PharmacoEconomics (2022) 40:1043–1067
https://doi.org/10.1007/s40273-022-01177-z

SYSTEMATIC REVIEW

Are We Agreed? Self- Versus Proxy-Reporting of Paediatric Health-Related Quality of Life (HRQoL) Using Generic Preference-Based Measures: A Systematic Review and Meta-Analysis

Diana Khanna[1] · Jyoti Khadka[1,2] · Christine Mpundu-Kaambwa[1] · Kiri Lay[1] · Remo Russo[3,4] · Julie Ratcliffe[1] on behalf of The Quality of Life in Kids: Key Evidence to Strengthen Decisions in Australia (QUOKKA) Project Team

Accepted: 11 July 2022 / Published online: 23 August 2022

Abstract

Objective The aim of this study was to examine the level of agreement between self- and proxy-reporting of health-related quality of life (HRQoL) in children (under 18 years of age) using generic preference-based measures.

Methods A systematic review of primary studies that reported agreement statistics for self and proxy assessments of overall and/or dimension-level paediatric HRQoL using generic preference-based measures was conducted. Where available, data on intraclass correlation coefficients (ICCs) were extracted to summarise overall agreement levels, and Cohen's kappa was used to describe agreement across domains. A meta-analysis was also performed to synthesise studies and estimate the level of agreement between self- and proxy-reported paediatric overall and domain-level HRQoL.

Results Of the 30 studies included, 25 reported inter-rater agreement for overall utilities, while 17 reported domain-specific agreement. Seven generic preference-based measures were identified as having been applied: Health Utilities Index (HUI) Mark 2 and 3, EQ-5D measures, Child Health Utility 9 Dimensions (CHU9D), and the Quality of Well-Being (QWB) scale. A total of 45 dyad samples were included, with a total pooled sample of 3084 children and 3300 proxies. Most of the identified studies reported a poor inter-rater agreement for the overall HRQoL using ICCs. In contrast to more observable HRQoL domains relating to physical health and functioning, the inter-rater agreement was low for psychosocial-related domains, e.g., 'emotion' and 'cognition' attributes of both HUI2 and HUI3, and 'feeling worried, sad, or unhappy' and 'having pain or discomfort' domains of the EQ-5D. Parents demonstrated a higher level of agreement with children relative to health professionals. Child self- and proxy-reports of HRQoL showed lower agreement in cancer-related studies than in non-cancer-related studies. The overall ICC from the meta-analysis was estimated to be 0.49 (95% confidence interval 0.34–0.61) with poor inter-rater agreement.

Conclusion This study provides evidence from a systematic review of studies reporting dyad assessments to demonstrate the discrepancies in inter-rater agreement between child and proxy reporting of overall and domain-level paediatric HRQoL using generic preference-based measures. Further research to drive the inclusion of children in self-reporting their own HRQoL wherever possible and limiting the reliance on proxy reporting of children's HRQoL is warranted.

✉ Diana Khanna
khan0420@flinders.edu.au

Extended author information available on the last page of the article

Key Points for Decision Makers

The application of child-specific preference-based measures enables the calculation of utilities for cost utility analysis of health technologies targeted for paediatric populations.

Proxy reports (e.g., parent/guardian or a health professional), used in lieu of child self-reports in circumstances when self-reports are not feasible, can often diverge from the child's assessment of their own HRQoL.

This review examined the agreement between the child self- and proxy-reported overall and domain-level HRQoL using generic preference-based measures.

In general, the inter-rater agreement was poor for overall utilities across the measure/s applied and/or the context of the application. In addition, the agreement between children and proxy respondents within the domains of the respective measures was lower for psychosocial-related attributes compared with physical attributes.

1 Introduction

Evidence from economic evaluation is increasingly being utilised by regulatory bodies such as the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia and the National Institute for Health and Care Excellence (NICE) in parts of the UK to evaluate the cost effectiveness of health technologies targeted for paediatric populations [1]. PBAC, for example, considers evidence derived from measures of health-related quality of life (HRQoL) when recommending medicines eligible for government subsidies under the Pharmaceutical Benefits Scheme (PBS) [2]. Economic evaluations involving cost-utility analysis (CUA) have become the most prevalent approach for providing health economic evidence to assess the cost effectiveness of new health technologies for adult and paediatric populations. Within CUA, outcomes are most typically presented as quality-adjusted life-years (QALYs). The QALY combines 'utility' indexed on a 0–1 scale (where 0 is equivalent to being dead and 1 is equivalent to full health) and length of life into a single generic measure of health outcome, thereby facilitating comparisons of the health gains generated from alternative interventions [3, 4].

The application of child-specific preference-based measures enables the derivation of utilities (preference weights) for incorporating into CUA of health technologies targeted for paediatric populations [5]. In a previous review of validated measures, Chen and Ratcliffe identified nine generic preference-based measures that have been applied to measure and value HRQoL in children and adolescents: Quality of Well-Being Scale (QWB), Health Utilities Index Mark 2 (HUI2), Health Utilities Index Mark 3 (HUI3), Sixteen-dimensional measure of health-related quality of life (HRQoL) [16D], Seventeen-dimensional measure of HRQoL (17D), Assessment of Quality of Life 6-Dimension (AQoL-6D) Adolescent, Child Health Utility 9 Dimensions (CHU9D), EQ-5D Youth version (EQ-5D-Y) and Adolescent Health Utility Measure (AHUM). Preference-based measures comprise two main components: a descriptive system for measuring HRQoL, and a preference-based scoring algorithm for generating utilities. The descriptive systems of the identified nine generic preference-based measures that have been applied to measure and value HRQoL in children and adolescents differ in the content, type, absolute number of HRQoL dimensions (domains/attributes) and/or response levels included. Similarly, the preference weighted scoring algorithms (value sets) for these measures also differ according to the methods used to generate the value set, e.g., time-trade off (TTO), standard gamble (SG) or discrete choice experiments (DCEs) and the population from whom the value set was derived, e.g. adults or young people [3].

Ideally, the individual themselves should be the principal source of information about their own HRQoL [1]; however, self-assessment of HRQoL is challenging in the paediatric population. According to the Professional Society for Health Economics and Outcomes Research (ISPOR) Good Research Practices Patient-Reported Outcomes (PRO) Task Force Report, there is insufficient evidence to determine whether self-reporting of HRQoL by children under 8 years of age is reliable or valid [6]. Furthermore, older children with conditions associated with neurodevelopmental delays may be unable to self-assess their own HRQoL due to limited cognitive abilities. Such circumstances may require relying on an adult proxy such as a parent/guardian or a health professional to assess the child's HRQoL [7].

It is well-documented that proxy assessments of HRQoL in any population group tend to differ from self-assessments, with proxy assessors typically reporting lower HRQoL than the person themselves [1, 6, 8, 9]. Two previous systematic reviews by Khadka et al. and Jiang et al. of child self- and proxy-reported child utilities found that utilities tended to differ, with proxies often underestimating the child's HRQoL [10, 11]. In child populations, there is some evidence to indicate that proxy assessment of the child's HRQOL may be influenced by external factors, e.g. mother's assessment of the child's HRQoL may be influenced by their own HRQoL [12].

In their systematic review, Jiang et al. examined the difference in self- and proxy-reported utilities [11]. Child

HRQoL ratings obtained by two different observers, the child self and the proxy, are likely to differ owing to the differences in their perspectives. Therefore, it is also important to determine the extent to which the two raters agree or assign the same rating for an item being measured, i.e., to report inter-rater agreement measures that estimate the strength of agreement between raters [13, 14]. This systematic review sought to add to the existing evidence by focusing on reported measures of agreement in child and proxy assessments of paediatric HRQoL using established generic preference-based measures, highlighting individual domain-level differences in agreement, in addition to overall utilities. This study also presents the methods and findings from a meta-analysis of reported agreement statistics to provide an overall indication of the extent of agreement in child self and proxy assessments of paediatric HRQoL according to the available evidence.

2 Methods

2.1 Search Strategy

The literature search strategy was adapted from a previous study undertaken by Khadka et al., and the search keywords were reproduced [10]. The time frame covered by the previous search was from inception to 30 July 2017. To reflect the latest publications during the 4-year period since the initial search undertaken by Khadka and colleagues, this review incorporated peer-reviewed articles published in electronic journals between 30 June 2017 and 19 May 2021. The online databases searched included PubMed, The Cochrane Library, Web of Science, EconLit, Embase, PsycINFO and CINAHL (via EBSCOhost). Key words such as 'utility', 'quality-adjusted life years', 'children', 'adolescents', and 'preference-based measure of HRQoL' as well as related Medical Subject Headings (MeSH) terms were used for the systematic literature search. A detailed account of the search terms and the strategy is presented in Appendix 1 (see electronic supplementary material [ESM]). The identified studies were screened using the web-based systematic review software Covidence [15]. This review is registered with the International Prospective Register of Systematic Reviews (PROSPERO; registration number CRD42021256815). The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement guidelines were used for reporting this review (Appendix 2, see ESM) [16].

2.2 Inclusion and Exclusion Criteria

All studies published in English with full-text availability were included. Eligible studies included primary studies applying generic preference-based measures to derive health utilities amenable to QALY calculations in a paediatric population as assessed by the child (from hereon, *child* or *children* refer to all school-age children and adolescents, i.e., between 5 and 18 years of age unless stated otherwise) and proxy dyads. Inclusion criteria were studies reporting the agreement level for overall and/or domain-level paediatric HRQoL by both children and the proxies reporting on behalf of the children. Those studies that reported the paediatric health state utilities as assessed by child (self) and proxy respondents but did not include the agreement statistics were excluded. Additionally, as this systematic review focused on studies applying generic preference-based measures to derive health utilities, primary studies conducted among the paediatric populations were excluded if the utilities were obtained (1) directly using SG, TTO and VAS, or (2) indirectly using condition-specific (as opposed to generic) HRQoL measures.

2.3 Article Screening

Article screening was carried out in three steps. In the first step, two independent reviewers (DK and KL) screened the titles and abstracts based on the inclusion and exclusion criteria. Records with conflicting decisions were deferred to a third reviewer to reach a consensus. Articles selected at the screening stage were then included for a full-text review in the second step. The same two reviewers reviewed all the articles included in this stage. Simultaneously, two other reviewers (JK and CMK) independently assessed 10% of the articles in total to confirm the decisions of the former pair of reviewers. Following a discussion with the initial reviewing pair and the other reviewers (JR, JK, CMK) to reach a consensus, full-text articles that met the criteria were included. In the final step of this process, all eligible articles were subsequently consolidated and information relevant to the study was extracted.

2.4 Data Extraction

Data extraction was performed by the first author (DK). Each article was assessed to retrieve the following information: bibliographic details, geographic setting, study design, health state experienced, the generic preference-based measure used, target sample size, age range of the children included, sample gender composition, proxy type and sample size, mode of administration for both individuals in the dyad, statistical test(s) that report the overall and/or domain level of agreement between self- and proxy-reported HRQoL, and any reported methodological concerns. A Microsoft Excel (Version 2019; Microsoft Corporation, Redmond, WA, USA) database was used to enter and store the extracted data.

2.4.1 Extraction and Interpretation of Agreement Statistics

Inter-rater agreement is the degree to which the assessments of two or more individuals (raters) are identical using the same measure and assessing the same subject. There are multiple methods to measure inter-rater agreement based on the type of variable (continuous or categorical) and the number of raters. Agreement measures such as the intraclass correlation coefficient (ICC), Cohen's kappa (κ), Bland–Altman plots, percentage agreement and Gwet's agreement coefficient (AC1) assess the degree to which the assessments by the individual raters are identical or in agreement based on the type of data (e.g., nominal or continuous) [14, 17]. Correlation coefficients, also commonly reported to indicate agreement, determine the linear relationship between two continuous variables (Pearson's product-moment correlation or Pearson's r) or two ranked variables (Spearman's rho) [18].

