categories: not at all, a little, quite a bit, very much. If a fatigue scale consists only of this question the patient obtains one of four possible values. On a 0 to 100 metric, the following values are assigned to the response categories: 0 (not at all), 33 (a little), 67 (quite a bit), and 100 (very much). As a result, the measurable change between two time points for this patient cannot be less than 33 points. If the response categories are coded zero, one, two, and three, the smallest measurable change is one. Irrespective of the coding of the response categories, a change that is smaller than the distance between two adjacent categories might be meaningful but cannot be detected by a single-item scale with its limited granularity. The number of possible values of a scale increases with the number of questions comprising this scale and with the number of response options. For computer-adaptive questionnaires and static short-forms (fixed length questionnaires) created from item banks, the number of items should be set to provide sufficient granularity (i.e., sufficiently small steps between scores to measure changes of a magnitude considered meaningful). For example, if a computer-adaptive questionnaire is set to ask patients four questions (with four response categories each), this allows for a larger number of different obtainable scores (i.e., higher granularity) than a computer-adaptive questionnaire with only three items. A questionnaire with more questions may be more suitable to measure small but meaningful changes in an individual patient than the single-item scale mentioned above.

2c. Reporting of PRO score interpretation thresholds

CMCrep1_GEN

Statement: a statement should be provided detailing what type of methodology the patient- and group-level PRO score interpretation thresholds are based on (e.g., anchor-based, or distribution-based).

Explanation: to allow for a better understanding of what an interpretation threshold value for a patient- or group-level PRO score means and represents, it is necessary to understand how this value has been established. Therefore, when using a threshold, it is important to provide basic information on the methodology underlying the patient- or group-level threshold. In addition, if available, a literature reference providing more detailed information should be provided. If thresholds based on a similar methodology are used for multiple PRO scales, they may be described in single summary statement.

Example: a trial protocol may state that “a within-patient change of at least 10 points on the fatigue scale of questionnaire X between baseline and a subsequent assessment time point is considered a meaningful within-patient change (response) in the context of a responder analysis”. The threshold of 10 points has been established based on a patient-reported anchor reflecting perceived change over time” [*add reference to study demonstrating this association*].

CMCrep2_GEN

Statement: if an anchor-based PRO score interpretation threshold for patient- or group-level data is used, the anchor(s) should be reported (e.g., patient-reported change, or clinician-reported performance status).

Explanation: to allow for a better understanding of what a PRO score interpretation threshold value means and represents, it is necessary to understand how the value has been established. Therefore, basic information should be provided when using a threshold. If thresholds based on a similar methodology are used for multiple PRO scales, they may be described in a single summary statement.

Example: a trial protocol may state that a within-patient change of at least 10 points on the fatigue scale and of at least 5 points on the pain scale of questionnaire X between baseline and a subsequent assessment time point is considered as a meaningful within-patient change (response) in the context of a responder analysis. These thresholds have been established based on a patient-reported anchor reflecting perceived change over time.

CMCrep3_GEN

Statement: if an anchor-based PRO score interpretation threshold for patient- or group-level data is used, the statistical methods for determining the threshold should be clearly stated.

Explanation: to allow for a better understanding of what a patient- or group-level PRO score interpretation threshold value means and represents, it is necessary to understand how the

value has been established. Therefore, basic information should be provided when using a PRO score interpretation threshold (e.g. type of statistical model, or method for triangulation). If thresholds based on a similar methodology are used for multiple PRO scales, they may be described in single summary statement.

Example: a trial protocol may state that a within-patient change of at least 10 points on a fatigue scale between baseline and a subsequent assessment time point is considered as a response (meaningful within-patient change) in the context of a responder analysis. The threshold of 10 points has been established based on a patient-reported anchor reflecting perceived change over time *using a linear regression model*.

CMCrep4_GEN

Statement: if a distribution-based PRO score interpretation threshold for patient- or group-level data is used, the underlying statistic should be clearly stated (e.g., effect size).

Explanation: to allow for a better understanding of what the interpretation threshold of a PRO score means and represents, it is necessary to understand how the value has been established. Therefore, basic information should be provided when using such a threshold. If thresholds based on a similar methodology are used for multiple PRO scales, they may be described in a single summary statement.

Example: a trial protocol may state that a within-patient change of at least 10 points on the fatigue scale of questionnaire X between baseline and a subsequent assessment time point is considered as a response (meaningful within-patient change) in the context of a responder analysis.

This threshold of 10 points corresponds to *0.5 standard deviations* of the fatigue scale in a population with similar disease characteristics [*add reference to study establishing this value*]. If the threshold was established based on a distribution-based estimate alone, it may not necessarily be considered meaningful.

In addition, standard error of measurement (SEM)-based thresholds and other statistics that combine reliability and variability, commonly included under distribution-based thresholds, should be explicitly distinguished from other distribution-based thresholds, such as those based on effect sizes. This is because SEM-based thresholds represent a distinct concept that relates to the uncertainty of an estimate and the smallest detectable change.

CMCrep5_GEN

Statement: if a threshold for interpreting PRO scores for patient- or group-level data has been established using methods other than anchor- or distribution-based methods (e.g., based on qualitative interviews), a description of the methods used should be provided.

Explanation: to allow for a better understanding of what a PRO score interpretation threshold means and represents, it is necessary to understand how the value has been established. This is especially relevant if the threshold has been established using a method that is less

commonly used than anchor- or distribution-based methods. If thresholds for multiple PRO scales have been established with a similar methodology, the methodology for all of these scales may be described in single summary statement.

Example: a trial protocol may state that a within-patient change of at least 10 points on the fatigue scale of questionnaire X between baseline and a subsequent assessment time point is considered as a response (meaningful within-patient change) in the context of a responder analysis. A change of 10 points has been found to be meaningful to patients in *qualitative interviews using case vignettes* [add reference to study establishing this value].

CMCrep6_GEN

Statement: a short description of the patient population(s) in which the PRO score interpretation thresholds for patient- or group-level data were established should be given (e.g., diagnosis, or treatment type).

Explanation: to facilitate understanding of the applicability of an interpretation threshold for PRO scores in a specific trial population, it is helpful to know in what population the value has been established. If thresholds for multiple PRO scales have been established in the same population, this can be described in a single summary statement

Example: a trial protocol may state that a within-patient change of at least 10 points on the fatigue scale of questionnaire X between baseline and a subsequent assessment time point is considered as a response (meaningful within-patient change) in the context of a responder analysis. The change of 10 points has been found to be meaningful to *breast cancer patients undergoing chemotherapy* in qualitative interviews using case vignettes [add reference to study establishing this value].

3. Study design considerations

Design1_GEN*

Statement: it is important to identify beforehand which patient and disease characteristics are expected to be associated with the primary and key secondary endpoints. These should be considered in the analysis, and, therefore, must be recorded and reported. If there is a known core set of variables in the disease domain, efforts should be made to collect, evaluate, and report them all.

Explanation RCT*: the estimation of a causal effect requires appropriate adjustment for possible confounding factors. Therefore, it is important that these variables are available during the analysis to ensure proper and efficient treatment arm comparison. Even in a randomised setting, differences at baseline can occur or a selected analysis set may need to be used.

Recording a core set of variables facilitates the comparison of the results of RCTs to external data sources. The core set of variables enables adjusting for corresponding potential

confounding in the analysis phase when making such an external comparison. Although the set of key variables is likely disease-dependant, a core set of variables typically includes, but is not limited to, age, sex, disease stage, treatment history and comorbidities. Availability of baseline or pre-treatment PRO data may help adjust in part for between-patient variations.

Explanation SAT*: the estimation of a causal effect requires appropriate adjustment for possible confounding factors. Therefore, it is important that these variables are available both in the single arm study and when external control data are used, in the external control data to reduce bias in any comparison. Variables needed to perform subgroup analyses, to handle missing data or intercurrent events also need to be collected.

A core set of variables is a set of possibly prognostic or predictive variables that is meant to be collected within every study in a particular disease domain. Recording a core set of variables facilitates the comparison of the results of single arm studies to other data sources. For example, the control arm of a previously conducted RCT could be used as historical control data. The core set of variables enables adjusting for corresponding potential confounding in the analysis phase when comparing between the groups. Although the set of key variables is likely disease-dependant, a core set of variables typically includes, but is not limited to, age, sex, disease stage, treatment history and comorbidities. Availability of baseline or pre-treatment PRO data may help adjust in part for between-patient variations.

Design2_GEN

Statement: PRO data collection methods should be aligned with the strategies established at the start of the trial to handle ICEs.

Explanation: when the strategy is to include PRO measurements even after an ICE in the analysis, all efforts should be made to continue PRO data collection until death or a pre-specified time point after treatment discontinuation. If this is not feasible, it should be justified beforehand in the protocol. During data collection, reasons for missing PRO data should be recorded.

A main incentive for patients to complete questionnaires is the principle that all data they provide will lead to a meaningful contribution to the actual results. It is understood that not all data may be equally relevant for all objectives.

If the PRO objective limits all analyses to pre-progression data, then PRO data after progression should not be collected. Limitations of the PRO objectives to pre-progression data may be intentional as part of the objective. However, it may also be justified by practical constraints such as when approaching patients post-progression may be logistically difficult.

Design3_GEN*

Statement RCT*: the timing of the PRO assessments and the duration of their associated time windows should be predetermined and not depend on post-baseline events related to the disease. Furthermore, the timing of the PRO assessments should be comparable or accounted for between treatment arms.

During the trial, it is advised to adhere as much as possible to these assessment time points.

Explanation RCT*: the schedule for PRO assessments should be established in advance. When selecting appropriate time points for assessment, it is, however, important to consider the natural history of the disease/progression, the hypothesised impact of therapy over time, the expected treatment side effects, the mode and schedule of treatment administration, and practical considerations such as alignment of assessments with clinic visits and recall period of the PRO measures (so that PROs and clinical data can be collected at the same time).

In cancer trials, PRO assessment times should remain as structured as possible and not depend on disease-related events. Asking patients to respond to a questionnaire only when they experience a disease- or treatment-related event, such as diarrhoea, is not recommended, as doing so biases the comparison between treatment arms.

This approach also complicates issues with missing data, as absences may result from either the questionnaire not being completed or the trigger event not occurring.

The PRO assessments should be completed according to the protocol at that specific time point. Visits may be delayed or brought forward for practical reasons or not occurred on the planned date. It is therefore recommended, to provide time windows for each PRO assessment. These windows are pre-specified time intervals around the planned date where a PRO assessment would still be considered valid and accurately represent the intended assessment date. Additionally, the recall period of the PRO assessment should be taken into account to ensure comparability between treatment arms.

Statement SAT*: the timing of the PRO assessments and the duration of their associated time windows should be predetermined and not depend on post-baseline events related to the disease. Furthermore, when there is an external control group, the timing of the PRO assessments should be comparable or accounted for between groups.

During the trial, it is advised to adhere as much as possible to these assessment time points.

Explanation SAT*: the schedule for PRO assessments should be established in advance. When selecting appropriate time points for assessment, it is, however, important to consider the natural history of the disease/progression, the hypothesised impact of therapy over time, the expected treatment side effects, the mode and schedule of treatment administration, and practical considerations such as alignment of assessments with clinic visits and recall period of the PRO measures (so that PROs and clinical data can be collected at the same time).

In cancer trials, PRO assessment times should remain as structured as possible and not depend on disease-related events. Asking patients to respond to a questionnaire only when they experience a disease- or treatment-related event, such as diarrhoea, is not recommended, as doing so biases the comparison with external control data.

This approach also complicates issues with missing data, as absences may result from either the questionnaire not being completed or the trigger event not occurring.

The PRO assessments should be completed according to the protocol at that specific time point. Visits may be delayed or brought forward for practical reasons or not occurred

on the planned date. It is therefore recommended, to provide time windows for each PRO assessment. These windows are pre-specified time intervals around the planned date where a PRO assessment would still be considered valid and accurately represent the intended assessment date. Additionally, when comparing timing of the PRO measure with external control data, the recall period for the PRO assessment should be taken into account to ensure comparability.

Design4_GEN

Statement: it is recommended to collect and report the reasons and frequencies for missingness and incorporate these in the sensitivity analyses.

Explanation: it is rarely appropriate to assume that the association between missingness and the PRO score is the same, whatever the cause for the missingness. This is especially relevant in cancer trials, where the disease's aggressive nature and the toxicity of the treatment can lead to missing PRO data for different reasons. Each reason for dropout may be related to the PRO score in a different way.

For example, if PRO assessments are missing due to a lack of a translated questionnaire, this can reasonably be assumed to be non-informative of the PRO endpoint, whilst missing PRO assessments due to the patient being too ill to fill in the questionnaire could reasonably be assumed to be informative of the PRO endpoint. These two different reasons for missing data could, therefore, be accounted for in the sensitivity analysis in a different way.

Collecting information about the reason for missing data may, however, not be feasible in all cases, for instance when using remote data collection of at-home diaries. During the design phase, it is important to implement strategies that reduce the occurrence of missing data and to evaluate how missing data might correlate with PRO scores.

Methods that could be used are, for example, multiple imputation and tipping point analysis. Missingness at a given time point (yes, no) can be a function of key baseline demographic and clinical characteristics, which can be analysed with a multiple logistic regression model.

4. External comparison

ExComp1_SAT

Statement: the criteria for selection of the patients in the single arm study and in the external comparison group must be accurately described. All relevant baseline variables, source and data quality of the comparison sample, timing of measurement, and PRO measures being used should be described for both groups.

Explanation: in single arm trials, external information for comparisons under other treatments is inevitably collected outside the trial. To ensure a fair comparison, it is important to detail not only the enrolment eligibility criteria of the single arm trial, but also all the descriptive

characteristics of the external comparison group. This is to assess whether a suitable comparison between the two populations may be possible in terms of known prognostic factors for the PRO.

ExComp2_SAT

Statement: if selecting an external control group, hypothesised comparability limitations should be defined and the planned approach for addressing them appropriately described.

Explanation: comparability issues are:

- Characteristics of the study population, both for the single arm study and control setting
- PROs being used
- The timing and frequency of PROMs
- Variability in PROs, both within and between patients
- Duration of the follow-up
- How ICEs were defined and how data were collected after ICEs
- Differences in patients' perceptions of PROs, such as cultural variations in pain evaluation
- Whether the external control setting is blinded or un-blinded
- The time period during which PROs were collected.

All comparisons in a single arm study should be carried out with appropriate caution to avoid potential bias due to lack of randomisation and blinding of treatment. Comparison with an external group may not be reliable if the timing and frequencies of PRO measurements largely differ between the studies (for instance, measured every month versus measured every six months). Additionally, if the external control group consists of patients from a different cultural background than the population of the single arm trial, one should be aware that expression for pain, symptoms, or HRQoL could be different. Any deviations should be justified.

5. Analyses consideration

Anal1_GEN*

Statement RCT*: both completion rates and available data rates should be reported for each assessment time point, in both the confirmatory and descriptive setting. The completion rate is calculated by setting the denominator to the expected number of assessments at that time point, defined as the number of patients scheduled for a PRO measurement at that time point according to the protocol. The available data rate is calculated by setting the denominator to the number of patients randomised in the trial. For both the completion rate and available data rate, the numerator is set to the number of patients who completed the PRO assessment at that specific time point. Any deviations should be justified.

Explanation RCT*: in order to clearly show the amount of missing data, both the completion rates and available data rates should be reported for each assessment time point and per treatment arm, in both confirmatory and descriptive settings.

The completion rate can be interpreted as the proportion of patients with a scheduled PRO assessment at time t who complete the PRO assessment. The available data rate can be interpreted as the proportion of all patients randomised in the trial completing the PRO assessment at time t . Any deviations should be justified.

To accurately calculate completion rates, a distinction should be made between instances where relevant data could not be collected (resulting in missing data) and cases where data was deliberately not collected or used due to an ICE. When a PRO observation is missing, the corresponding patient is removed from the numerator of the completion rate. When a PRO observation is not collected because of the chosen ICE strategy, the corresponding patient is removed from both the numerator and denominator of the completion rate.

The denominator would exclude death since patients who died cannot be expected to provide PRO assessments. Death is included in the denominator when calculating available data rates.

Example: in a cancer RCT where PROs are to be collected at the start of the 6th cycle of chemotherapy, the scenario for the control arm is as follows:

Criterion (control arm):	Number of patients:
Randomised	200
Deceased before 6th cycle	20
Lost-to-follow up before 6th cycle	5
Discontinued treatment before 6th cycle	75
Started the 6th cycle	100
Provided PRO data	80

Completion rate=number of patients who provided PRO data/expected number of assessments at that time point=80/100=80%

Available data rate=number of patients who provided PRO data/patients randomised in the trial=80/200=40%.

