Quality of patient- and proxy-reported outcomes for children with impairment of the upper extremity: a systematic review using the COSMIN methodology

JPR Kalle, TFF Saris, IN Sierevelt, D Eygendaal, CJA van Bergen

Orthopedics

Background

- **Patient-reported outcome measures** (**PROMs**) have become of significant importance in patient evaluation.
- Ill-advised application of PROMs: incorrect interpretation of results, making misleading recommendations for clinical practice
- **Adequately selecting** the appropriate instrument for the appropriate study population and purpose: **integral part** of pediatric orthopedic research and clinical practice

Aim/hypothesis

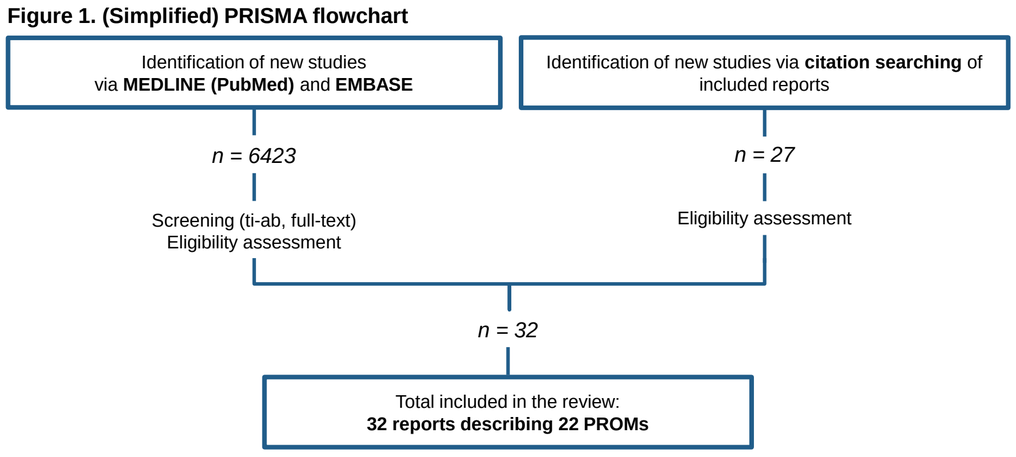
1. Comprehensive overview of **PROMs** targeted at **children with impairment of the upper limb**
2. Critically appraise and summarize the quality of their **measurement properties** by applying the COnsensus-based Standards for selection of health Measurement INstruments (**COSMIN**) methodology

Methods

- **Eligibility criteria**
  - *Inclusion criteria*
    - Full-text original article is available
    - Studies which evaluate one or more **measurement properties**/the **development**/the **interpretability** of a patient-reported and/or parent-proxy-reported questionnaire of **any language**
    - **Generic** and/or **disease-specific questionnaires**
    - **Children** (**0-18 years old**) with an **orthopedic diagnosis** in the **upper extremity** region
  - *Exclusion criteria*
    - Studies in which the patient-reported and/or parent-proxy-reported questionnaire is **only used as an outcome measurement instrument** (e.g., randomized controlled trials, longitudinal studies)
    - Questionnaires that evaluate the **use of prostheses** by children (0-18 years old)
- **Search**
  - Databases: MEDLINE (via PubMed) and EMBASE
  - Timeframe: 2000-2021
- **Data extraction and appraisal**
  1. **Evaluation of study methodological quality**
     - 'COSMIN Risk of Bias checklist' for rating studies on validity, reliability and responsiveness
     - 'COSMIN methodology for evaluating content validity' for rating studies on PROM development
  2. **Data extraction**
     - Data on the characteristics of the included study populations, of the studied PROMs and results of each study on a measurement property (e.g., Cronbach's alpha)
  3. **Assessment of measurement properties**
     - Result of each study: rated against the 'updated criteria for good measurement properties'
  4. **Qualitative synthesis of evidence**
     - Summarizing individual results on a measurement property of a PROM → rating the summarized result
     - Grading the quality of evidence using a modified GRADE-approach

Results

Figure 1. (Simplified) PRISMA flowchart



Conclusion

This review provides a comprehensive overview of currently available PROMs for evaluating children with impairment of the upper limb. Based on our findings, **none of the PROMs** demonstrated **sufficient evidence** on their measurement properties **to justify recommending the use** of these instruments.