It is important to note that in statistical analysis, correlation coefficients (e.g., Pearson's r) are considered as suboptimal measures of inter-rater agreement. They only provide a measure of the strength of a linear association between scores by raters and may indicate strong correlations even in the presence of a significant difference between the HRQoL assessments if the scores by both raters vary similarly. As a result, correlation coefficients may over- or underestimate the true level of agreement and inaccurately reflect the degree of agreement between raters [14, 18–20]. Inter-rater agreement is also often estimated using the percentage agreement approach [20]. However, percentage agreement does not correct for the level of agreement resulting from a random decision made by the raters. Cohen's kappa accounts for this random agreement and is more robust [21]. Therefore, percentage agreement is excluded from this review as a measure of child and proxy agreement. Only two studies reported the inter-rater agreement using the Bland–Altman plot and were thus not included in this review.

Thus, in the present study, to examine the concordance in the paediatric HRQoL obtained by self and proxy reports, we treat the ICC and kappa values as primary evidence. In addition, the results of the correlation coefficients, both Pearson's r and Spearman's rho, are presented as supplementary evidence.

ICC's can take a value between 0 and 1, whereas kappa and correlation coefficient statistics range from -1 to 1. Values for ICCs < 0.5 indicate poor agreement between raters, whereas values between 0.5 and 0.75, 0.75 and 0.9, and > 0.9 indicate moderate, good, and excellent agreement, respectively [22]. Spearman's correlation coefficients with a value < 0.20 represent no correlation, values between 0.20 and 0.35 represent weak correlation, values between 0.35 and 0.50 represent moderate correlation, and values ≥ 0.50 represent strong correlation [23]. Pearson's r coefficients are interpreted using Cohen's conventions. The correlation is small if the coefficient is 0.30 or less, medium if it is 0.50 or less, and large if it is > 0.50 [24]. Cohen's kappa and Gwet's AC1 have similarly defined thresholds, with classifications defined as slight (poor), fair, moderate, substantial (good) and almost perfect (very good) correlation for values ≤ 0.2, 0.4, 0.6, 0.8 and 1, respectively [17, 25].

2.5 Data Synthesis and Analysis

The estimates of the agreement level between child self- and proxy-reported HRQoL were described using a textual approach in the form of a narrative synthesis [26, 27]. Several studies did not report the mean age of participating children in the dyad, and hence only the age range was analysed. Studies that included children with cancer along with other chronic illnesses were identified as non-cancer-related studies. Caregivers reporting as proxies on behalf of children were grouped under parents. When the type of correlation was not mentioned in the study, it was assumed to be Pearson's r.

A meta-analysis was performed on a subset of the studies to synthesise the quantitative information and estimate the overall and domain-level agreement between child self- and proxy-reported HRQoL. To obtain an average estimate of inter-rater agreement, we synthesised the ICCs for overall utilities as they are reported on a continuous scale. Similarly, considering the ordinal nature of the responses within the attributes, kappa statistic was used to estimate the domain-level inter-rater agreement. Studies reporting only the correlation coefficients were excluded from the meta-analysis.

The meta-analysis was conducted using Stata 16.1 (Stata Corp LLC, College Station, TX, USA). Since the assumption of homogeneity is not reasonable for the present data due to the diverse nature of the target samples in consideration, we used a random-effects model to allow for between-study variability in effect sizes. The weights were estimated using a restricted maximum likelihood (REML) method [28]. A Fisher's z-transformation was applied to obtain an approximately normal sampling distribution in order to calculate the 95% confidence intervals (CIs) for each ICC for the overall utilities. The z-scores were then transformed back into correlations for ease of interpretation [29].

For the domain level meta-analysis, the standard errors (se) for kappa values ($\widehat{\kappa}$) were calculated using the following formula (Eq. 1):

$$\mathrm{se}_{\kappa} = \sqrt{\frac{p(1-p)}{n(1-p_{\mathrm{c}})^2}}, \tag{1}$$

where p is the observed percentage agreement, n is the number of rater pairs and p_{c} is the agreement expected by chance.

However, since no study reported the values for p_c, but did report p and $\hat{\kappa}$, p_c was calculated as shown in Eq. (2) [30]:

$$p_c = \sqrt{\frac{p - \hat{\kappa}}{1 - \hat{\kappa}}}. \quad (2)$$

A forest plot was used to depict the results of the meta-analysis (overall agreement). Heterogeneity was assessed using a forest plot as well as Cochran's test of homogeneity (Q statistic) and the I^2 statistic. Each sample was considered unique if any of the following variables relevant to the analysis were unique: type of proxy, measure, health condition, or age group composition (i.e., if children below 8 years of age were included in the sample). An exploratory meta-analysis (assuming a random-effects model) was conducted to estimate the moderation by these variables. A random-effect meta-regression was used to supplement the findings of the meta-analysis, as the studies were not considered sufficiently similar for a fixed-effects model [31]. The sample was also considered to be unique if the same sample was examined in a different time period for longitudinal studies. Publication bias was evaluated using funnel plots and a regression-based funnel plot asymmetry test.

2.6 Risk of Bias and Quality Assessment

Two independent reviewers (DK and JK) appraised the quality and suitability of the included studies. The overall reporting quality score was calculated using a checklist for quantitative studies as given by Kmet et al., and was used to assess the risk of bias [32]. From each of the selected articles that met the inclusion criteria, we extracted information for 14 quality indicator variables (details provided in ESM Appendix 3). Two points were assigned to each of these variables if they were appropriately reported in the article, one if the item was incompletely reported, and none if not reported at all. The sum of all the points indicated the overall reporting quality score of the article, with 28 being the maximum. The summary scores were rescaled between 0 and 1, with 1 denoting the highest quality. If the item was not applicable to a particular study, scores were adjusted by excluding the total possible scores of those items from the summary score. The minimum threshold for inclusion of studies based on quality scores was set at 0.6. The results of a sensitivity analysis carried out using the criteria by Papaioannou and colleagues to confirm the conclusions from the former appraisal are reported in Appendix 4 (see ESM) [33].

3 Results

3.1 Search Results

A PRISMA flow diagram illustrates the selection process (Fig. 1). An extensive literature search of seven databases was conducted using the search strategy described above. 43,522 records published between 30 June 2017 and 19 May 2021 were identified and were subsequently imported into Covidence; 19,309 records were deduplicated by Covidence, leaving 24,213 records for title and abstract screening. Of these, the vast majority (23,547) were excluded. Reasons for exclusion were (1) non-primary studies; (2) non-paediatric target population; (3) no health state utilities reported; (4) inaccessible articles; and (5) English was not the main language of publication. Subsequently, 666 records were included in the full-text review stage. At this stage, in addition to the previously specified exclusion criteria, studies were excluded if agreement statistics between the child self- and proxy-reported health state utilities and/or at domain level were not reported. In total, 30[1] studies fully met the inclusion criteria and were thus included in the final review.

3.2 Main Characteristics of the Studies

Table 1 presents an overview of the studies included in this systematic review. All the studies appraised for quality of reporting were of high quality, scoring 0.7 and over. The following study designs were employed: cross-sectional (83%), longitudinal (23%), and case-control (3%). HRQoL measures applied to obtain health state utilities either independently or in combination with other measures included the HUI3 (57%), EQ-5D measures (EQ-5D-Y-3L, EQ-5D-Y-5L, EQ-5D-3L, and the EQ VAS; 37%), HUI2 (33%), CHU9D (7%), and the QWB scale (3%). Cancer or history of cancer was the most common condition for which HRQoL was assessed (27%), predominantly blood and brain malignancies. Some studies (30%) also included children from the general population as the target sample or as the comparator/control group. The proxy respondent was exclusively a parent (mother, father, or a caregiver) in most of the identified studies (83%). Several studies (17%) used health professionals (nurses, physicians, and physiotherapists) or teachers as proxies, together with parents. The only exception was the study by Barr et al., which used only nurses and

[1] The two papers by Glaser et al., i.e. 'Standardized quantitative assessment of brain tumor survivors treated within clinical trials in childhood' [36] and 'Applicability of the Health Utilities Index to a population of childhood survivors of central nervous system tumours in the U.K.' [37], were published in two different journals but used the same sample to report different results. To prevent double counting, these two papers were considered as one.

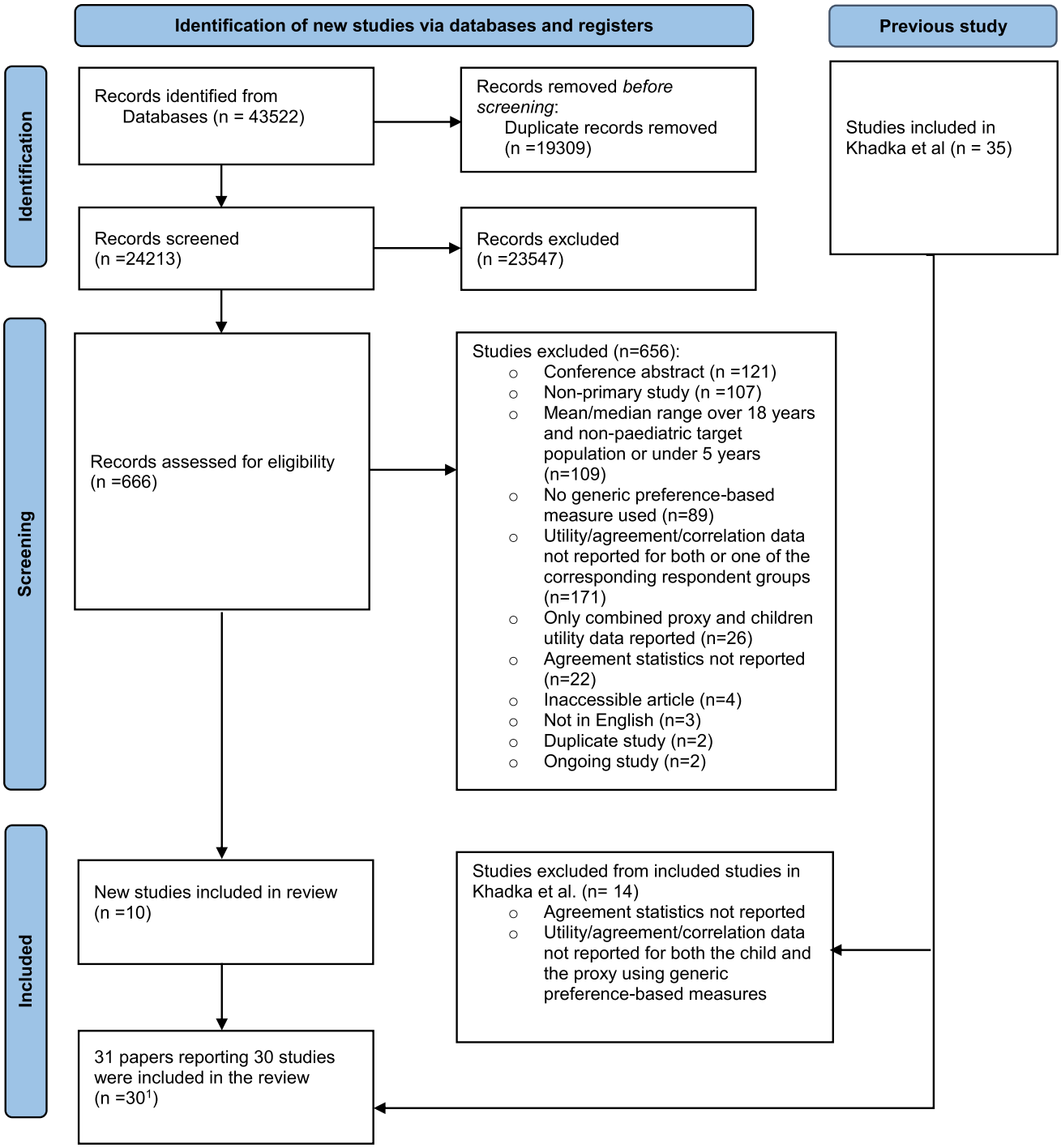

Fig. 1 Literature search flow diagram using the PRISMA checklist. *PRISMA* Preferred Reporting Items for Systematic Reviews and Meta-Analyses. [1]Thirty studies were included in the final review. The two papers by Glaser et al., i.e. 'Standardized quantitative assessment of brain tumor survivors treated within clinical trials in childhood' [36] and 'Applicability of the Health Utilities Index to a population of childhood survivors of central nervous system tumours in the UK' [37], were published in two different journals but used the same sample to report different results. To prevent double counting, these two papers were considered as one

physicians for proxy-reported utilities using HUI2 and 3 in cancer survivors [34]. Each study administered the proxy version of the measures adopting a proxy/proxy perspective, except one [35], which used a proxy/patient perspective (asking the proxy to rate the child's HRQoL from the child's perspective).