Statement SAT*: both completion rates and available data rates should be reported for each assessment time point. The completion rate is calculated by setting the denominator to the expected number of assessments at that time point, defined as the number of patients scheduled for a PRO measurement at that time point according to the protocol. The available data rate is calculated by setting the denominator to the number of patients included in the trial. For both the completion rate and available data rate, the numerator is set to the number of patients who completed the PRO assessment at that specific time point. Any deviations should be justified.

Explanation SAT*: in order to clearly show the amount of missing data, both the completion rates and available data rates should be reported for each assessment time point, in both confirmatory and descriptive settings.

The completion rate can be interpreted as the proportion of patients with a scheduled PRO assessment at time t who complete the PRO assessment. The available data rate can be interpreted as the proportion of all patients included in the trial completing the PRO assessment at time t . Any deviations should be justified.

To accurately calculate completion rates, a distinction should be made between instances where relevant data could not be collected (resulting in missing data) and cases where data was deliberately not collected or used due to an ICE. When a PRO observation is missing, the corresponding patient is removed from the numerator of the completion rate. When a PRO observation is not collected because of the chosen ICE strategy, the corresponding patient is removed from both the numerator and denominator of the completion rate.

The denominator would exclude death since patients who died cannot be expected to provide PRO assessments. Death is included in the denominator when calculating available data rates.

Example: in a cancer single arm trial where PROs are to be collected at the start of the 6th cycle of chemotherapy, the scenario is as follows:

Criterion:	Number of patients:
Included in the trial	200
Deceased before 6th cycle	20
Lost-to-follow up before 6th cycle	5
Discontinued treatment before 6th cycle	75
Started the 6th cycle	100
Provided PRO data	80

Completion rate=number of patients who provided PRO data/expected number of assessments at that time point=80/100=80%

Available data rate=number of patients who provided PRO data/patients randomised in the trial=80/200=40%.

Anal2_RCT

Statement: analysing only patients who completed all planned PRO assessments (complete-case analysis) will potentially bias the estimate of the PRO treatment effect.

Explanation: including only patients without missing values will result in a loss of information and, therefore, a loss in the study's statistical power. The complete cases from a selected subsample will be representative of the trial population only under missing completely at random (MCAR) assumption. This is unrealistic in most scenarios, thus the estimates obtained will be biased.

Example: patients who did not complete all their expected PRO assessments are excluded from the analysis; as a result, the analysis population is likely to be a selected subgroup that may no longer be representative of the overall trial population. If patients with more symptoms stop filling in the PRO measure, the results would reflect only those with fewer problems/better health, and not the reality of the total patient population.

5a. Assumptions

Assump1_GEN

Statement: model diagnostics should be used to check the model assumptions, where feasible, and investigate the presence of outlying and influential observations.

Explanation: when analysing PRO data, the assumptions underlying both the statistical model and, if relevant, the test for the significance of the treatment effect should be clearly stated and verified. Incorrect assumptions can render the results invalid. Additionally, the diagnostics should be used to identify the presence of any outlying observations (i.e., an observation for which the predicted value based on the model is significantly different from the observed value) and influential observations (i.e., data points that greatly affect the model's fit).

Assumptions that are often encountered are independence of the observations, normality of the residual errors of a general linear mixed model for a continuous outcome, and proportional hazards assumption for a proportional hazards model, amongst others. Some assumptions, such as the independence of the observations, are related to the design of the study and cannot be checked using model diagnostics. Other assumptions, such as the normality of the residual errors, are related to the statistical model.

Whilst not all assumptions can be easily checked, reasonable efforts should be made to assess their plausibility, where feasible. In case no formal test exists for a particular assumption, indirect evidence (such as graphical methods) can be explored.

Assump2_SAT

Statement: when assessing PRO in single arm studies, one should be aware that methods such as linear mixed models (LMMs) or generalised linear mixed models (GLMMs) implicitly impute values of expected outcomes after death, when, in reality, PRO values cease to exist after death.

Explanation: in GLMMs, the PROMs after death are considered to be missing at random. In reality, PROMs cease to exist after death.

GLMMs assume that the patterns over time for patients can be extrapolated and used at a later stage to replace missing data. This hypothetical strategy assumes that the patterns over time after death evolve similarly to those of patients with observed values. Hence these methods implicitly impute values after death. One may see this as estimating change in a hypothetical world where no patients die during the study.

5b. Main analyses

AnalMain1_GEN*

Statement RCT*: when performing a descriptive analysis, summaries of the observed data should be reported for each treatment arm by assessment time, with, ideally, a variability measure (e.g., standard deviation, variance or interquartile range) and all estimates should have a measure of error (standard error or confidence interval). A formal comparison between treatment arms is not recommended. Inferential statistics are not part of such an analysis.

Explanation RCT*: the goal of a descriptive analysis is to summarise the observed data. In most trials, the tolerability profiles are best addressed by applying descriptive statistical methods to the data.

When interpreting summary measures from later time points that are based on data from a smaller number of patients, caution is crucial due to the potential limitations. If the data are sparse and affected by selection bias, it is recommended not to rely on the summary measures.

The ICE strategy should be clearly specified, such as strategies where summaries are limited to specific patients (alive, on treatment and/or progression-free) at the different assessment time points according to the study objective.

When appropriate, a distribution of the occurrence of relevant ICE for each treatment arm and for each time point could be provided.

Statement SAT*: when performing a descriptive analysis, summaries of the observed data should be reported by assessment time, with, ideally, a variability measure (e.g., standard deviation, variance or interquartile range) and all estimates should have a measure of error (standard error or confidence interval). Inferential statistics are not part of such an analysis.

Explanation SAT*: the goal of a descriptive analysis is to summarise the observed data. In most trials, the tolerability profiles are best addressed by applying descriptive statistical methods to the data.

When interpreting summary measures from later time points that are based on data from a smaller number of patients, caution is crucial due to the potential limitations. If the data are sparse and affected by selection bias, it is recommended not to rely on the summary measures.

The ICE strategy should be clearly specified, such as strategies where summaries are limited to specific patients (alive, on treatment and/or progression-free) at the different assessment time points according to the study objective.

When appropriate, a distribution of the occurrence of relevant ICE for each time point could be provided.

AnalMain2_GEN*

Statement: when the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective), it is not recommended to analyse data at each time point separately using multiple cross-sectional analyses of the magnitude of PRO (change) score or proportion of responders.

Explanation RCT*: when the PRO assessments take place at predefined time points with repeated assessments expected per patient over the course of the study, the data can be analysed using a multiple cross-sectional univariate analysis performed at each time point separately. This is a valid approach if the focus is on assessing the treatment effect at specific time points. However, several caveats apply. First, there is a considerable loss of information by considering cross-sectional analyses instead of modelling the full longitudinal profiles, which results in reduced statistical power at each time point. Second, when analysing time points one at a time, it becomes challenging to fully take into account missing data. Third, simple univariate tests like the two-sample t-test are based on the assumption that the scores are identically and independently distributed. This assumption is often untenable and too restrictive.

Additionally, multiple statistical testing will inflate the type I error rate. Finally, performing multiple cross-sectional univariate analysis for each time point does not enable researchers to study evolutions in PRO values over time, drawing longitudinal interpretation, nor does it allow them to consider the correlation between different observations of the same patient.

Explanation SAT*: when the PRO assessments take place at predefined time points with repeated assessments expected per patient over the course of the study, the data can be analysed using a multiple cross-sectional univariate analysis performed at each time point separately. This is a valid approach if the focus is on assessing the treatment effect at specific time points. However, several caveats apply. First, there is a considerable loss of information by considering cross-sectional analyses instead of modelling the full longitudinal profiles, which results in reduced statistical power at each time point. Second, when analysing time points one at a time, it becomes challenging to fully take into account missing data.

Additionally, multiple statistical testing will inflate the type I error rate. Finally, performing multiple cross-sectional univariate analysis for each time point does not enable researchers to study evolutions in PRO values over time, drawing longitudinal interpretation, nor does it allow them to consider the correlation between different observations of the same patient.

AnalMain3_GEN

Statement: when performing a responder analysis or an analysis of continuous PROs at a pre-specified time point of interest, a longitudinal analysis that considers the data's repeated measurement structure should be conducted, in line with the defined estimand of interest.

Explanation: when the endpoint is magnitude of PRO (change) score, with time included as a discrete variable, and repeated assessments have been made per patient over the course of

the study, a linear regression model can be applied. This model accounts for the correlation between a patient's measurements through a residual covariance matrix.

When the endpoint is binary (success/failure), the probability of being a responder over time can be modelled using a longitudinal model with a logit link-function. One can choose, for example, to use a marginal model using generalised estimating equations (GEE), or to use a hierarchical model such as a GLMM using maximum likelihood.

However, when using LMMs or GLMMs, one should be aware that these methods implicitly impute values after death, while in reality PRO values after death do not exist.

AnalMain4_GEN*

Statement: when performing a responder analysis or an analysis of the magnitude of PRO (change) score, it is recommended to include time as a categorical variable as this requires fewer assumptions compared to a model using time as a continuous variable. If applying models using time as a continuous variable, their underlying assumptions need to be justified.

Explanation RCT*: in analyses focused on the proportion of responders or on the magnitude of PRO (change) score, when patients have been assessed multiple times throughout the study, it is important to consider the timing of these assessments in relation to the baseline anchor time, such as time of randomisation.

Assessment time should be included as a discrete variable. This approach ensures that no assumptions are made on the relationship between time and the outcome variable, even though it requires the use of more parameters. The requirements are that the assessment schedule is pre-specified and similar between treatment arms (in frequency and timing) for all included patients and that the sample size is sufficiently large compared to the number of assessments.

When the requirements are not met (e.g., in a rare cancer trial with a small sample size and frequent PRO assessments), time could be included as a continuous variable. In this case, the model should be sufficiently flexible in order to accurately represent the relationship between the outcome and time. Any assumption made on the functional relationship between time and the PRO variable of interest should be justified.

If the data shows a linear relationship between the PRO variable of interest and time, time could be included as a continuous variable.

Explanation SAT*: in analyses focused on the proportion of responders or on the magnitude of PRO (change) score, when patients have been assessed multiple times throughout the study, it is important to consider the timing of these assessments in relation to the baseline anchor time, such as time of the start of treatment.

Assessment time should be included as a discrete variable. This approach ensures that no assumptions are made on the relationship between time and the outcome variable, even

though it requires the use of more parameters. The requirements are that the assessment schedule is pre-specified and similar between treatment groups (in frequency and timing) for all included patients and that the sample size is sufficiently large compared to the number of assessments.

When the requirements are not met (e.g., in a rare cancer trial with a small sample size and frequent PRO assessments), time could be included as a continuous variable. In this case, the model should be sufficiently flexible in order to accurately represent the relationship between the outcome and time. Any assumption made on the functional relationship between time and the PRO variable of interest should be justified.

If the data shows a linear relationship between the PRO variable of interest and time, time could be included as a continuous variable.

AnalMain5_GEN*

Statement: when performing a time-to-event analysis on PRO data, it is recommended to apply statistical methods that incorporate the interval-censored nature of the data.

Explanation RCT*: in practice, when PROs are used to define an event, the exact time of the event is unknown, as the PRO assessments are performed at pre-specified periodic assessment time points. PRO data, therefore, are interval-censored, as the event time is only known to fall within a particular interval (i.e., between two pre-specified assessment time points). Right-censoring occurs when a participant's survival time is known to exceed a certain value (equivalent to single right-point imputation). This could lead to bias in estimating the effect of the treatment and may result in reduced statistical power. The extent of this bias depends on the frequency of measurements, as it relates to the ratio of the length of time between assessments and the anticipated duration until an event occurs. Secondly, different assessment schemes between treatment arms can create substantial bias.

The standard time-to-event analysis assumes that censoring is independent of the event.

With this in mind, reasons for censoring should be explored and reported.

Whilst most non-PRO time-to-event endpoints in RCTs (e.g. overall survival, progression-free survival) have been analysed using right-censoring techniques, the underlying assumptions are rarely verified. PRO data collection, which is dependent on patient participation, is even more susceptible to deviations from right-censoring assumptions due to higher rates of missing data and deviations from a fixed assessment schedule. Moreover, this is an issue that is not corrected by the randomisation process as mechanisms causing deviations to the assessment schedule may differ across the different arms.

Overall, it is therefore advised to use methods for interval-censored data when PRO data are obtained at pre-specified assessment time points.

Explanation SAT*: in practice, when PROs are used to define an event, the exact time of the event is unknown, as the PRO assessments are performed at pre-specified periodic assessment time points. PRO data, therefore, are interval-censored, as the event time is

only known to fall within a particular interval (i.e., between two pre-specified assessment time points). Right-censoring occurs when a participant's survival time is known to exceed a certain value (equivalent to single right-point imputation). This could lead to bias in estimating the effect of the treatment and may result in reduced statistical power. The extent of this bias depends on the frequency of measurements, as it relates to the ratio of the length of time between assessments and the anticipated duration until an event occurs. Secondly, different assessment schemes in the single arm study and external control data can create substantial bias.

The standard time-to-event analysis assumes that censoring is independent of the event.

With this in mind, reasons for censoring should be explored and reported. Overall, it is therefore advised to use methods for interval-censored data when PRO data are obtained at pre-specified assessment time points.

AnalMain6_GEN

Statement: when a longitudinal model is fitted, appropriate correction for the baseline value of the PRO variable should be considered.

Explanation: in a longitudinal study, there are two different approaches depending on the parameter of interest.

Approach A

Under the first approach, the interest is to estimate a single average (or contrasts between such averages from different groups). The single average can be expressed as the original outcome or as a change from the baseline. The estimate can be obtained through a cross-sectional model (i.e., ANOVA or ANCOVA) or, more appropriately, a longitudinal model to estimate the average at the time point of interest. If these models are used, correcting for the baseline outcome is recommended, as this may account for unwanted variability between subjects in terms of their change from baseline or at the time point of interest, thereby reducing standard errors.

Approach B

Under the second approach, the parameter of interest is the longitudinal trend (i.e., slope) or the contrast of trends between randomised groups. When the change from the baseline is modelled, correcting for the baseline will not affect the slope's estimate or standard error. Therefore, correcting for baseline is not required. However, when the original outcome is being modelled, adjusting for baseline would imply that the outcome at time $t=0$ is not included in the response vector, as it cannot be a response and a covariate simultaneously. Given the interest in estimating slopes, and since correction for the baseline outcome would not affect the results, such a correction for the baseline outcome would imply that baseline is no longer used in the analysis hence resulting in a loss of efficiency.

Example: in a randomised longitudinal study where the main objective is to compare the improvement in a specific domain (e.g., pain) over time between two treatment arms,

PROs are collected at baseline (i.e., before treatment administration) and then at months one, two, three and four. The analysis of interest could be to compare the pain scores between the two treatment arms at month four. This can be achieved by fitting a linear mixed model to the study data and calculating the difference in average pain scores at month four between the two treatment arms. In this case, it is recommended to use the baseline pain score as a covariate in the model regardless of whether the endpoint is the pain score itself or the change from the baseline pain score to account for the baseline variability.

If the research interest is to compare the time trends (i.e., slopes) in pain between the two treatment arms, a linear mixed model can be fitted. In this scenario:

- If the model uses changes from baseline values, there is no advantage in also correcting for the baseline value.
- If the model uses the actual pain scores at each time point, the baseline score should be included as part of the response vector, not as a covariate in the model to make maximum use of the data. In this case, it is not recommended to use the baseline pain score as a covariate in the model.

Therefore, the choice to correct for baseline variables in the statistical model depends on the nature of the research question, as illustrated in this hypothetical clinical trial scenario.

AnalMain7_GEN*

Statement: using single imputation techniques, complete-case analysis or available case analysis to handle missing data is generally not recommended. A justification should be given if these approaches are used.

Explanation RCT*: analyses could be heavily impacted by the assumptions regarding the type of missing data, i.e., whether or not missingness can be considered to occur randomly. In longitudinal studies, attrition bias could occur, i.e., some participants are more likely to drop out than others. Reasons for missingness should be explored. Several statistical techniques exist to deal with missing data.

When missing data are non-negligible, simple techniques such as single imputation, complete-case analysis (only including patients with no missing data), or available case analysis (only including patients with no missing data at the time point of interest) are generally not recommended due to potential bias and loss of information. In the case of a binary endpoint (success/failure), single imputation is often implemented by considering patients with missing data as non-responders. For a continuous endpoint, single imputation is often implemented by imputing the last observed PRO score (last observation carried forward [LOCF]) or the mean of the observed PRO scores of the patient (unconditional mean imputation).

Although these approaches are straightforward, the results from most simple imputation techniques are only valid under assumptions which are often untenable and too restrictive. For example, LOCF assumes patients' PRO score will no longer change from last observed score; considering patients with missing data as non-responders assumes the missingness is due to adverse PRO scores. These methods may lead to biased estimates. Secondly, the use

of single imputation values ignores the uncertainty associated with the missingness, resulting in an overestimation of the precision of the treatment effect.

