EORTC
European Organisation for Research and Treatment of Cancer

Improving standards of health-related quality of life and patient reported outcomes analysis: a SISAQOL initiative

Corneel Coens

Lead Statistician

Quality of Life Department, EORTC HQ

on behalf of the SISAQOL consortium

The future of cancer therapy

SISAQOL Consortium

Ethan Basch[1]
Andrew Bottomley[2]
Melanie Calvert[3]
Alicyn Campbell[4]
Charles Cleeland[5]
Kim Cocks[6]
Corneel Coens[2]
Laurence Collette[2]
David Collingridge[7]
Nancy Devlin[8]
Lien Dorme[2]
Amylou Dueck[9]
Hans-Henning Flechtner[10]
Carolyn Gotay[11]
Eva Greimel[12]
Ingolf Griebsch[13]
Mogens Grønvold[14]
Jean-Francois Hamel[15]
Laura Lee Johnson[16]
Madeleine King[17]
Paul Kluetz[16]

Michael Koller[18]
Daniel C Malone[19]
Francesca Martinelli[2]
Sandra A Mitchell[20]
Jammbe Z Musoro[2]
Daniel O' Connor[21]
Kathy Oliver[22]
Madeline Pe[2]
Elisabeth Piault-Louis[4]
Martine Piccart[23]
Francisco Pimentel[24]
Chantal Quinten[25]
Jaap C Reijneveld[26]
Christoph Schürmann[27]
Jeff Sloan[28]
Ashley Wilder Smith[20]
Katherine M Soltys[29]
Rajeshwari Sridhara[16]
Martin Taphoorn[30]
Galina Velikova[31]

[1] Lineberger Comprehensive Cancer Center; University of North Carolina, USA
[2] European Organisation for Research and Treatment of Cancer, Belgium
[3] Centre for Patient Reported Outcomes Research, University of Birmingham, UK
[4] Genentech, San Francisco, USA
[5] Dept. Of Symptom Research, MD Anderson Cancer Center; University of Texas, USA
[6] Adelphi Values, UK
[7] The Lancet Oncology, UK
[8] Office of Health Economics, UK
[9] Alliance Statistics and Data Center; Mayo Clinic, Arizona, USA
[10] Clinic for Child and Adolescent Psychiatry and Psycohtherapy; University of Magdeburg, Germany
[11] School of Population and Public Health; University of British Columbia, Canada
[12] Dept. Of Obstetrics and Gynecology; Medical University of Graz, Austria
[13] Boehringer-Ingelheim, Germany
[14] Dept. Of Public Health; Bispebjerg Hospital, University of Copenhagen, Denmark
[15] Methodology and Biostatistics dept.; University Hospital of Angers, France
[16] US Food and Drug Administration, USA
[17] School of Psychology and Sydney Medical School; University of Sydney, Australia
[18] Center for Clinical Studies; University Hospital Regensburg, Germany
[19] College of Pharmacy, University of Arizona, USA
[20] National Cancer Institute, Bethesda, USA
[21] Medicines and Healthcare products Regulatory Agency, UK
[22] International Brain Tumour Alliance, UK
[23] Institut Jules Bordet; Université Libre de Bruxelles, Belgium
[24] Blueclinical Phase I, Centro de Estudos e Investigação em Saúde da Universidade de Coimbra, Portugal
[25] European Centre for Disease Prevention and Control, Sweden
[26] VU University Medical Center, Dept. of Neurology & Brain Tumor Center, The Netherlands
[27] Institute for Quality and Efficiency in Health Care, Germany
[28] Alliance Statistics and Data Center; Mayo Clinic, Rochester, USA
[29] Health Canada, Canada
[30] Leiden University/Haaglanden Medical Center, The Netherlands
[31] Leeds Institute of Cancer and Pathology; University of Leeds, St. James's Hospital, UK

