Quality of Life Research (2023) 32:1703–1716
https://doi.org/10.1007/s11136-023-03344-x

Comparative evaluation of the health utilities index mark 3 and the short form 6D: evidence from an individual participant data meta-analysis of very preterm and very low birthweight adults

Corneliu Bolbocean[1] · Peter J. Anderson[2,3] · Peter Bartmann[4] · Jeanie L. Y. Cheong[3,5,6] · Lex W. Doyle[3,5,6,7] · Dieter Wolke[8] · Stavros Petrou[1]

Accepted: 9 January 2023 / Published online: 27 January 2023

Abstract

Background The most appropriate preference-based health-related quality of life (HRQoL) instruments for trials or research studies that ascertain the consequences of individuals born very preterm and/or low birthweight (VP/VLBW) are not known. Agreement between the HUI3 and SF-6D multi-attribute utility measures have not been previously investigated for VP/VLBW and normal birthweight or term-born controls. This study examined the agreement between the outputs of the HUI3 and SF-6D measures among adults born VP/VLBW and normal birthweight or term born controls.

Methods We used two prospective cohorts of individuals born VP/VLBW and controls contributing to the 'Research on European Children and Adults Born Preterm' (RECAP) consortium which assessed HRQoL using two preference-based measures. The combined dataset of individual participant data (IPD) included 407 adult VP/VLBW survivors and 367 controls, ranging in age from 18 to 26 years. Bland–Altman plots, intra-class correlation coefficients, and generalized linear mixed models in a one-step approach were used to examine agreement between the measures.

Results There was significant discordance between the HUI3 and SF-6D multi-attribute utility measures in the VP/VLBW sample, controls, and in the combined samples. Agreement between the HUI3 and SF-6D multi-attribute utility measures was weaker in controls compared with VP/VLBW individuals.

Conclusions and relevance The HUI3 and SF-6D each provide unique information on different aspects of health status across the groups. The HUI3 better captures preterm-related changes to HRQoL in adulthood compared to SF-6D. Studies focused on measuring physical or cognitive aspects of health will likely benefit from using the HUI3 instead of the SF-6D, regardless of gestational age at birth and birthweight status.

Keywords HUI3 · SF-6D · HRQoL · Health utilities · Very preterm birth · Very low birth weight

Introduction

Health-related quality of life (HRQoL) is an important aid for evaluating clinical and policy interventions [1–3] and can be defined as "how well a person functions in their life and his or her perceived well-being in physical, mental, and social domains of health" [4]. Functioning refers to an individual's ability to carry out some pre-defined activities; however, well-being is understood as an individual's subjective feeling(s) [4]. HRQoL measures that are accompanied by preference-based value sets generate utility scores that reflect preferences for health states on a cardinal scale where 0 represents being dead and 1 represents full health [1]. Preference-based HRQoL measures are widely recommended for use in publicly funded health care systems because of their role in cost-utility analysis, which in forms reimbursement, regulatory, and pricing mechanisms [3, 5–8]. Furthermore, they are increasingly used as outcome measures in clinical trials and patient care [9–12] across a wide spectrum of conditions and environments [1–3, 13–15] partly because they are highly correlated with widely used health metrics, including morbidity, mortality, and healthcare costs [2, 3, 16]. However, there is discordance between the preference-based HRQoL measures that are recommended for use by health technology assessment agencies in different

✉ Corneliu Bolbocean
Corneliu.Bolbocean@phc.ox.ac.uk

Extended author information available on the last page of the article

jurisdictions [17]. Guidelines for selecting preference-based HRQoL instruments for randomized trials and observational studies are lacking.

Preference-based HRQoL measures are made up of descriptive systems and accompanying valuation systems. The descriptive system defines HRQoL across a number of health states and the valuation system is a mathematical construct for scoring each possible health state described by the measure. The valuation or scoring system generates utility scores that reflect population preferences for living in a particular health state. While these utility scores are indexed on a cardinal scale where 1 indicates full health and 0 represents death, negative values are theoretically possible and represent health states considered worse than death [1–3, 18].

The choice of a preference-based HRQoL measure is a critical decision because of downstream consequences related to cost-utility analysis, its use for deriving quality-adjusted life-years (QALYs), and subsequent resource allocation decisions [19]. Thus, evidence comparing the performance of preference-based HRQoL measures is needed to justify the selection of the most appropriate assessment tool [20]. Furthermore, researchers attempting to measure health outcomes face a trade-off on whether to include a single or multiple HRQoL instruments in their studies. The latter option is not always possible because of budgetary and time constraints as well as evidence for lower completion rates when multiple instruments are used [21–23].