However, it is important to distinguish this recommendation from the composite strategy used to address ICEs when defining the PRO variable of interest. Within this estimand framework, the imputation of a pre-specified value is based on the relevant PRO objective.

Single imputation of missing data could be reasonable in specific cases where there is a clear justification. For example, imputation of extreme PRO scores or (non-) responses after a certain event, when there is a clear justification, could be one component of the sensitivity analysis. This may be relevant in the context of tolerability analyses if the missingness is assumed to be related to a patient's health status. This recommendation does not concern imputation as part of the instrument scoring instructions.

As an alternative, multiple imputation techniques can be considered to address missing data. These techniques generate several completed datasets and combine the results from these datasets, thereby better accounting for the uncertainty associated with missing data.

When the PRO objective is to inform safety and tolerability, an available case analysis may be considered when summarising PRO scores descriptively at each assessment time point by treatment arm. This approach is only justifiable with high PRO completion rates and no evidence of selective missingness to avoid misinterpretation of PRO score results. The number of patients expected to complete the PRO measure and the number of patients who did not complete the PRO measure at each assessment time point and by treatment arm should be reported along with reasons for non-completion.

Explanation SAT*: analyses could be heavily impacted by the assumptions regarding the type of missing data, i.e., whether or not missingness can be considered to occur randomly. In longitudinal studies, attrition bias could occur, i.e., some participants are more likely to drop out than others. Reasons for missingness should be explored. Several statistical techniques exist to deal with missing data.

When missing data are non-negligible, simple techniques such as single imputation, complete-case analysis (only including patients with no missing data), or available case analysis (only including patients with no missing data at the time point of interest) are generally not recommended due to potential bias and loss of information. In the case of a binary endpoint (success/failure), single imputation is often implemented by considering patients with missing data as non-responders. For a continuous endpoint, single imputation is often implemented by imputing the last observed PRO score (last observation carried forward [LOCF]) or the mean of the observed PRO scores of the patient (unconditional mean imputation).

Although these approaches are straightforward, the results from most simple imputation techniques are only valid under assumptions which are often untenable and too restrictive. For example, LOCF assumes patients' PRO score will no longer change from last observed score; considering patients with missing data as non-responders assumes the missingness is due to adverse PRO scores. These methods may lead to biased estimates. Secondly, the use of single imputation values ignores the uncertainty associated with the missingness, resulting in an overestimation of the precision of the treatment effect.

However, it is important to distinguish this recommendation from the composite strategy used to address ICEs when defining the PRO variable of interest. Within this estimand framework, the imputation of a pre-specified value is based on the relevant PRO objective.

Single imputation of missing data could be reasonable in specific cases where there is a clear justification. For example, imputation of extreme PRO scores or (non-) responses after a certain event, when there is a clear justification, could be one component of the sensitivity analysis. This may be relevant in the context of tolerability analyses if the missingness is assumed to be related to a patient's health status. This recommendation does not concern imputation as part of the instrument scoring instructions.

As an alternative, multiple imputation techniques can be considered to address missing data. These techniques generate several completed datasets and combine the results from these datasets, thereby better accounting for the uncertainty associated with missing data.

When the PRO objective is to inform safety and tolerability, an available case analysis may be considered when summarising PRO scores descriptively at each assessment time point. This approach is only justifiable with high PRO completion rates and no evidence of selective missingness to avoid misinterpretation of PRO score results. The number of patients expected to complete the PRO measure and the number of patients who did not complete the PRO measure at each assessment time point should be reported along with reasons for non-completion.

AnalMain8_RCT

Statement: when the goal of the PRO objective is to draw conclusions about clinical benefit (confirmatory objective), the main PRO endpoint is an overall effect (i.e., average response over time), and it is assumed that the average PRO score is an adequate summary of each patient's PRO longitudinal profile over time, longitudinal statistical approaches rather than two-step summary approaches are recommended. The underlying ICE strategy, particularly for death, and missing data assumptions should be described.

Explanation: a two-step summary approach works by first summarising data per patient into a single statistic (within-patient summary). Whilst in a second step, these within-patient statistics are averaged by treatment arm. Common analytical challenges stem from the reduction of each patient's longitudinal data into a single summary measure. As a result, differences between treatment groups in terms of PRO follow-up times, ceiling or floor effects, different ICE and missing data patterns will affect the treatment effect estimates. When using a two-step summary approach, it is therefore important to identify which of these factors are relevant and describe how those are addressed in the summary statistic. As an example, if a maximum value of all reported scores during treatment exposure is used as the within-patient summary statistic, then this summary implies the use of a while-on-treatment strategy. The endpoint and population-level summary need to be carefully selected and justified vis-à-vis the trial objective.

LMMs (with time as a continuous variable) and other models can be used to estimate and compare the average effect over time (i.e. overall effect) under certain assumptions. The

average effect of a given treatment over time can be assessed through longitudinal models. LMM, GLMM and GEE are the main approaches to model longitudinal data. These models are appropriate when dealing with non-informative missing data and unbalanced data (i.e., unequal number of measurements per subject and measurements not taken at fixed time points). Selecting an approach depends on the structure of the mean, the missing data mechanism and the level of analysis that is of interest.

AnalMain9_SAT

Statement: if the study only includes patients with specific high or low PRO scores, such as high pain scores, and uses the same PRO domain as the outcome, the effects of regression to the mean should be considered in the PRO analysis, or should be acknowledged when discussing the findings.

Explanation: due to regression to the mean, when patients are selected with high pain scores, the mean pain score at the next visit is expected to be lower, irrespective of the treatment effect. Not accounting for this would lead to a biased treatment effect.

During the trial design phase, it is important to carefully consider regression to the mean. This includes, selecting an appropriate PRO for the intended population and enrolment criteria in the trial design, and throughout the analysis, interpretation, and reporting stages of results.

The effect of regression to the mean can be estimated by a quantitative bias analysis, which requires estimates of both the intra- and inter-individual variance. These estimates can be derived from repeated pre-treatment measurements or from external data. If this data is not available or sufficiently reliable, the potential bias due to regression to the mean should be discussed when reporting the results.

5c. Sensitivity/supplementary analyses

AnalSens1_GEN*

Statement: the overall PRO analysis strategy should include a main PRO analysis supported by sensitivity and/or supplementary analyses (if necessary).

Explanation RCT*: an overall PRO analysis strategy should be designed to answer the main research questions and consist of a predefined main analysis/es to reach conclusions. It is important to check the sensitivity of the main conclusions against the limitations of the data assumptions and different data analysis approaches.

Sensitivity analyses are conducted to ensure that the main findings from PROs are not substantially affected (or remain consistent) when different assumptions or methods are used in the analysis. Supplementary analyses are conducted in addition to the main analysis and sensitivity analysis to better understand the effects of the treatment.

Example: suppose the PRO objective is to consider levels of physical functioning at month six between two treatment arms. However, a number of patients did not have PRO data at month six because they were too ill to fill out the questionnaire. A primary analysis was planned with specific assumptions on how to handle these missing data. As a sensitivity analysis, another analysis method with different missing data assumptions was used to understand the robustness of the findings from the primary analysis.

Explanation SAT*: an overall PRO analysis strategy should be designed to answer the main research questions and consist of a predefined main analysis/es to reach conclusions. It is important to check the sensitivity of the main conclusions against the limitations of the data assumptions and different data analysis approaches.

Sensitivity analyses are conducted to ensure that the main findings from PROs are not substantially affected (or remain consistent) when different assumptions or methods are used in the analysis. Supplementary analyses are conducted in addition to the main analysis and sensitivity analysis to better understand the effects of the treatment.

Example: suppose the PRO objective is to consider levels of physical functioning at month six. However, a number of patients did not have PRO data at month six because they were too ill to fill out the questionnaire. A primary analysis was planned with specific assumptions on how to handle these missing data. As a sensitivity analysis, another analysis method with different missing data assumptions was used to understand the robustness of the findings from the primary analysis.

AnalSens2_GEN

Statement: when performing a confirmatory analysis, supplementary analyses could be conducted to provide additional insights into the understanding of the treatment effect.

Explanation: supplementary analyses are conducted in addition to the main analysis with the intent to provide additional insights into the understanding of the treatment effect.

A distinction should be made between sensitivity analyses and supplementary analyses. Sensitivity analyses are used to assess the robustness of the estimator by varying the assumptions about, for instance, the missing data mechanism. On the other hand, supplementary analyses provide additional results by considering different estimands through alternative ICE strategies, changes in population-level summaries definition or subgroup analyses. Relevant supplementary analyses should be pre-specified in the SAP as this is important for strengthening the analysis. Post-hoc analyses may be considered to augment (not replace) the pre-specified analyses.

AnalSens3_GEN

Statement: it is recommended to perform sensitivity analyses to assess the robustness of results with varying plausible assumptions and methods for handling missing data and its impact on the conclusions for the primary and key secondary PRO study objectives. Each sensitivity analysis should be designed to assess the effect on the results of the particular assumptions and methods used to account for the missing data.

Explanation RCT*: sensitivity analysis should be used to assess how varying the assumptions about the missing data mechanism impacts conclusions of interest. When the missingness mechanism is assumed to be MAR (missing at random), standard estimation methods can be used, but the assumption of MAR against MNAR (missing not at random) itself cannot be tested on the available data. As PRO data collection depends on the voluntary participation of the patient, MNAR is more plausible as patients' health status may directly or indirectly influence the completion of PRO assessments. Therefore, different scenarios in the direction of MNAR should be considered. All MNAR models are based on assumptions that cannot be verified based on the observed data, yet the plausibility of the models can be formulated in terms of clinical arguments. A large number of different methods exist to perform a sensitivity analysis, such as multiple imputation in the pattern-mixture framework, local influence approaches, shared-parameter models, etc. Each sensitivity analysis should be designed to assess the effect on the results of the particular assumptions made to account for the missing data. If the primary analysis, for example, assumes the same missingness mechanism to hold across all randomised treatment arms, then multiple imputation using an imputation model that differs across treatment arms can be used as a sensitivity analysis. Key sensitivity analyses should be pre-specified in the SAP.

In sensitivity analyses, the information on deviation from the assumptions should be clearly described. The deviations should be plausible and support interpretability.

Explanation SAT*: sensitivity analysis should be used to assess how varying the assumptions about the missing data mechanism impacts conclusions of interest. When the missingness mechanism is assumed to be MAR (missing at random), standard estimation methods can be used, but the assumption of MAR against MNAR (missing not at random) itself cannot be tested on the available data. As PRO data collection depends on the voluntary participation of the patient, MNAR is more plausible as patients' health status may directly or indirectly influence the completion of PRO assessments. Therefore, different scenarios in the direction of MNAR should be considered. All MNAR models are based on assumptions that cannot be verified based on the observed data, yet the plausibility of the models can be formulated in terms of clinical arguments. A large number of different methods exist to perform a sensitivity analysis, such as multiple imputation in the pattern-mixture framework, local influence approaches, shared-parameter models, etc. Each sensitivity analysis should be designed to assess the effect on the results of the particular assumptions made to account for the missing data. Key sensitivity analyses should be pre-specified in the statistical analysis plan.

In sensitivity analyses, the information on deviation from the assumptions should be clearly described. The deviations should be plausible and support interpretability.

AnalSens4_SAT

Statement: handling of missing data should be distinguished from values observed but not utilised in the analysis due to the strategy of handling an ICE. Reasons for missing PRO measurements should be collected and the implications for those missing PRO values need to be carefully explored. Care should be taken that the assumptions about the distribution of the missing data are consistent with the defined estimand.

Explanation: missing at random (MAR) implies that the probability of missing data depends only on observed information. For example, PRO scores at a certain time point are more likely to be missing if the previously observed PRO score of a patient was low. Often, analyses should begin with the assumption of MAR, considering methods such as multiple imputation, maximum likelihood estimation models or reweighting. It is important to notice that imputation models and models for missingness can differ from those used in the main analysis. These methods are allowed to use information occurring in the future (such as the time of death). To make the MAR assumption more plausible, the imputation and missingness models should include as much information as possible about the reasons for missing data. For example, PRO scores often change in the months leading up to death. Therefore, when handling missing data in this period, it is important to use the time period between measurement and death in the imputation or in the missingness models.

5d. Results presentation and interpretation

AnalPres1_GEN

Statement: if the objective of the PRO is to evaluate treatment tolerability and the clinical focus is only on PROs while-on-treatment, PRO scores may be analysed at pre-specified time points using the subset of those still on treatment (while-on-treatment strategy). Alongside this analysis, the percentage of the study population who discontinued treatment, and the reasons for doing so, should be provided.

Explanation: in certain situations, only the PRO scores of patients still under treatment are of interest, for instance, in PRO scores that assess treatment side effects, such as nausea. To accurately interpret the PRO scores among patients still on treatment, it is important to take into account the percentage of patients who discontinued treatment and their reasons for doing so.

However, if there are notable tolerability issues after treatment discontinuation, such as long-term toxicities, the while-on-treatment estimand is not suitable and a different estimand strategy should be chosen.

AnalPres2_SAT

Statement: PRO scores may be analysed at a pre-specified time point using the subset of those still alive at a pre-specified time point. This should be accompanied by the percentage of patients included in the study who were alive at that time point (while-alive strategy).

Explanation: addressing death as an ICE should be aligned with the estimand of interest. The while-alive strategy is generally the preferred approach to address death when the aim is to describe PROs over time in a single arm study. This approach requires two sets of outcomes: first, descriptive statistics about death, such as an estimated percentage of patients still alive

at a certain time point (time-to-event outcome); and second, the PRO score (continuous or ordinal outcome) while alive.

For example, one can present the mean PRO at month six or the mean change in a PRO from the baseline until six months after initiation of treatment in the subgroup of patients still alive at month six, together with the probability of the patient still being alive at month six. The survival probability and the PRO score while alive can both be used for counselling patients and for shared decision-making between patients and HCPs.

AnalPres3_SAT

Statement: when making statements about a treatment in a single arm study based on changes over time (e.g., mean change from baseline, or change in proportions of responders, or time until improvement/worsening), consideration should be given to reasons other than the treatment that would result in y PROs changes over time. These factors include natural variations over time (e.g., due to disease worsening), response shift and the impact of concomitant therapies and comorbidities.

Explanation: patients' HRQoL or pain scores can vary over time for reasons unrelated to the treatment. This variation might be due to disease progression, which might result in declining HRQoL. Additionally, improvements in a respondent's HRQoL scores could stem from a change in internal standards, such as the patient learning better disease management strategies, or a shift in value prioritisation, for instance, where aspects spending time with family may compensate for the pain experienced. This phenomenon is known as a response shift. Furthermore, in un-blinded studies, patients might report improvements after receiving a novel therapeutic drug, irrespective of the drug's actual efficacy. Not taking this into account would lead to a biased (over- or under-estimated) treatment effect.

6. Results communication and visualisation

The statements for visualising PRO data from cancer clinical trials for a general audience can be used as follows:

- 1) Scientific figures can be adapted for a general audience by first creating them according to the scientific recommendation statements, and then modifying them for a plain language audience. When certain plain versions do not have corresponding scientific graphic types, such as Kaplan-Meier curves and forest plots, consider alternative visualisations, such as bar charts, to present the data effectively.
- 2) When corresponding scientific figures are not available, focus on the intended message for a general audience, using the statements on plain figures as a framework for the presentation of the figure. Simplified figures should align with the statements for scientific figures (e.g. a scientific statement says that graphs should not include different directionalities, and this should also apply to plain graphs).

An overview of the graph types is presented in Appendix 3.

6a. Scientific figure types

VizSciType1_GEN

Statement: bar charts are suitable representations for scientific figure versions in two cases: (i) to visualize the observed/estimated difference between treatment groups for a predefined PRO domain, and/or (ii) to show the observed/estimated data for each treatment group for a predefined PRO domain.

Explanation: the results for the treatment arms can be displayed as bars and their difference can be indicated numerically above/below the corresponding bars (e.g., as a delta). Alternatively, the difference itself can be shown as a bar. However, each PRO endpoint should be reported in separate bar charts. For instance, all magnitude of change endpoints should be grouped in one bar chart, and all time-to-event endpoints should be shown in another bar chart.

One can present up to six bars in a single bar chart. More bars could be included if well justified. If data is available and it corresponds to the predefined endpoint, provide graphs for different time points.

Present the results numerically, placing these in a way that makes it clear to which bar the number refers. In practice, depending on the length of the bar, place the numbers inside or outside the bar.

Data such as the proportion of patients for several time points can be depicted using several bar charts or one grouped bar chart. Include a definition of the reported proportion categories, e.g., in the caption. Percentages of categories should be indicated numerically and placed in a way that makes it clear to which bar they refer. In practice, depending on the length of the bar, place the numbers inside or outside the bar.

