**How to select outcome measurement instruments for outcomes included in a ‘Core Outcome Set’ – a practical guideline**

Cecilia AC Prinsen1\*, Sunita Vohra2, Michael R Rose3, Maarten Boers1,4, Peter Tugwell5, Mike Clarke6, Paula R Williamson7, Caroline B Terwee1

1VU University Medical Center, Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, The Netherlands (c.prinsen@vumc.nl; cb.terwee@vumc.nl; m.boers@vumc.nl)

2Department of Pediatrics, Faculty of Medicine and Dentistry, and School of Public Health, and Women’s and Children’s Health Research Institute, University of Alberta, Canada (svohra@ualberta.ca)

3Department of Neurology, King’s College Hospital, London, United Kingdom (m.r.rose@kcl.ac.uk)

4 Amsterdam Rheumatology & immunology Center, Amsterdam, The Netherlands

5 Department of Medicine, University of Ottawa, Ottawa, Canada (Tugwell.BB@uOttawa.ca)

6 Queen’s University Belfast, Northern Ireland Network for Trials Methodology Research, Institute of Clinical Sciences, Royal Hospitals, Belfast, UK (m.clarke@qub.ac.uk)

7 Department of Biostatistics, University of Liverpool, Liverpool, UK (p.r.williamson@liverpool.ac.uk)

\*Corresponding author

**Address for correspondence:**

C.A.C. Prinsen, VU University Medical Center, Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, De Boelelaan 1089a, 1081 HV Amsterdam, The Netherlands, Telephone: +31 20 4445903, Facsimile: +31 20 4448181, Email: c.prinsen@vumc.nl

***Abstract***

Background: In cooperation with the Core Outcome Measures in Effectiveness Trials (COMET) Initiative, the **CO**nsensus-based **S**tandards for the selection of health **M**easurement **In**struments (COSMIN) initiative aimed to develop a guideline on how to select outcome measurement instruments for outcomes (i.e. constructs or domains) included in a Core Outcome Set (COS). A COS is an agreed minimum set of outcomes that should be measured and reported in all clinical trials of a specific disease or trial population.

**Methods:** Informed by a literature review to identify potentially relevant tasks on instrument selection, a Delphi study was performed among a panel of international experts, representing diverse stakeholders. In three consecutive rounds, panelists were asked to rate the importance of different tasks in the selection of outcome measurement instruments, to justify their choices, and to add other relevant tasks. Consensus was defined as being achieved when ≥70% of the panelists agreed and when <15% of the panelists disagreed.

**Results:** Of the 481 invited experts, 120 agreed to participate of whom 95 (79%) completed the first Delphi questionnaire. We reached consensus on four main steps in the selection of outcome measurement instruments for COS: Step 1) conceptual considerations; Step 2) finding existing outcome measurement instruments, by means of a systematic review and/or a literature search; Step 3) quality assessment of outcome measurement instruments, by means of the evaluation of the measurement properties and feasibility aspects of outcome measurement instruments; and Step 4) generic recommendations on the selection of outcome measurement instruments for outcomes included in a COS (consensus ranged from 70 to 99%).

**Conclusions:** This study resulted in a consensus-based guideline on the methods for selecting outcome measurement instruments for outcomes included in a COS. This guideline can be used by COS developers in defining *how* to measure core outcomes.

**Key words**

COMET, Core Outcome Set, COSMIN, Delphi study, guideline, instrument selection, outcomes research, outcome measurement instrument

**BACKGROUND**

There is a lack of consensus with regard to the selection of outcomes (i.e. constructs or domains) and outcome measurement instruments for clinical trials.[1] As a result, different outcomes are assessed and a variety of instruments (e.g. assessments by health professionals, biomarkers, clinical rating scales, imaging tests, laboratory tests, patient questionnaires, and performance-based tests) measure the same outcome, causing inconsistencies in reporting and difficulties in comparing and combining the findings in systematic reviews and meta-analyses.[2,3] In addition, the quality of outcome measurement instruments varies considerably, and it is usually not apparent that the most reliable and valid instrument has been selected. Standardization of the selection of outcomes and outcome measurement instruments is needed.

