Supplementary File S3. Tools to assess measurement properties.

[1]**COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) (2010, 2018)**		
1.	Reliability	The degree to which an instrument is free from random error.
1.1.	Internal consistency	The degree of the interrelatedness among the items. *In COSMIN (2018) internal consistency is derived from internal structure evaluation.*
1.2.	Reliability	Scores for patients who have not changed are the same for repeated measurement under several conditions
1.3.	Measurement error	The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
2.	Validity	The degree to which a Health Related-Patient Reported Outcome (HR-PRO) instrument measures the construct(s) it purports to measure. *Concept with major changes in COSMIN (2018) the definition and classification changed to content, structural, cross-cultural validity/measurement invariance, criterion, and hypothesis testing for construct validity (convergent, discriminative or known groups)*
2.1.	Content (including Face validity)	The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured (or looks as though the items are an adequate reflection)
2.2.	Construct (Structural, Hypothesis, Cross-cultural)	The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured. Scores of an HR-PRO instrument are consistent with hypotheses. Performance of the items on a translated or culturally adapted HR-PRO instrument is an adequate reflection of the performance of the items of the original version of the HR-PRO instrument
2.3.	Criterion	The degree to which the scores of an HR-PRO instrument are an adequate reflection of a "gold standard"
3.	Responsiveness	The instrument's ability to detect change over time in the construct to be measured
4.	Interpretability	The degree to which one can assign easily understood meaning to an instrument's quantitative scores. *A complementary attribute, not a measurement property in COSMIN (2018), plus feasibility*

[2] **Quality Criteria for Measurement Properties (Terwee et al. 2007)**		
1.	Content validity	The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire
2.	Internal consistency	The extent to which items in a (sub)scale are inter correlated, thus measuring the same construct
3.	Criterion validity	The extent to which scores on a particular questionnaire relate to a gold standard
4.	Construct validity	The extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured
5.	Reproducibility	
5.1.	Agreement	The extent which the scores on repeated measures are close to each other (absolute measurement error)
5.2.	Reliability	The extent to which patients can be distinguish from each other (relative measurement error)
6.	Responsiveness	The ability of a questionnaire to detect clinically important changes over time
7.	Floor and ceiling effects	The number of respondents who achieved the lowest or highest possible score
8.	Interpretability	The degree to which one can assign qualitative meaning to quantitative scores

[3]**Attributes and Criteria to assess Health Status and Quality of Life Instruments (1996, 2002)**		
1.	Conceptual and measurement model	The rationale for a description of the concepts and the populations that a measure is intended to assess and the expected relationship between these concepts
2.	Reliability	The degree to which an instrument is free from random error
2.1.	Internal consistency	The precision of a scale, homogeneity (inter correlations) of items at one point in time
2.2.	Reproducibility	Stability of an instrument over time (test-retest) and inter-rater agreement
3.	Validity	The degree to which the instrument measures what it purports to measure.
3.1.	Content validity	The domain of an instrument is appropriate relative to its intended use
3.2.	Construct-related validity	Interpretation of scores based on theoretical implications associated with the construct to be measured
3.3.	Criterion-related validity	The extent to which scores of the instrument are related to a criterion measure (gold standard).
4.	Responsiveness	The instrument's ability to detect change overtime
5.	Interpretability	The degree to which one can assign easily understood meaning to an instrument's quantitative scores
6.	Respondent and administrative burden	The time, effort, and other demands placed on those to whom the instrument is administered (respondent burden) or on those who administer the instrument (administrative burden)
7.	Administration/Accessible forms	Data collection method, including self-report, interviewer-administered, trained observer rating, computer-assisted interviewer-administered, performance-based measures. Accommodations (e.g. Braille)
8.	Cultural and language adaptations	Assessment of conceptual and linguistic equivalence.

[1] Prinsen C, Mokkink L, Bouter L, et al. COSMIN guideline for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018;0(0):1-11. doi:10.1007/s11136-018-1798-3. Mokkink L, Terwee C, Patrick D, 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(7):737-745. doi:10.1016/j.jclinepi.2010.02.006. [2] Terwee C, Bot S, de Boer M, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42. doi:10.1016/j.jclinepi.2006.03.012. [3] Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: Development of scientific review criteria. Clin Ther. 1996;18(5):979–92. Aaronson N, Alonso J, Burnam A, et al. Assessing health status and quality-of-life instruments and review criteria. Qual Life Res. 2002;11(3):193-215.