Disclaimer

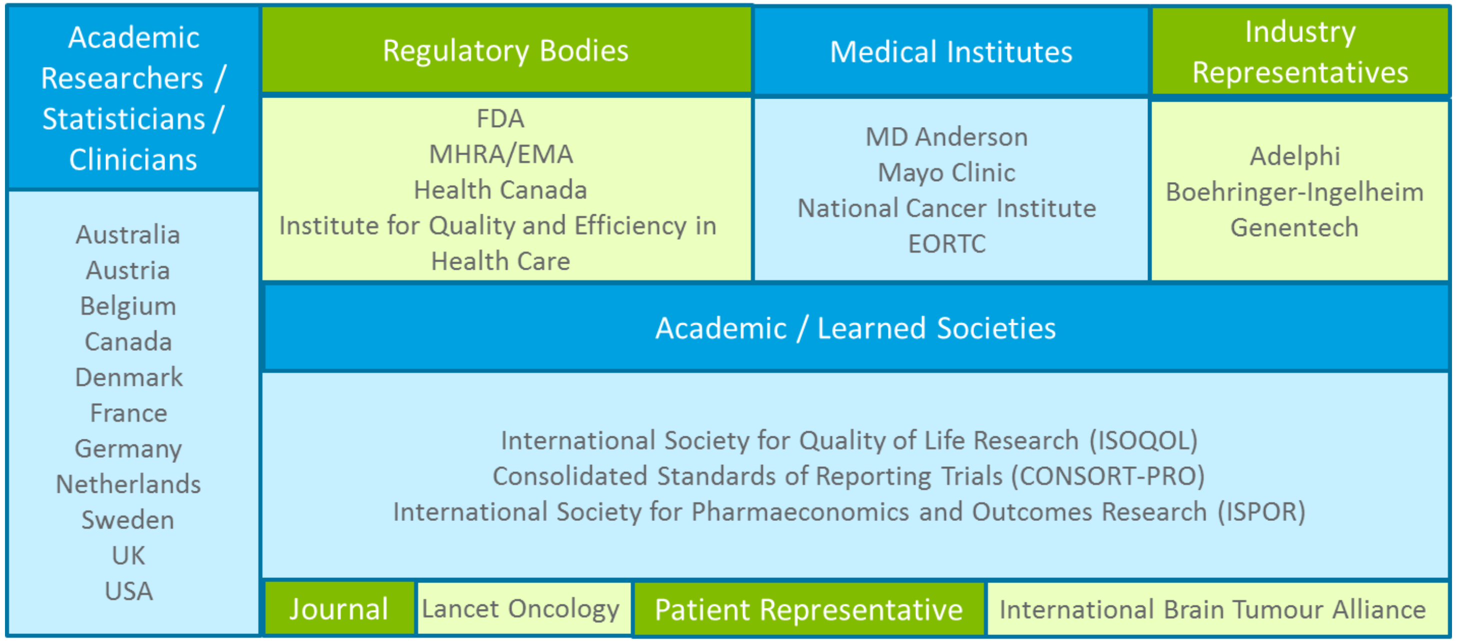
The views here reflect that of the individual authors and should not be construed to represent official views or policies of the US Food and Drug Administration, US National Cancer Institute, Medicines and Healthcare products Regulatory Agency, Institute for Quality and Efficiency in Health Care, Germany or Health Canada.

Acknowledgements

EORTC received an unrestricted education grant from Boehringer Ingelheim GmbH to initiate the SISAQOL work, with additional financial support provided by Genentech, Roche and the EORTC Cancer Research Fund

What is SISAQOL

- SISAQOL
  - **S**etting **I**nternational **S**tandards in **A**nalyzing Patient-Reported Outcomes and **Q**uality **o**f **L**ife Endpoints Data
  - International multi-stakeholder consortium with shared interest in improving the standards for the **statistical analysis** of Patient-Reported Outcomes (PRO)
  - Current Focus: randomized clinical trials (RCT) in oncology.