The current project is a joint initiative between the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative[4] and the Core Outcome Measures in Effectiveness Trials (COMET) initiative.[5] COSMIN aims to improve the selection of measurement instruments, and has developed methodological standards for studies on the measurement properties of measurement instruments.[6] COMET aims to facilitate the development and application of agreed standardized sets of outcomes, also known as ‘core outcome sets’ (COS). A COS is an agreed minimum set of outcomes that should be measured and reported in all clinical trials of a specific disease or trial population. It is a recommendation of what should be measured and reported in all clinical trials.[7]

Once the COS is defined, it is then important to achieve consensus on *how* these outcomes should be measured, i.e. which outcome measurement instruments should be selected. In the selection of outcome measurement instruments, a number of tasks need to be performed. For example, a literature search to find potentially relevant outcome measurement instruments, and a quality assessment to evaluate the (methodological) quality of the available instruments. However, no guidelines are currently available to support outcome measurement instrument selection in a standardized and rigorous way.[8]

The primary aim of this study was to develop a guideline to select outcome measurement instruments for outcomes included in a COS. However a COS is not usually specific for any given clinical trial. A clinical trial may impose additional requirements for selecting outcome measurement instruments perhaps relating to feasibility or sensitivity. We therefore had a secondary aim of investigating whether the methods for selecting outcome measurement instruments for a COS are similar to the methods for selecting outcome measurement instruments for individual clinical trials.

**METHODS**

As details on the methods and design have been published previously,[9] this section is restricted to a summary.

*Study design*

A Delphi study was performed to achieve consensus on relevant tasks that need to be performed in the process of selecting outcome measurement instruments for outcomes (i.e. constructs or domains) included in a COS. The resulting guideline is based on the results of the Delphi study. Also, existing methodology that has been developed by COSMIN for performing systematic reviews of outcome measurement instruments were used to support the guideline,[4] as well as methodology that stems from the Outcome Measures in Rheumatology (OMERACT) Filter 2.0 and the OMERACT handbook for developing COS for rheumatic diseases,[10,11] and the Primary Outcomes Reporting in Trials (PORTal) initiative which looks at primary outcomes reported in adult and pediatric clinical trials.[12] These other sources of evidence were used to expand on items to a level not discussed in the Delphi study.

*Literature review*

To inform the Delphi study, a literature review was performed to identify existing studies that provide guidance on outcome measurement instrument selection. A health research librarian conducted an electronic literature search in November 2012 in MEDLINE, EMBASE, PsycINFO, and Cinahl.

Inclusion criteria: studies that were guidelines, meta-analyses, review articles, or systematic reviews, and study protocols that developed or applied methodology for selecting outcomes or outcome measurement instruments to be used in clinical trials. Exclusion criteria: studies that discussed ‘how to measure’ rather than ‘how to select’ outcomes or outcome measurement instruments for use in clinical trials; and studies that aimed to evaluate the measurement properties of outcome measurement instruments.

All search strategies are presented in Additional file 1.

*Development of the Delphi questionnaire*

The potentially relevant tasks on instrument selection identified from the literature review were included in the Delphi questionnaire. Questions were formulated on the relevance of each of the tasks, for example: ‘Should COS developers agree upon the target population before starting to search for outcome measurement instruments?’. Response options included ‘highly recommended’, ‘desirable’, ‘not relevant’, and ‘not my expertise’. Free text boxes were included after each question to facilitate comments.

*Selection of experts*

Experts that were identified from the literature review, as well as experts who participated in a previous COSMIN Delphi study[13] were invited to participate. A ‘snowball sampling’ approach was used to identify other potential experts. We found no guidelines for sample sizes of Delphi studies but in general, having more panelists will facilitate acceptance and implementation of the guideline.[14] Based on our previous experiences with Delphi studies,[6,13,15,16] we anticipated a response rate between 30-40%. We therefore invited all 481 previously identified experts to participate.

*Delphi rounds*

The Delphi study was planned to consist of three questionnaire rounds in order to achieve consensus.[17] Panelists were asked to anonymously rate the relevance of different tasks on outcome measurement instrument selection. They were encouraged to justify their choices and to add other possibly relevant tasks. Subsequently, panelists were asked for their opinion on whether the methods for selecting outcome measurement instruments for a COS are similar to the methods for selecting outcome measurement instruments for individual clinical trials.

