A Systematic Review of Generic Multidimensional Patient-Reported Outcome Measures for Children, Part II: Evaluation of Psychometric Performance of English-Language Versions in a General Population

Astrid Janssens, PhD[1],*, Morwenna Rogers, MSc[1], Jo Thompson Coon, PhD[1], Karen Allen, MSc[1], Colin Green, PhD[1], Crispin Jenkinson, DPhil[2], Alan Tennant, PhD[3], Stuart Logan, MSc, F.R.C.P.C.H.[1], Christopher Morris, DPhil[1]

[1]University of Exeter Medical School, University of Exeter, Exeter, UK; [2]Nuffield Department of Population Health, University of Oxford, Oxford, UK; [3]Department of Rehabilitation Medicine, University of Leeds, Leeds, UK

A B S T R A C T

Objectives: The objectives of this systematic review were 1) to identify studies that assess the psychometric performance of the English-language version of 35 generic multidimensional patient-reported outcome measures (PROMs) for children and young people in general populations and evaluate their quality and 2) to summarize the psychometric properties of each PROM. **Methods:** MEDLINE, EMBASE, and PsycINFO were searched. The methodological quality of the articles was assessed using the COnsensus-based Standards for selection of health Measurement INstruments checklist. For each PROM, extracted evidence of content validity, construct validity, internal consistency, test-retest reliability, proxy reliability, responsiveness, and precision was judged against standardized reference criteria. **Results:** We found no evidence for 14 PROMs. For the remaining 21 PROMs, 90 studies were identified. The methodological quality of most studies was fair. Quality was generally rated higher in more recent studies. Not reporting how

missing data were handled was the most common reason for downgrading the quality. None of the 21 PROMs has had all psychometric properties evaluated; data on construct validity and internal consistency were most frequently reported. **Conclusions:** Overall, consistent positive findings for at least five psychometric properties were found for Child Health and Illness Profile, Healthy Pathways, KIDSCREEN, and Multi-dimensional Student Life Satisfaction Scale. None of the PROMs had been evaluated for responsiveness to detect change in general populations. Further well-designed studies with transparent reporting of methods and results are required.

Keywords: children and young people, measurement properties, patient-reported outcomes, review.

Copyright © 2015, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

Introduction

Patient-reported outcome measures (PROMs in the United Kingdom and patient-reported outcomes in the United States) are increasingly advocated for use in clinical trials [1,2] and as key performance indicators for evaluating health systems [3]. PROMs can be domain-specific, and focus on particular aspects of health (e.g., mental health or physical functioning), or be multidimensional instruments with subscales that assess different aspects of health. Some PROMs are condition-specific, designed for use by people with a particular diagnosis; other PROMs are generic and appropriate for anyone to report their health. Generic PROMs can be used across people with a range of health conditions, which is particularly useful when no condition-specific measure is available, or when comparisons

are made between the health of subgroups of people and findings from general population surveys [4].

When selecting PROMs for a specific purpose, it is necessary to examine both what is being assessed and how robust (valid and reliable) is the measurement. Language and cultural issues can affect how people interpret and/or respond to questions; hence, one cannot simply assume that PROMs perform consistently across languages and cultures [5,6]. Therefore, for example, the Food and Drug Administration guidance on PROMs recommends that evidence be provided of the process used to test measurement properties across different languages and cultures [1].

This article reports the results of a systematic review and critical evaluation of the literature on the measurement properties of PROMs for children and young people up to 18 years old. We focused on evaluations of English-language versions of

generic multidimensional PROMs for children to take account of methodological developments and any evidence published since previous reviews [7–9]. A new quality evaluation tool, the COnsensus-based Standards for the Selection of health status Measurement INstruments (COSMIN) system, has been developed to standardize the assessment of methodological quality of measurement studies [10–12]. In a related article, we have documented a systematic search and descriptive review of generic multidimensional PROMs for children, identifying 35 PROMs. In this study, we sought to identify and critically appraise studies that have assessed the psychometric performance of these PROMs, and to describe available evidence for the psychometric properties of each PROM.