Academic Researchers / Statisticians / Clinicians

Australia
Austria
Belgium
Canada
Denmark
France
Germany
Netherlands
Sweden
UK
USA

Regulatory Bodies

FDA
MHRA/EMA
Health Canada
Institute for Quality and Efficiency in Health Care

Medical Institutes

MD Anderson
Mayo Clinic
National Cancer Institute
EORTC

Industry Representatives

Adelphi
Boehringer-Ingelheim
Genentech

Academic / Learned Societies

International Society for Quality of Life Research (ISOQOL)
Consolidated Standards of Reporting Trials (CONSORT-PRO)
International Society for Pharmaeconomics and Outcomes Research (ISPOR)

Journal: Lancet Oncology

Patient Representative: International Brain Tumour Alliance

What is the issue?

A common PRO objective:

“Treatment A will improve physical functioning relative to Treatment B”

Which statistical method would be appropriate to test this objective?

t-test, linear regression, ANOVA, repeated measures ANOVA, Mann-Whitney, linear mixed model, generalised estimating equation, joint longitudinal model, pattern mixture model, log-rank test, Cox proportional hazards, Chi-square test, Fisher’s exact test, Cochran-Mantel Haenszel test, logistic mixed model, area under the curve… and many more…

Hamel et al. EJC 2017. “A systematic review of the quality of statistical methods employed for analysing quality of life data in cancer randomised controlled trials”

What is the solution?

PRO data in oncology trials have specific characteristics:

- Multidimensional
- Longitudinal
- Missing data (likely informative)

Such data require appropriate analysis procedures, which are rarely implemented in a standardised manner.
Methodologically appropriate and consistent approach is needed.

Major hurdles are:

- PRO objectives need to be clearly defined.
- Terminology is not consistent

Taxonomy of research objectives

Consensus: Clearly state the broad PRO research objectives intention for each PRO domain(s)/item(s) of interest:

- **Treatment efficacy / clinical benefit:** confirmatory objective therefore conclusions regarding comparisons between treatment arms can be drawn.
  - *a-priori* hypothesis needed
  - Statistical test - correction for multiple testing needed
  - Conclusions regarding comparisons between treatment arms

- **Describe patient experience**: Exploratory/descriptive objective therefore only presentation of findings but no comparative conclusions between treatment arms can be drawn
  - No *a-priori* hypothesis needed
  - Descriptive / exploratory - multiple testing is not an issue
  - No comparisons between treatment arms

Taxonomy of research objectives

Consensus: Pre-specifying superiority, equivalence and non-inferiority: clearly state for each objective PRO domain/item of interest will be used to provide evidence for:

- **Superiority**
- **Equivalence**
- **Non-inferiority**

A non-significant superiority result should not be interpreted as evidence of equivalence or non-inferiority.

Taxonomy of research objectives

Consensus: Valid PRO objectives at the within-individual / within-treatment level are the following:

Treatment efficacy / Clinical benefit

- **Improvement**
  - Time to improvement
  - Proportion of patients with improvement at time *t*
  - Intensity of improvement at time *t*