Clinical implication

The **absence of studies on content validity** for any of the included PROMs is especially worrisome, as this implies it is **currently unknown** if the questionnaires used in pediatric orthopedic research and clinical practice **adequately reflect the construct** they intend to measure. This uncovers **a need for high-quality development** and **validation studies**, and especially studies on content validity, for PROMs utilized in pediatric orthopedics.

Results *(continued)*

Table 1. Synthesized evidence

PROM (refs)	Measurement property	Summarized result	Overall rating*	Quality of evidence§
ABILHAND-Kids (33)(40)(41)(47)(23)(42)	*Structural validity*	INFIT mean square range 0.66-1.18; OUTFIT mean square range 0.45-1.55	–	Moderate
	Internal consistency	Person separation reliability coefficient 0.94	?	
	Reliability	ICC range = 0.81-0.91	+	Moderate
	Measurement error	SEM = 1.7; SDD_{95} = 6.7; SDD_{95}/range = 0.16; SEM = 1.9; SDD_{95} = 4.8; SDD/range = 0.11; LOA = -2.06-1.40	?	
	Construct validity	9 out of 20 hypotheses confirmed	±	
	Responsiveness	RM ANOVA F = 29.89, p<0.001; Effect size T1vsT2 = 0.916, T2vsT3 = 0.158; Correlation changes measured by PEDI and ABILHAND-Kids Spearman r = 0.430, p = 0.003; Correlation changes measured by AHA and ABILHAND-Kids Pearson r = –0.104, p = 0.493	?	
ABILHAND-Kids (Ukrainian version) (27)	*Structural validity*	Standardized residuals range = -2.19-1.58	–	Moderate
	Internal consistency	Person separation index = 0.95	?	
	Cross-cultural validity	3 major DIF's were observed across countries (Ukrainian versus Belgian cohort)	–	Moderate
ABILHAND-Kids (Danish version) (28)	*Structural validity*	TLI = 0.98; CFI = 0.98; RMSEA = 0.07; SRMR = 0.07 Fit residuals (z) range = -2.178-2.170	–	Moderate
	Internal consistency	Cronbach's alpha = 0.96	?	
	Measurement invariance	1 non-uniform DIF was observed across age groups	–	Moderate
	Reliability	ICC2.1 = 0.97 (95% CI 0.95-0.98)	+	High
	Measurement error	SE = 0.5; LOAs range: –4.8-5.5; SDC = 5.15 points	?	
ABILHAND-Kids (Turkish version) (29)	*Structural validity*	Residual (z) range = -1.636-1.934	?	
	Internal consistency	Cronbach's alpha = 0.94	?	
	Measurement invariance	No DIF was observed	+	Very low
	Reliability	ICC = 0.98 (95% CI 0.98-1.00)	+	Very low
	Construct validity	2 out of 2 hypotheses confirmed	+	High
ABILHAND-Kids (Arabic version) (30)	*Structural validity*	Unidimensionality T-Tests (CI): 6.08% significant tests (lower limit of 95% CI = 2.60); Fit residual range = -2.06-2.01	–	Moderate
	Internal consistency	Person separation index = 0.93	?	
	Measurement invariance	No DIF was observed	+	Very low
	Reliability	$ICC_{agreement}$ = 0.98 (95% CI 0.97-0.99)	+	Very low
	Measurement error	$SEM_{agreement}$ = 0.24; MDC_{95} = 0.68	?	
	Construct validity	6 out of 7 hypotheses confirmed	+	Moderate
ABILHAND-Kids (Persian version) (31)	*Structural validity*	χ^2 probability = 0.40; PCA on the residuals, first residual factor accounts for 13% of the observed variance; Standardized residuals range = -1.34-1.60	+	Low
	Internal consistency	Cronbach's alpha = 0.963	+	Moderate
	Cross-cultural validity	2 major DIF's were observed across countries	–	Very low
	Measurement invariance	No DIF was observed	+	Very low
	Reliability	$ICC_{agreement}$ = 0.7 (CI 95% 0.33–0.85)	+	Very low
	Measurement error	SEM for CP measure = 11.21% (1.16 logits, raw score of 2.21); SDC for CP measure = 31.07% (3.21 logits, raw score of 6.13)	?	
	Construct validity	1 out of 1 hypothesis confirmed	+	Very low
ChARM (36)	*Structural validity*	Unidimensionality T-Tests (CI): 8% significant tests, lower limit of 95% CI = 4.6; Fit residuals range = -1.603-1.484	–	Moderate
	Internal consistency	Cronbach's alpha = 0.95	?	
	Construct validity	1 out of 1 hypothesis confirmed	+	Low
CHEQ (34)(37)(48)	*Structural validity*	Rasch analyses showed misfits (INFIT mean square >1.5 and/or Z-standardized values <-2 or >2) for several items of all three subscales	?	