The measures were either administered by a trained interviewer (50%) or self-completed by the children (47%).

Table 1 An overview of the included studies

Description	No. of studies
Total studies included	30
Child-specific preference-based measures used	
HUI2	10
HUI3	17
EQ-5D-Youth, EQ-5D and EQ VAS	11
CHU9D	2
QWB	1
Health conditioned studied	
Cancer or history of cancer	8
Other health conditions (including general health)	22
Child proxy pairs (with some studies using more than one proxy type)	
Child/parent	29
Child/health professionals (nurses, physicians, physiotherapists) or teachers	5
Self-mode of administration for child in the age range	
6–7 years	3
8 years and above	11
Interviewer mode of administration for child in the age range	
6–7 years	6
8 years and above	10
Level of agreement statistics reported	
For overall utilities	25
For attribute-level utilities	17

HUI2 Health Utilities Index Mark 2, *HUI3* Health Utilities Index Mark 3, *VAS* visual analogue scale, *CHU9D* Child Health Utility 9 Dimensions, *QWB* Quality of Well-Being scale

One study used both an interviewer administration mode for children below 8 years of age and self-completion for the older children [36, 37]. The majority of the studies (83%) reported the inter-rater agreement for overall utilities. Five studies only reported the domain-level agreement [35, 38–41]. When reported, ICCs were slightly more commonly represented (60%) than correlation coefficients in measuring the overall child/proxy agreement level. Cohen's kappa (59%) was the most frequently used measure of agreement at the attribute level, followed by ICC (18%) and Gwet's AC1 (12%).

A summary of the included studies is presented in Tables 2 and 3 grouped into cancer- and non-cancer-related conditions, respectively. All the included studies were published between 1994 and 2021 and used primary data to obtain child health state utilities by employing generic preference-based measures. Majority of the studies were published in North America (USA and Canada; 33%) and Europe (UK, Spain, Netherlands, and Germany; 33%), followed by Asia (Thailand, Japan, Hong Kong, and China; 17%). Forty-five unique dyad samples based on the proxy type were included in the studies, with a total pooled sample of 3084 children and 3300 proxies. The age range for children in the included studies was between 5 and 18 years. Eight studies reported children younger than 8 years of age completing a self-report questionnaire either independently or with some assistance [35–37, 40, 42–46].

3.3 Proxy/Child Agreement

Table 4 presents a summary of reported agreement statistics for overall utilities using ICCs or correlation coefficients, i.e., Pearson's *r* and Spearman's rho. The studies used all the identified measures except for the EQ-5D-Y-5L, and employed both caregivers and health professionals as proxies. The sample size of the dyad ranged from 11 [45] to 654 [47]. From a total of 26 studies (58 samples), 12 studies reported only the ICCs [34, 42, 43, 46–54], and three studies reported ICCs alongside the correlation coefficients [36, 37, 55, 56]. Six studies reported only Spearman's rho [45, 57–61], whereas four studies reported only Pearson's *r* [44, 62–64]. Details of the included studies reporting the domain-level agreement statistics are presented in Table 5. The domain-level agreement was reported for 17 studies (40 samples), of which 10 studies used Cohen's kappa [34–41, 46, 47, 51], three studies used ICC [42, 43, 49], and two used Gwet's AC1 [53, 54]. No study reported the domain-level agreement for the CHU9D and QWB measures.

Table 2 Details of the cancer studies that reported dyad self and proxy HRQoL using preference-based quality-of-life instruments

Author Country Year	Health state experienced	Mean/median age (range) of the child in the dyad (where available) or study	Child sample (male %) included in the dyad (where available) or study	Proxy type (*n*)	Measure	Administration mode child	Quality score
Barr et al. [34] Canada 1999	Cancer survivors: CNS tumours	13.5 (9.5–17.9)	15 (46.3)	Nurses (15), physicians (12)	HUI2/3	Self-administered	0.7
Glaser et al. [36, 37] UK 1999	Cancer survivors: CNS tumours	10.5 (6.0–16.0)	28	Physiotherapist (30), parents (30), physicians (27)	HUI2/3	Self and Interviewer administered	0.85
Sung et al. [56] Canada 2004	Cancer patients: rheumatic diseases, haemophilia, conditions requiring bone marrow transplant	13.7 (12.0–18.0)	22 (55)	Parents	HUI2	Self-administered	0.9
Fu et al. [42] El Salvador, Honduras, Nicaragua, Panama 2006	Cancer survivors: leukaemia, lymphoma, renal tumours, germ cell tumours, retinoblastoma, malignant bone tumours, CNS tumours, sympathetic nervous system tumours, soft tissue sarcomas, carcinomas, others	12.8 (5–25.8)	211 (52.6)	Parents (180), physicians (201)	HUI2/3	Interviewer administered	0.8
Banks et al. [48] Canada 2008	Cancer including leukaemia, lymphoma, and brain tumour	9.5 (10.0–18.0)	11 (65)	Parents (22)	HUI2/3	Self-administered	0.85
Fluchel et al. [43] Uruguay 2008	Cancer survivors: ALL, brain tumours, Wilms tumour, retinoblastoma, Hodgkin disease, non-Hodgkin lymphoma, acute myeloid leukaemia, rhabdomyosarcoma, neuroblastoma, Ewing sarcoma, ovarian sarcoma, osteogenic sarcoma	13.6 (7.0–28.0)	95 (49.5)	Parents (95)	HUI3	Interviewer administered	0.95
	General health (control)	12.2 (8.0–17.0)	96 (33.3)	Parents (91)	HUI3	Interviewer administered	
Penn et al. [59] UK 2011	Cancer patients: brain tumour	12.4 (8.0–17.6)	29 (48.3)	Parents (29)	HUI3	Interviewer administered	0.85
	General health (control)	10.7 (8.0–18.9)	32 (50)	Parents (32)	HUI3	Interviewer administered	

Table 2 (continued)

Author Country Year	Health state experienced	Mean/median age (range) of the child in the dyad (where available) or study	Child sample (male %) included in the dyad (where available) or study	Proxy type (*n*)	Measure	Administration mode child	Quality score
Zhou et al. [54] China 2021	Haematological malignancies	10.5 (8.0–17.0)	96 (64.6)	Caregiver (96)	EQ-5D-3L-Y/VAS, EQ-5D-5L-Y	Interviewer administered	0.95

HRQoL health-related quality of life, *HUI2* Health Utilities Index Mark 2, *HUI3* Health Utilities Index Mark 3, *VAS* visual analogue scale, *CNS* central nervous system, *ALL* acute lymphoblastic leukaemia

3.3.1 Inter-Rater Agreement Based on the Type of Measure

HUI2 and 3 The inter-rater agreement between children and proxies for nine studies as indicated by the ICCs was poor for overall utilities [34, 36, 37, 42, 43, 48–50, 55, 56]. The overall ICC for HUI2 was slightly higher than that of HUI3. In contrast to HUI2, which showed good to excellent agreement for the overall utilities for one-quarter of the samples in the studies, the agreement using HUI3 was moderate at best. The correlation coefficients obtained from 10 studies indicated moderate associations between child self and proxy reports [36, 37, 44, 45, 55–60, 63].

Across the HUI2 attributes of 'emotion', 'cognition' and 'pain', the overall kappa values indicated fair agreement for those domains with a moderate agreement for 'sensation'. Overall, the kappa values suggested a substantial agreement for 'mobility', the highest level of agreement among all attributes, and a moderate agreement for 'self-care' between the child/proxy dyad [34, 36, 37, 39]. The lowest kappa values were reported for 'emotion' and 'cognition' in the assessment of HRQoL by children and proxies. For the 'pain' attribute, both slight and substantial levels of agreement were reported equally among the samples.

For HUI3, the overall agreement using kappa values was fair for 'cognition', 'emotion', 'speech' and 'pain'; moderate for 'hearing', 'dexterity' and 'ambulation'; and substantial for 'vision' [36–39, 41]. Similar to HUI2, the lowest agreement between children and proxies for HUI3 attributes was reported for 'emotion' and 'cognition'. In contrast, high kappa values were frequently reported for the attributes of 'vision', 'ambulation' and 'dexterity', with the agreement level ranging from substantial to almost perfect.

The ICC values demonstrated a poor agreement for subjective domains ('emotion', 'cognition', and 'pain') with some even reporting negative values. The agreement was between good to moderate for the observable domains of sensation, mobility, self-care, vision, hearing, and dexterity, with the notable exception of 'ambulation' and 'speech', which showed poor inter-rater agreement [42, 43, 49]. The agreement within the 'ambulation' and 'speech' attributes was moderate only in one instance between cancer survivors and their parents [43].

EQ-5D measures and the EQ VAS None of the studies reported the ICCs for the overall utilities or the summary scores using EQ-5D measures. Of the six studies reporting the ICCs for the EQ VAS scores, the majority showed poor agreement between child/proxy dyads [46, 47, 51–54]. However, an improvement in the inter-rater agreement was noted from baseline to follow-up [51, 54]. Kappa statistics reported for five studies indicated, on average, fair agreement between children and parents for all domains of EQ-5D [35, 40, 46, 47, 51]. The agreement was the lowest for the 'feeling worried, sad, or unhappy' and 'having pain or discomfort'

Table 3 Details of the studies with health conditions other than cancer that reported dyad self and proxy HRQoL using preference-based quality-of-life instruments

Author Country Year	Health state experienced	Mean/median age (range) of the child in the dyad (where available) or study	Child sample (male %) included in the dyad (where available) or study	Proxy type (*n*)	Measure (proxy measure)	Administration mode child	Quality score
Czyzewski et al. [62] USA 1994	Cystic fibrosis	(12–17.9)	55	Parents (199)	QWB	Self-administered	0.8
Verrips et al. [38] Netherlands 2001	Very low birth weight (VLBW): Mail Telephone Face-to-face Repeat mail	14.2 (14.0) 14.3 (14.0) 14.3 (14.0) 14.2 (14.0)	486 (49) 100 (54) 103 (51) 203 (52)	Parents (481) Parents (100) Parents (103) Parents (203)	HUI3 HUI3 HUI3 HUI3	Self-administered Self-administered Self-administered Self-administered	0.85
Brunner et al. [55] Canada 2003	Musculoskeletal disorders	9 (8.0–18.0)	55	Parents (68)	HUI3	Interviewer administered	0.8
Jelsma and Ramma [35] South Africa 2010	Children with functional impairment General health (control)	(7.0–12.0) (7.0–12.0)	61 (74) 567 (45)	Mother (57) Mother (530)	EQ-5D-Y/VAS (EQ-5D-Y Proxy 2) EQ-5D-Y/VAS	Self-administered Self-administered	0.85
Belfort et al. [57] Germany 2016	Overweight or obese General health (control)	10.3 (8.0–17.0) 11.5 (8.0–18.0)	76 (52.6)	Parents (63)	HUI3	Interviewer administered	0.95
Lee et al. [58] USA 2011	Type 1 diabetes mellitus. Complications: hypertension, hypercholesterolaemia, cardiovascular disease, renal disease, neurological disease, retinopathy	13.7 (8.0–18.0)	231 (48.5)	Parents (223)	HUI3	Interviewer administered	0.95
Morrow et al. [39] Australia 2012	Chronic illness: any cancer, cystic fibrosis, type 1 diabetes, cerebral palsy (GMFCS V), any chronic neurological condition, liver transplant, inflammatory bowel disease, chronic kidney disease, autism	12.2 (12.0–18.0)	69 (54.2)	Parents (129) Physicians (34)	HUI2/3	Self-administered	0.85
Rhodes et al. [60] USA 2012	Obesity; type 2 diabetes mellitus; prediabetes; insulin resistance	15.5 (12.0–18.0)	108	Parents (108)	HUI3	Interviewer administered	0.85

Table 3 (continued)