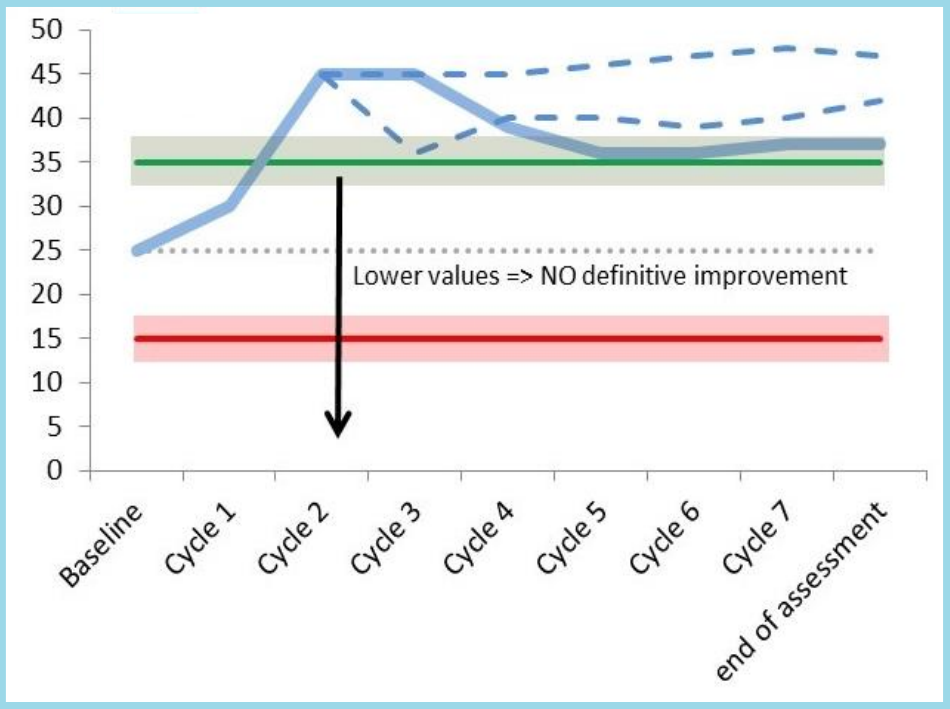
- **Worsening**
  - Time to worsening
  - Proportion of patients with worsening at time *t*
  - Intensity of worsening at time *t*

- **(End of) Maintenance**
  - Time to (end) of maintenance
  - Proportion of patients with maintenance at time *t*

- **Overall effect**
  - Overall PRO score over time (e.g., assessed by overall means, area under the curve, best / worst response)

Definition of Improvement

Consensus: Definition of improvement + duration

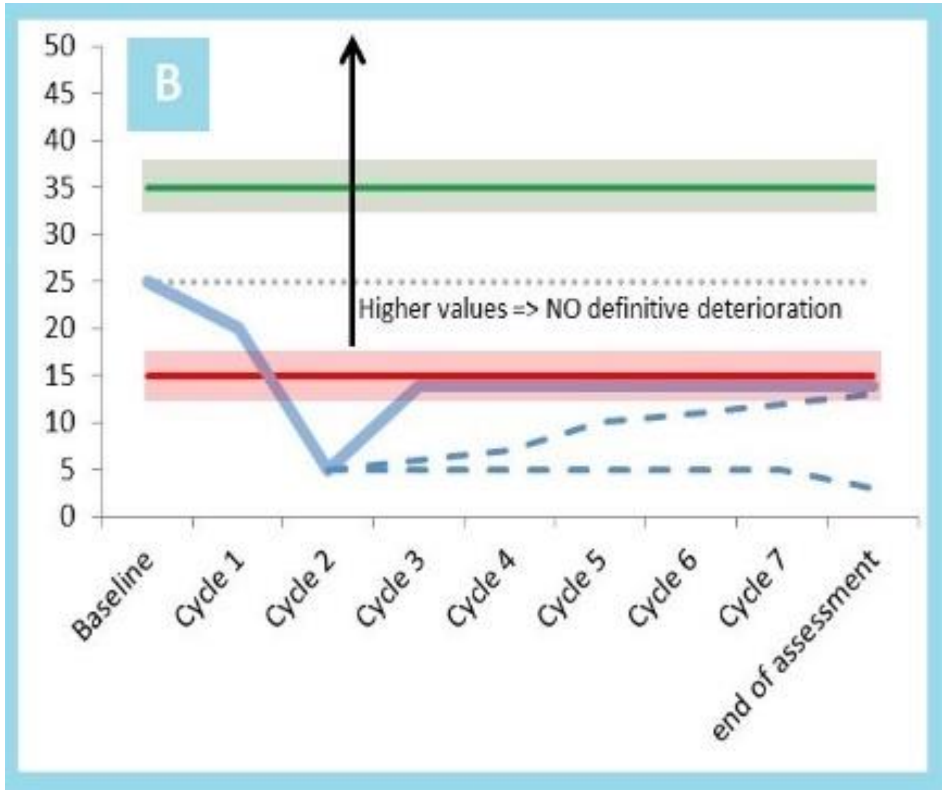


Improvement: Change from baseline that reaches a pre-defined improvement threshold level (post-baseline improvement).