VizSciType2_GEN

Statement: line graphs are suitable representations for scientific figure versions to visualize continuous data over time, such as the magnitude of change for a PRO domain over time for each treatment arm.

Explanation: one can present up to four lines in a single line chart. More lines could be included in well justified cases. Use differentiating line symbols, such as symbols at each time point (e.g., a triangle for arm A, a rectangle for arm B), and/or different line styles (e.g., solid, thick, dashed, dotted) for treatment groups.

If predefined, highlight which time points significantly differ from each other. For statistical significance, you may mark these time points with an asterisk, whilst choosing a different marking method for meaningful changes/differences (e.g., a hashtag).

It is important to note that line charts can be misleading and give the impression of an inappropriate linearity between the data points shown, for instance in cases when there are very long intervals between the measurement points. Displaying connected dots might

be perceived as a trend for the population although the underlying sample might differ substantially between baseline and the end of treatment. If there is a high likelihood of these misinterpretations occurring, it is recommended to not connect the point estimates/ measurement points.

VizSciType3_GEN

Statement: Kaplan-Meier curves are recommended for scientific figure versions to visualize the probability of the occurrence of the PRO event (such as deterioration or improvement over time for each treatment arm in the context of right-censored time-to-event endpoints).

Explanation: when implementing Kaplan-Meier curves for time-to-event endpoints of PRO domains, one should follow the prespecified and applied methodology when including statistical details, such as showing the applied censoring method.

It is important to note that this recommendation is applicable to right-censored time-to-event endpoints. If the data is interval-censored, the Turnbull plot is the appropriate option since using a Kaplan-Meier curve will result in a biased visualisation.

Information that can be included into Kaplan-Meier curves

Between-arm details

The applicable effect estimate, such as a relative risk measure, including the confidence interval, should be indicated numerically within the figure.

Within-arm details

Display of the numbers at risk beneath the x-axis per arm for all time points. Optionally, the number of censored cases, number of events and number of patients without an event per arm can be added along the time axis. Additional data such as the overall number at risk, number of events, number of censored cases and the median time-to-event including confidence intervals can be presented per arm using a table placed in blank space of the graph.

Shading of confidence intervals

Consider that using shading to represent confidence intervals around the curves can add visual clutter. Therefore, shading should only be added if the attrition rate between the baseline and the last follow-up is considerably high. In such cases, confidence intervals (shading) can help to better understand the loss of information over time due to high attrition.

Other visual interpretation aids

Dashed lines representing the median time-to-event, and tick marks on the curves indicating censored cases, can be used to facilitate interpretation.

Definitions

The caption should include a definition of the event being investigated. If death is investigated as an event (e.g., time-to-deterioration of fatigue, or death), the suffix “survival” should be used throughout the figure, including the y-axis label and the captions, as well as in the text (e.g., fatigue-free survival). Additionally, to further facilitate interpretation, the caption could include information about the time points shown (such as whether they relate to an observed assessment schedule or a time window approach).

VizSciType4_GEN

Statement: forest plots are suitable representations for scientific figure versions to visualize multiple effect sizes.

Explanation: when implementing a forest plot for multiple PRO domains:

- provide clear labels on the horizontal axis to indicate which direction favours which study intervention. Provide a clear label for the applicable effect estimates on the horizontal axis (e.g., whether hazard ratios, risk ratios or risk differences are depicted), and provide a vertical null effect line. The vertical axis should indicate the relevant PRO domains of interest.
- use numerical data to report the number of events, relevant population level summary measures (e.g., hazard ratio, mean difference) and its corresponding variability measure (e.g., confidence intervals).

Within each forest plot, only effect sizes relying on the same analysis method/procedure should be depicted. When there are variations in the analysis components, such as ways of handling missing data or different thresholds defining the event for one specific PRO domain, provide a separate forest plot graph for each variation. In other words, do not combine variations for different PRO domains within one forest plot.

If thresholds for meaningful between-arm differences are applied, they must reflect whether a superiority, equivalence or non-inferiority objective is being investigated. These thresholds can be represented using vertical (dashed) lines.

Forest plots are not appropriate when investigating only a limited number of PRO domains. An acceptable minimum number could be three PRO domains.

6b. Considerations applicable to all scientific figures

VizSci1_GEN

Statement: figures depicting the main results should correspond to the pre-specified PRO domain(s) and time frame of the trial's PRO objective, as analysed according to the SAP. Additional figures may also represent exploratory/descriptive PRO objectives.

Explanation: the content of a figure should match the relevant PRO domain and its time point/time frame. This should be based on the study's PRO objective(s) and analysed according to the SAP. Specify in the figure title and axis label whether the data presented is observed or modelled (e.g., by mentioning "least square mean change" for modelled data).

Including several PRO domains in a single figure may lead to overcrowding and make it difficult to interpret the results. If it is necessary to include more than one PRO domain in the figure, it is important to ensure that the figure remains clearly understandable to

readers. This can be achieved by limiting the number of PRO domains included, or by applying a pre-specified rule to determine the order of the PRO domains/scales/items included in the figure.

Additional figures that represent exploratory/descriptive PRO objectives need to clearly indicate their exploratory or descriptive nature. This information can be included in the title or the caption of the figure.

VizSci2_GEN

Statement: the scaling applied to the graphs presenting PRO data should reflect the full PRO score range to ensure the data is represented without distortion. If a different scaling or only a selected range is used, the reasons for this choice should be justified.

Explanation: to ensure the data is represented without distortion, the chosen scaling applied to the graphs should reflect the full PRO score range. When multiple graphs are created based on the same PRO score or measure, it is crucial to use consistent scaling across all graphs whenever possible. This is particularly important when graphs are presented close together, as it prevents misleading comparisons. In justified cases, using a different scaling may be acceptable, for example, when relevant effects are not visible using the full range of scores/answer possibilities/percentages, or when a specific range is relevant. However, it is important to clearly indicate that the scaling shown reflects a selected section of the full range. For computer adaptive testing, adaptive scaling strategies should be used.

VizSci3_GEN

Statement: information on statistical significance (such as p -values and confidence intervals) should only be included in figures when they correspond to a predefined hypothesis, such as those presenting confirmatory PRO objectives.

Explanation: information on statistical significance should only be provided when appropriate, i.e., if the PRO objective is confirmatory. P -values should be positioned to clearly indicate what they refer to. To highlight statistically significant results, consider using formatting elements such as bold or italic fonts for p -values or adding symbols (e.g., an asterisk).

Results that are not controlled for multiplicity are considered nominal and not statistically significant.

If results from a statistical test ideally used in the confirmatory setting (e.g., p -values, confidence intervals) are used for exploratory or descriptive purposes, do not provide any information that may lead readers to conclude that the descriptive results provide the same level of evidence as confirmatory results. It should be clearly indicated that these results are for exploratory/descriptive purposes only and not for confirmatory purposes.

VizSci4_GEN

Statement: indicators should be included to show the directionality of PRO scores. The directionality of the PRO scores presented should adhere to the scoring and interpretation guidelines specific to the PRO measure used.

Explanation: for scoring and directionality information, the guidelines that apply to the selected PROM should be followed (as provided by the authors of the PROM). Use clear directionality indicators, such as clearly labelled arrows, explanatory sentences, or descriptive labels along the y-axis when data supporting their placement do exist, at least the extremes (e.g., “none” and “severe” or “very poor” and “very high”). Using a scoring direction of the given PROM other than the one recommended by the PROM developer should be avoided.

VizSci5_GEN

Statement: when using normative values, reference values or any other external data to support the presentation of PRO results, it is important to clearly and accurately indicate and label the data and its sources.

Explanation: if in line with the study objective and provided that appropriate reference data are available, norms or data from comparison populations can be included. Provide information on the external sources used in both the methods section and in the figure itself (e.g., in the legend/caption).

Normative data could be added, such as horizontal lines. Consider the trade-off between added interpretive value and greater visual complexity, particularly if the primary focus is comparing treatments.

VizSci6_GEN

Statement: figures should include clear, descriptive titles and labels to facilitate understanding.

Explanation: use clear, descriptive and appropriate titles and labels for each figure, such as “physical functioning in treatments A and B in the first six months of trial”. Whenever possible, avoid using double negatives as these are difficult to interpret. Use harmonised and consistent labels within each figure, across all figures, and throughout the manuscript.

Consider using labels for:

- the axes (x-axis and y-axis)
- directionality indicators (which direction indicates better/worse scores)
- curves/lines/bars
- numerical data (e.g., as percentages, numbers, raw scores)
- PRO domains/scales, symptoms
- investigated treatments (including the substance name of treatments; do not use brand names)

VizSci7_GEN

Statement: figures should provide information on sample sizes, ICEs and missing data.

Explanation: it is recommended to indicate:

- sample size: how many patients contributed data to the statistical analysis
- ICEs: how many patients were excluded from the analysis population due to the ICE strategy and
- missing data: how many patients were excluded due to missing data.

Present these numbers for each treatment arm or subgroup, each domain or symptom and each time point, e.g., as rows below the x-axis. In principle, the sum of these numbers should add up to the overall sample size of the analysis population, such as the intention-to-treat population, which should be indicated as well.

To improve the readability of the primary graph(s), consider presenting detailed reports of the frequencies of applicable categories of ICEs and/or missing data separately. This can be done using a stacked bar chart or a table.

The following should be included in the figure caption:

- A note clarifying the categories of ICEs included in this row that led to the exclusion of available data from the analysis population, such as death, progression, treatment discontinuation, start of new treatment, protocol deviation;
- A note clarifying the categories of “missing data” included in this row, such as administrative failure;
- A note on whether a separate reporting on ICEs and missing data is provided in a different chart or table.

The examples below illustrate how different strategies for handling ICEs can impact the actual number of ICEs that are ultimately listed in the graph.

In cases where the graph presents an analysis based on the while-on-treatment strategy, observations after the ICE “treatment discontinuation” might be excluded from the analysis although they are not missing. If the graph represents the confirmatory PRO analysis based on a treatment policy strategy not taking into account ICEs, the number of ICEs could be zero.

VizSci8_GEN

Statement: figures that provide information based on PRO score interpretation thresholds (meaningful changes and/or differences) should include the applied threshold.

Explanation: if relevant, indicate the applied threshold for meaningful changes/differences (e.g., with a note in the figure legend or caption).

Information on meaningful changes/differences can be provided visually and explained in the figure legend or in the caption. Visual elements can be, for example, a horizontal dashed threshold line, symbols (e.g., a hashtag), colour coding, highlighted text (e.g., text in italics or bold). Visual elements are preferred over text explanations to indicate meaningful changes/differences. Avoid using asterisks to indicate meaningful changes/differences as asterisks are commonly associated with statistical significance.

VizSci9_GEN

Statement: when presenting results for multiple PRO domains in one graph, it is not recommended to mix PRO domains that have different directionality.

Explanation: provide separate figures or clear separation within one figure for PRO domains with different directionality. Different panels within the same figure might depict different directionalities, but, within the panel the directionality should be consistent. Changing directionality of given PRO measures should only be considered in well justified cases, in order to maintain consistency.

VizSci10_GEN

Statement: figures visualizing results of confirmatory PRO objectives should provide information on statistical results that correspond to the predefined SAP and performed analyses.

Explanation: to facilitate appropriate interpretation of the figure, the following statistical details could be included in the figure, if applicable (non-exhaustive list):

- A numerical and/or visual representation of confidence intervals, which could be displayed using whiskers or shading;
- A numerical indication of p-values.

Pre-specified analysis results should be presented regardless of their statistical significance.

VizSci11_GEN

Statement: a figure should include the technical, numerical and statistical details and/or the terms necessary for its interpretation.

Explanation: figures should include the minimal amount of technical, numerical and statistical details necessary for their correct interpretation - such as *p*-values, confidence intervals, norms, delta, variability measures, sample sizes, intercurrent events (ICEs), and missing data. For the plain figure versions, further reduce, if possible, these technical details whilst still ensuring that the figures are easy to understand and to interpret.

VizSci12_GEN

Statement: visualisation of distributions should be provided for a scientific audience, when it is relevant for data interpretation, model assumptions or other reasons.

Explanation: if it supports the interpretation of the data or model assumptions, distributions should be presented to a scientific audience. Examples of commonly used options for the visual representation of distributions (requiring at least ordinal scaled data) are box plots, violin plots, histograms, cumulative distribution function curves or probability density function curves. Since there is no evidence supporting the use of any of these options, the users themselves must justify the choice of the type of graph, based on the properties of the distribution, the messages to be conveyed and the level of expertise of the scientific target audience.

VizSci13_GEN

Statement: the types of graphs recommended to visualize confirmatory PRO objectives can also be used for exploratory/descriptive PRO objectives. However, when these figures are used for exploratory/descriptive PRO objectives, information that may refer to any kind of formal statistical testing should not be included without clear indication of their exploratory/descriptive purpose.

Explanation: when the types of graphs recommended for visualisation of confirmatory PRO objectives (e.g. bar charts, line graphs, pie charts, icon arrays, Kaplan-Meier curves, forest plots) are used for exploratory/descriptive purposes, do not provide any information that may lead readers to conclude that the exploratory/descriptive results provide the same level of evidence as confirmatory results. Statistical test results ideally used in the confirmatory setting (such as p -values or confidence intervals) should be avoided. If it is necessary to use these statistical test results, it should be clearly indicated that these results are for exploratory/descriptive purposes only and not for confirmatory purposes.

6c. Plain figure types

Plain figure versions are a powerful tool for effectively communicating PRO data results from cancer clinical trials. They enable patients to better comprehend scientific papers, empowering them to engage actively with their healthcare journey. Clinicians can use these figures to facilitate discussions with patients, enhancing understanding of treatment outcomes and informed decision-making. Health journalists can also rely on plain figures to improve the clarity and accessibility of their writing, ensuring accurate and informative reporting on medical advancements and research findings.

Recommended graph types for plain versions

VizPlainType1_GEN

Statement: bar charts are suitable representations for *plain figure* versions in two cases: (i) to visualize the observed/estimated difference between treatment groups for a predefined PRO domain and/or (ii) to show the observed/estimated data for each treatment group for a predefined PRO domain.

Explanation: the results of the treatment arms can be shown as bars, with their difference indicated numerically above or below the corresponding bars. Alternatively, the difference itself can be shown as a bar. However, each PRO endpoint should be reported in separate bar charts. For instance, all magnitude of change endpoints should be grouped in one bar chart, and all time-to-event endpoints should be shown in another bar chart.

One can present up to six bars in a single bar chart. More bars could be included in well justified cases. If data is available and it corresponds to the predefined endpoint, provide graphs for different time points.

Present the results numerically, placing these in a way that makes it clear to which bar the number refers. In practice, depending on the length of the bar, place the numbers inside or outside the bar.

Proportion of patients for several time points can be depicted using several bar charts or a grouped bar chart. Include a definition of the reported proportion categories, e.g., in the caption. Percentages of depicted categories should be indicated numerically and placed in a way that makes it clear to which bar the number refers. In practice, depending on the length of the bar, place the numbers inside or outside the bar.

VizPlainType2_GEN

Statement: pie charts are recommended for *plain figure* versions for visualisation of proportion of patients (e.g., improvement, stable state or worsening) at a specific time point.

Explanation: according to available evidence, this recommendation only refers to the presentation of proportions of patients with a maximum of three categories (e.g., improvement, stable state, deterioration). Hence, do not use pie charts to depict frequencies of response categories of PRO measures with more than three response categories. More categories should only be shown in justified cases, as interpreting more than three categories might be harder. Use pie charts to highlight specific PRO domains at a specific time point. More than one time point might be presented, if these time points correspond to the predefined endpoint. Limit the number of time points presented to a minimum and avoid presenting multiple pie charts and/or pie charts for a large number of PRO domains.

Indicate the proportions (percentages) numerically on the pie chart. Include a definition of the depicted categories. Colour coding can be used to distinguish depicted categories. Use colours that are appropriate for individuals with colour vision impairments and that print well in grayscale.

VizPlainType3_GEN

Statement: icon arrays can be used for *plain figure* versions for visualisation of proportion of patients (e.g., improvement, stable state or deterioration) at a specific time point.

Explanation: use icon arrays to highlight specific PRO domains/scales at a specific time point. More than one time point might be presented if it corresponds to the predefined endpoint. Limit the number of time points presented to a minimum, and avoid presenting multiple icon arrays or icon arrays for a large number of PRO domains/scales.

Use common icons such as stick figures to represent people. Ensure that the icons used are culturally, ethnically and sex-appropriate, in order to accurately represent the study population.