Author Country Year	Health state experienced	Mean/median age (range) of the child in the dyad (where available) or study	Child sample (male %) included in the dyad (where available) or study	Proxy type (*n*)	Measure (proxy measure)	Administration mode child	Quality score
Ungar et al. [49] Canada 2012	Asthma	10.9 (8.0–17.0)	91 (55)	Parents (91)	HUI2/3	Interviewer administered	1
Kulpeng et al. [44] Thailand 2013	Common pneumococcal infections and sequelae: meningitis, bacteraemia, pneumonia, AOM, hearing loss, chronic lung disease, epilepsy, MMR, SMR, and MR combined with epilepsy	10 (7.0–14.0)	74	Caregiver (74)	HUI2/3, EQ-5D	Interviewer administered	0.85
Wolke et al. [41] Germany 2013	VLBW/VP	13 (13.0)	260 (52)	Parents (260)	HUI3	Self-administered	0.85
	General health (control)	13 (13.0)	282 (49)	Parents (282)	HUI3	Self-administered	
Gusi et al. [40] Spain 2014	General health	(6.0–17.0)	442	Mother (442) Father (266)	EQ-5D-Y	Self and interviewer administered	0.9
Sims-Williams [63] Uganda 2017	Open spina bifida; associated complications	(10.0–14.0)	66 (56)	Caregiver (66)	HUI3	Interviewer administered	0.95
Bharij et al. [50] USA 2017	Paediatric liver transplant recipients	13.6 (12.0–21.7)	108 (44.4)	Parents (108)	HUI2/3, CHU9D	Interviewer administered	0.9
Bray et al. [45] UK 2017	Long-term mobility impairment: cerebral palsy, hemiplegia, muscular dystrophy	(6.0–18.0)	13 (61.5)	Parents (13)	HUI2/3, EQ-5D-Y/VAS	Self-administered	0.9
Perez Sousa et al. [46] Spain 2017	Cerebral palsy	10.9 (6.0–17.0)	62 (65.4)	Mother (62)	EQ-5D-Y/VAS	Interviewer administered	0.95
Perez Sousa et al. [51] Spain 2018	Obesity: exercise Obesity: control	9.6 (6.0–14.0) 8.7 (6.0–13.0)	106 (55) 45 (47)	Parents (106) Parents (45)	EQ-5D-Y/VAS EQ-5D-Y/VAS	Interviewer administered Interviewer administered	0.95
van Summeren et al. [52] The Netherlands 2018	Functional constipation	10 (8.0–18.0)	56 (43)	Parents (56)	EQ VAS	Self-administered	0.95

Table 3 (continued)

Author Country Year	Health state experienced	Mean/median age (range) of the child in the dyad (where available) or study	Child sample (male %) included in the dyad (where available) or study	Proxy type (*n*)	Measure (proxy measure)	Administration mode child	Quality score
Rogers et al. [64] Netherlands 2019	Dental caries	11 (11.0)	486 (48)	Parents (486)	CHU9D (NL)	Self-administered	1
Shiroiwa et al. [47] Japan 2019	General health	11 (8.0–15.0)	654 (50)	Parents (654)	EQ-5D-Y/VAS	Self-administered	0.9
Sinlapamongkolkul et al. [61] Thailand 2020	Thalassaemia	9.1 (8.0–18.0)	85 (54)	Parents (85)	EQ VAS	Self and interviewer administered	0.95
Lin et al. [53] Hong Kong 2020	Adolescent/Juvenile idiopathic sclerosis (AIS/JIS)	14 (10.0–12.0)	125 (9.4)	Caregiver (125)	EQ-5D-Y/VAS	Self-administered	0.95

HRQoL health-related quality of life, *HUI2* Health Utilities Index Mark 2, *HUI3* Health Utilities Index Mark 3, *EQ-5D-Y* EQ-5D Youth version, *VAS* visual analogue scale, *CHU9D* Child Health Utility 9 Dimensions, *CHU9D (NL)* CHU9D Dutch version, *QWB* Quality of Well-Being scale, *GMFCS* Gross Motor Function Classification System, *AOM* acute otitis media, *MMR* mild mental retardation, *SMR* severe mental retardation, *MR* mental retardation, *VLBW* very low birth weight, *VP* very preterm, *AIS* adolescent idiopathic scoliosis, *JIS* juvenile idiopathic scoliosis

domains, followed by, 'doing usual activities', 'looking after myself' and the highest for 'walking about'.

The inter-rater agreement between children and proxies within the EQ-5D domains using Gwet's AC1 ranged from moderate to very good [53, 54]. Children and adolescents with haematological malignancies were assessed using both 3L and 5L versions of the EQ-5D-Y in the study by Zhou et al. They found moderate to good agreement between the self- and caregiver-reported HRQoL for the five dimensions. The agreement improved from baseline to follow-up for all except the 'having pain or discomfort' domain in the 3L version and the 'walking about' and 'looking after myself' domains in the 5L version. However, no significant difference between the 3L and 5L versions was reported [54]. Among children with Adolescent/Juvenile idiopathic scoliosis (AIS/JIS), Lin et al. showed very good agreement with the caregivers in all domains except the 'having pain or discomfort' and 'feeling worried, sad, or unhappy' domains [53].

CHU9D and QWB The only study that reported the ICC using CHU9D showed moderate inter-rater agreement [50]. Using a large sample of 384 child/parent dyads, Rogers et al. reported a weak but significant correlation between the child self and proxy reports using CHU9D [64]. In their study, Czyzewski et al. reported a moderate correlation between the self- and proxy-reported utilities using QWB [62].

3.3.2 Inter-Rater Agreement Based on the Type of Proxy

Both types of proxies (parents and health professionals) showed poor inter-rater agreement, although parents showed higher agreement overall, regardless of measures and/or health conditions. All studies using health professionals as proxies assessed the HRQoL of children with cancer or child cancer survivors. Among these, Fluchel and colleagues used physicians and teachers as proxies for the children in the control group with no health condition [43]. A negative ICC (− 0.31, 95% CI − 0.22 to 0.262) was noted, indicating poor inter-rater agreement between the pair [43]. Only one study showed good to excellent agreement between cancer survivors and health professionals (nurses and physicians) using HUI2 [34]. Glaser and colleagues compared the inter-rater agreement between children with a history of cancer and their parents, physicians, and physiotherapists. Both the agreement (ICC) and correlation (Pearson's *r*) values were better for parents, closely followed by physiotherapists, and worst for physicians [36, 37]. In the study by Ungar et al., the authors found a poor inter-rater agreement when children and parents reported paediatric HRQoL separately using the HUI2 and 3; however, the agreement was found to be statistically significant and moderate using a consensus-based dyad approach [49].

Table 4 Details of the included studies of level of agreement by overall utilities between self- and proxy-reported HRQoL using preference-based quality-of-life instruments

Authors (intervention)	Measure	Proxy type	Sample size dyad	Correlation test	Correlation coefficient (*p* value)	95% CI
Barr et al. [34]	HUI2	Nurses Physicians	15 12	ICC	0.85 0.95	
Glaser et al. [36]	HUI2	Physiotherapist Parents Physicians	25 24 19	ICC	0.4 0.57 0.15	
Glaser et al. [37]	HUI2	Physiotherapist Parents Physicians	25 24 19	Pearson	0.54 (< 0.01) 0.59 (< 0.01) 0.37 (0.12)	
Sung et al. [56]	HUI2	Parents	19	ICC Spearman	0.11 (0.3) 0.14	− 0.35, 0.53 − 0.34, 0.55
	HUI3	Parents	19	ICC Spearman	− 0.01 0.11	− 0.45, 0.44 0.35, 0.55
Fu et al. [42]	HUI2	Parents Physicians	120 156	ICC	0.389 0.379	0.227, 0.531 0.237, 0.506
	HUI3	Parents Physicians	156 166	ICC	0.433 0.341	0.297, 0.552 0.200, 0.469
Banks et al. [48]	HUI2	Parents	11	ICC	0.74	0.29, 0.92
	HUI3	Parents	11	ICC	0.42	− 0.21, 0.80
Fluchel et al. [43]	HUI3	Parents Physicians	92 91	ICC	0.3087 0.066	0.1125, 0.4818 − 0.1402, 0.2669
Fluchel et al. [43] (control)	HUI3	Physicians/teachers	89	ICC	− 0.3103	− 0.4857, − 0.1106
Penn et al. [59]	HUI3	Parents	21	Spearman	0.76 (< 0.001)	
Penn et al. [59] (control)	HUI3	Parents	22	Spearman	0.31	
Zhou et al. [54] (baseline)	EQ VAS	Caregiver	96	ICC	0.22	
Zhou et al. [54] (follow-up)	EQ VAS	Caregiver	96	Yes	0.556	
Czyzewski et al. [62]	QWB	Parents	55	Pearson	0.39	
Brunner et al. [55]	HUI3	Parents	45	ICC Pearson	0.43 0.57	
Belfort et al. [57] (overall)	HUI3	Parents	63	Spearman	0.47 (0.0002)	
Lee et al. [58]	HUI3	Parents	223	Spearman	0.34	0.22, 0.45
Rhodes et al. [60]	HUI3	Parents	96	Spearman	0.24 (< 0.05)	
Ungar et al. [49]	HUI2	Parents	72	ICC	0.021	− 0.22, 0.262
	HUI3	Parents	75	ICC	0.169	− 0.070, 0.389
Ungar et al. [49] (Dyad)	HUI2	Parent with child	72	ICC	0.545 (< 0.0001)	0.360, 0.689
	HUI3	Parent with child	75	ICC	0.735 (< 0.0001)	0.611, 0.824
Kulpeng et al. [44]	HUI2	Caregiver	74	Pearson	0.58 (< 0.05)	
	HUI3	Caregiver	74	Pearson	0.67 (< 0.05)	
	EQ-5D	Caregiver	74	Pearson	0.77 (< 0.05)	
	EQ VAS	Caregiver	74	Pearson	0.5 (< 0.05)	
Sims-Williams et al. [63]	HUI3	Caregiver	62	Pearson	0.848	
Bharij et al. [50]	HUI2	Parents	61	ICC	0.9 (< 0.001)	
	HUI3	Parents	60	ICC	0.75 (< 0.001)	
	CHU9D	Parents	96	ICC	0.69 (< 0.001)	
Bray et al. [45]	HUI2	Parents	13	Spearman	0.728 (0.005)	
	HUI3	Parents	13	Spearman	0.842 (< 0.001)	
	EQ-5D-Y	Parents	11	Spearman	0.665 (0.026)	
	EQ VAS	Parents	13	Spearman	0.545 (0.054)	
Perez Sousa et al. [46]	EQ VAS	Mother Father	62 62	ICC	0.389 (0.029) 0.581 (0.962)	
Perez Sousa et al. [51] (overall: baseline)	EQ VAS	Parents	151	ICC	0.5 (< 0.0001)	

Table 4 (continued)

Authors (intervention)	Measure	Proxy type	Sample size dyad	Correlation test	Correlation coefficient (*p* value)	95% CI
Perez Sousa et al. [51] (overall: follow-up: post treatment)	EQ VAS	Parents	151	ICC	0.7 (< 0.0001)	
van Summeren et al. [52]	EQ VAS	Parents	56	ICC	0.78	0.65, 0.87
Rogers et al. [64]	CHU9D	Parents	184	Pearson	0.156 (0.02)	
Rogers et al. [64] (control)	CHU9D	Parents	302	Pearson	0.183 (0.01)	
Rogers et al. [64] (overall)	CHU9D	Parents	386	Pearson	0.183 (< 0.001)	
Shiroiwa et al. [47]	EQ VAS	Parents	654	ICC	0.06	
Sinlapamongkolkul et al. [61]	EQ VAS	Caregiver	85	Spearman	0.334 (0.001)	
Lin et al. [53] (overall)	EQ VAS	Caregiver	125	Yes	0.29	

HRQoL health-related quality of life, *CI* confidence interval, *HUI2* Health Utilities Index Mark 2, *HUI3* Health Utilities Index Mark 3, *EQ-5D-Y* EQ-5D Youth version, *VAS* visual analogue scale, *CHU9D* Child Health Utility 9 Dimensions, *QWB* Quality of Well-Being scale, *ICC* intraclass correlation coefficient

The agreement between children and physiotherapists was generally low with the exception of one study where physiotherapists reported higher agreement than parents and physicians within the HUI3 attributes of 'vision' and 'speech' [36, 37]. Overall, physicians reported excellent agreement when assessing the functional attributes, e.g., 'mobility' and 'ambulation', whereas the subjective attributes of 'emotion', 'pain' and 'cognition' lacked sufficient agreement [36, 37, 39, 42, 43].