Consensus was defined as being achieved when at least 70% of the panelists agreed with a task (i.e. highly recommended or desirable) with no opposing arguments provided, and when <15% of the panelists disagreed with a task (i.e. not relevant). Tasks on which such consensus was reached were included in the guideline and panelists were not asked to vote for these tasks again. When at least 50% of the panelists disagreed with a task (i.e. not relevant) and when no strong arguments in favor of this task were given, we excluded the task from the guideline. Tasks with an indeterminate response were taken to the subsequent round. When consensus was not reached after the third round, the need for a fourth questionnaire round was considered by the Delphi steering committee (CP, SV, MR, CT).

*Data analysis*

Data were analyzed both quantitatively (absolute values, percentages) and qualitatively (listings of the comments and suggestions given by the panelists). Based on the responses given in the first round, including the comments given in the free text boxes, new proposals were formulated. Response options included ‘strongly agree’, ‘agree’, ‘no opinion’, ‘disagree’, and ‘strongly disagree’. Additionally, new questions that arose based on the comments given, were formulated and were marked as ‘new questions’. Panelists were asked to rate their agreement on the given proposals and the relevance of the new tasks in the second round. The results of the second round were then again be analysed for consensus following the same procedure as for the first round.[9]

*Ethics*

As this project does not involve patients or study subjects as defined by the Dutch Medical Research in Human Subjects Act (WMO), the study was exempted from ethical approval in The Netherlands and similarly in the UK. Ethical approval was needed, and was obtained, from the Health Research Ethics Board of the University of Alberta, Canada (reference number: Pro00048898). Since our study sample consisted of experts (defined for these purposes as people who have a credibility relating to the target audience as indicated by, for example, authorship of multiple frequently cited publications in this field) and not patients or vulnerable subjects, consent to participate was implied through return of the questionnaire.

**RESULTS**

*Study population*

A total of 481 experts were invited to participate. Delivery failed to 41 recipients and four out of office notifications were received concerning long term absence. A total of 120/436 panelists (28%) accepted the invitation. 95/120 panelists (79%), from 14 different countries, completed the first questionnaire (Table 1). 65/95 (68%) completed the 2nd questionnaire, and 76/95 (80%) completed the 3rd questionnaire.

[Table 1]

*Delphi rounds*

In the first round, panelists were asked to rate 78 questions. Consensus was reached on 58 questions (74%).

In the second round, panelists were asked to re-rate 20 questions on which no consensus was achieved in the first round. In addition, 19 new questions were formulated based on the additional comments invited in the first round. For 2/19 new questions, ≥70% consensus was not reached (67%[[1]](#endnote-1) and 48%[[2]](#endnote-2) respectively). For 7/19 questions, consensus was reached (range 71% to 84%) but ≥15% of panelists disagreed. In reviewing the panelist’s comments on these items, it was clear that for a total of eight questions we were too restrictive in our formulations, too brief in the descriptions of the tasks, or that certain tasks might not be applicable in all circumstances.

In the third round, panelists were provided with eight new formulations, instead of questions, of the paragraph for potential inclusion in the guideline intending to address nuances applicable to specific situations. For example: in the first round it was suggested that the selection of outcome measurement instruments should always be guided by a review of the face validity of an instrument. In the second round, panelists were asked if COS developers themselves should assess the face validity of an outcome measurement instrument to be included in a COS. 84% of the panelists agreed, however, 16% of the panelists (strongly) disagreed. It was argued that only if no face validity assessment is reported in the literature, COS developers should do it themselves. In the third round, we proposed the following recommendation for the guideline: “It is recommended that, in case no face validity assessment is reported in the literature, COS developers assess the face validity of an outcome measurement instrument to be included in a COS”. On all eight formulations of the paragraph for potential inclusion in the guideline consensus was reached (range 81% to 93%), but ≥15% of the panelists disagreed on three of these formulations (15%, 15%, and 19% respectively). As no opposing arguments were provided against these three formulations, the Steering Committee decided to include all eight proposed formulations in the guideline.

We reached consensus on four main steps in the selection of outcome measurement instruments for COS (Table 2). Each of these four steps includes a variety of tasks.