Clearly indicate what each icon represents, such as absolute numbers or percentages. Indicate numerically the percentages of depicted categories, specifying what the number refers to (e.g., "XY % of patients with/without improvement/stable state/deterioration"). Colour coding can be used to distinguish depicted categories. Use colours that are appropriate for individuals with colour vision impairments and that print well in grayscale.

Be aware that although icon arrays are usually well received, some people might find them difficult to interpret. Common misinterpretations can include thinking that each depicted icon represents one patient.

VizPlainType4_GEN

Statement: line graphs are suitable representations for *plain figure* versions to visualize continuous data over time (e.g., the magnitude of change for the predefined PRO domains over time for each treatment arm).

Explanation: a line chart can display up to four lines. If well justified, more lines may be included. Use unique line symbols (i.e., symbols at each time point, such as a triangle for arm A, a rectangle for arm B) and/or different line styles (such as solid, thick, dashed, dotted) for each treatment group.

If predefined, highlight which time points significantly differ from each other. For statistical significance, you may mark these time points with an asterisk, whilst choosing a different marking method for meaningful changes/differences (e.g., a hashtag).

Line charts can be misleading and give the impression of an inappropriate linearity between the data points shown, for instance in cases when there are very long intervals between the measurement points. Displaying connected dots might be perceived as a trend for the population although the underlying sample might differ substantially between baseline and the end of treatment. If there is a high likelihood of these misinterpretations occurring, it is recommended to not connect the point estimates/measurement points.

VizPlainType5_GEN

Statement: to ensure that the general audience easily understands the plain language graphs, it is recommended that they are reviewed by at least one representative of the intended audience to assess their clarity.

Explanation: when reporting trial results, the patient representatives recommend that the plain language graphs are reviewed alongside the plain language summaries by representatives of their respective audiences, including patient organisations, patients, nurses, doctors.

6d. Considerations applicable to all plain figures

VizPlain1_GEN

Statement: for *plain figure* versions, technical, numerical, and statistical details and/or terms used should be kept to a minimum. Explanations of the meaning of these details and/or terms should always be included.

Explanation: any technical, numerical, statistical details and/or terms used should be clearly explained. These details and terms include, but are not restricted to, *p*-values, confidence intervals, norms, delta, variability measures, indicated sample sizes, ICEs and missing data, meaningful changes/differences. This can be done by providing plain language definitions as well as plain language explanations of what these details mean in the respective figures. Explanations could be added in the caption or in a separate note/appendix.

VizPlain2_GEN

Statement: *plain figure* versions should only have one predefined PRO domain per graph. If more PRO domains need to be included in one graph, the number of PRO domains included should be kept to a minimum.

Explanation: using separate graphs for each predefined PRO domain might facilitate understanding. For instance, instead of creating one overall graph with six bars/lines for three different PRO domains, consider creating three separate graphs with two bars/lines each.

This approach could lead to a high number of graphs. Furthermore, having a limited number of PRO domains within one single graph makes comparisons more difficult unless the graphs are displayed close together. One should consider the trade-offs between simplicity and the graph's ability to convey detailed information and facilitate comparisons.

VizPlain3_GEN

Statement: for *plain figure* versions, use common, everyday language.

Explanation: simplified, plain language should be used throughout the figure and the caption.

If technical, statistical and numerical information cannot be avoided, define and explain the terms. Make every possible effort to convey the messages using context- and audience-appropriate terms.

VizPlain4_GEN

Statement: for *plain figure* versions addressing confirmatory PRO objectives, information on statistical significance should be indicated using highlighting instead of numerical indication of p -values.

Explanation: numerically indicating p -values can be confusing for both patients and clinicians alike. Highlighting significant results may be more effective to convey data. If analysed, use formatting elements such as bold or italics fonts or add symbols for statistical significance (e.g., an asterisk). If statistical significance is reported by highlighting, there should be a simple explanation of what the highlighting and non-highlighting means in the specific case. To clarify where the comprehensive data is available, the legend for each figure should include a reference (within the trial report, to a publication; within a publication, to the figure or paragraph where the p -values are numerically reported). Care must be taken to avoid biased reporting. The PRO domains selected should be related to the main research question of the study. Selected PRO domains should be shown regardless of whether or not they are considered statistically significant.

General good advice for creating quality illustrations (not consensus-based)

The following general good advice was collated to highlight key additional considerations for creating quality graphic representations of PRO results. This content, unlike other recommendation statements, did not undergo the consensus process; however, it was subject to review, discussion, and acceptance by the SISAQOL-IMI consensus members. These practical issues are particularly important for interpretation by patients and their next of kin.

Colours: colours can be used to highlight certain details/data in the graph, create structure and facilitate interpretation. The graph should still be interpretable when shown in a greyscale version.

- Use colours purposefully and as sparingly as possible.
- Colour coding can be used to differentiate between variables such as treatment groups, responder categories or to highlight differences (e.g., better-worse, significant-non-significant). It can be used to distinguish scores reaching or not reaching a predefined threshold.
- Choose strongly contrasting colours so that they can be easily distinguished.
- Check whether the used colours appear on printouts as intended, as there might be differences in how colours appear on computer monitors and paper.
- If colours are used for a graph, ensure that they are colour vision impairment friendly and can also be clearly interpreted in greyscale/black and white. You may consider using the Viridis colour palette. Online tools are available to check what your graph looks like to colour-blind readers (<https://www.color-blindness.com/coblis-color-blindness-simulator/>).

A summary could accompany each graph to explain its content, enhancing accessibility for visually impaired users relying on text-to-speech tools. The language of the summary—scientific or plain—can be tailored to suit the intended audience.

Highlighting: highlighting of elements (e.g., using bold fonts, colours, asterisks) in the graph should be used purposefully and as sparingly as possible.

Figure caption: the figure caption should be concise and follow a clear structure.

Careful consideration should be given to what information needs to be included in the caption to make the graph understandable.

- Only information relevant to the interpretation of the graph should be included.
- Ensure that all information included is fitted to the target audience (e.g., scientific audience or plain audience).
- The caption should follow a clear structure, which can be obtained by grouping the information content-wise (e.g., have a descriptive figure title first; second, further information and explanations relevant to the graph; third, definitions of abbreviations used; fourth, references to other sources of information, if applicable).

Readability: the readability of all elements of a figure needs to be ensured. This includes choosing clear and plain fonts (e.g., sans serif fonts like Arial or Helvetica) and the largest possible font size.

Consistency: ensure consistency within and across graphs

- Design elements like fonts, font sizes and colours should be used consistently throughout individual graphs and across different graphs.
- Colour code-related information should be used consistently.
- A consistent order of data should be maintained across graphs, including the presentation of treatment arms in the legends, sample sizes, and other graph features.
- Language and terminology should be used consistently throughout the manuscript, the graphs and the caption.

Clutter: clutter should be avoided; only elements with a meaning should be used. If the meaning of an element is not immediately evident, this needs to be clearly explained.

References

- U.S. Food & Drug Administration (FDA). Non-inferiority clinical trials [Internet]. 2016 [Content current as of: 24 August 2018]. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidancedocuments/non-inferiority-clinical-trials>
- Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG, CONSORT Group. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA. 2012;308(24):2594–2604. <https://doi.org/10.1001/jama.2012.87802>
- European Medicines Agency (EMA). Choice of a non-inferiority margin – scientific guideline [Internet]. 2005 [Current effective version: 27 July 2005]. Available from: <https://www.ema.europa.eu/en/choice-non-inferiority-margin-scientific-guideline>

5 SISAQOL-IMI

recommendations according to the matrix in the interactive table



In this chapter, the recommendations are listed according to the matrix of the interactive table (Table 3). The table includes 30 cells, each representing a different study design, objective, and PRO endpoint of interest. The content of each cell can be accessed by clicking the interactive link. Users with a limited or unstable internet connection can generate printable PDFs by clicking on each cell for easy offline access.

Table 3. Number of recommendations according to design, objective and patient-reported variable of interest

PRO variable of interest	Randomised controlled trials (number of recommendations in each cell)			Single arm trials (number of recommendations in each cell)	
	Confirmatory objective		Descriptive objective	Confirmatory objective Superiority	Descriptive objective
	Superiority	Equivalence/ non-inferiority			
Magnitude of PRO (change) score at time <i>t</i>	68	65	59	79	74
Responder with PRO improvement at time <i>t</i>	70	65	61	81	76
Responder with PRO worsening at time <i>t</i>	70	65	61	81	76
Time to PRO improvement	64	60	57	71	69
Time to PRO worsening	65	61	58	73	71
Overall mean or median PRO scores over a specified time	38	37	35	49	47

There were 146 recommendations, but the recommendations may be included in multiple cells.
PRO: patient-reported outcomes

6. How the recommendations were developed



This chapter describes the processes for developing consensus-based recommendations by SISAQOL-IMI. The scientific work was organised in dedicated work packages (see [Chapter 7, Figure 4](#)).

The cross-cutting work packages played a crucial role in driving the project's success, overseeing essential activities such as project management, conducting the consensus survey, facilitating consensus meetings and re-voting processes, and coordinating the language review. Their work also included critical interactions with patient representatives and producing the final consensus reports. These efforts were key to ensuring the credibility of the guidelines, maintaining transparency throughout the process, and ultimately delivering high-quality, well-validated outcomes.

The framework for organising the statements

The SISAQOL-IMI built on the previous work carried out by the SISAQOL (Coens et al., 2020).

The previous SISAQOL developed recommendations for standardising the analysis and interpretation of PRO data in cancer RCTs. A key element common to all efforts to harmonise of evidence generation in clinical trials is well-defined research objectives. The SISAQOL taxonomy categorised PRO objectives and the different PRO variables of interest.

Read more about the SISAQOL taxonomy

Taxonomy of patient-reported outcome objectives

A common denominator to harmonise evidence collection in clinical trials is well defined research objectives. The SISAQOL taxonomy categorises patient-reported outcome (PRO) objectives by several defining attributes: the broad objective (confirmatory or

descriptive/ explorative), the between-group objective (superiority or equivalence/ non-inferiority), the within-group assumption (clinical benefit/ or exploratory) and the type of PRO variable of interest (e.g., magnitude of change, proportion of patients).

Each SISAQOL-IMI recommendation was organised using this taxonomy, cross-tabulating the attributes that define a PRO objective, as illustrated in the interactive table. The considerations related to the SISAQOL taxonomy of PRO objectives are outlined below.

○ **Confirmatory PRO objectives**

When a PRO is used to demonstrate treatment efficacy, clinical benefit or tolerability by providing formal comparative conclusions between treatment groups, confirmatory objective rules apply. An *a priori* hypothesis must be established for each PRO, which will be statistically tested upon the conclusion of the trial. If multiple PRO or multiple assessment points are considered, adjustments for multiple testing are needed.

Types of comparison: superiority or equivalence/ non-inferiority

When developing a confirmatory PRO objective, the type of comparison should be clearly specified for each PRO of interest. Comparisons may have either a superiority or an equivalence/non-inferiority objective. Design and analysis techniques for superiority differ from those for equivalence or non-inferiority. Non-significant *p*-values should not be used as evidence that the two treatment groups are similar (equivalent) or not worse (non-inferior).

A superiority PRO objective aims to demonstrate that, for the pre-specified PRO domain, the treatment group is superior to the reference group by a clinically relevant treatment effect size. This effect size should be pre-defined in the protocol, and the PRO score interpretation threshold for the relevant PROM should be applied. The trial design should allow for unbiased and sufficiently powered testing to reject the hypothesis of no treatment effect.

An equivalence or non-inferiority PRO objective aims to show that, for the pre-specified PRO domain, the treatment group is similar (equivalent) or not worse (non-inferior) than the reference group by a pre-specified, clinically relevant margin. These margins must be pre-specified in the protocol. The trial should be designed to enable unbiased and adequately powered testing to reject the hypothesis of non-equivalence or inferiority.

The choice of effect size (superiority) and margins (equivalence or non-inferiority) should be tailored to the PROM and clinical context, with justification based on both clinical and statistical grounds. Trials may combine these, but the protocol should clearly specify which PROM domains or items will be tested for superiority, equivalence, or non-inferiority.

○ **Descriptive (exploratory) PRO objectives**

When a PRO domain is used to capture the patient perspective during the trial—such as assessing tolerability, describing patients’ health-related quality of life, or exploring PRO data to inform future studies—the rules for descriptive or exploratory objectives apply. In these cases, an *a priori* hypothesis is not required. However, these outcomes cannot support comparative conclusions or serve as definitive evidence for treatment efficacy. Findings should be reported as either descriptive (i.e., summarising estimates with or without confidence intervals, without statistical testing), or exploratory (i.e., choice of hypothesis might be data-driven, with possible statistical testing), but not as evidence of treatment efficacy.

This framework constituted the template for organising the SISAQOL-IMI recommendations as provided in the interactive table and in Chapter 4 of this guidebook. With this framework, the users can focus on the statements applicable to their specific situation (i.e., to the PRO objective that applies to their study).

Discussions with partners

All SISAQOL-IMI institutions (Appendix 4) were invited to be a member of work packages, based on their individual interests. From the project launch, all partners had to understand each other’s needs and come to shared expectations on the final consensus recommendations. The objective was for partners to collaboratively define the problem scope and goals, clarify key topics and issues, identify concerns and expectations, identify areas of potential agreement and disagreement, and agree on a shared work plan towards building consensus recommendations.

To facilitate this, SISAQOL-IMI held a kick-off consensus meeting where partners agreed upon a development plan to define goals and priorities for each scientific work package. Partners could provide feedback to inform the development plan and future work throughout the process.

Literature reviews and existing guidelines

Within each scientific work package, the researchers conducted literature reviews, and collected relevant guidelines/protocols to provide an overview of the current state of practice in the design, analysis, interpretation, and presentation of PRO endpoints (Liu et al., 2023).

Read more about why the literature reviews were performed

Literature reviews, including a review of existing guidance documents, were conducted to provide an overview of the current practices in designing, analysing, interpreting and presenting PRO endpoints. Partners contributed relevant literature, critical guidance documents, and insights based on their expertise. The literature included information from regulators, such as assessment reports of PRO data submitted for cancer drugs and conceptually relevant qualification guidance. Relevant literature was collated and summarised, and areas of agreement and disagreement identified. Additional information on the literature reviews conducted in each work package is available in Appendix 5, and for SATs in the publication by Liu and co-authors (Liu et al., 2023).

The group reviewed the initial proposals, reaching a consensus on the proposed recommendations based on the findings from the literature reviews. They also identified any gaps in the literature in relation to current practices. They outlined strategies to address evidence gaps, and the proposals were scheduled for further discussion at the next consensus meeting or sooner, if necessary. The members identified both areas of agreement and disagreement, with suggestions on how the final consensus recommendations could resolve these disagreements.

Patient involvement

Patients' involvement is critical in the development of cancer therapies, allowing for more impactful delivery of patient-relevant outcomes (Hoos et al., 2015). Patients bring a deep understanding of their disease and treatment experiences, offering valuable insights into key aspects such as research objectives, study design, and the interpretation of PRO findings.

Patient representatives were involved in all aspects of the SISAQOL-IMI activities, including the development, review, and consultation on proposed statements, communication materials, web-based tools, and the other final outputs. They also contributed to reviewing communication and educational materials and tools designed for patients, carers, and patient organisations.

Read more about patient involvement

The involvement of patient representatives allowed the recommendations to be formulated and discussed in parallel at different statistical and technical levels. Their role was to ensure that proposed statements met their needs and were clear, relevant and understandable to stakeholders without a statistical background. They also contributed to developing the plain language version of the glossary to improve understanding and to facilitate the implementation of the recommendations.

SISAQOL-IMI organised several workshops and meetings with patients to encourage their input. To enable patients to contribute effectively to the design and interpretation of PRO data, complex statistical language was explained, discussed and adapted into an accessible format.

Virtual meetings were held before the surveys were conducted to explain the statements and allow patient representatives to ask questions to the working group core team. After the surveys, where one or more patient representatives had expressed disagreement, an additional meeting reviewed statements to clarify reasons for disagreement and to explain challenging content.

Face-to-face workshops were also conducted prior to the consensus meetings with patient representatives, clinicians, and other members to discuss proposed statements and the implications for patients and their next of kin. These meetings prepared and encouraged patient representatives to actively participate in the general assembly discussions. Additional activities to enhance patient involvement and the format of the plain language version of the final recommendations were also discussed.

During the SISAQOL-IMI project, an educational workshop for patient advocates was held at the Workgroup of European Cancer Patient Advocacy Networks (WECAN) academy. The workshop provided background on SISAQOL-IMI and, through an interactive session, demonstrated how analytical methods can influence data interpretation. Special emphasis was placed on intercurrent events to help patient advocates critically review cancer clinical trial protocols.