Parents followed a similar suit and reported slight to fair agreement within the 'emotion and 'cognition' attributes of HUI2 and 3. In the assessment of 'emotion', the only exception was reported in a study of children with very low birth weight by Wolke et al., which showed moderate agreement with the parents in the study population [41]. Moreover, father/child pairs agreed only slightly within all domains of EQ-5D-Y. In comparison, a better agreement was reported with mothers for the domains 'walking about', 'doing usual activity' and 'having pain or discomfort' [46].

3.3.3 Inter-Rater Agreement Based on the Type of Condition

Within the cancer-related studies, children with a history of cancer showed a much better agreement (ICC 0.44, 95% CI 0.26–0.62) with the proxy reports than those with active cancer (ICC 0.34, 95% CI 0.04–0.64). In addition to the higher agreement level, correlations observed were also large for the former cohort (0.52, 95% CI 0.31–0.68), whereas cancer patients showed weak associations (0.40, 95% CI − 0.15 to 0.76) with the proxy reports of their HRQoL. It is unclear if cancer-related studies showed an overall lower agreement between the child self and proxy reports of HRQoL, than studies with conditions other than cancer. For instance, in a longitudinal study of cancer patients, Penn and colleagues found strong associations between the HUI3 generated overall utilities as reported by children and proxies in the study population, but weak correlations for those in the control group [59]. Conditions such as respiratory (asthma) and musculoskeletal diseases assessed using HUI2 and 3 showed poor inter-rater agreement between child self- and proxy-reported utilities [49, 55]. Using the EQ VAS, van Summeren and colleagues found good inter-rater agreement in children with functional constipation [52]. Additionally, in a longitudinal study of children with obesity, the agreement between children and parents for EQ VAS scores was found to be moderate at baseline and at follow-up [51]. Strong associations (Spearman's rho) were noted between the utilities reported by children with cerebral palsy, hemiplegia, and/or muscular dystrophy and their parents using both EQ-5D-Y and EQ VAS [45], while the correlation between children with thalassaemia and their caregivers using the EQ VAS was weak [61]. Kulpeng et al. also indicated a large correlation (Pearson's *r*) between self- and proxy-derived utilities using EQ-5D and EQ VAS in children with severe childhood infections [44].

The agreement and correlation between child self- and proxy-reported overall HRQoL observed between healthy children and proxies, including parents, physicians, and teachers, was, on average, low [43, 47]; however, evidence for the domain-level agreement was inconsistent. Kappa values in the study by Wolke et al., suggested moderate to almost perfect agreement between children with no specific health condition and parents across all HUI3 attributes [41]. In contrast, another study observed perfect agreement only within the 'hearing', 'ambulation', and 'dexterity' attributes, while the remaining attributes showed poor or no agreement [43]. Notably, this study used physicians/teachers as proxies rather than parents, which could potentially account for the contrasting findings. Similarly, one of the two studies using the EQ-5D-Y reported a moderate to almost perfect agreement across all domains except 'having pain or

Table 5 Details of the included studies' level of agreement by domains (attributes) between self- and proxy-reported HRQoL using preference-based quality-of-life instruments

Authors (intervention)	Measure	Proxy type	Attribute	Statistic reported	Agreement statistic (*p* value)	95% CI
Barr et al. [34]	HUI2	Nurses	Sensation	Cohen's kappa	0.05	
			Emotion		0.13	
			Cognition		0.54	
			Pain		0.71	
	HUI2	Physicians	Sensation	Cohen's kappa	0.42	
			Emotion		0.13	
			Cognition		0.37	
			Pain		0.73	
Fu et al. [42]	HUI2	Parents	Sensation	ICC	0.773	0.706, 0.826
			Mobility		0.67	0.584, 0.742
			Emotion		0.104	− 0.058, 0.262
			Cognition		0.121	− 0.026, 0.263
			Self-care		0.422	0.298, 0.532
			Pain		0.14	− 0.002, 0.277
	HUI2	Physicians	Sensation	ICC	0.829	0.778, 0.870
			Mobility		0.569	0.465, 0.657
			Emotion		0	− 0.143, 0.143
			Cognition		0.102	− 0.045, 0.245
			Self-care		0.754	0.686, 0.810
			Pain		0.08	− 0.063, 0.219
Morrow et al. [39]	HUI2	Parents	Sensation	Cohen's kappa	0.51	0.23, 0.78
			Mobility		0.59	0.31, 0.86
			Emotion		0.32	0.10, 0.53
			Cognition		0.29	0.35, 0.54
			Pain		0.44	0.23, 0.64
	HUI2	Physicians	Sensation	Cohen's kappa	0.27	− 0.26, 0.56
			Mobility		0.62	0.37, 0.88
			Emotion		0.18	− 0.03, 0.88
			Cognition		0.07	− 0.16, 0.30
			Pain		0.11	− 0.11, 0.34
	HUI3	Parents	Ambulation	Cohen's kappa	0.52	0.29, 0.77
			Dexterity		0.12	− 0.11, 0.34
			Emotion		0.27	0.04, 0.51
			Cognition		0.32	0.09, 0.55
			Pain		0.43	0.25, 0.62
	HUI3	Physicians	Ambulation	Cohen's kappa	0.56	0.31, 0.82
			Dexterity		0.11	− 0.12, 0.33
			Emotion		0.16	− 0.05, 0.37
			Cognition		0.05	− 0.11, 0.20
			Pain		0.36	0.17, 0.55
Glaser et al. [36]	HUI2	Physiotherapist	Sensation	Cohen's kappa	0.32	
			Mobility		NS	
			Emotion		0.37	
			Cognition		0.7	
			Self-care		0.43	
			Pain		NS	
	HUI2	Parents	Sensation	Cohen's kappa	0.54	
			Mobility		0.72	

Table 5 (continued)

Authors (intervention)	Measure	Proxy type	Attribute	Statistic reported	Agreement statistic (*p* value)	95% CI
			Emotion		0.37	
			Cognition		NS	
			Self-care		0.47	
			Pain		0.62	
	HUI2	Physicians	Sensation	Cohen's kappa	0.38	
			Mobility		0.77	
			Emotion		NS	
			Cognition		NS	
			Self-care		0.78	
			Pain		NS	
Glaser et al. [37]	HUI3	Physiotherapist	Vision	Cohen's kappa	0.62	
			Hearing		0.12	
			Speech		0.64	
			Ambulation		0.19	
			Dexterity		0.77	
			Emotion		0.4	
			Pain		0.33	
	HUI3	Parents	Vision	Cohen's kappa	0.62	
			Hearing		0.49	
			Speech		0.47	
			Ambulation		0.73	
			Dexterity		0.82	
			Emotion		0.28	
			Pain		0.56	
	HUI3	Physicians	Vision	Cohen's kappa	0.6	
			Hearing		0.67	
			Speech		0.14	
			Ambulation		0.77	
			Dexterity		0.48	
			Emotion		0.14	
			Pain		0.14	
Ungar et al. [49]	HUI2	Parents	Mobility	ICC	0.108	− 0.101, 0.308
			Emotion		0.065	− 0.155, 0.278
	HUI2	Parent with child	Mobility	ICC	0.713	0.593, 0.802
			Emotion		0.468	0.281, 0.621
Verrips et al. [38]: Mail	HUI3	Parents	Vision	Cohen's kappa	0.87	
			Hearing		0.33	
			Speech		0.23	
			Ambulation		0.66	
			Dexterity		0.63	
			Emotion		0.29	
			Cognition		0.36	
			Pain		0.43	
Verrips et al. [38]: Telephone	HUI3	Parents	Vision	Cohen's kappa	0.69	
			Speech		0.21	
			Ambulation		0.73	
			Dexterity		0.61	
			Emotion		0.2	
			Cognition		0.17	

Table 5 (continued)

Authors (intervention)	Measure	Proxy type	Attribute	Statistic reported	Agreement statistic (*p* value)	95% CI
			Pain		0.22	
Verrips et al. [38]: Face-to-face	HUI3	Parents	Vision	Cohen's kappa	0.75	
			Hearing		0	
			Speech		0.19	
			Ambulation		0.39	
			Dexterity		0.8	
			Emotion		0.07	
			Cognition		0.09	
			Pain		0.08	
Wolke et al. [41]	HUI3	Parents	Vision	Cohen's kappa	0.87	0.88, 0.86
			Hearing		0.59	0.59, 0.59
			Speech		0.22	0.22, 0.22
			Ambulation		0.78	0.78, 0.78
			Dexterity		0.67	0.68, 0.66
			Emotion		0.41	0.42, 0.4
			Cognition		0.32	0.32, 0.32
			Pain		0.48	0.49, 0.47
Wolke et al. [41]: General health (control)	HUI3	Parents	Vision	Cohen's kappa	0.82	0.81, 0.83
			Hearing		1	0.99, 1.01
			Speech		0.23	0.23, 0.23
			Dexterity		0.67	0.66, 0.68
			Emotion		0.37	0.36, 0.38
			Cognition		0.2	0.2, 0.2
			Pain		0.46	0.45, 0.47
Gusi et al. [40]			Pain or discomfort		0.68 (< 0.05)	
			Worried, sad, or unhappy		0.221 (< 0.05)	
Jelsma and Ramma [35]	EQ-5D-Y	Mother	Mobility	Cohen's kappa	0.15	
			Self-care		0.08	
			Doing usual activities		0.01	
			Pain or discomfort		0.2	
			Worried, sad, or unhappy		0.21	
Jelsma and Ramma [35]: General health (control)	EQ-5D-Y	Mother	Mobility	Cohen's kappa	0.6	
			Self-care		0.33	
			Doing usual activities		0.34	
			Pain or discomfort		0.41	
			Worried, sad, or unhappy		0.22	
Perez Sousa et al. [46]	EQ-5D-Y	Mother	Mobility	Cohen's kappa	0.713 (< 0.001)	
			Self-care		0.057 (0.536)	
			Doing usual activities		0.436 (< 0.001)	
			Pain or discomfort		0.128 (0.183)	
			Worried, sad, or unhappy		0.165 (0.14)	
	EQ-5D-Y	Father	Mobility	Cohen's kappa	0.042 (0.653)	
			Self-care		0.044 (0.622)	
			Doing usual activities		0.019 (0.841)	
			Pain or discomfort		0.067 (0.469)	
			Worried, sad, or unhappy		0.016 (0.854)	

Table 5 (continued)

Authors (intervention)	Measure	Proxy type	Attribute	Statistic reported	Agreement statistic (*p* value)	95% CI
Perez Sousa et al. [51]	EQ-5D-Y	Parents	Mobility	Cohen's kappa	0.51 (< 0.001)	
			Self-care		0.36 (< 0.001)	
			Doing usual activities		0.22 (< 0.001)	
			Pain or discomfort		0.27 (< 0.001)	
			Worried, sad, or unhappy		0.42 (< 0.001)	
Perez Sousa et al. [51]: control	EQ-5D-Y	Parents		Cohen's kappa	0.15 (0.03)	
			Self-care		0.13 (0.04)	
			Doing usual activities		0.09 (0.19)	
			Pain or discomfort		0.26 (< 0.001)	
			Worried, sad, or unhappy		0.37 (< 0.001)	
Shiroiwa et al. [47]	EQ-5D-Y	Parents	Mobility	Cohen's kappa	0.5	
			Self-care		0.91	
			Doing usual activities		0.78	
			Pain or discomfort		0.15	
			Worried, sad, or unhappy		0.12	

HRQoL health-related quality of life, *CI* confidence interval, *HUI2* Health Utilities Index Mark 2, *HUI3* Health Utilities Index Mark 3, *EQ-5D-Y* EQ-5D Youth version, *ICC* intraclass correlation coefficient, *NS* non-significant

discomfort' and 'feeling worried, sad or unhappy', while the other reported lower agreement ranging from slight to fair across all domains [35, 47].