[Table 2]

**Step 1. Conceptual considerations**

We reached 98-99% consensus that the first step in the selection of outcome measurement instruments is to agree in detail upon the construct (i.e. outcome or domain) to be measured[11] and the target population (e.g. age, gender, disease characteristics) (Table 2). This is a key task of the group developing the COS for which instruments are sought.

**Step 2. Finding existing outcome measurement instruments**

We reached 70-99% consensus that the second step is to find existing outcome measurement instruments. With the intention to search for all existing instruments, three sources of information can be used: 1) systematic reviews, 2) literature searches, and 3) other sources, considered as optional (Table 2). The COSMIN guideline for systematic reviews of outcome measurement instruments recommends that those searching the literature for all instruments do not use search terms to cover ‘type of measurement instrument’ because a wide variety of terminology is used (e.g. instruments are also termed measures, methods, questionnaires, tests, etc). This variety of terms that have been used in the original articles can lead to a high risk of missing relevant studies.[4] There is, however, one exception for patient-reported outcome measures (PROMs): for these a comprehensive PROM filter, developed for PubMed by the Patient Reported Outcomes Measurement Group of the University of Oxford, can be used. This search filter is available through the COSMIN website.[18] In all other cases it is recommened to only use search terms for ‘construct’, ‘population’, and ‘measurement properties’ in the search for all validated instruments.[4]

**Step 3. Quality assessment of outcome measurement instruments**

We reached 70-97% consensus that the third step in the selection of outcome measurement instruments is quality assessment of the available instruments. According to COSMIN, this includes two distinctive parts: 1) evaluation of the methodological quality of the included studies by using the COSMIN checklist,[6] and 2) evaluation of the quality of the measurement instruments (i.e. their measurement properties and feasibility aspects) by applying quality criteria for good measurement properties (Table 2).[19]

Following the COSMIN taxonomy on which international consensus was reached,[6,13] the following measurement properties were considered relevant: content validity (including face validity), structural validity, internal consistency, reliability, measurement error, construct validity, cross-cultural validity, criterion validity, and responsiveness (Table 3). Consensus was achieved on quality criteria for all measurement properties (Table 4). The quality assessment applies to all different types of outcome measurement instruments, such as assessments by health professionals, biomarkers, clinical rating scales, imaging tests, laboratory tests, patient questionnaires, and performance-based tests, and the applicable measurement properties should be evaluated.

In the evaluation of the measurement properties of the measurement instruments potentially included in a COS, COSMIN recommends a predefined order of importance of evaluating the measurement properties: 1) content validity; 2) internal structure (i.e. structural validity and internal consistency, and/or IRT/Rasch model fit); and where applicable 3) the remaining measurement properties (i.e., reliability, measurement error, construct validity, cross-cultural validity, criterion validity, and responsiveness). Content validity is considered to be the most important measurement property of a measurement instrument because if it is unclear what the instrument is actually measuring, the assessment of the other measurement properties is not valuable. If the content validity of a measurement instrument is poor or unknown, the instrument will not be further considered in the selection process. Subsequently, the internal structure (i.e. internal consistency and structural validity) should be evaluated. In case there is evidence that the internal structure of an measurement instrument is poor, the measurement instrument will not be further considered, i.e. the other measurement properties (including reliability, measurement error, construct validity, cross-cultural validity, criterion validity, and responsiveness) will not be further evaluated.[4]

[Tables 3 and 4]

To come to a conclusion about the overall quality of a measurement instrument, an overall evaluation of the measurement instrument should be constructed, based on all available evidence.[20] This can be done by a best evidence synthesis, where levels of evidence should be applied to each measurement property, taking into account the number of studies, the methodological quality of the studies, and the consistency of the results of the measurement properties (Table 5).[4]

[Table 5]

We reached 77-97% consensus that COS developers should take feasibility aspects into consideration in the selection of outcome measurement instruments for a COS (Table 6).

[Table 6]

**Step 4. Generic recommendations on the selection of outcome measurement instruments for a COS**

We reached 81-90% consensus on three generic recommendations concerning the final decision making on including an outcome measurement instrument in a COS: 1) it is recommended to select only one outcome measurement instrument for each outcome (i.e. constructs or domains) in a COS, which will enhance the comparability of clinical trials; 2) it is recommended that an outcome measurement instrument can be provisionally included in a COS if there is at least strong evidence[[3]](#endnote-3) for good[[4]](#endnote-4) content validity and good internal consistency (or evidence for test-retest or inter-rater reliability), and if the instrument is feasible; and 3) it is recommended that COS developers use a consensus procedure to get final agreement on the selected outcome measurement instruments included in a COS among relevant stakeholders, including patients (Table 2).