A patient workshop on PRO score interpretation thresholds presented specific examples of using these thresholds for different types of endpoints. Patient representatives contributed valuable insights, such as the importance of selecting meaningful outcome parameters, and using normative data to evaluate treatment burden alongside assessment of meaningful change.

A webpage was developed for patients and their representatives to share information about the SISAQOL-IMI project with the wider public. Videos were used to explain the consensus process and clarify complex terms and concepts. Consortium members summarised the Consortium's activities and achievements in short, plain language format.

The patient organisation Myeloma Patients Europe (MPE), representing the Workgroup of European Cancer Patient Advocacy Networks (WECAN), led efforts to ensure patient involvement and dissemination, aiming to promote wide and clear communication of SISAQOL-IMI activities and results.

Read more about the involvement of Myeloma Patients Europe

Myeloma Patients Europe (MPE) was appointed to represent patients on behalf of the Workgroup of European Cancer Patient Advocacy Networks (WECAN), an umbrella organisation representing 21 patient advocacy groups active in Europe. The participating individuals brought expertise from a wide range of areas, though some representatives changed over time.

MPE ensured timely, international promotion of the project's objectives, activities, and outcomes to all stakeholders involved in the clinical trial process. They also worked to maximise the project's visibility and impact within the community, reaching relevant stakeholders, end users, patient groups and the broader public.

Additionally, checklists and tutorial videos in plain language rendered the recommendations particularly relevant for patients accessible to the general public. A plain language version of the glossary was also included in the scientific documents (the guidebook and the interactive table) to explain technical concepts and terms clearly.

Developing the SISAQOL-IMI recommendations: a consensus-based approach

The generation of SISAQOL-IMI consensus recommendations followed a multi-step approach with a set of prespecified rules to ensure all participants were fairly involved. It was critical that all consensus recommendations were of “a minimum of acceptable, but still of high methodological quality, and feasible to implement”.

In total, five sets of consensus processes (including the kick-off and closing meeting) were held. The first consensus process focused on prioritising concepts (Appendix 6) (M Pe et al., 2023). The second and third consensus processes focused on recommendations related to RCT, SATs and PRO score interpretation thresholds, with the third process also including recommendations on how to communicate PRO results. The fourth consensus process focused on final updates to recommendations for RCT, SATs, and PRO score interpretation thresholds. Each consensus process followed a consistent pattern every year, starting in November/December and concluding with the final report in September of the following year.

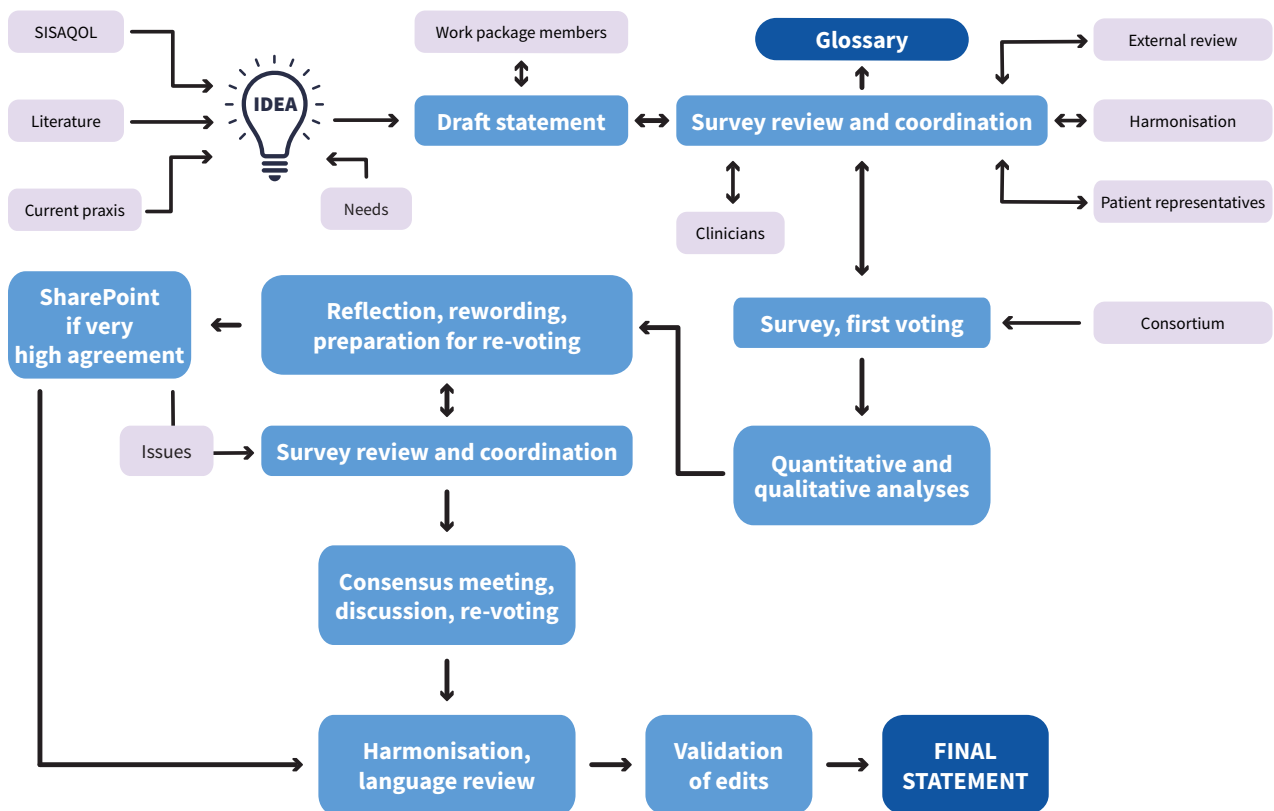


Figure 2. Overview of the consensus process

Source: Authors' own elaboration

To guarantee maximum transparency, a SharePoint site was created and made available exclusively to Consortium members. Each work package had dedicated sections for sharing documents and presentations. During the consensus processes, all relevant documents, including updates and corrections, were accessible to all members. Pre-published detailed timelines encouraged active involvement. Additionally, all members received personal e-mails inviting them to provide feedback on the final outputs.

Preparations for the consensus survey

The core teams of the consensus process developed standardised templates and guidelines for drafting statements. Proposals for new statements were submitted each December and reviewed by patient representatives, clinicians, and external Consortium members, along with the draft survey. Based on their feedback, WP leaders had another opportunity to revise the statements as needed.

Read more about preparations for the consensus survey

Each year, the scientific work packages (WPs) submitted new statements to the core consensus teams, along with important context and clinical examples for each statement. Two clinicians on the consensus core team conducted a qualitative review of the statements and examples, followed by discussions with patient representatives. Comments were then sent back to the WP leaders for text updates as needed. Draft copies of the consensus surveys were reviewed by the consensus core team and one or two external colleagues, who provided valuable inputs on statements, examples, scoring and layout. Based on this feedback, WP leaders updated the statements as needed.

Voting rules and the two-step voting process for the SISAQOL-IMI consensus survey

The Consortium agreed upon a set of voting rules where each of the 41 SISAQOL-IMI organisations had one vote. A statement was ratified if at least two-thirds agreed on the statement. The participants were encouraged to provide comments justifying their response.

To prepare the Consortium for the coming survey, a PDF of the voting text was sent out to all participating organisations two weeks before the online voting. All organisations voted digitally using the software Turning Point®. Comments, proposed changes, and the final proposal for the second vote were shared on slides sent out two-three weeks ahead of consensus meetings. During these meetings, the updated statements were presented, discussed, and re-voted on via a voting application, with results immediately visible to all participants.

Read more about voting rules and the two-step voting process for the SISAQOL-IMI consensus survey

The Consortium agreed upon a set of voting rules.

- Each SISAQOL-IMI organisation had one vote.
- For each statement, the possible options for agreement or disagreement were: “strongly agree”, “somewhat agree”, “neither agree nor disagree”, “somewhat disagree”, or “strongly disagree”.
- Two additional response options were available: “this statement is not applicable to my organisation” and “don’t know”.
- Respondents were encouraged to provide comments if they selected “neither agree nor disagree”, “somewhat disagree”, “strongly disagree”, “this statement is not

applicable to my organisation” or “don’t know”. Comments could also be provided when responding “strongly agree” or “somewhat agree”.

- In the analyses, “strongly agree” and “somewhat agree” were counted as agreement, while “neither agree nor disagree”, “somewhat disagree”, and “strongly disagree” were counted as no agreement.
- If the responses were “this statement is not applicable to my organisation” or “don’t know” were used, this was considered “abstaining”, and the organisation was excluded from the vote count (and was removed from the denominator).
- A statement was ratified if at least two-thirds of non-abstaining organisations agreed on it.
- If a ratified statement lacked two-thirds consensus within all stakeholder groups, comments from disagreeing organisations were analysed and discussed. When only one or few stakeholder groups disagreed with a statement, their reasons were carefully reviewed (comments were analysed and discussed and any misunderstandings were clarified), and possible compromises were explored.

The two-step voting process

First voting

To prepare the Consortium for the upcoming survey, the voting text was shared with all participating organisations two to three weeks before the online voting. Voting was conducted digitally using the Turning Point® Software. Reminders were carefully pre-scheduled and sent out as needed to ensure the timelines were adhered to. The consensus core group analysed both quantitative and qualitative results from the online survey, and tables with results were shared with the scientific work packages, who were asked to provide responses and propose changes if necessary. For issues specific to certain stakeholder groups, dialogue meetings were arranged to reach a better understanding and consensus. All comments, proposed changes, and the final proposal for the second voting were compiled into slides shared with all organisations several weeks before the consensus meetings.

Second voting

During the consensus meetings, the updated statements were presented, discussed and re-voted using a voting application, providing immediate results visible to all participants. Most statements received a very high level of approval. In a few cases, statements were reformulated during the meeting, requiring a third or fourth vote to achieve a higher consensus level. Very few statements were rejected; some were reformulated and revisited in the consensus process the following year.

Following each consensus meeting, the resulting statements were posted on SharePoint, together with the minutes from the meeting. This final step offered the last opportunity for input from the Consortium to correct any misunderstandings or mistakes.

Diverging views document

Achieving 100% agreement among Consortium members was not always possible when developing recommendations. Even with a two-thirds majority, in some cases, stakeholders voiced remaining concerns or differing views. This is common in large groups, and the goal was to address these views whilst preserving consensus. To avoid formal disclaimers from organisations that could affect the future use of the SISAQOL-IMI recommendations, the Steering Committee proposed a document outlining principles for managing divergent views within the Consortium (see Appendix 8).

Read more about the diverging views document

The SISAQOL-IMI diverging views document established some principles to address divergent views within the Consortium.

- Specific disclaimers for individual organisation were not included within individual recommendations. However, to accommodate varying perspectives and organisation needs within the Consortium, the following general disclaimer will be included in publications:

“This publication reflects the views of the individual authors and should not be construed to represent official views or policies of the European Medicines Agency (EMA), the US Food and Drug Administration (FDA), the US National Cancer Institute (NCI), the Medicines and Healthcare products Regulatory Agency (MHRA), the Institute for Quality and Efficiency in Health Care (IQWiG), Health Canada, the Norwegian Medicines Agency (NOMA), the American Society of Clinical Oncology (ASCO), the European Society for Medical Oncology (ESMO) or any other institution, organisation, or entity.”

- A preamble to the SISAQOL-IMI recommendations is included stating that reaching 100% agreement among Consortium members was not always possible. Readers should therefore consider that individual organisations may hold differing views on specific recommendations, reflecting their institutional or stakeholder perspective. SISAQOL-IMI recommend that readers consult relevant organisations or stakeholders when developing their clinical programmes.
- For each accepted recommendation statement where individual organisations expressed concerns or alternative views, it was noted that, while consensus was reached (with the percentage agreement reported), other perspectives exist. These relevant views were included under “considerations” alongside the accepted recommendation statement (Appendix 7).

Additionally, a table showing percentage agreement by stakeholder group will inform users about any stakeholder who held differing positions on a given recommendation statement (Appendix 8).

- Any remaining substantive concerns were noted under “considerations”, with the text for each recommendation statement reviewed and approved by all organisations.

Harmonising of statements between work packages

There was a substantial overlap between statements relevant to RCTs and statements relevant to SATs. Harmonisation meetings were held to identify statements relevant to both groups. These statements were adopted either with identical wording or adapted as needed (see Appendix 7). The statements also underwent a thorough editorial review, further detailed below (see section on language review). The statements were reviewed for clarity, readability and ease of understanding, and clinicians and patient representatives reviewed the clinical examples.

Read more about the harmonising of statements between work packages

Throughout the project, the SISAQOL-IMI prioritised harmonisation of statements, examples explanations, and all other text in the final deliverables, including the consistent use of glossary terms. Before the consensus surveys, efforts focused on avoiding unnecessary variations in language and definitions. During the preparations, native English-speaking Consortium members conducted extensive wordsmithing and informal language review to ensure consistent use of language, style, terms and acronyms.

A separate process was followed for the examples and explanations. The text was included in the surveys, and institutions were invited to provide comments and suggestions for improvements. Following this, the text was updated, published on SharePoint, and institutions could again offer comments and propose further changes. A second update was then followed by another review for language consistency.

Before the fourth consensus survey, all statements related to either randomised controlled trials or single arm trials were reviewed and discussed to determine their relevance to both groups. If applicable, modifications or minor changes were made, and they were included in the subsequent consensus survey.

The final harmonisation process was completed after the survey, before completing the “final outputs”.

Finally, a professional editor and proof-reader was contracted to perform a thorough language review for all statements, examples and explanations and all final output documents of the consortium. All proposed edits and modifications were submitted to the Consortium for approval.

Participating organisations and individuals

All 41 SISAQOL-IMI organisations participated in the consensus process, representing a range of stakeholder perspectives, including those of academic institutions, industry, non-profit/cancer organisations, small to mid-size enterprises/contract research organisations, regulatory bodies, HTAs, and patient representatives (see how the SISAQOL-IMI was organised in Chapter 7).

The SISAQOL-IMI glossary

The glossary was created to ensure consistent interpretation of all relevant terms used in the recommendations and to maintain harmonised terminology across the statements, explanations, and examples and all future SISAQOL-IMI documents (Pe et al., 2023; Joseph et al., in manuscript). The glossary defines all key terms and acronyms used in Consortium reports and publications. The final consensus-based glossary is comprehensive, containing 227 terms with both scientific and plain language definitions. The glossary is integrated in the interactive table and in the online version of this guidebook and can also be accessed as a separate document.

Read more about the development process and results of the glossary

The glossary defines all important terms and acronyms used in reports and publications from the Consortium, providing both scientific and plain language definitions.

Development process

To facilitate discussion among the stakeholders and to ensure harmonised terminology across the work packages' (WPs) recommendation statements, an extensive glossary was developed alongside the statements (Pe et al., 2023). A dedicated team was formed to coordinate this process by proposing scientific definitions for relevant terms identified in project documents and reports, including the proposed statements and examples and explanations. This approach aimed to ensure that the process of developing the glossary remained dynamic and consensus-based throughout the project.

In the early stages of development, the Consortium discussed whether to include only PRO-specific terms or also relevant generic terms. Determining a clear "cut-off" for terms deemed too generic proved challenging. Ultimately, it was agreed to include terms essential to the SISAQOL-IMI documents, making the recommendations more accessible to end-users.

A template for definitions was published on SharePoint and revised multiple times based on feedback from Consortium members, who reviewed and commented on identified terms, proposed definitions, and added terms. To streamline the process, each WP appointed a focal point for the glossary. The Consortium agreed to use a pre-defined hierarchy of well-recognised dictionaries in to define terms (see Appendix 9). WP leaders and the focal points provided feedback on the proposed scientific definitions, while patient representatives and members of MPE proposed plain language definitions, with support from expert members as needed.

Results

The glossary supported the harmonisation of the recommendations terminology, examples and explanations, and was included in the independent validation and language review. This consensus-based glossary contributes to a better understanding and smoother implementation of the SISAQOL-IMI recommendations.

The initial glossary version, published on SharePoint, comprised 98 terms and was gradually expanded and updated throughout the project. The final version now includes 227 terms, each with scientific and plain language definitions.

External validation and review

The SISAQOL-IMI project used different processes to ensure that the final outputs were relevant and of high quality.