Duration of improvement: Improvement is maintained if follow-up assessments remain at or is higher than the improvement threshold (taken from definitive improvement definition). Improvement is discontinued once a follow-up assessment is below the improvement threshold (taken from the transient improvement definition).

Definition of Worsening

Consensus: Definition of worsening + duration

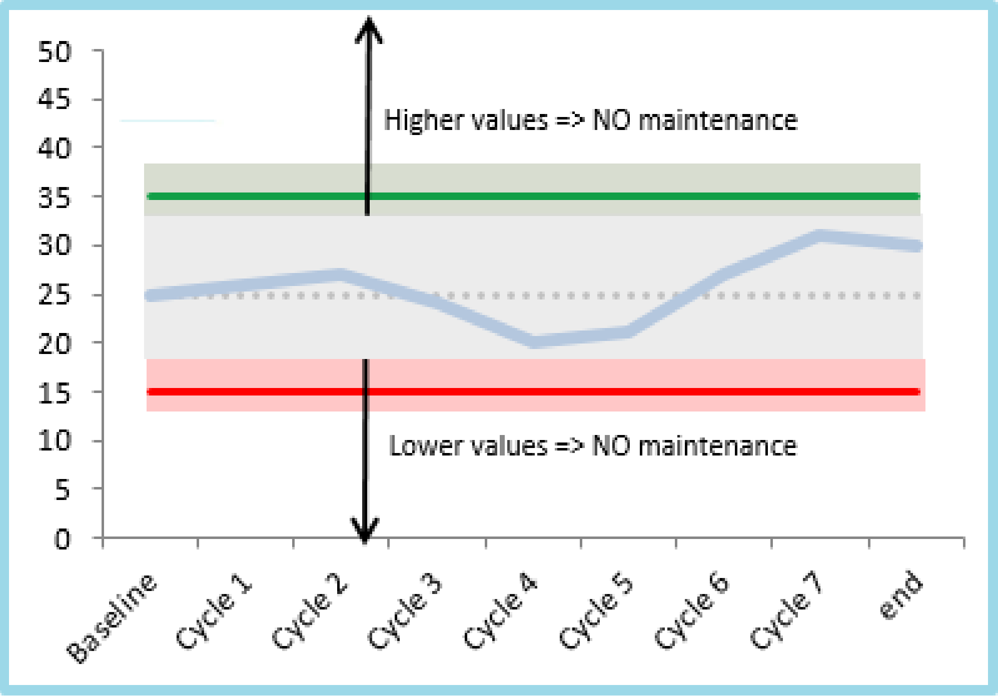


Worsening: Change from baseline that reaches a pre-defined worsening threshold level (post-baseline worsening).

Duration of worsening: Worsening is maintained if follow-up assessments remain at or is lower than the worsening threshold (taken from definitive worsening definition). Worsening is discontinued once a follow-up assessment is above the worsening threshold (taken from the transient worsening definition).

Definition of Maintenance

Consensus: Definition of maintenance + duration



Maintenance: No change from baseline or change from baseline is within the pre-defined baseline margin.

Duration of maintenance: Maintenance lasts as long as follow-up assessments remain at the baseline pre-defined margin. Maintenance is discontinued once the follow-up assessment leaves the pre-defined baseline margin (and reaches the improvement threshold or the deterioration threshold).

Recommending Statistical Methods

Consensus: Essential/Highly desirable statistical features for analyzing PRO data

Essential

- Perform a statistical test between two samples
- Be clinically relevant (the treatment effect can be expressed in the PRO scale unit)

Highly desirable

- Adjust for covariates, including baseline PRO score
- Handle missing data with least restrictions
- Ability to handle clustered data (repeated assessments)

Implication: adjust for relevant covariates and baseline PRO in the primary analysis

Time to event

Consensus: Method for Time to Event

For evaluating time to event outcomes, it is recommended to use Cox proportional hazards.