Following the OMERACT Handbook, the next phase of research needs to be more explicit on what categories of stakeholders should be considered (patients, public, practitioner, press, policymaker, program manager, professor, payer) and what the minimum requirements are for consensus.[10,11]

In addition, we reached 95% consensus that, in general, the methods for the selection of outcome measurement instruments for a COS are considered to be similar to the methods for selecting instruments for individual clinical trials. However, as in practice it may not be feasible to perform all these steps for a clinical trial, trialists can then chose to use those instruments that are included in the COS.

The four main steps, including their tasks, were included in the final guideline that can be found in Appendix 1.

**DISCUSSION**

The present guideline on methods for selecting outcome measurement instruments can be used by COS developers in defining ‘how’ to measure the core outcomes (i.e. constructs or domains) that are included in a COS. The guideline is based on the results of the Delphi study, the methodology derived from the COSMIN initiative, and recommendations from OMERACT.[11] With this stepwise approach, we intend to optimize the methodology of selecting outcome measurement instruments for a COS. The field of COS development is relatively new but rapidly growing; COMET maintains a database with the aim of including all registered and ongoing initiatives on COS development, including, for example, the Harmonising Outcome Measures for Eczema (HOME), and the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) initiatives. Currently, this database includes 326 papers relating to 272 COS, of which, 65 are ongoing.[5,21] Other examples of the potential impact of COS are that the National Institute for Health Research’s (a UK research funding body) Health Technology Assessment programme, requires COS to be considered in the funding applications of clinical trials, and that Cochrane and Grading of Recommendations Assessment, Development and Evaluation (GRADE) are encouraging the use of COS in reviews and clinical practice guidelines. We believe that methodology guidelines should be based on the agreed methodology so as to deliver high quality COS that can be used in future clinical trials and other research. Using high quality COSs will ultimately improve the conduct and reporting of clinical trials, enhance the value of evidence synthesis by reducing heterogeneity between trials, and may reduce outcome reporting bias. COSs reflect the best evidence at the time. However, as the field of COS development is continuously evolving (e.g. existing instruments are further tested and new ones are being developed), the outcome measurement instruments included in the COS might be reconsidered and/or replaced in light of new evidence.

There may be good reasons for COS developers to deviate from the guideline. For example, OMERACT want responsiveness to be assessed before inclusion in a provisional core set, whereas we reached consensus for at least strong evidence for good content validity and for good internal consistency. Another example is that, although a Cronbach’s alpha of >0.95 usually indicates item redundancy, there may be good reasons to retain certain potentially redundant items in a questionnaire. Also, we realize that in practice not all steps might be feasible within a given timeframe or budget. We recommend that COS developers should decide what is feasible in their time frame and within their budget.

Although the methods for the selection of outcome measurement instruments for a COS are considered to be similar to the methods for selecting instruments for individual clinical trials, it was argued that a higher standard for selecting instruments for a COS may be justified. Furthermore, it may not be feasible to perform all these steps for a clinical trial. This underlines the importance of the development of COS, as trialists can then chose to use those instruments that are included in the COS. When the primary outcome of a clinical trial is not a core outcome, the COS still needs to be measured. However, trialists could apply these recommendations to select the instrument for their primary outcome.

We acknowledge the limitations that might arise because of the relatively low response rate to the initial invitation of our Delphi study. As the results of Delphi studies in general are highly dependent upon the composition of the panel, we aimed to include a sample of experts who represent diverse disciplines, institutes and organizations and reflect the population that is intended to use a guideline for outcome measurement instrument selection. However, it is difficult to examine the representativeness of the panelists as it is impossible to draw a random sample from all experts. Experts were therefore selected non-systematically, which may be considered as a limitation of our Delphi study. Another limitation of our study is that we did not include patient research partners in the Delphi process. We acknowledge that, herewith, we may have omitted their contribution to the selection of outcome measurement instruments.