The Health Utilities Index Mark 3 (HUI3) and Short Form 6D (SF-6D) are two widely used preference-based HRQoL measures that are anchored on a cardinal scale (with 0 = dead and 1 = full health) and generate utility scores that reflect population preferences for health states as well as for estimating QALYs for cost-utility analysis purposes [1–3, 18]. HRQoL measures accompanied by preference-based value sets are often referred to as multi-attribute utility instruments in the literature [1–3, 18]. Scoring algorithms for these measures have been derived based on nationally representative, community-based samples from different jurisdictions, such as Canada (HUI3) and the United Kingdom (SF-6D) [2, 3].

A recent review from health technology assessment (HTA) agencies regarding the preferred choice of preference-based HRQoL measures for cost-effectiveness based decision-making identified that, out of thirty-four guidelines, twenty-one recommended either the SF-6D ($n = 11$) or HUI2 or HUI3 ($n = 10$) instruments [17]. There is limited evidence for concurrent health status assessments using both the HUI3 and SF-6D instruments. However, at the same time, there is little consensus about the head-to-head performance of preference-based HRQoL measures across psychometric criteria.

Individuals born very preterm (VP; < 32 weeks' gestation) or at very low birthweight (VLBW; < 1500 g) are at high-risk of adverse functional, neurodevelopmental and behavioral outcomes [24–28] and their HRQoL is frequently examined because of increasing rates of preterm birth worldwide [13, 29, 30]. However, the agreement and discrepancies between the outputs of the HUI3 and SF-6D instruments have not been evaluated in head-to-head comparisons in a sample of VP or VLBW individuals.

This constrains efforts to enhance comparability and standardization of findings across different VP/VLBW studies, as well as reduces transparency and reproducibility of outcomes research in this area. It is important to determine the most appropriate HRQoL instruments for individuals born VP/VLBW because preterm birth and low birthweight represent a growing public health concern. Increasing VP/VLBW rates coupled with improvements in survival rates place increased pressures on healthcare budgets worldwide [13, 29, 30]. At the same time, the evidence regarding agreement between these measures in general population samples is opaque.

The first study that described (dis)agreement between the outputs of the HUI3 and SF-6D was published over twenty years ago [3][1]; however, subsequent studies struggled to provide conclusive evidence and explain the source(s) of (dis)agreement. One explanation is that existing studies generally have not evaluated concurrent agreement of the HUI3 and SF-6D instruments in general population samples [31–39], since most studies recruited participants with specific conditions within clinical settings, such as tertiary care [31–39] or primary care [35]. Thus, existing findings may not be generalizable across general population contexts. To our knowledge, only one study has assessed levels of agreement between the HUI3 and SF-6D measures [40] among healthy individuals. However, given the overall study design employed, the evidence related to levels of agreement of the HUI3 and SF-6D measures within the general population is not conclusive [1]. Furthermore, results from other studies were based on patients recruited into clinical trials [31, 32, 38], which are prone to experimental design limitations [41, 42]. Finally, the majority of studies that assessed agreement between the outputs of the HUI3 and SF-6D were

[1] This study recruited online participants who were registered with a panel company and who may have differed from the general population and narrowly defined as a healthy individual a respondent who indicated absence of a chronic disease using a controversial cut-off point (above 70 on an arbitrary numerical health scale, where 0 represented death and 100 represented best possible health). It excluded individuals who were too ill to complete the survey and did not harmonize the socio-economic factors associated with HRQoL across different countries. The authors were not able to verify the health status of participants or to control for clinical characteristics derived from medical records. Finally, given that participants self-selected into the study, its results might be biased.

Table 1 Background characteristics of cohorts

	BLS (Germany)				VICS (Australia)			
	HUI3 at 26 years		SF-6D at 26 years		HUI3 at 18 years		SF-6D at 18 years	
	VPT/VLBW	Controls	VP/VLBW	Controls	EP/ELBW	Controls	EP/ELBW	Controls
Number completing MAUI	231	224	231	226	186	137	180	143
Age at assessment Mean (SD)	26.3	26.3	26.3	26.3	17.9	18.1	18.0	18.1
	(0.68)	(0.69)	(0.69)	(0.70)	(0.78)	(0.88)	(0.79)	(0.87)
GA at birth, Mean (SD)	30.6	39.7	30.6	39.7	26.7	39.0	26.7	39.2
	(2.19)	(1.18)	(2.18)	(1.18)	(2.10)	(1.43)	(2.10)	(1.44)
Birth weight, Mean (SD)	1330	3360	1330	3362	887	3419	886	3422
	(320)	(448)	(319)	(447)	(155)	(468)	(155)	(463)
Sex, n (%) male	125	105	124	105	84	60	82	60
	(54.1)	(46.9)	(53.7)	(46.5)	(45.2)	(43.8)	(45.6)	(42.0)
Study name	Bavarian longitudinal study				Victorian infant collaborative study			
Birth Year	1985–1986				1991–1992			
Eligibility Criteria VP/VPBW	VPT\VLBW (GA < 32wk or BW < 1500 g)				EPT\ELBW (GA < 28wk or BW < 1000 g)			
Controls	Recruited in the same obstetric hospitals				Normal birth weight, contemporaneously recruited			

MAUI multi-attribute utility instrument, *VP* very preterm (< 32 weeks GA), *VLBW* very low birth weight (< 1500-g birth weight), *EP* extremely preterm (< 26-week GA for EPICure and < 28-week GA for VICS), *ELBW* extremely low birth weight (< 1000 - birth weight), *SD* standard deviation

limited to one country or geographic region [31–39] and thus may have limited external validity.