The Cox PH outperformed the log-rank test for these two criteria:

- ✓ Clinical relevance of results
- ✓ Adjustment for covariates, including baseline

Note: most recommendations ruled in favor of (semi)-parametric tests over non-parametric ones. Acceptable for PRO due to bounded data.

Cautionary note:

- When using Cox PH test, the proportional hazards assumption should be checked. If this assumption is not met, we recommend employing the log-rank test, but taking note that this statistical test does not address clinical relevance.
- General assumptions of time-to-event analysis must hold. Most notable: event time and censoring time should be independent.

Intensity of event at time *t*

Consensus: Method for intensity of event at time t

For evaluating intensity of event at time *t*, it is recommended to use linear mixed models (time as discrete).

The linear mixed model (time as discrete) has the advantage in:

- ✓ Adjustment for covariates, including baseline
- ✓ Handling of missing data
- ✓ Takes into account repeated data

While requiring fewer assumptions to be made *a priori* (e.g., regarding the relationship between time & outcome variable) than more complex mixed models extensions.

Cautionary note:

- Analysis strategy: fit a LMM to the data THEN obtain test estimate for specific time *t*.
  - General recommendations for fitting LMMs to be provided.
- Suitable if a study has a limited number of follow-up assessments.
- General assumption of linear mixed models hold
  - MAR assumption: provides unbiased estimate of the treatment effect that would have been observed if missing data is dependent on known (and observed factors).

Proportion of patients with event at time *t*

No consensus: Method for proportion of patients at time t

Based on the evaluation criteria, **logistic mixed model** could be recommended for this research objective.

- ✓ Adjustment for covariates, including baseline
- ✓ Handling of missing data
- ✓ Takes into account repeated data
- ✓ Extension of the linear mixed model to address binary data at time t

However, the consortium felt that there was uncertainty about the practical application of these models.
Recommendations for fitting LogMMs to be provided.

For cross-sectional outcomes: (Cochran) Mantel-Haenszel test out performed other tests for these two criteria:

- ✓ Clinical relevance of results
- ✓ Adjustment for covariates, including baseline (stratification is possible)

Cautionary note:

- (Cochran) Mantel-Haenszel test is sensitive to missing data and will only provide valid inference when missing data are MCAR.
- It is also a statistical technique that was designed for independent observations and does not take into account the repeated assessments of the PRO data

Overall PRO score over time

***Under discussion*: evaluating overall PRO score over time**

Two-step analysis:

- Summarizing a single PRO domain into a single score over a given time period.
- Comparative test on summary score between two arms.

Recommendations for summary measures are difficult as there are few standardized summary measures available and their interpretation is debatable.

Two-step analysis remains sensitive to missing data and will only provide valid inference when missing data are MCAR.

- Min/max especially sensitive to missing data.
- Handling of missing data can be done on the summary level or on the analysis level.

Note: Clinical relevant thresholds (Minimal Important Differences) need to be derived on the between-patient level (not on the within-patient level) to be appplicable.

Standardizing Terminology

Consensus: PRO data is missing iff data would be meaningful for the analysis of a given research objective but were not available for any reason.

Consequence:

- Not all unobserved assessments are considered as missing data.
- Missingness depends on the objective, ie. within a trial several missing data rates are possible.
- Data is meaningful for analysis if it reduces the sample size (non-informative missing data), distorts the treatment estimate (informative missing data) or both.

PRO study population ≠ PRO analysis population.

- PRO study population: all patients who consented to and were eligible to participate in the PRO data collection (ITC: intention-to collect population).
- PRO analysis population: all patients that will be included in the primary PRO analysis.