	Internal consistency	Three CHEQ subscales: Person separation reliability coefficient range = 0.89-0.94	?	
	Reliability	Opening questions: 'performing the activity independently' average κ = 0.63, 'using the affected hand as support or to grasp' average κ = 0.57; Three CHEQ subscales: average ICC 0.87–0.91	+	Very low
	Construct validity	2 out of 2 hypotheses confirmed	+	Very low
CHQ (49)	*Construct validity*	No hypotheses were defined a priori	?	
CHSQ (38)	*Structural validity*	'Leisure and play domain': INFIT mean square range = 0.8-1.5, INFIT Zstd range = -1.6-2.8; OUTFIT mean square range = 0.7-1.5, OUTFIT Zstd range = -1.7-1.8 'School/education domain': INFIT mean square range = 0.7-1.2, INFIT Zstd range = -2.6-1.1; OUTFIT mean square range = 0.6-1.1, OUTFIT Zstd range = -2.1-0.4 'Activities of daily living domain': INFIT mean square range = 0.7-1.2, INFIT Zstd range = -1.6-1.3; OUTFIT mean square range = 0.5-1.4, OUTFIT Zstd range = -1.4-0.8	?	
	Internal consistency	Three CHSQ domains: Person reliability coefficient range = 0.67-0.75	?	
	Cross-cultural validity	7 items with DIF by cultural difference (Australian versus Taiwanese cohort)	–	Very low
	Construct validity	5 out of 7 hypotheses confirmed	±	
CHSQ (Turkish version) (32)	*Internal consistency*	Three CHSQ-TR subscales: Cronbach's alpha range = 0.83-0.86	?	
	Reliability	Three CHSQ-TR subscales; ICC range = 0.98-0.99	+	Low
	Construct validity	1 out of 1 hypothesis confirmed	+	Moderate
DHI (43)	*Internal consistency*	Cronbach's alpha range = 0.83-0.94	?	
	Reliability	ICC range = 0.84-0.93	+	Very low
	Construct validity	No hypotheses were defined a priori	?	
HUH (35)(44)	*Structural validity*	INFIT mean square range = 0.78-1.39; OUTFIT mean square range = 0.71-1.36	?	
	Internal consistency	Cronbach's alpha = 0.941	?	
	Reliability	ICC = 0.89 (95% IC 0.81–0.93)	+	Very low
	Measurement error	SEM (logits) = 0.599; $SDC_{individual}$ (logits) = 1.66; SDC_{group} (logits) = 0.22	?	
	Construct validity	7 out of 7 hypotheses confirmed	+	High
IMAL (50)	*Internal consistency*	Two IMAL subscales: Cronbach's alpha range = 0.94-0.95	?	
	Reliability	Two IMAL subscales: Spearman's correlation range = 0.64-0.70	?	
	Measurement error	'How Often' scale: SEM = 0.66 'How Well scale: SEM = 0.61	?	
	Construct validity	No hypotheses were defined a priori	?	
PEDI self-care domain (51)	*Construct validity*	1 out of 2 hypotheses confirmed	±	
PODCI (52)(24)(53)(49)(54)	*Construct validity*	11 out of 11 predefined hypotheses confirmed; for several analyses hypotheses could not be defined a priori	±	
	Responsiveness	4 out of 6 hypotheses confirmed	±	
PODCI (v2.0) (25)	*Internal consistency*	Cronbach's alpha range = 0.82-0.93	?	
	Construct validity	No hypotheses were defined a priori	?	
	Responsiveness	Moderate-large SRM (0.38-1.27)/effect size (0.32-1.37) for UE function, mobility, pain/comfort, happiness, global function; SRM 0.12/effect size 0.14 for sports/physical	?	
PODCI (v2.0; Dutch version) (26)	*Internal consistency*	Cronbach's alpha range = 0.161-0.928	?	
	Reliability	4 subscales and total score: ICC = 0.636–0.972 (p <0.025) 'Pain and comfort'-subscale: ICC = 0.022 (p = 0.476)	–	Very low
	Construct validity	2 out of 2 hypotheses confirmed	+	Very low
	Responsiveness	No hypotheses were defined a priori	?	
PROMIS – Upper Extremity item bank (short form) (55)	*Construct validity*	3 out of 3 hypotheses confirmed	+	Very low
PROMIS – Upper Extremity item bank (CAT) (55)	*Construct validity*	3 out of 3 hypotheses confirmed	+	Very low
QuickDASH (56)	*Internal consistency*	Cronbach's alpha = 0.91	?	
	Construct validity	Results in line with 1 hypothesis	+	Low
Revised PMAL (39)	*Structural validity*	'How Often' scale: EU associated with the first PCA contrast = 2.6 'How Well' scale: EU associated with the first PCA contrast = 2.5	?	
	Internal consistency	Two rPMAL subscales: Person reliability index range = 0.89-0.90	?	
	Reliability	Two rPMAL subscales: ICC range = 0.93-0.94	+	Very low
	Construct validity	2 out of 2 hypotheses confirmed	+	Very low