3.4 Meta-Analysis Results

In the following, results for the meta-analysis are provided for studies that reported the ICC (95% CI) for the overall utilities and Cohen's kappa for the domain-level HRQoL. Nine studies were included in the analysis to estimate the ICC for overall utilities elicited using child-specific generic preference-based measures [34, 36, 37, 42, 43, 48–50, 55, 56]. Six studies that reported the ICCs for EQ VAS scores were excluded as there is some debate in the literature about VAS scores and the extent to which they can be interpreted as utilities [46, 47, 51–54]. Kappa statistics for the domain-level agreement were reported for 10 studies employing HUI2 and 3 (five studies) [34, 36–39, 41] and EQ-5D-Y (five studies) [35, 40, 46, 47, 51]. However, since four of five studies using EQ-5D-Y did not report the standard errors of the kappa values or the percentage agreement values, the EQ-5D measure was excluded from the domain-level meta-analysis of agreement.

3.4.1 Inter-Rater Agreement for Overall Utilities

The overall ICC for all 24 samples using HUI2 and 3 with CHU9D was 0.49 (0.34–0.61) and without CHU9D was 0.48 (0.32–0.61). Figure 2 depicts the study-specific and overall estimates of ICC, their respective 95% CIs and the study weights (%). The test for homogeneity resulted in a Q test statistic of 196.18 ($p < 0.001$). The heterogeneity in the studies was high ($I^2 = 91\%$) due to the presence of high variability between studies.

Exploratory moderators such as type of measure, health condition, proxy, and the age composition of the children in the sample were used to potentially explain this heterogeneity. The moderators were categorised according to the (1) type of measure used—HUI2 (12 samples) or HUI3 (11 samples) or CHU9D (1 sample); (2) health condition assessed—cancer- (15 samples) or non-cancer-related (9 samples); (3) type of proxy used—parent/caregiver (16 samples) or health professional/teacher (8 samples); and (4) lower age limit of the sample—below 8 years (10 samples) or 8 years and above (14 samples).

HUI3 had an estimated ICC of 0.37 (0.18–0.53), much lower than HUI2, which had an estimated ICC of 0.58 (0.34–0.75). The overall ICC for cancer-related samples was 0.43 (0.27–0.57), whereas for samples with conditions other than cancer, including general health, it was 0.54 (0.28–0.73). The ICC estimate for parent proxies was 0.49 (0.31–0.63), whereas for health professionals it was only marginally lower at 0.47 (0.11–0.72). Samples that also included younger children had an ICC of 0.39 (0.33–0.44), which was lower than the ICC of 0.5 (0.44–0.56) with older children. However, none of the group differences were statistically significant and therefore did not suggest moderation by any of the included variables.

The results of the meta-regression showed that none of the explanatory variables were statistically significant, thus showing no significant differences in child and proxy agreement according to the type of measure, health condition

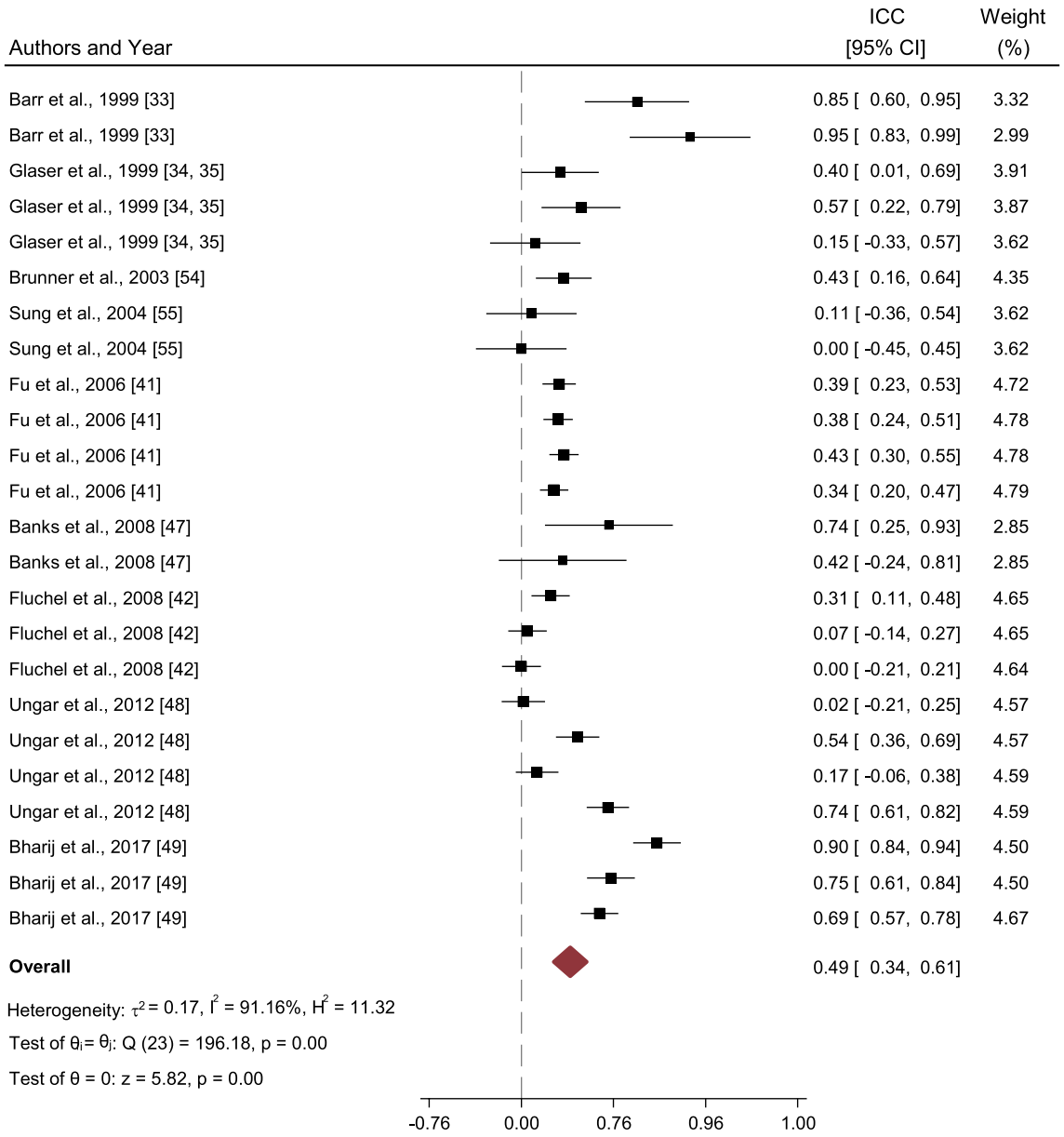

Fig. 2 Summary of the interrater reliability across studies. The forrest plot depicts the study-specific and overall estimates of ICCs, their respective 95% CIs and the study weight (%) for 24 studies obtained using a random effects model. *ICCs* intraclass correlation coefficients, *CIs* confidence intervals

experienced, proxy type and the inclusion of children below 8 years in the sample. The funnel plot and the funnel-plot test for asymmetry ($p = 0.133$) did not suggest any publication bias.

3.4.2 Inter-Rater Agreement for Domain-Level Health-Related Quality of Life

The estimated kappa and its 95% CI for HUI2 and 3 attributes is summarised in Table 6. In total, 36 samples for HUI2 and 68 samples for HUI3 were synthesised for the meta-analysis. The estimated kappa values for HUI2 attributes of 'emotion' (0.25), 'cognition' (0.3) and 'pain' (0.38), and the HUI3 attributes of 'cognition' (0.23), 'emotion' (0.27), 'speech' (0.3) and 'pain' (0.36) were the lowest. In contrast, there was higher agreement for the more easily observable physical- or function-related attributes such as 'mobility' (0.61) for HUI2 and 'ambulation' (0.64), 'dexterity' (0.65) and 'vision' (0.78) for HUI3. The heterogeneity was lower for HUI2 studies ($I^2 = 75\%$) than for HUI3 studies ($I^2 = 90\%$). Although no small-study bias was present in the

analysis of HUI3 samples ($p = 0.327$), there was a possibility of such a bias using the HUI2 samples ($p = 0.003$).

4 Discussion

To our knowledge, this is the first study to comprehensively examine the evidence relating to the level of agreement between child- and proxy-reported paediatric HRQoL using generic preference-based measures across health conditions. This study systematically reviewed the papers reporting agreement measures to describe the inter-rater agreement in the assessment of paediatric HRQoL by child self and proxy reports.

Thirty studies were identified that reported the agreement statistics between child self- and proxy-reported overall and/or domain-level HRQoL. Most of these studies showed poor inter-rater agreement for overall utilities. At the domain level, there were some important differences common to all the generic preference-based measures. In particular, the agreement between children and proxy respondents was weaker for psychosocial-related HRQoL domains and stronger for physical HRQoL domains. No studies that reported agreement measures between self- and proxy-reported overall utilities over time were identified. This is an important omission as repeated HRQoL assessments over time form critical inputs for the calculation of QALYs for CUA. Divergences in self- and proxy-reported childhood utilities over time may impact, potentially substantially, upon the results of economic evaluations and regulatory decision making for the recommendation of new pharmaceuticals/medical technologies.

It is unclear if the preference-based measure/s applied in the identified studies have any influence on the level of agreement between self- and proxy-reported paediatric HRQoL. In this review, we found a greater agreement with HUI2 than HUI3. There are two main differences between the measures. First, the two measures differ in their response levels. HUI3 has 5–6 response levels whereas HUI2 has 3–5 [65]. Intuitively, a higher inter-rater agreement would be expected with measures with fewer response levels if the inter-rater agreement depended on the response levels within the measure. However, a study evaluating the child and proxy agreement using the EQ-5D-Y-3L and -5L versions found a higher agreement with the five-response-level version than with three [66]. Second, HUI2 and HUI3 have different underlying constructs for the attributes with the same name. For example, in HUI2 the 'emotion' attribute assesses distress and anxiety, while the HUI3 frames 'emotion' in terms of happiness rather than depression [65]. Currently, there is insufficient evidence to investigate whether the discrepancy reflects this difference or is a coincidental finding.

Table 6 Domain (attribute)-level overall kappa estimates with their 95% CIs for HUI2 and 3

Attribute	Agreement ($\hat{\kappa}$)	Lower 95% CI	Upper 95% CI
HUI2			
Self-care	0.576	0.347	0.806
Cognition	0.296	0.088	0.505
Emotion	0.250	0.158	0.342
Mobility	0.615	0.463	0.767
Pain	0.385	0.148	0.622
Sensation	0.409	0.306	0.512
HUI3			
Ambulation	0.641	0.535	0.747
Cognition	0.229	0.145	0.313
Dexterity	0.646	0.541	0.751
Emotion	0.272	0.190	0.353
Hearing	0.497	0.232	0.762
Pain	0.361	0.265	0.457
Speech	0.300	0.174	0.427
Vision	0.782	0.713	0.850

CIs confidence intervals, *HUI2* Health Utilities Index Mark 2, *HUI3* Health Utilities Index Mark 3, $\hat{\kappa}$ estimated kappa value

The agreement for EQ VAS was lower than for the EQ-5D-Y domains. This may be attributed to the fact that the VAS and the domain-level responses are elicited using different response scales. The VAS has a response scale from 0 to 100, whereas each of the five domains are described using a 3- or 5-level response scale [3]. Hence, a higher discrepancy may be expected with VAS due to the much larger range for its response scale.

Proxy type used was found to have some influence on the level of agreement between self- and proxy-reported paediatric HRQoL. The findings of HRQoL studies conducted in a paediatric oncology setting suggest that the information obtained from the child, the parent and the health professional are generally complementary and valid [67]. However, Sprangers and Aronson concluded that health professionals generally tend to underestimate the pain and also, conversely, the overall HRQoL of the individual [68]. While able to accurately assess the patient's physical condition, health professionals often failed to consider the emotional and social components of HRQoL [69]. In line with previous studies in adult cancer patients where agreement was higher with close companions, the child/parent agreement in this review was also found to be higher compared with child/health professional agreement [70]. Moreover, mothers demonstrated a higher agreement than fathers. This gender disparity may be associated with their degree of involvement in childcare [71].