Standardizing Terminology

Missing data rates:

- The available data rate (a fixed denominator rate):

$$\frac{Nbr\ of\ patients\ submitting\ valid\ PRO\ assessment\ at\ time\ t}{Number\ of\ patients\ in\ PRO\ study\ population}$$

- The completion rate (a variable denominator rate):

$$\frac{Nbr\ of\ patients\ submitting\ valid\ PRO\ assessment\ at\ time\ t}{Nbr\ of\ patients\ on\ PRO\ assessment\ at\ time\ t}$$

Note : the denominator of the completion rate depends on the research question.

Missing data

Discussions are still ongoing. Preliminary conclusions are:

- Missing data should be minimized prospectively.

- Capturing the reasons for missing PRO assessments is important.
  - Impact of missing data depends on the reasons/mechanism for missing data.
  - Justifying strategies for intercurrent events.
  - Standardizing reasons

- Missing data implies unverifiable assumptions during the analysis.

- Missing data and scoring algorithm:
  - Missing data approach at the item- and scale-level should be specified a priori.
  - Item-level missing data should be handled according to the scoring algorithm of the instrument (when available).

Missing data

Primary statistical analysis approach:

- Critical assessment of missing data reasons and rates (by arm and time point) should be undertaken.
- Use all available data. Approaches that require ignoring missing data and only performing analysis with patients with complete data are not recommended (e.g., complete case analysis)
- Explicit imputation is not recommended unless justified within the context of the clinical trial.
- Sensitivity analysis should be specified a priori within the protocol/statistical analysis plan. At least two different approaches to handle missing data are recommended to assess the impact of missing data across various assumptions.
    - If the results are consistent with the primary analysis, this provides some assurance that the missing data did not have an important effect on the study conclusions.
    - If they produce inconsistent results, their implications for the conclusions of the trial must be discussed.