ICC = intraclass correlation coefficient, SEM = standard error of measurement, SDD = smallest detectable difference, LOA = limits of agreement, DIF = differential item functioning, TLI = Tucker Lewis index, CFI = Comparative fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual, MDC = minimal detectable change, SDC = smallest detectable change, PCA = Principal Component Analysis, SRM = standard response mean; ChARM = Children's Arm Rehabilitation Measure, CHEQ = Children's Hand-use Experience Questionnaire, CHQ = Child Health Questionnaire, CHSQ = Children's Hand-Skills ability Questionnaire, DHI = Duruöz Hand Index, HUH = Hand-Use-at-Home questionnaire, IMAL = Infant Motor Activity Log, PEDI = Pediatric Evaluation of Disability Inventory, PODCI = Pediatric Outcomes Data Collection Instrument, PROMIS = Patient-Reported Outcomes Measurement Information System, CAT = computer-adaptive test, DASH = Disabilities of the Arm, Shoulder and Hand, PMAL = Pediatric Motor Activity Log

* The results of the different studies on a particular measurement property of a PROM were qualitatively summarized and then rated against the updated criteria for good measurement properties: – = insufficient; + = sufficient; ± = inconsistent; ? = indeterminate

§ The quality of the evidence was graded by using a modified GRADE approach

Table 2. Characteristics of the included studies

PROM	Ref	*n*	Age Mean (SD, range) yr	Gender % female	Disease	Country	Language
ABILHAND-Kids	(33)	113	10 (6-15)	41%	CP	Belgium	French
	(40)	20	8.7 (2.9, 4-12)	50%	ULRD	The Netherlands	English
	(41)	20	7.6 (2.4, 4-12)	25%	RD	The Netherlands	English
	(47)	16	13 (2.3, 9-17)	56%	Spastic, unilateral CP	Germany	German (translational process not documented)
	(23)	52	9.1 (1.9, 6-12)		Unilateral, spastic CP	USA, The Netherlands, Belgium	English
	(42)	27	10 (4)	41%	Unilateral or bilateral CP	The Netherlands	English
ABILHAND-Kids (Ukrainian version)	(27)	113	10.3 (2.9, 6-16)	40%	CP	Ukraine	Ukrainian
ABILHAND-Kids (Danish version)	(28)	150	10 (2.7, 6-15)	40.7%	CP	Denmark	Danish
ABILHAND-Kids (Turkish version)	(29)	109	9.3 (2.9, 6-15)	43%	CP	Turkey	Turkish
ABILHAND-Kids (Arabic version)	(30)	154	7.4 (2.9)	45.5%	CP	Saudi Arabia	Arabic
ABILHAND-Kids (Persian version)	(31)	50	7.9 (2.2, 6-15)	40%	CP	Iran	Persian
ChARM	(36)	148	10.1 (3.3, 4.7-16.9)	39%	CP	UK	English
CHEQ	(34)	86	12 (3)	51%	Unilateral CP, OBPP, ULRD	Sweden	English