The level of inter-rater agreement decreases with more severe conditions [69]. A recent study in paediatric

patients found that the agreement between children and caregivers was higher when their condition improved compared with when they were ill [66]. We found that cancer-related cohorts had a lower overall agreement than cohorts with or without health conditions other than cancer. Interestingly, a low inter-rater agreement was seen between children with no obvious health conditions and their parents. One study showed worse correlations between parents and healthy children than children with a history of cancer [43]. These findings should be explored in more detail to determine whether this is a demonstrable trend. Self and proxy agreement data in the assessment of mental illnesses remains scarce. Studies have examined HRQoL in children with mental or behavioural disorders using preference-based measures, but none have assessed the level of child/proxy agreement [72, 73].

Self-report using the EQ-5D-Y has been prescribed for children aged 8 years and older [3]. The use of HUI2/3 was not recommended for self-report in children under 12 years of age [65]; however, studies have reportedly used these measures for self-completion in children younger than the recommended age group [35, 45, 48]. The minimum age at which children can reliably and accurately self-report has not been conclusively identified yet and is likely to be influenced by a variety of factors (including the reading and comprehension abilities of the child, the measure/s being applied and the mode of completion) [6]. There also remains a gap in the literature exploring the potential for differential levels of agreement between proxies and children by age groups. A previous study in a sample of children aged 8–18 years has shown that agreement decreases with age [74]. In this review, one study reported the agreement statistics (Gwet's AC1) for children (10–12 years) and adolescents (13–15 years) separately. In both groups, the correlation between child self- and proxy-reported domain-level HRQoL was strong and positive, with a marginally stronger association reported between adolescents and caregivers than children and caregivers [53]. Due to these inconsistent findings, further research is needed to determine if an age differential exists in the level of child/proxy agreement.

We found that 33% of the studies reported only the correlation coefficients that were synthesised to describe the inter-rater agreement in this review. The difference between agreement and correlation has been addressed in literature [19]. However, until recently, standalone correlation coefficients have been employed to assess agreement between child self and proxy report [75]. Correlation and agreement both measure the strength of association between two the variables of interest; however, the key difference is that agreement coefficients, in addition, account for the absolute agreement between the raters. Correlations may be high even if the ratings are not equal but only vary similarly. On the other hand, a perfect agreement would imply that all ratings, by each rater, are the same [14, 18]. Thus, correlation coefficients, if used, presented along with agreement statistics may provide a more comprehensive picture of the level of agreement.

This study has several limitations that are important to highlight. The inter-rater agreement for overall utilities and for the respective domains was quantitatively examined for only HUI2 and 3 for the following reasons. (1) HUI measures were widely used among the studies included in this analysis, with HUI3 being the most dominant. (2) Despite its relatively wide application, the majority of the identified studies using the EQ-5D-Y did not report the overall utilities, potentially due to the absence of an established preference-based scoring algorithm for the EQ-5D-Y to date. When reported, only the correlation (using Pearson's *r* or Spearman's rho) between the child self- and proxy-reported utilities was examined. While agreement was reported for the EQ VAS scores, they were not pooled due to paucity of evidence demonstrating the comparability of the VAS scores with the index scores. The EQ VAS scores were therefore not included in the meta-analysis. Furthermore, due to a lack of studies reporting the domain-level agreement between self and proxy reports of paediatric HRQoL, along with percentage agreement, the meta-analysis of the EQ-5D-Y domains was not feasible. (3) The analysis of the agreement level using the CHU9D and the QWB was also limited due to inadequate reporting of agreement statistics. Interpretation of the results of the meta-analysis is bounded by the presence of high heterogeneity between studies, which could not be explained by the subgroup analysis. Furthermore, due to practical resource constraints, we were only able to include articles published in the English language.

5 Conclusion

This systematic review summarising the agreement between child self and proxy rating of HRQoL using established generic preference-based measures generally found a poor inter-rater agreement. Convergence with child self-rating was more likely in the proxy assessment of paediatric HRQoL within domains with observable attributes e.g., physical health domains, than with less-observable attributes e.g., psychosocial domains. Further research to drive the inclusion of children in self-reporting their own HRQoL wherever possible and limiting the reliance on proxy reporting of children's HRQoL is warranted.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40273-022-01177-z.

Acknowledgements The authors would like to thank Shannon Brown and Pawel Skuza from Flinders University. Shannon Brown provided helpful advice on strategies for the literature search, and Pawel Skuza provided statistical consultation.

QUOKKA Project Team—Collaborating investigators: Nancy Devlin, Richard Norman, Rosalie Viney, Julie Ratcliffe, Kim Dalziel, Brendan Mulhern, Harriet Hiscock, Deborah Street, Gang Chen; Research fellows and PhD students: Tessa Peasgood, Cate Bailey, Christine Mpundu-Kaambwa, Alice Yu, Mina Bahrampour, Renee Jones, Rachel O'Loughlin, Yiting Luo, Alex van Heusden, Xiuqin Xiong, Diana Khanna, Ashwini De Silva.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Data availability statement All data generated and/or analysed during this study are included in this published article (and its supplementary information file).

Author contributions JR, JK, CM-K and DK conceptualised this study. DK led the systematic review with contributions from JK, CM-K and JR. DK and KL performed abstract and title screening. DK, KL, CM-K, JK and JR screened the full text of the articles for inclusion. DK wrote the first draft. JR, JK and CM-K provided feedback on the first draft and agreed on the final draft. All authors reviewed and approved the final amendments. DK and JR act as guarantors of the review.

Funding DK is supported by a PhD scholarship awarded from a project funded by the Department of Health, Australian Government, National Health and Medical Research Council (grant number MRF1200816). Project title'Quality of Life in Kids: Key Evidence to Strengthen Decisions in Australia (QUOKKA)'. The project aims to improve how quality of life is measured and valued within paediatric populations. The funder was not involved in developing this protocol.

Conflict of interest Diana Khanna, Jyoti Khadka, Christine Mpundu-Kaambwa, Kiri Lay, Remo Russo and Julie Ratcliffe declare no conflicts of interest.

References

1. Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2016.
2. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (version 5.0). Australian Government, Department of Health; 2016.
3. Chen G, Ratcliffe J. A review of the development and application of generic multi-attribute utility instruments for paediatric populations. Pharmacoeconomics. 2015;33(10):1013–28. https://doi.org/10.1007/s40273-015-0286-7.
4. Virgili G, Koleva D, Garattini L, Banzi R, Gensini GF. Utilities and QALYs in health economic evaluations: glossary and introduction. Intern Emerg Med. 2010;5(4):349–52. https://doi.org/10.1007/s11739-010-0420-7.
5. Connolly MA, Johnson JA. Measuring quality of life in paediatric patients. Pharmacoeconomics. 1999;16(6):605–25. https://doi.org/10.2165/00019053-199916060-00002.
6. Matza LS, Patrick DL, Riley AW, Alexander JJ, Rajmil L, Pleil AM, et al. Pediatric patient-reported outcome instruments for research to support medical product labeling: report of the ISPOR PRO good research practices for the assessment of children and adolescents task force. Value Health. 2013;16(4):461–79. https://doi.org/10.1016/j.jval.2013.04.004.
7. Petrou S. Methodological issues raised by preference-based approaches to measuring the health status of children. Health Econ. 2003;12(8):697–702. https://doi.org/10.1002/hec.775.
8. Kwon J, Kim SW, Ungar WJ, Tsiplova K, Madan J, Petrou S. A systematic review and meta-analysis of childhood health utilities. Med Decis Mak. 2017;38(3):277–305. https://doi.org/10.1177/0272989X17732990.
9. Roddenberry A, Renk K. Quality of life in pediatric cancer patients: the relationships among parents' characteristics, children's characteristics, and informant concordance. J Child Fam Stud. 2008;17(3):402–26. https://doi.org/10.1007/s10826-007-9155-0.
10. Khadka J, Kwon J, Petrou S, Lancsar E, Ratcliffe J. Mind the (inter-rater) gap. An investigation of self-reported versus proxy-reported assessments in the derivation of childhood utility values for economic evaluation: a systematic review. Soc Sci Med. 2019;240: 112543. https://doi.org/10.1016/j.socscimed.2019.112543.
11. Jiang M, Ma Y, Li M, Meng R, Ma A, Chen P. A comparison of self-reported and proxy-reported health utilities in children: a systematic review and meta-analysis. Health Qual Life Outcomes 2021;19(1):45. https://doi.org/10.1186/s12955-021-01677-0.
12. Otero S, Eiser C, Wright N, Butler G. Implications of parent and child quality of life assessments for decisions about growth hormone treatment in eligible children. Child Care Health Dev. 2013;39(6):782–8.
13. Shoukri MM. Measurement of agreement. Wiley StatsRef: Statistics Reference Online. © 2014, John Wiley & Sons, Ltd
14. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002.
15. Babineau J. Product review: covidence (systematic review software). J Can Health Libr Assoc Journal de l'Association des Bibliothèques de la Santé du Canada. 2014;35(2):68–71. https://doi.org/10.5596/c14-016.
16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev. 2021;10(1):89. https://doi.org/10.1186/s13643-021-01626-4.
17. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted