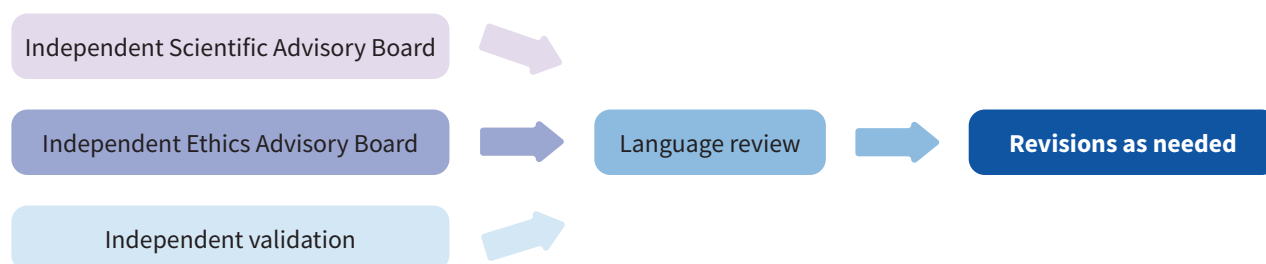


Figure 3. External validation and review

Source: Authors' own elaboration

Independent Scientific Advisory Board

An Independent Scientific Advisory Board provided critically reviewed the project's scientific quality and the methodological quality of the recommendations developed in the scientific work packages. Their input and subsequent dialogue helped clarify confusing and conflicting elements, significantly improving the final outputs.

Read more about the Independent Scientific Advisory Board

Though not involved in internal work package (WP) discussions, the Independent Scientific Advisory Board (ISAB) participated in all consensus meetings and was granted access to all internal documents. ISAB presented their review with questions and recommendations at two consensus meetings. WP leaders responded to the

initial report, and made adjustments as needed. This dynamic process culminated in a final ISAB report. The ISAB mid-term report reviewed all draft statements as part of the independent review processes and provided feedback to the WPs. They generally offered multiple specific comments to each WP. Highlights from this report, along with WP leaders' responses illustrating ISAB's inputs, are included in Appendix 10. No major revisions were required.

Independent Ethics Advisory Board

The Independent Ethics Advisory Board ensured the project maintained strong ethical standards throughout its duration. They verified that the use of data from closed and published trials fully complied with General Data Protection Regulation, and ensured rigorous ethical oversight was upheld.

Read more about the Independent Ethics Advisory Board

SISAQOL-IMI's Independent Ethics Advisory Board (IEAB) ensured compliance with ethical regulations by actively participating in project meetings, remaining available for ongoing advice on ethics, privacy, transparency, and publication policies. With at least one IEAB representative present at both virtual and in-person meetings, the board closely monitored activities and provided feedback. No complaints from or major ethical concerns were raised by Consortium members, reflecting a shared commitment to upholding high ethical standards.

Independent validation

The validation process involved two steps: 1) cognitive interviews and 2) independent validation of the preliminary statements by individuals outside SISAQOL-IMI ("blind members").

During the third consensus process, 17 cognitive debriefing interviews were conducted to assess the understanding of the SISAQOL-IMI statements. The scientific work packages reviewed the feedback and incorporated suggestions as appropriate into the revised statements.

Subsequently, a larger independent validation was conducted. This process validated the preliminary statements regarding study protocol writing, development of SAPs for PROs, and visualisation and presentation of PROs for both RCTs and SATs.

Read more about the independent validation process and results

Case studies were used to validate the SISAQOL-IMI recommendations, ensuring they were clear and feasible to implement in protocols and statistical analysis plans in cancer clinical trials.

Validation process

Validation was conducted by a group of independent experts (“blind members”) separate from those who designed these recommendations. These blind members tested and validated the initial set of recommendations using real-life case studies applying them to the study protocol and statistical analysis plans.

To assess the feasibility of using the recommendations and templates, three main questions were addressed:

1. Given the design for a new clinical trial, can we develop a protocol with a well-defined PRO objective fitting the research question, which can subsequently aid in the development of the SAP?
2. Given a protocol with a well-defined PRO objective, can we reproduce comparable SAPs?
3. Given PRO results based on the planned statistical analyses, how can the results be best visualised?

The validation process involved two steps. The first step was semi-structured interviews intended to evaluate whether the recommendations were understandable and logical to individuals outside the Consortium. Fifteen interviews were conducted with 17 participants, including eight statisticians, one clinician, four quality-of-life specialists and four other experts from seven academic institutions, two industry partners, seven representatives from regulatory/health technology assessment bodies, and one representative from a non-profit cancer organisation.

In each interview, participants discussed the recommendations. Most recommendations were interpreted as intended, though some difficulties arose mostly due to unfamiliar terminology or complex concepts. Longer recommendations were often perceived as challenging because of their information density. Participants suggested -adding explanations for terms and concepts in the glossary and in the explanation section of the recommendation. Insights from these interviews were shared with the teams developing the recommendations, who used this feedback to revise the statements. The results from these interviews were provided to the scientific work package leaders for consideration in re-voting.

As a second step, independent validation members (referred to as “blind members”) were identified within SISAQOL-IMI partner organisations and outside the Consortium. The blind members validated the preliminary recommendations by 1) developing the PRO part of a study protocol with a clearly defined PRO objective; 2) creating a SAP for

the PROs based on the study protocol provided, performing the relevant PRO analyses; and, lastly, 3) suggesting how best to visualise the obtained PRO results.

After completing these tasks, the documents produced by the blind members were evaluated against 'reference' documents created according to pre-specified criteria by the relevant SISAQOL-IMI work package. To identify similarities and differences between the reference case study and the results produced by the blind members, each blind member participated in a semi-structured interview or a focus group discussion. This helped uncover any difficulties encountered, aspects that were helpful, and areas for improvement.

Results of the independent validation

The independent validation results indicated that, overall, the recommendations were beneficial in drafting protocols and SAPs. The recommendations enhanced the understanding of PROs and increased confidence in protocol and SAP development. However, concerning practicality, the validation results suggested that the extensive list of statements may pose challenges during implementation, and generated suggestions to make the recommendations more accessible, informative, and user-friendly. The aim of this process was to streamline the implementation and encourage greater uptake by stakeholders by:

- providing a structured overview aligned with research objectives through a web tool, such as an interactive table, to help users focus on recommendations pertinent to their current endpoint.
- Providing tutorial sessions or webinars to guide users through the recommendations and offer practical insights, including real-world examples.
- Implementing information buttons or other design features within the web tool to display explanations and examples for each recommendation.
- Including dedicated sections with examples to clarify recommendations and facilitate implementation.
- Incorporating a comprehensive glossary to address complex terminology and improve the overall user experience.

The validation exercise also yielded specific feedback on certain recommendation statements. For instance, challenges were identified regarding the definition of intercurrent event strategies and a question on the treatment policy strategy for disease progression, as PRO data collection often stops at disease progression. It was suggested to include a disclaimer section recommending alternative approaches if PRO data is not collected in cases of disease progression or treatment discontinuation.

Statements indicating the avoidance of specific techniques (e.g. "avoid complete case analysis") can provide recommended strategies.

Concerns were also raised regarding the amount of information required for reporting on PRO score interpretation thresholds. Suggestions were made to differentiate between

recommendations that must be included in the protocol versus considerations for selecting appropriate PRO score interpretation thresholds, which do not need to be included in the protocol.

Agreements, potential problems and gaps were identified, with proposed solutions to address these issues. The WP leaders used the feedback from the independent validation to revise and improve the statements, examples and explanations as needed.

EMA qualification opinion

An extensive report, including the final RCT and PRO interpretation threshold statements, was submitted to the EMA for a qualification process. This work assisted SISAQOL-IMI in compiling the recommendation statements into the SISAQOL-IMI matrix of endpoints and trial PRO objectives.

EMA reviewed and gave feedback on the preliminary outputs of the SISAQOL-IMI work.

The EMA qualification fostered productive discussions on PROs with the Committee for Medicinal Products for Human Use of the European Medicines Agency (Silva M. et al., 2023). Although EMA concluded that SISAQOL-IMI recommendations fall outside the scope of the current qualification process due to the programme's evolving nature, the dialogue was valuable. Both parties explored alternative approaches to improve the use of PROs.

Language review

As a result of the consensus process, the documents describing the statements, examples and explanations, and other documents were increasingly voluminous. There was a need for a comprehensive language review to ease readability, to harmonise the writing style, and to ensure the consistent use of terms and descriptions. A freelance, independent copy editor and proofreader was hired to improve and substantially shorten the text. The main purpose of the language review was to eliminate redundancies, cross-check definitions used across statements and against the glossary, and ensure that the wording and writing style were as aligned as possible across all of the statements. To carefully maintain the wording agreed upon throughout the process, changes to the statements were minimal and were validated by work package leaders. Where possible, the explanations and examples accompanying the statements were condensed to avoid repetitions, and harmonised between work packages. A final set of changes was presented to work package leaders during a dedicated harmonisation meeting, and the final text was published on SharePoint for review by the Consortium. The final versions are easier to understand, less likely to be misinterpreted, and therefore more likely to be used as intended.

Read more about the language review

It was recognised early on that the project's final deliverables required more than a basic proofreading service, and a professional proofreader and editor was hired. The language editing followed a multi-step systematic approach. Initially, each statement and its examples and explanations were reviewed to remove redundant text, improve language, and correct spelling and grammar mistakes. Subsequently, the editor compared recommendations within and between various work packages to identify inconsistencies, contradictions, or unwanted variability in the text. Relevant words, terms, and definitions for the SISAQOL-IMI glossary were identified, ensuring consistent usage, with proposals for new terms or changes to existing ones as needed. Work package leaders thoroughly reviewed and validated all relevant edits.

To ensure consistent use of terminology across all of the outputs, the editor also reviewed the executive summaries, the final text of the interactive table, the guidebook, and the final publication.

References

- Coens, C., Pe, M., Dueck, A. C., Sloan, J., Basch, E., Calvert, M., et al. (2020). International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomized controlled trials: Recommendations of the SISAQOL Consortium. *The Lancet Oncology*, 21(2), e83–e96. [https://doi.org/10.1016/S1470-2045\(19\)30790-9](https://doi.org/10.1016/S1470-2045(19)30790-9)
- Hoos A, Anderson J, Boutin M, Dewulf L, Geissler J, Johnston G, Joos A, Metcalf M, Regnante J, Sargeant I, Schneider RF, Todaro V, Tougas G. Partnering with patients in the development and lifecycle of medicines: a call for action. *Ther Innov Regul Sci*. 2015 Nov;49(6):929-939. <https://doi.org/10.1177/2168479015580384>. PMID: 26539338; PMCID: PMC4616907
- Joseph K. L. H., Martinelli F., Lisa M. Wintner, M. L., ten Seldam, S., Claus, V., Belančić, A., et al. (manuscript in preparation). The SISAQOL-IMI Glossary: Promoting harmonised use of terminology for patient-reported outcomes in cancer clinical trials.
- Liu L, Choi J, Musoro JZ, Sauerbrei W, Amdal CD, Alanya A, et al. Single-arm studies involving patient-reported outcome data in oncology: a literature review on current practice. *Lancet Oncol*. 2023 May;24(5):e197-e206. [https://doi.org/10.1016/S1470-2045\(23\)00110-9](https://doi.org/10.1016/S1470-2045(23)00110-9)
- Pe M, Alanya A, Falk RS, Amdal CD, Bjordal K, Chang J, et al. Setting international standards in analyzing patient-reported outcomes and quality of life endpoints in cancer clinical trials—Innovative Medicines Initiative (SISAQOL-IMI): stakeholder views, objectives, and procedures. *Lancet Oncol*. 2023 Jun;24(6):e270-e283. [https://doi.org/10.1016/S1470-2045\(23\)00137-6](https://doi.org/10.1016/S1470-2045(23)00137-6)
- Silva M, Moseley J, Vetter T, et al. Patient-reported, observer-reported and performance outcomes in qualification procedures at the European Medicines Agency 2013–2018. *Br J Clin Pharmacol*. 2024 Jan;90(1):299-312. <https://doi.org/10.1111/bcp.15907>

7. How the SISAQOL-IMI was organised



SISAQOL-IMI international multi-stakeholder Consortium

The SISAQOL-IMI was an international multi-stakeholder Consortium designed as such to ensure that its consensus recommendations (i) addressed diverse stakeholder needs, (ii) maintained high methodological quality, (iii) were interpretable to both stakeholders with statistical and non-statistical backgrounds, including those without scientific knowledge, and (iv) were based on broad international consensus (Pe et al., 2023). The organisations involved in SISAQOL-IMI are listed in Appendix 4.

There was broad participation from patient representatives, 17 academic institutions, eight non-profit organisations representing small and large institutions across Europe and the United States (Appendix 4). Members also included representatives from international PRO initiatives like SPIRIT PRO, PROTEUS and STRATOS. Regulatory bodies, including FDA, EMA, and HTA representatives also participated. Five large international pharmaceutical companies and four representatives from small to mid-size enterprises (SMEs)/contract research organisations (CROs) were also actively involved.

[Read more about representation of stakeholders here](#)

Clinicians and patient representatives were included as Consortium partners to ensure that the proposed consensus recommendations were accessible to those without a statistical background. Each work package had dedicated patient representatives and was co-chaired by academia and pharmaceutical industry representatives.

Methodological experts provided comprehensive guidance on the design, analysis, interpretation, and presentation of patient-reported outcome data, ensuring that the strengths and limitations of various statistical approaches were clearly outlined, with transparent records of agreements and disagreements.

Management structure

The project's management structure consisted of the General Assembly, the Steering Committee, the Project Coordination team, the Independent Ethics Advisory Board, and the Independent Scientific Advisory Board. The members of the Independent Ethics Advisory Board and the Independent Scientific Advisory Board are listed in Appendix 11.

Five General Assemblies/consensus meetings were held, each focused on a specific milestone.

Table 4. Milestones of the General Assembly meetings

General Assembly meetings	Year	Milestones
1	2021	Defined the goals, priority of patient-reported outcome (PRO) objectives and identified expectations
2	2022	Ratification of the first set of recommendations for cancer randomised controlled trials, single arm trials and clinical meaningful change/ PRO score interpretation thresholds
3	2023	Ratification of the updated and expanded version of recommendations for cancer randomised controlled trials, single arm trials, visualisation and presentation of PRO results and for clinical meaningful change/ PRO score interpretation thresholds
4	2024	Ratifications of the final version of recommendations for cancer randomised controlled trials, single arm trials, visualisation and presentation of PRO results and for clinical meaningful change/ PRO score interpretation thresholds
5	2025	Ratification of the final output and sustainability plan

[Read more about the management structure here](#)

General Assembly

The General Assembly (GA) was the primary decision-making body, with each participating organisation (hereafter SISAQOL-IMI participant) represented by one voting member.

Co-chaired by the European Organisation for Research and Treatment of Cancer (EORTC) and Boehringer Ingelheim (BI), the GA was responsible for approving consensus recommendations and major strategic plans, confirming consensus meeting outcomes and endorsing the Consortium's composition. Members were encouraged to attend

in person, though virtual participation was always possible. SISAQOL-IMI participants were allowed to invite experts or qualified persons from their institute to attend the GA meetings as non-voting advisers to facilitate interaction. GA meetings operated under specific voting rules and quorum requirements set out in the Consortium agreement.

Project coordination team

The project coordination team comprised the coordinator (EORTC) and project leader (BI). They were supported by a deputy coordinator (UoL), a deputy project leader (Merck KGaA), a scientific coordinator (EORTC), a project manager (EORTC) and a finance-administrative manager (EORTC). Given their expertise, a patient representative and regulatory representatives from the European Medicines Agency were also involved. Among the tasks of the coordination team were communicating within the Consortium and with external partners, addressing project issues, ensuring the safeguard of data throughout the process and leading broader communication and dissemination efforts.

Steering Committee

Chaired by the coordinator (EORTC) and project leader (BI), the Steering Committee (SC) included WP leaders (both industry and academic co-leads), the Project Coordination team, a regulatory representative (EMA – non-voting) and a public representative (C-Path – non-voting). The SC met monthly to monitor project milestones, facilitate cooperation among work packages (WPs), and, where necessary, prepare amendments to the workplan for discussion by the GA. The SC was mandated to request specific actions or reports from the project coordination team and WP leaders to resolve outstanding issues.

If necessary, the SC could also establish task forces with specialised expertise to address specific areas requiring attention. Additionally, experts from the Consortium could be invited to participate in SC meetings as necessary.

The work package leaders

Each WP leader coordinated their respective WPs activities, ensured cross-WPs collaboration, and provided updates to the SC and GA. WP leaders ensured engagement of WP members, addressed their concerns, and ultimately ensured that members endorsed outputs.

The Independent Ethics Advisory Board

The Independent Ethics Advisory Board (IEAB) brought experts on ethics in clinical trials, medicine, and data protection to safeguard the ethical integrity of the project throughout its life cycle. IEAB placed particular emphasis on data protection from closed and published cancer clinical trials, particularly vis-à-vis compliance with the General Data Protection Regulation and patient confidentiality. EORTC had the expertise in applying General Data Protection Regulation rules, in relation to the case studies used to validate the recommendations. EORTC, holding legal responsibility for data used, appointed a data protection officer. Consortium members received

ongoing training in ethical practices relevant to this project. Members of the IEAB are listed in Appendix 10.