Methods

Search Strategy

A separate search strategy was created for each of the 35 PROMs. MEDLINE, EMBASE, and PsycINFO were searched (via OvidSP) between July 18 and September 5, 2012, using three groups of terms: 1) name(s) and standard acronym of the PROM, 2) terms to describe children and young people, and 3) psychometric terms. No language or date limits were applied to the search. An illustration of the search strategy as used in EMBASE for one PROM (EuroQol 5D Youth [EQ-5D-Y]) can be seen in Data 1 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2015.01.004. Individual search strategies for the remaining PROMs can be supplied on request.

Backwards citation chasing (one generation) was carried out using all reference lists from articles included in the review. Forward citation chasing was carried out between January 28 and February 6, 2013, using Science Citation Index and Social Science Citation Index (via Web of Knowledge) for the key reference(s) of each of the selected PROMs. Developers of PROMs for which no published peer-reviewed articles were found were contacted to verify that we had not missed any eligible articles.

Inclusion and Exclusion Criteria

Articles were selected when written in English and reporting on a study that 1) was specifically designed to evaluate the psychometric properties of a selected PROM using an English-language version of the questionnaire, 2) was conducted in a general population of children up to 18 years old, and 3) published in a peer-reviewed journal. Articles were excluded if 1) the PROM was used as a criterion standard to test another instrument, 2) less than 10% of the study population was younger than 18 years, and 3) the study targeted children and young people with a specific condition or illness.

Study Selection

Titles and abstracts of records were screened against the eligibility criteria by one reviewer (A.J.); 10% were checked by a second reviewer (C.M.), with disagreements resolved by discussion with a third (C.J.) where necessary. The full text of any potentially relevant article was retrieved and screened using the same procedure.

Assessment of Methodological Quality of Included Articles

For each article, the methodological quality of the study and the completeness of the report were assessed using the COSMIN checklist (Table 1) [12]. This checklist consists of nine boxes with methodological standards for how each measurement property should be assessed [13]. Each item is rated on a four-point scale (poor, fair, good, or excellent); an overall score for each methodological quality is determined by a "worst-score counts" procedure.

Table 1 – Appraisal of psychometric properties and indicative criteria.

Psychometric property	Indicative criteria
Content validity	Clear conceptual framework consistent with stated purpose of measurement
	Qualitative research with potential respondents
Construct validity	Structural validity from factor analysis
	Post hoc tests of unidimensionality by Rasch analysis
	Hypothesis testing, with a priori hypotheses about direction and magnitude of expected effect sizes
	Tests for differential item and scale functioning between sex, age groups, and different diagnoses
Reproducibility	Test-retest reliability: ICC >0.7 adequate, >0.9 excellent.
	Proxy reliability: Child and parent-reported reliability ICC > 0.7
Internal consistency	Cronbach α coefficient >0.7 and <0.9
Precision	Assessment of measurement error; floor or ceiling effects $<15\%$; evidence provided by Rasch analysis and/or interval-level scaling
Responsiveness	Longitudinal data about change in scores with reference to hypotheses, measurement error, minimal important difference

ICC, intraclass correlation coefficient.

The checklist was administered by one reviewer (C.M./A.T.), and a 10% sample was rated by a second (A.J./C.M.). Any discrepancies were resolved by discussion, or with the involvement of a third reviewer (C.J.), where necessary.

Data Extraction

For each article describing a study evaluating the psychometric performance of an eligible PROM, the following descriptive data were extracted: instrument version, first author name, publication year, study aim, study population, number of participants, age range, mean age, and setting or country where the study was conducted. Data were extracted by one reviewer (K.A.), and 50% were checked by a second (A.J.), with disagreements resolved by discussion with a third (C.M.), where necessary.

For each version of a PROM, evidence of the following psychometric properties was extracted: content validity (theoretical framework and/or qualitative research), construct validity (structural validity and hypothesis testing), internal consistency, test-retest reliability, proxy reliability, responsiveness, and precision. Data were extracted by one reviewer (K.A./A.J./A.T.) and checked by a second (A.J./K.A./A.J.), with disagreements resolved by discussion with a third (C.M.), where necessary.

Appraisal and Summary of Evidence for Psychometric Performance

Evidence of performance was summarized by psychometric property and judged using standardized reference criteria and thresholds (Table 1). We included an appraisal of validity, reliability, responsiveness, and precision [4]. These data were