with personality disorder samples. BMC Med Res Methodol. 2013;13:61. https://doi.org/10.1186/1471-2288-13-61.
18. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. Perspect Clin Res. 2017;8(4):187–91. https://doi.org/10.4103/picr.PICR_123_17.
19. Liu J, Tang W, Chen G, Lu Y, Feng C, Tu XM. Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Arch Psychiatry. 2016;28(2):115–20. https://doi.org/10.11919/j.issn.1002-0829.216045.
20. Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Pract Assess Res Eval. 2004;9:4.
21. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104.
22. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
23. Juniper EF, Guyatt GH, Jaeschke R. How to develop and validate a new health-related quality of life instrument. In: Spilker B, editor. Quality of life and pharamacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1996. p. 49–56.
24. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New York: Academic Press; 1988. p. 82.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
26. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. 2020;368: l6890. https://doi.org/10.1136/bmj.l6890.
27. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC methods programme. Version. 2006;1:b92. https://doi.org/10.13140/2.1.1018.4643.
28. Raudenbush SW. Analyzing effect sizes: random-effects models. The handbook of research synthesis and meta-analysis. 2nd ed. New York: Russell Sage Foundation; 2009. p. 295–315.
29. Sánchez-Meca J, López-López JA, López-Pina JA. Some recommended statistical analytic practices when reliability generalization studies are conducted. Br J Math Stat Psychol. 2013;66(3):402–25. https://doi.org/10.1111/j.2044-8317.2012.02057.x.
30. Sun S. Meta-analysis of Cohen's kappa. Health Serv Outcomes Res Method. 2011;11(3):145–63. https://doi.org/10.1007/s10742-011-0077-3.
31. Tufanaru C, Munn Z, Stephenson M, Aromataris E. Fixed or random effects meta-analysis? Common methodological issues in systematic reviews of effectiveness. Int J Evid Based Healthc. 2015;13(3):196–207. https://doi.org/10.1097/xeb.0000000000000065.
32. Kmet LM, Cook LS, Lee RC. Standard quality assessment criteria for evaluating primary research papers from a variety of fields. ERA. 2004. https://doi.org/10.7939/R37M04F16.
33. Papaioannou D, Brazier J, Paisley S. Systematic searching and selection of health state utility values from the literature. Value Health. 2013;16(4):686–95. https://doi.org/10.1016/j.jval.2013.02.017.
34. Barr RD, Simpson T, Whitton A, Rush B, Furlong W, Feeny DH. Health-related quality of life in survivors of tumours of the central nervous system in childhood–a preference-based approach to measurement in a cross-sectional study. Eur J Cancer. 1999;35(2):248–55. https://doi.org/10.1016/s0959-8049(98)00366-9.
35. Jelsma J, Ramma L. How do children at special schools and their parents perceive their HRQoL compared to children at open schools? Health Qual Life Outcomes. 2010;8:72–72. https://doi.org/10.1186/1477-7525-8-72.
36. Glaser A, Kennedy C, Punt J, Walker D. Standardized quantitative assessment of brain tumor survivors treated within clinical trials in childhood. Int J Cancer Suppl. 1999;12:77–82. https://doi.org/10.1002/(sici)1097-0215(1999)83:12+%3c77::aid-ijc14%3e3.0.co;2-x.
37. Glaser AW, Furlong W, Walker DA, Fielding K, Davies K, Feeny DH, et al. Applicability of the Health Utilities Index to a population of childhood survivors of central nervous system tumours in the UK. Eur J Cancer. 1999;35(2):256–61. https://doi.org/10.1016/s0959-8049(98)00367-0.
38. Verrips GH, Stuifbergen MC, den Ouden AL, Bonsel GJ, Gemke RJ, Paneth N, et al. Measuring health status using the Health Utilities Index: agreement between raters and between modalities of administration. J Clin Epidemiol. 2001;54(5):475–81. https://doi.org/10.1016/s0895-4356(00)00317-6.
39. Morrow AM, Hayen A, Quine S, Scheinberg A, Craig JC. A comparison of doctors', parents' and children's reports of health states and health-related quality of life in children with chronic conditions. Child Care Health Dev. 2012;38(2):186–95. https://doi.org/10.1111/j.1365-2214.2011.01240.x.
40. Gusi N, Perez-Sousa MA, Gozalo-Delgado M, Olivares PR. Validity and reliability of the spanish EQ-5D-Y proxy version [in Spanish]. An Pediatr (Barc). 2014;81(4):212–9. https://doi.org/10.1016/j.anpedi.2013.11.028.
41. Wolke D, Chernova J, Eryigit-Madzwamuse S, Samara M, Zwierzynska K, Petrou S. Self and parent perspectives on health-related quality of life of adolescents born very preterm. J Pediatr. 2013;163(4):1020-6.e2. https://doi.org/10.1016/j.jpeds.2013.04.030.
42. Fu L, Talsma D, Baez F, Bonilla M, Moreno B, Ah-Chu M, et al. Measurement of health-related quality of life in survivors of cancer in childhood in Central America: feasibility, reliability, and validity. J Pediatr Hematol Oncol. 2006;28(6):331–41. https://doi.org/10.1097/00043426-200606000-00003.
43. Fluchel M, Horsman JR, Furlong W, Castillo L, Alfonz Y, Barr RD. Self and proxy-reported health status and health-related quality of life in survivors of childhood cancer in Uruguay. Pediatr Blood Cancer. 2008;50(4):838–43. https://doi.org/10.1002/pbc.21299.
44. Kulpeng W, Sornsrivichai V, Chongsuvivatwong V, Rattanavipapong W, Leelahavarong P, Cairns J, et al. Variation of health-related quality of life assessed by caregivers and patients affected by severe childhood infections. BMC Pediatr. 2013;13:122. https://doi.org/10.1186/1471-2431-13-122.
45. Bray N, Noyes J, Harris N, Edwards RT. Measuring the health-related quality of life of children with impaired mobility: examining correlation and agreement between children and parent proxies. BMC Res Notes. 2017;10(1):377. https://doi.org/10.1186/s13104-017-2683-9.
46. Perez Sousa M, Olivares Sánchez-Toledo PR, Gusi FN. Parent-child discrepancy in the assessment of health- related quality of life using the EQ-5D-Y questionnaire. Arch Argent Pediatr. 2017;115(6):541–6. https://doi.org/10.5546/aap.2017.eng.541.
47. Shiroiwa T, Fukuda T, Shimozuma K. Psychometric properties of the Japanese version of the EQ-5D-Y by self-report and proxy-report: reliability and construct validity. Qual Life Res. 2019;28(11):3093–105. https://doi.org/10.1007/s11136-019-02238-1.
48. Banks BA, Barrowman NJ, Klaassen R. Health-related quality of life: changes in children undergoing chemotherapy. J Pediatr Hematol Oncol. 2008;30(4):292–7. https://doi.org/10.1097/MPH.0b013e3181647bda.
49. Ungar WJ, Boydell K, Dell S, Feldman BM, Marshall D, Willan A, et al. A parent-child dyad approach to the assessment of health

status and health-related quality of life in children with asthma. Pharmacoeconomics. 2012;30(8):697–712. https://doi.org/10.2165/11597890-000000000-00000.
50. Bharij A, Neighbors K, Alonso EM, Mohammad S. Health utility and quality of life in pediatric liver transplant recipients. Pediatr Transpl. 2020;24(4): e13720. https://doi.org/10.1111/petr.13720.
51. Perez-Sousa MA, Olivares PR, Garcia-Hermoso A, Gusi N. Does anthropometric and fitness parameters mediate the effect of exercise on the HRQoL of overweight and obese children/adolescents? Qual Life Res. 2018;27(9):2305–12. https://doi.org/10.1007/s11136-018-1893-5.
52. van Summeren J, Klunder JW, Holtman GA, Kollen BJ, Berger MY, Dekker JH. Parent-child agreement on health-related quality of life in children with functional constipation in primary care. J Pediatr Gastroenterol Nutr. 2018;67(6):726–31. https://doi.org/10.1097/mpg.0000000000002124.
53. Lin J, Wong CKH, Cheung PWH, Luo N, Cheung JPY. Feasibility of proxy-reported EQ-5D-3L-Y and its agreement in self-reported EQ-5D-3L-Y for patients with adolescent idiopathic scoliosis. Spine (Phila Pa 1976). 2020;45(13):E799-e807. https://doi.org/10.1097/brs.0000000000003431.
54. Zhou W, Shen A, Yang Z, Wang P, Wu B, Herdman M, et al. Patient-caregiver agreement and test-retest reliability of the EQ-5D-Y-3L and EQ-5D-Y-5L in paediatric patients with haematological malignancies. Eur J Health Econ. 2021;22(7):1103–13. https://doi.org/10.1007/s10198-021-01309-w.
55. Brunner HI, Maker D, Grundland B, Young NL, Blanchette V, Stain AM, et al. Preference-based measurement of health-related quality of life (HRQL) in children with chronic musculoskeletal disorders (MSKDs). Med Decis Mak. 2003;23(4):314–22. https://doi.org/10.1177/0272989x03256008.
56. Sung L, Young NL, Greenberg ML, McLimont M, Samanta T, Wong J, et al. Health-related quality of life (HRQL) scores reported from parents and their children with chronic illness differed depending on utility elicitation method. J Clin Epidemiol. 2004;57(11):1161–6. https://doi.org/10.1016/j.jclinepi.2004.05.003.
57. Belfort MB, Zupancic JA, Riera KM, Turner JH, Prosser LA. Health state preferences associated with weight status in children and adolescents. BMC Pediatr. 2011;11:12. https://doi.org/10.1186/1471-2431-11-12.
58. Lee JM, Rhee K, O'Grady MJ, Basu A, Winn A, John P, et al. Health utilities for children and adults with type 1 diabetes. Med Care. 2011;49(10):924–31. https://doi.org/10.1097/MLR.0b013e318216592c.
59. Penn A, Lowis SP, Stevens MC, Shortman RI, Hunt LP, McCarter RJ, et al. A detailed prospective longitudinal assessment of health status in children with brain tumors in the first year after diagnosis. J Pediatr Hematol Oncol. 2011;33(8):592–9. https://doi.org/10.1097/MPH.0b013e31821388c0.
60. Rhodes ET, Prosser LA, Lieu TA, Songer TJ, Ludwig DS, Laffel LM. Preferences for type 2 diabetes health states among adolescents with or at risk of type 2 diabetes mellitus. Pediatr Diabetes. 2011;12(8):724–32. https://doi.org/10.1111/j.1399-5448.2011.00772.x.
61. Sinlapamongkolkul P, Surapolchai P. Health-related quality of life in Thai children with thalassemia as evaluated by PedsQL and EQ-5D-Y: a single-center experience. Mediterr J Hematol Infect Dis. 2020;12(1): e2020036. https://doi.org/10.4084/mjhid.2020.036.
62. Czyzewski DI, Mariotto MJ, Bartholomew LK, LeCompte SH, Sockrider MM. Measurement of quality of well being in a child and adolescent cystic fibrosis population. Med Care. 1994;32(9):965–72. https://doi.org/10.1097/00005650-199409000-00007.
63. Sims-Williams HJ, Sims-Williams HP, Mbabazi Kabachelor E, Warf BC. Quality of life among children with spina bifida in Uganda. Arch Dis Child. 2017;102(11):1057–61. https://doi.org/10.1136/archdischild-2016-312307.
64. Rogers HJ, Vermaire JH, Gilchrist F, Schuller AA. The relationship between caries-specific quality of life and generic wellbeing in a Dutch pediatric population. Dent J (Basel). 2019. https://doi.org/10.3390/dj7030067.
65. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. Health Qual Life Outcomes. 2003;1:54. https://doi.org/10.1186/1477-7525-1-54.
66. Fitriana TS, Purba FD, Stolk E, Busschbach JJV. EQ-5D-Y-3L and EQ-5D-Y-5L proxy report: psychometric performance and agreement with self-report. Health Qual Life Outcomes. 2022;20(1):88. https://doi.org/10.1186/s12955-022-01996-w.
67. Pickard AS, Knight SJ. Proxy evaluation of health-related quality of life: a conceptual framework for understanding multiple proxy perspectives. Med Care. 2005;43(5):493–9. https://doi.org/10.1097/01.mlr.0000160419.27642.a8.
68. Sprangers MAG, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol. 1992;45(7):743–60. https://doi.org/10.1016/0895-4356(92)90052-O.
69. Weinfurt KP, Trucco SM, Willke RJ, Schulman KA. Measuring agreement between patient and proxy responses to multidimensional health-related quality-of-life measures in clinical trials. An application of psychometric profile analysis. J Clin Epidemiol. 2002;55(6):608–18. https://doi.org/10.1016/s0895-4356(02)00392-x.
70. Sneeuw KC, Aaronson NK, Sprangers MA, Detmar SB, Wever LD, Schornagel JH. Value of caregiver ratings in evaluating the quality of life of patients with cancer. J Clin Oncol. 1997;15(3):1206–17. https://doi.org/10.1200/JCO.1997.15.3.1206.
71. Raley S, Bianchi SM, Wang W. When do fathers care? Mothers' economic contribution and fathers' involvement in child care. AJS Am J Sociol. 2012;117(5):1422–59. https://doi.org/10.1086/663354.
72. Creswell C, Violato M, Fairbanks H, White E, Parkinson M, Abitabile G, et al. Clinical outcomes and cost-effectiveness of brief guided parent-delivered cognitive behavioural therapy and solution-focused brief therapy for treatment of childhood anxiety disorders: a randomised controlled trial. Lancet Psychiatry. 2017;4(7):529–39. https://doi.org/10.1016/s2215-0366(17)30149-9.
73. Vermeulen KM, Jansen DE, Knorth EJ, Buskens E, Reijneveld SA. Cost-effectiveness of multisystemic therapy versus usual treatment for young people with antisocial problems. Crim Behav Ment Health. 2017;27(1):89–102. https://doi.org/10.1002/cbm.1988.
74. Rajmil L, López AR, López-Aguilà S, Alonso J. Parent-child agreement on health-related quality of life (HRQOL): a longitudinal study. Health Qual Life Outcomes. 2013;11:101. https://doi.org/10.1186/1477-7525-11-101.
75. Abraham S, Edginton E, Cottrell D, Tubeuf S. Measuring health-related quality of life measures in children: lessons from a pilot study. Res Psychother. 2022. https://doi.org/10.4081/ripppo.2022.581.

Authors and Affiliations

Diana Khanna[1] · Jyoti Khadka[1,2] · Christine Mpundu-Kaambwa[1] · Kiri Lay[1] · Remo Russo[3,4] · Julie Ratcliffe[1] on behalf of The Quality of Life in Kids: Key Evidence to Strengthen Decisions in Australia (QUOKKA) Project Team

Jyoti Khadka
jyoti.khadka@flinders.edu.au

Christine Mpundu-Kaambwa
christine.mpundu-kaambwa@flinders.edu.au

Julie Ratcliffe
julie.ratcliffe@flinders.edu.au

[1] Health and Social Care Economics Group, Caring Futures Institute, College of Nursing and Health Sciences, Flinders University, Adelaide, SA, Australia

[2] Registry of Senior Australians, Healthy Ageing Research Consortium, South Australian Health and Medical Research Institute, Adelaide, SA, Australia

[3] Department of Paediatric Rehabilitation, Women's and Children's Hospital, Adelaide, SA, Australia

[4] Faculty of Health Sciences, School of Medicine, Flinders University, Adelaide, SA, Australia