Independent Scientific Advisory Board

The Independent Scientific Advisory Board (ISAB) conducted independent reviews of the project's scientific quality. They were also responsible for ensuring the methodological quality of the consensus recommendations developed in each of the scientific work packages. The ISAB attended GA meetings as non-voting observers, providing inputs as necessary, and issued mid-term progress reports. Members of the ISAB are listed in Appendix 10.

Work package structure

The SISAQOL-IMI was organised into five scientific work packages, and three cross-cutting work packages, ensuring a collaborative approach to analysing PRO data in cancer clinical trials. (Figure 4)

[Read more about each work package](#)

WP1: Management and coordination

WP1 supported overall scientific, operational, and administrative coordination, including administrative and financial management, adherence to ethical standards, data protection regulations and internal communication.

WP2: Methodological work for cancer randomised controlled trials

WP2 established methodological recommendations for cancer randomised controlled trials (RCTs), assessing the strengths and limitations of existing methodological practices.

WP3: Feasibility of developing recommendations for non-randomised controlled trials, in particular single arm trials

WP3 developed methodologies for single arm trials (SATs), identifying valid PRO objectives and criteria for analysing PRO findings, whilst addressing bias in open-label studies and adapting WP2 recommendations to SATs where applicable, and vice-versa.

WP4: Communication tools for PRO findings from cancer clinical trials

WP4 developed templates for visualising and presenting PRO data identified for RCTs and SATs, focusing on integrating key metrics like sample size numbers on intercurrent events and missing data. Visualisation options with established scientific validation were selected, including pie charts, bar charts, and line graphs. (Synder et al., 2022; PROTEUS).

WP5: Independent validation of preliminary recommendations

WP5 developed a database of case studies including access to closed cancer clinical trials for both the pilot and independent validation studies. They coordinated blind teams to validate recommendations and communication tools, ensuring robust testing methodologies.

WP6: Develop international recommendations for the terminology and definitions of PRO score interpretation thresholds in cancer clinical trials

WP6 harmonised terminologies and definitions related to clinically meaningful change in cancer clinical trials (initially kept as an umbrella term but later replaced with the new term: PRO score interpretation thresholds), and clearly differentiating between group-level and individual-level change, and between group differences and change over time. WP6 matched terminologies and definitions to PRO objectives for RCTs and SATs, ensured collaboration with other work packages, and identified best practices for developing PRO objectives using relevant thresholds for PRO score interpretations.

WP7: Develop international recommendations for the design, analysis, interpretation and presentation of PROs in cancer clinical trials

WP7 developed a consensus process to ensure recommendations for the design, analysis, interpretation and presentation of PRO data represented all stakeholders, produced a glossary, and adapted final recommendations into scientific versions (interactive table and guidebook) and a plain language format for optimal accessibility.

WP8: Patient engagement, dissemination strategies and education programmes/workshops

WP8 focused on communication strategies, educational programmes and workshops, and tools for healthcare professionals, patients, and the general public to enhance understanding of PRO measures in cancer clinical trials.

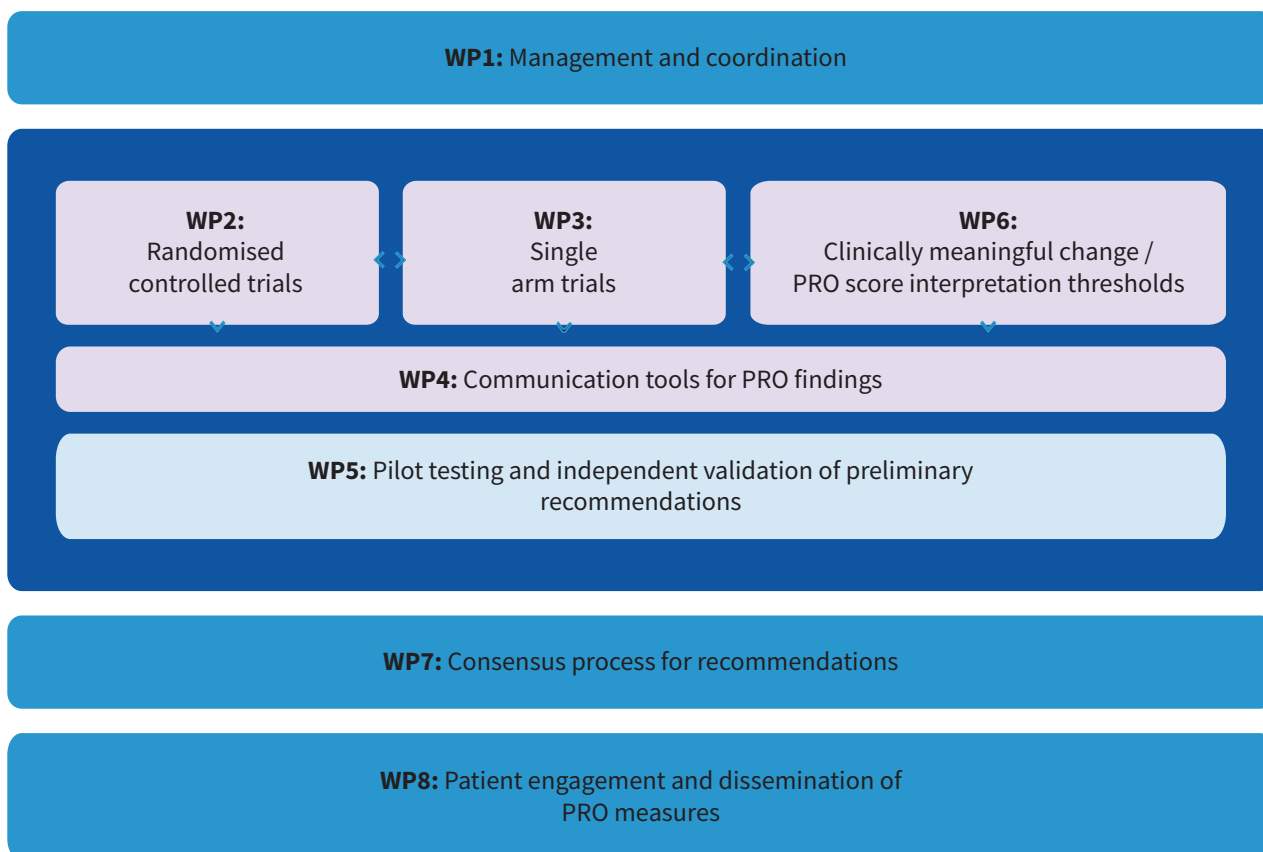


Figure 4. Organisation of work packages

Source: Authors' own elaboration

Each institution could have one or more participants in the Consortium, with members able to volunteer for various work packages. The development of statements in the scientific work packages is detailed in Appendix 5, while information on the work package on communication tools is provided in Appendix 3.

References

Pe M, Alanya A, Falk RS, Amdal CD, Bjordal K, Chang J, Cislo P, Coens C, Dirven L, Speck RM, Fitzgerald K, Galinsky J, Giesinger JM, Holzner B, Le Cessie S, O'Connor D, Oliver K, Pawar V, Quinten C, et al. Setting international standards in analyzing patient-reported outcomes and quality of life endpoints in cancer clinical trials—Innovative Medicines Initiative (SISAQOL-IMI): stakeholder views, objectives, and procedures. *Lancet Oncol.* 2023 Jun;24(6):e270-e283. [https://doi.org/10.1016/S1470-2045\(23\)00157-2](https://doi.org/10.1016/S1470-2045(23)00157-2)

8. Lessons learned from the consensus project



International consensus processes involving diverse stakeholder groups and renowned experts are challenging but highly rewarding. The enthusiasm and participation of world-leading scientific experts, many with decades of experience and involvement in developing other international guidelines, were essential. Strong participation from patients and their representatives was a key aspect, ensuring the relevance of the recommendations and supporting future implementation.

The project provided a unique platform for dialogue through biannual meetings—one virtual and one in-person—where all participants had equal opportunities to pose questions and provide input, independent of experiences or scientific status. Digital meetings proved valuable, both during the COVID-19 pandemic and for participants facing travel restrictions. Digital meetings also facilitated multiple smaller meetings between collaborators within and across work packages. This ongoing dialogue was crucial for the consensus process.

The in-person meeting was essential to deepen connections, encourage open input, and foster a collaborative spirit. It allowed discussions from virtual meetings to progress naturally, with informal conversations during coffee breaks and dinners helping to bridge disagreements. This face-to-face interaction was invaluable for building trust and shared commitment to the project's goals.

In addition to developing RCT and SAT recommendations, there were extensive discussions regarding clinically meaningful changes over time and differences between groups. Several proposed statements and explanations and examples required modifications before receiving final approval. To address the existing heterogeneity in terminology and concepts related to thresholds for interpreting differences and changes in PRO scores, the Consortium established clear terms and definitions to differentiate the various types of thresholds for interpreting individual patient PRO scores from those used for group-level interpretations, such as a mean difference between two trial arms at a specific time point.

Significant effort was devoted to developing recommendations and best practices for visualising PROM results. Discussions focused on balancing the need for details with avoiding overly cluttered figures and ensuring the illustrations were understandable to non-specialists. The recommendations

underwent a formal independent validation process, including qualitative interviews and testing within a test protocol and SAP. This process allowed each work package to review and refine their recommendations based on the validation results before finalising them.

The scientific efforts of the different work packages were inspiring, leading to several current and future publications. These included literature reviews (Liu et al., 2023), validation of statements using case studies (Thomassen et al., 2023, and Glossary development Joseph et al., in manuscript). Further publications are planned for 2025, including but not limited to the use of external controls in SATs and addressing missing data (WP3) and death as intercurrent event (WP2).

From the outset, ensuring true consensus rather than a “majority wins” approach was crucial. This required careful consideration of the views of specific stakeholder groups or institutions that represented a dissenting minority when recommendations were approved. For those groups sceptical of one or more recommendations, focused discussions were held to reach an acceptable compromise. Resolving misunderstandings and disagreements as soon as they arose improved adaptation and helped achieve a final consensus. All work packages adhered to this approach throughout the recommendations development process. A formal “diverging views document” document was approved and made available, instead of including disclaimers from institutions on many recommendations.

A few important issues required further consideration. One key question was whether the guidance should be mandatory (“must”) or simply advisory (“should”). Members opted for the latter, including a clause stating “any deviations should be justified”. Additionally, there was a question of whether to take a pragmatic approach, by providing guidelines with a high probability of implementation (“p-should”), or to raise scientific standards by encouraging researchers to obtain higher quality work by providing the ideal solutions to many of the questions (“i-should”)? An important motivation for the project was to demonstrate that PROs could and should meet the same scientific standards as other clinical endpoints. For some Consortium members, achieving the latter option was important. PROs have often been undervalued in terms of validity, reliability, and applicability of the methods. The SISAQOL-IMI agreed that all consensus recommendations meet minimal acceptable standards while maintaining high methodological quality, acceptable to all stakeholder groups, and feasible for implementation. In cases where consensus could not be reached, SISAQOL-IMI provided clear justifications for the minimum standards of methodological quality. As a compromise, a clause stating, “any deviations should be justified” was included in statements where the ideal situation was often not feasible.

Finally, extensive discussions took place regarding intercurrent events, focusing on both definitions and how to handle various types of intercurrent events. Addressing issues such as death and dropouts proved particularly challenging. Once again, the diverse backgrounds of participants enriched these discussions.

A project of this scale is highly resource-intensive. Strict timelines were established up to 12 to 18 months in advance to ensure the availability of participants with busy schedules. Strong

encouragement, along with close follow-up via emails and calendar invites, was necessary to keep everyone engaged and on track.

Early in the project, it became evident that some disagreements stemmed from misunderstandings about the terms and definitions used in the documents. As a result, the Consortium grew interested in creating a comprehensive glossary, recognising that consensus required a shared understanding of terminology. The glossary, containing over 200 terms, was thoroughly discussed, revised and reviewed multiple times before reaching final approval.

Another key issue was the need to harmonise statements and recommendations for both RCTs and SATs, as many recommendations applied to both oncology research settings. Where possible, the same text was used for both (“adopted”), or slightly modified (“adapted”). This process was supported by a thorough professional language review and copy-editing, resulting in a final text with consistent structure and terminology, now easier to read and understand.

The success of a project of this magnitude depends heavily on the level of implementation achieved. Thus, the extensive sustainability plan and planned training activities are key to reaching the project’s goals. Continuous interest and support from stakeholders provide a strong foundation for this effort.

References

- Liu L, et al. Literature review on patient-reported outcomes in oncology: recommendations for the future. *J Clin Oncol*. 2023 May;41(15):1379-1390. <https://doi.org/10.1200/JCO.23.00013>
- Pe M, et al. Setting international standards in analyzing patient-reported outcomes and quality of life endpoints in cancer clinical trials—Innovative Medicines Initiative (SISAQOL-IMI): stakeholder views, objectives, and procedures. *Lancet Oncol*. 2023 Jun;24(6):e270-e283. [https://doi.org/10.1016/S1470-2045\(23\)00157-2](https://doi.org/10.1016/S1470-2045(23)00157-2)
- Thomassen D, Roychoudhury S, Amdal CD, Reynders D, Musoro JZ, Sauerbrei W, Goetghebeur E, le Cessie S, et al. The role of the estimand framework in the analysis of patient-reported outcomes in single-arm trials: a case study in oncology. *BMC Med Res Methodol*. 2024;24(1):29069. <https://doi.org/10.1186/s12874-024-02408-x>



9. The future

Sustainability and further plans

The outputs of the SISAQOL-IMI project will be sustained by the SISAQOL-IMI network and coordinated by EORTC. The recommendations may be revised in the future based on feedback, critique, and new evidence. While no direct extension of the SISAQOL-IMI project is currently planned, the recommendations may be expanded upon and updated to address new topics through potential future follow-up projects, should the need arise.

The Consortium welcome comments and suggestions. Please feel free to contact us through the website at www.sisaqol-imi.org. Additionally, information on trainings and educational tools related to the SISAQOL-IMI recommendations are available on the website.

The SISAQOL-IMI will develop a clinical trial protocol template and a SAP templates to exemplify how the SISAQOL-IMI recommendations can be integrated into these essential trial documents. These templates will systematically incorporate PRO elements, presenting them in a logical sequence with example text aligned with the SISAQOL-IMI guidelines. The templates will be available for access on the project website.

Glossary – sustainability plan

Updating the glossary after the publication of the recommendations is a key aspect of the sustainability efforts, helping to ensure that the recommendations remain harmonised, complete, and practical.

The glossary will be revised regularly, based on user feedback, with updates scheduled, for example, once a year. All relevant information about the updates, such as the dates of the updates and which terms or definitions have been revised, will be transparently communicated to users.

The team will develop a user-friendly platform for providing feedback on terminology and definitions to ensure the glossary remains current and practical for end users. Several reference sources were used to create the glossary, such as the ‘Glossary of ICH terms and definitions’ (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use [ICH], 2023; ICH, 2024). Keeping the SISAQOL-IMI glossary aligned with updates from these reference sources will also be crucial to maintain its accuracy and relevance.

References

- International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). *Glossary of ICH terms and definitions (Version 3) [Internet]*. 2023 [cited YYYY MMM DD]. Available from: <https://www.ich.org/page/glossary-of-ich-terms-and-definitions>
- International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). *Glossary of ICH terms and definitions (Version 6) [Internet]*. 2024 [cited YYYY MMM DD]. Available from: <https://www.ich.org/page/glossary-of-ich-terms-and-definitions>



10. List of tables, figures and appendices

Table 1. Overview of available guidance on PRO in clinical research	Page 15
Table 2. The SISAQOL-IMI analytical framework	Page 31
Table 3. Number of recommendations according to design, objective and patient-reported variable of interest	Page 105
Table 4. Milestones of the General Assembly meetings	Page 125
Figure 1. Structure of the interactive table	Page 23
Figure 2. Overview of the consensus process	Page 112
Figure 3. External validation and review	Page 118
Figure 4. Organisation of work packages	Page 129
Appendix 1 Harmonised statements between RCTs and SATs	
Appendix 2 Original statements for each PRO variable of interest presented as condensed summary statements	
Appendix 3 Allocation of graph types to PRO endpoints	
Appendix 4 List of participating organisations in SISAQOL-IMI by stakeholder group	
Appendix 5 Development of statements in the scientific work packages	
Appendix 6 SISAQOL-IMI priority of objectives for statements development	
Appendix 7 Overall divergent views table	
Appendix 8 Table of diverging views	
Appendix 9 Hierarchical list of references/sources used to define the SISAQOL-IMI glossary terms	
Appendix 10 ISAB Midterm Report highlights	
Appendix 11 Members of the Independent Ethics and Scientific Advisory Boards	

The Appendices are available separately on SISAQOL-IMI's [website](#).