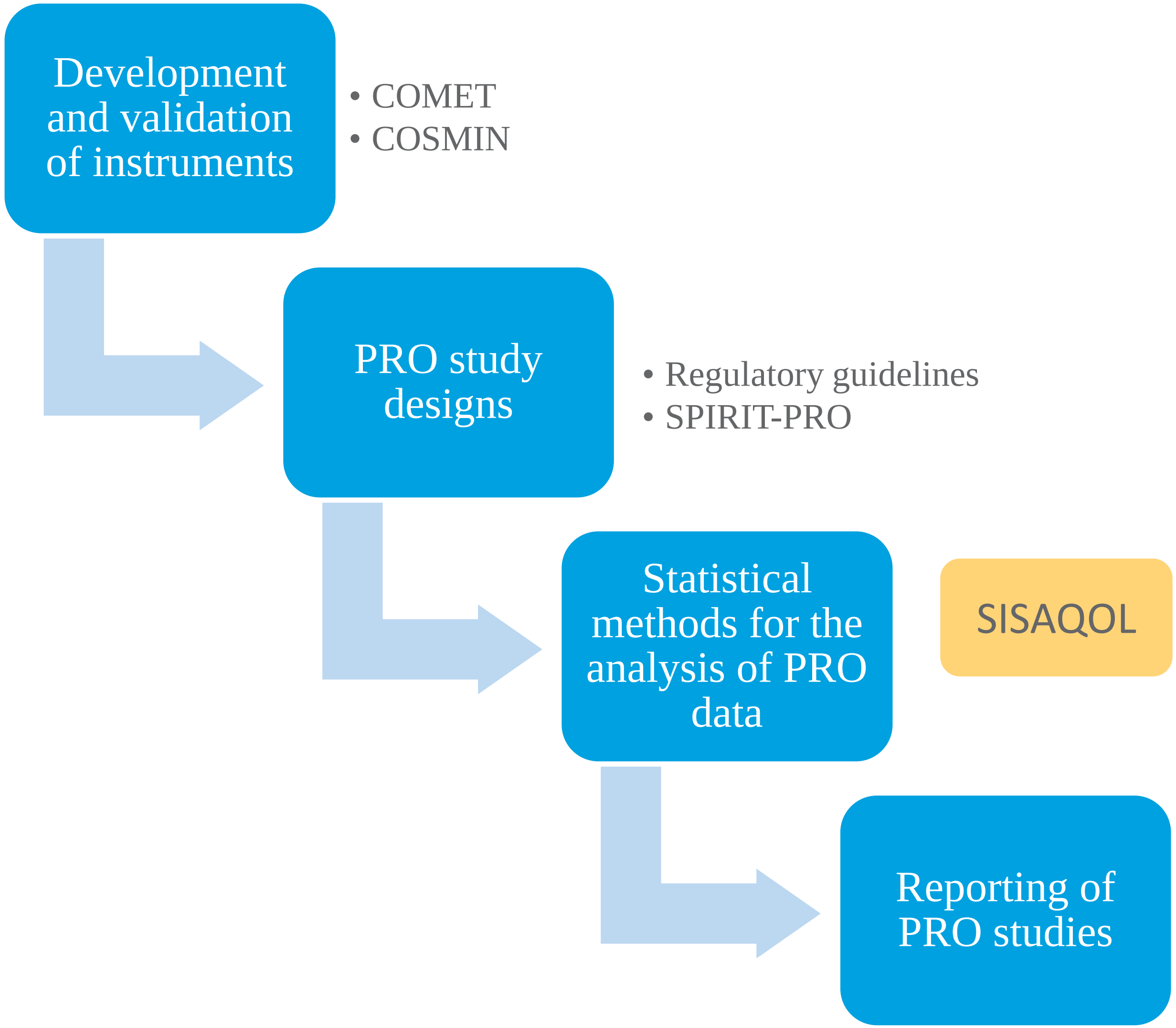
Concluding thought

Matching research objectives with statistical methods

- When developing research objectives and designing the trial, we need to think about **minimizing missing data**.
  - There is no panacea to analyze trial data with substantial amounts of missing data.
  - No analysis method recovers the potential for robust treatment comparisons derived from follow-up of all randomized patients (Little et al., 2012)

- Recommendation for ordinal outcome data needs further work and experience
  - Debate on how robust the linear mixed model is to violations of statistical assumptions (e.g., non-normal distribution of data)
    - Many researchers treat ordinal data the same way as continuous data
  - Potential use of generalized linear mixed model (extension of linear mixed models for ordinal outcome), but it is more complex. It's application in actual practice is unclear
    - Different kinds of logit functions that can be used
    - How to interpret the results?

Conclusions (2)



Development and validation of instruments

- COMET
- COSMIN

PRO study designs

- Regulatory guidelines
- SPIRIT-PRO

Statistical methods for the analysis of PRO data

SISAQOL

Reporting of PRO studies

- CONSORT-PRO
- PROTEUS

Thank you!

Questions? Suggestions?

	Draw conclusions on treatment efficacy / clinical benefit (Confirmatory Objective)		**Describe patient experience (Exploratory / Descriptive Objective)**
Within-treatment arms assumption *(longitudinal design: applies to both short-term and long-term)*	*Between treatment arms objective*		
	Superiority	*Equivalence / Non-inferiority*	
1. Improvement/worsening (event)	-		
a. Time to event	- Cox PH	- Cox PH	- Cox PH
b. Proportion of patients with event at time *t*	- LogMM / CMH	- LogMM / CMH	- LogMM / CMH
c. Intensity of event at time *t*	- LMM	- LMM	- LMM
2. Maintenance			
a. Time to (end of) maintenance	- Cox PH	- Cox PH	- Cox PH
b. Proportion of patients with maintenance at time *t*	- LogMM / CMH	- LogMM / CMH	- LogMM / CMH
c. Intensity of maintenance at time *t*	- Not applicable	- Not applicable	- Not applicable
3. Overall effects			
a. Overall PRO score over time	- TBD (2-step analysis)	- TBD (2-step analysis)	LMM (time as discrete / continuous)