Supplementary File S3. Continue

[4]Health Status Measures in Economic Evaluation (1999, 2017)		
1.	Practicality	Time to complete the instrument. Response rate. Rate of completion
2.	Reliability	The degree to which an instrument is free from random error
2.1.	Test-retest	Ability to reproduce results over repeated measurements with the minimum amount of random error
2.2.	Inter-rater	Reliability between places of administration
3.	Validity	Dimensions covered. Items relevant for population. Ability of an instrument to reflect known or expected differences and changes in health to reflect preferences.
3.1.	Descriptive validity (Content, Face, Construct)	
3.2.	Valuation	Values used. Main assumptions of the model and how well the preferences of the patients and decision makers are likely to conform to these assumptions.
3.3.	Empirical	Evidence regarding whether or not a measure could generate values which reflect people's preferences using revealed preferences; stated preferences or hypothetical preferences as criteria
[5] Guidance for Industry patient-reported outcomes measures (2006, 2009)		
1.	Conceptual model	Conceptual framework.
2.	Administration/Accessible forms	Data collection method, including self-report or interviewer, format and scoring. Adaptations for children and adolescents, patients cognitively impaired, or unable to communicate, culture and language subgroups
3.	Respondent/Administrator Burden	Length, formatting, font size, instructions for items, privacy, time, need for physical support in responding.
4.	Reliability	
4.1.	Test retest	Stability of scores over time when no change has occurred in the concept of interest
4.2.	Internal consistency	Whether the items in a domain are inter correlated, as evidenced by an internal consistency statistic
4.3.	Inter interviewer reproducibility	Agreement between responses when the PRO is administered by two or more different interviewers
5.	Validity	
5.1.	Content validity	Whether items and response options are relevant and are comprehensive measures of the domain or concept
5.2.	Construct validity (Hypotheses testing, including discriminant, convergent, known groups validity)	Ability to measure the concept. Whether relationships among items, domains, and concepts conform to what is predicted by the conceptual framework for the PRO instrument itself and its validation hypotheses
6.	Criterion	Scores of a PRO instrument are related to a known gold standard. When the gold standard is not possible to be evaluated, criterion measure assesses sensitivity specificity, and predictive values
7.	Responsiveness. Ability to detect change	Evidence that the instrument is equally sensitive to gains and losses in the measurement concept and to change at all points within the entire range expected for the clinical trial population
[6] Evaluating patient-based outcomes measures for use in clinical trials (1998) (Fitzpatrick's criteria)		
1.	Reliability	The extent to which the instrument is free from random error and may be considered as the amount of a score that is a signal rather than noise
1.1.	Internal consistency	The extent to which individual items in a questionnaire scale measure the same construct (homogeneity of items in the scale)
1.2.	Reproducibility (test retest)	Whether and instrument yields the same results on repeated applications, when respondents have not changed on the domain being measured. Stability of the questionnaire over time
2.	Validity	The extent to which it measures what it purports to measure
2.1.	Criterion and Predictive validity	When a new measure correlates with other measures generally accepted as a more accurate variable. When the new measure correlates with future values of the criterion variable
2.2.	Face and content validity	Face validity refers to what an item appears to measure based on tis manifest content. Content validity refers to how well a measurement battery covers important parts of the health components to be measured
2.3.	Construct validity	A health status measure is intended to assess a postulated underlying construct.
2.3.1.	Convergent validity	Correlations are expected to be strongest with the most related constructs
2.3.2.	Discriminant validity	Correlations are expected to be weakest with most distally related constructs
2.3.3.	Internal structure	A set of assumed relationships between underlying constructs
2.3.4	Validity for specific purposes	Measures need to be assessed for health status, personal preferences and utilities, and social values.
3.	Responsiveness (sensitivity to change)	Ability to detect changes over time. Effect size, sensitivity and specificity of scores.
4.	Precision	How precise are the distinctions between levels of health and illness (sensitivity). Format categories.
5.	Interpretability	How meaningful are the scores from an instrument
6.	Acceptability	Evidence of acceptability is associated with high response rates. Respondent burden.
7.	Cultural applicability	Rigorous translation can by itself establish the appropriateness of an instrument
8.	Feasibility	Impact of different patient-based outcome measures upon staff and researchers. Administrator burden.
[7]International Classification of Functioning (ICF) & International Classification of Functioning for Children and Youth (ICFCY)) (2019)		
1.	Content validity	Health and Health-related domains.