***Conclusions***

This consensus-based guideline on the methods for selecting outcome measurement instruments to be included in a COS can be used by COS developers and clinical trialists to define *how* to measure core outcomes (i.e. constructs or domains) for any diseases or other condition in health and social care.

***List of abbreviations***

COMET: Core Outcome Measures in Effectiveness Trials Initiative

COS: Core Outcome Set

COSMIN: COnsensus-based Standards for the selection of health Measurement INstruments

OMERACT: Outcome Measures in Rheumatology

***Competing interests***

Dr CB Terwee developed the COSMIN checklist. Dr CB Terwee and Dr CAC Prinsen are members of the COSMIN steering committee.

***Authors contributions***

CP conceptualized the study protocol, developed the Delphi questionnaire, coordinated the Delphi study, analyzed the data, wrote the manuscript, and reviewed it for important intellectual content. SV reviewed the study protocol, reviewed the Delphi questionnaire, and reviewed the manuscript for important intellectual content. MR reviewed the study protocol, reviewed the Delphi questionnaire, and reviewed the manuscript for important intellectual content. MB reviewed and revised the manuscript for important intellectual content. PT reviewed and revised the manuscript for important intellectual content. MC reviewed and revised the manuscript for important intellectual content. PW reviewed and revised the manuscript for important intellectual content. CT conceptualized the study protocol, developed the Delphi questionnaire, analyzed the data, wrote the manuscript, and reviewed it for important intellectual content. All authors have given final approval of the version to be published and agreed to be accountable for all aspects of the work.

***Acknowledgements***

We thank all panelists for their time spent completing all three Delphi questionnaires, including the following who accepted our invitation to be mentioned in the acknowledgement section:

J Alonso, Felix Angst, Dorcas Beaton, Anna Beurskens, Andrea E Bialocerkowski, Jane M Blazeby, Maarten Boers, Peter J Bower, Peter Bragge, Merel Brehm, Zeeshan Butt, Patrina HY Caldwell, Sarah L Chamlin, An-Wen Chan, Alessandro Chiarotto, Kaj S Christensen, Tammy J Clifford, Natalie J Collins, Megan Davidson, Aileen M Davis, Jill Dawson, Marcel P Dijkers, Fiona L Dobson, Antonio Escobar, Ricardo M Fernandes, Montse Ferrer, Carlos G Forero, Robert Frouf, Margreth Grotle, Jan Hartvigsen, Kirstie L Haywood, Bridget E Hill, Albere JA Koke, Ida J Korfage, Jan Kottner, Johanna H van der Lee, Sarah McIntyre, Joerg J Meerpohl, Lidwine B Mokkink, Jennifer L Moore, Catherine J Morgan, Kelly K O'Brien, Amy C Plint, Marcel WM Post, Kevin Pottie, Cecilia AC Prinsen, Carlijn van Randeraad-van der Zee, Bernd Reuschenbach, Ewa M Roos, Michael R Rose, Anne WS Rutjes, Jochen Schmitt, Beverley J Shea, Lee S Simon, Peter M Smith, Claire Snyder, Sally Spencer, Mirjam AG Sprangers, David L Streiner, Nicholas F Taylor, Caroline B Terwee, Kim S Thomas, Peter Tugwell, Peter Vedsted, Henrica CW de Vet, Ester Villalonga-Olives, Sunita Vohra, Angela C Webster, Anne Whitehead, Gavin Williams, Hywel C Williams, Paula R Williamson.

The research has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013] under grant agreement n° [305081].

***Reference List***

 1. Chalmers I, Glasziou P: **Avoidable waste in the production and reporting of research evidence.** *Lancet* 2009, **374:** 86-89.

 2. Williamson PR, Altman DG, Blazeby JM, Clarke M, Devane D, Gargon E *et al*.: **Developing core outcome sets for clinical trials: issues to consider.** *Trials* 2012, **13:** 132.

 3. Adams D., Liu Y., Bhaloo Z., Hansraj N., Hartling L., Vohra S.: **Primary Outcomes Reporting in Trials (PORTal): a systematic review of pediatric randomized controlled trials.** *(accepted by J Clin Epidemiol)* 2016.