	(37)	242	9.8 (3.4)	43%	Unilateral CP	Australia, UK, Israel, Italy, the Netherlands, Sweden	English, Hebrew, Italian, Dutch, Swedish (translational process not documented)
	(48)	34	12.1 (3.9)	47%	Unilateral CP	Sweden	Swedish (translational process not documented)
CHQ	(49)	18	11.6 (10-17)	72%	NBPP	USA	English
CHSQ	(38)	123	7.17 (2.57)	28.5%	Various known disabilities (e.g., cerebral palsy, brachial plexus birth palsy)	Australia, Taiwan	English, Taiwan Chinese
CHSQ (Turkish version)	(32)	112	7.39 (2.51, 3-12)	39%	Hemiplegic CP	Turkey	Turkish
DHI	(43)	23	10.87 (2.8, 7-16)	39.2%	Unilateral CP	Turkey	English
HUH	(35)	322	Unilateral CP group: 6.5 (2.2, 3.0-10.8) NBPP group: 6.8 (2.0, 3.0-10.4)	Unilateral CP: 52% NBPP: 50%	Unilateral CP, NBPP	The Netherlands	English
	(44)	260	NBPP group: median age 6.9 (3.0-10.5) UCP group: median age 6.4 (3.0-10.8)	NBPP: 52% UCP: 49%	NBPP or unilateral CP	The Netherlands	English
IMAL	(50)	66	1.14 (0.44)	52%	Hemiplegic/quadriplegic CP	USA	English
PEDI self-care domain	(51)	45	5.1 (3.6-6.8)	64%	OBPP	Canada	English
PODCI	(52)	23	5.6 (3.5-8.6)	61%	BPBP	USA	English
	(24)	23	6.3 (4.4-12.8)	70%	BPBP	USA	English
	(53)	150	5 (2-10)	55%	BPBP	USA	English
	(49)	18	11.6 (10-17)	72%	NBPP	USA	English
	(54)	109	- (-)	46%	Congenital upper limb differences	USA	English
PODCI (v2.0)	(25)	125	11 (2-18)	43.2%	Acute hand and wrist injuries	USA	English
PODCI (v2.0; Dutch version)	(26)	10	5.3 (2.4)	50%	NBPP	The Netherlands	Dutch
PROMIS – Upper Extremity item bank (short form, CAT)	(55)	32	11.4 (3.9)	41%	Congenital hand differences	USA	English
QuickDASH	(56)	149	- (8-18)	48%	Several types of upper extremity injuries	USA	English
Revised PMAL	(39)	61	4.5 (-)	39%	Spastic hemiplegic CP	Australia	English

SD = standard deviation, yr = year, CP = cerebral palsy, ULDR = upper limb reduction deficiencies, RD = radius deficiencies, OBPP = obstetric brachial plexus palsy, NBPP = neonatal brachial plexus palsy, BPBP = brachial plexus birth palsy, ChARM = Children's Arm Rehabilitation Measure, CHEQ = Children's Hand-use Experience Questionnaire, CHQ = Child Health Questionnaire, CHSQ = Children's Hand-Skills ability Questionnaire, DHI = Duruöz Hand Index, HUH = Hand-Use-at-Home questionnaire, IMAL = Infant Motor Activity Log, PEDI = Pediatric Evaluation of Disability Inventory, PODCI = Pediatric Outcomes Data Collection Instrument, PROMIS = Patient-Reported Outcomes Measurement Information System, CAT = computer-adaptive test, DASH = Disabilities of the Arm, Shoulder and Hand, PMAL = Pediatric Motor Activity Log