[4] Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. Health Technol Assess (Rockv). 1999;3(9). Brazier J, Ara R, Rowen D, Chevrou-Severac H. A Review of Generic Preference-Based Measures for Use in Cost-Effectiveness Models. Pharmacoeconomics. 2017;35(s1):21-31. doi:10.1007/s40273-017-0545-x..[5]Department of Health and Human Services. Guidance for Industry Patient-reported Outcome measures: Use in Medical Product Development to Support Labeling Claims: draft guidance. Health Qual Life Outcomes. 2006;20:1-20. doi:10.1186/1477-7525-4-79. Department of Health and Human Services. Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.; 2009. doi:10.1111/j.1524-4733.2009.00609.x.[6]Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating Patient-Based Outcome Measures for Use in Clinical Trials. Vol 2.; 1998. doi:9812244.[7]World Health Organization. International Classification of Functioning (ICF). www.who.int/classifications/icf/en/.

Supplementary File S3. Continue

[8]Evaluating Measures of Patient Reported Outcomes (EMPRO) (2008)		
1.	Conceptual and Measurement model	The rationale for description of the concept and the populations that a measure is intended to assess and the relationship between these concepts
2.	Reliability	The degree to which an instrument is free from random error
2.1.	Internal consistency	The precision of a scale, based on the homogeneity of the scale´s items at one point in time
2.2.	Reproducibility	The stability of an instrument over time (test retest) and inter-rater agreement at one point in time
3.	Validity (including content, criterion, hypotheses testing and construct)	The degree to which the instrument measures what it purports to measure
4.	Responsiveness	The ability to detect change over time
5.	Interpretability	The degree to which one can assign meaning to an instrument´s quantitative scores
6.	Burden (Respondent/Administrator burden)	Time, effort and other demands placed on the administration of the instrument
7.	Administration mode	Data collection method. For each mode of administration, the information about validity, reliability, responsiveness, interpretability and burden should be assessed.
8.	Cultural and language adaptations	Methods to achieve linguistic equivalence are adequately described and appropriate. Differences from the original are adequately described and appropriate.
[9]Spinal Cord Injury Criteria (2008, 2016)		
1.	Content	Description. Items. Scale development. Internal structure or subscales
2.	Administration/Accessible forms	Data collection method. Items, time, training, burden of administering. Disability adaptation (e.g. Braille)
3.	Reliability (test retest, internal consistency)	Degree to which an instrument is consistent or free from random error
4.	Criterion oriented validity (concurrent, predictive, discriminant, and clinical validity)	Scale predicts other measures of the same construct. Gold standard and/or sensitivity and specificity. Scale distinguish between scores and/ or groups. Clinical utility, also called prescriptive and consequential validity
5.	Responsiveness, sensivity to change	Evidence of change in expected direction using methods such as standardized effect sizes
6.	Floor and ceiling effects	Floor and ceiling issues can determine whether change is detected or obscured by the measure
7.	Population application (Applicability in SCI groups, languages, norms)	Description of use in people with spinal cord injury (vs other people). Information of norms are available. Available in other languages
[10] Criteria for Assessing the Tools of Disability Outcomes Research (2000) (Andresen´s Tool)		
1.	Conceptual model	Relevant domains are completely covered
2.	Norms, standard values	Published data (or public-domain data) are available for both general population and with disabilities
3.	Measurement model	Tool captures the detail and breadth of real differences among persons, includes floor/ceiling effects
4.	Instrument bias	In practical or statistical terms, individual questions (or scores) are biased for the population
5.	Respondent burden	Length and content are acceptable to the intended subjects
6.	Administrative burden	Ease to administer, score and interpret
7.	Reliability (test retest and internal consistency)	Instrument gives a consistent answer
8.	Validity (discriminant, convergent, structure)	The tool measures what it purports to measure. It distinguish among different levels of mobility
9.	Responsiveness	Instrument is sensitive to changes in interventions
10.	Administration/Accessible forms	Data collection method, as interviews, self-administration, computer surveys. Adaptations (e.g. Braille)
11.	Culture/language adaptations	Tested versions of the tool for subgroups (including ethnicity, gender, disability)
[11]CanChild Outcomes Measures (2004)		
1.	Focus. Purpose	Focus of measurement (using the International Classification of Functioning Framework, ICF). Rating attributes measured. List the primary purpose for which the scales have been designed (discriminative, predictive, evaluative, etc.). Describe population. Evaluation of the context
2.	Clinical utility	Clarity of instructions, format, time to complete the assessment, administration, scoring and interpretation. Specify whether formal training is required. Cost of the manual and score sheets.
3.	Scale construction	Item selection, weighting, level of measurement
4.	Standardization	Manual (published, specific procedures for administration, scoring) Norms.
5.	Reliability	
5.1.	Internal consistency	The degree of homogeneity of test items to the attribute being measured. Measured at one point in time
5.2.	Intra/Inter observer	Measures variation within an observer; measures variation between two or more observers
5.3.	Test retest	Measures variation in the test over a period of time
6.	Validity	
6.1.	Content	The instrument is comprehensive and fully represents the domain of the characteristics it claims to measure
6.2.	Construct	Measurements of the attribute conform to prior theoretical relationships among characteristics or individuals
6.3.	Criterion	Measurements obtained by the instrument agree with another more accurate instrument (gold standard)
6.4.	Responsiveness	Ability to detect minimal clinically important change over the time