 4. COnsensus-based Standards for the selection of health Measurement INstruments. COSMIN website. [accessed 25-4-2016]. Available from <http://www.cosmin.nl/>

 5. Core Outcome Measures in Effectiveness Trials. COMET website. [accessed 4-12-2014]. Available from <http://www.comet-initiative.org/>

 6. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL *et al*.: **The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes.** *J Clin Epidemiol* 2010, **63:** 737-745.

 7. Clarke M: **Standardising outcomes for clinical trials and systematic reviews.** *Trials* 2007, **8:** 39.

 8. Schmitt J, Apfelbacher C, Spuls PI, Thomas KS, Simpson EL, Furue M *et al*.: **The Harmonizing Outcome Measures for Eczema (HOME) Roadmap: a methodological framework to develop core sets of outcome measurements in dermatology.** *J Invest Dermatol* 2014, **135:** 24-30.

 9. Prinsen CAC, Vohra S, Rose MR, King-Jones S, Ishaque S, Bhaloo Z *et al*.: **Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'.** *Trials* 2014, **15:** 247.

 10. OMERACT Handbook. [accessed 25-4-2016]. Available from <http://www.omeract.org/pdf/OMERACT_Handbook.pdf>

 11. Boers M, Kirwan JR, Wells GA, Beaton DE, Gossec L, D'Agostino MA *et al*.: **Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0.** *J Clin Epidemiol* 2014, **67:** 745-753.

 12. Adams D., Liu Y., Bhaloo Z., Hansraj N., Hartling L., Vohra S.: **Primary Outcomes Reporting in Trials (PORTal): a systematic review of pediatric randomized controlled trials.** *(accepted by J Clin Epidemiol, 2016)* 2013.

 13. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL *et al*.: **The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study.** *Qual Life Res* 2010, **19:** 539-549.

 14. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J *et al*.: **Consensus development methods, and their use in clinical guideline development.** *Health Technol Assess* 1998, **2:** i-88.

 15. Chiarotto A, Deyo RA, Terwee CB, Boers M, Buchbinder R, Corbin TP *et al*.: **Core outcome domains for clinical trials in non-specific low back pain.** *Eur Spine J* 2015, **24:** 1127-1142.

 16. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM *et al*.: **The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus.** *J Clin Epidemiol* 1998, **51:** 1235-1241.

 17. Jones J, Hunter D: **Consensus methods for medical and health services research.** *BMJ* 1995, **311:** 376-380.

 18. PROM Bibliography. Patient Reported Outcomes Measurement Group database. [accessed 25-4-2016]. Available from <http://phi.uhce.ox.ac.uk/perl/phig/phidb_search.pl>

 19. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J *et al*.: **Quality criteria were proposed for measurement properties of health status questionnaires.** *J Clin Epidemiol* 2007, **60:** 34-42.

 20. The Standards for Educational and Psychological Testing. <http://www.apa.org/science/programs/testing/standards.aspx> [accessed 25-4-2016].

 21. Gargon E, Gurung B, Medley N, Altman DG, Blazeby JM, Clarke M *et al*.: **Choosing important health outcomes for comparative effectiveness research: a systematic review.** *PLoS One* 2014, **9:** e99111.

***Legend***

Table 1. Characteristics of the panelists

Table 2. Consensus on four main steps in the selection of outcome measurement instruments for COS, including their tasks

Table 3. Overview of all measurement properties, including their definitions

Table 4. Quality criteria for good measurement properties

Table 5. Levels of Evidence for the quality of the measurement instruments

Table 6. Overview of all feasibility aspects

Additional file 1. Search strategies for MEDLINE, EMBASE, PsycINFO and Cinahl

Appendix 1. Guideline for selecting outcome measurement instruments for outcomes included in a COS

1. Question: “The minimum standard for internal consistency of outcome measurement instruments to be included in a COS should be a Cronbach’s alpha between 0.70-0.90.” [↑](#endnote-ref-1)
2. Question: “If no outcome measurement instrument exists that meets the requirements for adequate measurement properties, it can be included in a COS ‘conditionally’. What should be the minimum condition before an instrument can be included in a COS?” [↑](#endnote-ref-2)
3. ‘Strong evidence’ is defined as consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality AND a total sample size of ≥100 patients (Table 5) [↑](#endnote-ref-3)
4. ‘Good’ is defined as a “+” rating according to the quality criteria for good measurement properties (Table 4) [↑](#endnote-ref-4)