[8]Valderas JM, Ferrer M, Mendívil J, et al. Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. Value Heal. 2008;11(4):700-708. doi:10.1111/j.1524-4733.2007.00309.x. [9]Johnston M V., Graves DE. Towards Guidelines for Evaluation of Measures: An Introduction With Application to Spinal Cord Injury. J Spinal Cord Med. 2016;31(1):13-26. doi:10.1080/10790268.2008.11753976. Spinal Cord. Spinal Cord Injury Rehabilitation Evidence. https://scireproject.com. [10]Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil. 2000;81(12 SUPPL. 2):15-20. doi:10.1053/apmr.2000.20619. [11]Law M. Outcome Measures Rating Form Guidelines.; 2004. Available from: https://www.canchild.ca/system/tenon/assets/attachments/000/000/371/original/measguid.pdf

Supplementary File S3. Continue

[12]Outcomes Measures in Rheumatology Clinical Trials (OMERACT) (2019)	
1. Truth	
1.1. Face validity (credibility)	Overall appropriateness of the method to be used for evaluation of the outcome, as assessed by the investigators and clinicians
1.2. Content validity (comprehensiveness)	Ability of the outcome measure to include or predict all those components of health status that are relevant to the intervention being assessed
1.3. Criterion validity (accuracy)	Ability of the outcome measure to reflect the best available estimate of the true clinical status of the patient. Comparison with the "gold standard"
1.4. Construct validity (convergent/divergent)	Ability of the outcome measure to match with the hypothesized expectations of the investigator when compared with other indirect assessments
2. Discrimination	
2.1. Sensitivity to change over time	Based on calculation of the standardized response mean (SRM) using repeated measures performed in a given population at 2 different time-points without therapeutic intervention
2.2. Discrimination capacity over treatment	Based on calculation of effect size (ES) in randomized controlled trials or SRM in open-label trials
2.3. Reliability (reproducibility)	Based on evaluation of intra- and interclass correlations
3. Feasibility	The measure's ease of use, cost-effectiveness, availability in different centres, and overall usefulness. Practicalities of using the instrument, as cost, burden, length, translations, equipment needs.
[13]Testing Standards (1999, 2014)	
1. Evidences of Validity	
1.1. Test Content	Themes, tasks, format of the items, wording, and processes of administration and scoring
1.2. Response Processes	Cognitive processes engaged in by test takers with consequences in the scores.
1.3. Internal Structure (Dimensionality, Differential item functioning)	The degree to which the relationships among test items and components conform to the construct on which the proposed test score interpretations are based including equivalence of scores among different populations.
1.4. Relations to other variables (Convergent, Discriminant, Criterion, nomological network including responsiveness)	The degree to which relationships with other variables are consistent with expectations derived from theory underlying the construct
1.5. Consequences of testing	Value judgement about unintended positive and negative consequences of test use
2. Reliability	*Revised Standard (2014) also includes Decision consistency/accuracy*
2.1. Internal consistency, Test- retest, Alternate forms, *Scorers Consistency, Decision consistency, Accuracy*	The degree to which an instrument is free from random error. The precision of a scale, homogeneity (inter correlations) of items. Replicability of the testing procedure.
3. Fairness	Characteristics of all individuals must be considered throughout all stages of development, administration, scoring, interpretation and use of test. *Revised Standards (2014) emphasize the role of the Fairness as a measurement property*
4. Scales, Norms and Score Comparability	Reference points should be documented based on population norms and/or expert criteria. Linking procedures devised to guarantee comparability of different measures of similar constructs should be described
5. Test development and revision	Tests and their supporting documents should be periodically reviewed. New forms such as those derived from translation to other languages should be thoroughly tested for equivalence

[12] OMERACT. Instrument selection for Core Outcome Measurement Sets. In: OMERACT Handbook [Internet]. 2019. Available from: https://omeracthandbook.org/handbook . [13]American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. American Educational Research Association.; 1999. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. American Educational Research Association; 2014.