Previous research called for comparative evaluations of the HUI3 and SF-6D measures across a diverse range of health conditions [43, 44]. Furthermore, research has advocated new comparative evaluation studies that use larger samples by maximizing their power and enhancing comparability when data across multiple cohort studies are combined [44]. To overcome the limitations associated with analyses restricted to a specific disease or disorder, conducted within limited clinical settings or within a single geographical region, the use of individual patient data analysis (IPD) consolidated over several geographically diverse cohorts offers advantages. This study uses IPD from European and Australian multi-site collaborative cohorts to inform the choice of the HUI3 and/or SF-6D measures for research studies that consider the consequences of VP/VLBW in adulthood as well as informs the cost-effectiveness of preventive or treatment interventions related to VP/VLBW status.

This study has the following aims: (a) to examine the agreement between the outputs of the HUI3 and SF-6D measures among adults born VP/VLBW and controls and to explain the sources of disagreement between instruments and (b) to provide useful information for the selection of preference-based HRQoL instruments for trials or research studies that ascertain the long-term consequences of VP/VLBW and birth at term or with normal birthweight.

Methods

Data

The following criteria have been utilized to identify relevant prospective cohorts: (1) have used two distinctive preference-based measures to assess HRQoL in adulthood (defined as ≥ 18 years [45]) amongst individuals born VP/VLBW, (2) included a comparison control group of term-born and/or normal birthweight individuals, and (3) contributed data to the RECAP consortium (www.recappreterm.eu), a database of cohorts of individuals born VP/VLBW. Two different and recent systematic reviews of preference-based HRQoL outcomes following preterm birth or low birthweight had identified eligible cohorts [46, 47]. The following two prospective cohort studies met the study inclusion criteria: The Bavarian Longitudinal Study (BLS) [48] and The Victorian Infant Collaborative Study (VICS) [49]. These two studies were designed to assess the associations of VP/VLBW status with various health outcomes [50] as well as received country-specific ethical approvals, including participants' written informed consent in adulthood.

Table 1 described the background eligibility criteria, age(s) at assessment, and the control groups for the BLS and VICS cohorts. Detailed descriptions of each participating cohort (the study's population, methodology, types of data and variables) have been previously published [48, 49]. All variables of interest across BLS and VICS were harmonized, meaning that an identical set of definitions, scaling methods,

and classification were applied to all variables across BLS and VICS cohorts.

Outcome measures

Participants' perceptions of their HRQoL were assessed using both the HUI3 and SF-12 [48, 49]. Study participants completed the unedited Health Utilities Index 15-item questionnaire for usual health status assessment, which was obtained from the Health Utilities Index developers and covers the HUI3 health status classification system. The HUI3 was developed to describe HRQoL in general population and clinical contexts and consists of eight attributes: ambulation, dexterity, cognition, vision, hearing, speech, emotion, and pain [51–53]. Within each attribute, the levels of function were scored on a 5- or 6-point scale ranging from optimal function to severe impairment. Responses within each of the eight attributes can be valued as single attribute utility (SAU) scores on a scale ranging from 0 and 1 [51]. Responses within each of the eight attributes can also be mapped onto an eight-attribute health status vector. Algorithms reflecting the preferences of the general public for the HUI3 health states can be used to convert responses to the measure's eight attributes into multiplicative multi-attribute utility scores. The Canadian algorithms [51–54] were applied in both cohorts, reflecting the preferences of 504 adults in the general population who were living in the city of Hamilton, Ontario, and who had previously been asked to value selected HUI3 health states using both visual analogue scaling and standard gamble techniques. HUI3 multi-attribute utility scores are valued on a cardinal scale ranging between -0.36 and 1.0, with -0.36 representing the worst possible HUI3 health state, 0.0 representing dead, and 1.0 representing full health [53, 54].

The SF-12 includes 12 of the 36 items contained within the SF-36. These have an identical dimension structure [55], and for each dimension, item responses are mapped onto a 0 to 100 scale. Responses to the SF-12 items were converted [56] into SF-6D multi-attribute utility scores using the UK SF-6D utility algorithms [55]. The SF-6D algorithms reduce the eight dimensions of the SF-36/12 to six by merging role limitations due to emotional and physical problems and eliminating general health perceptions. SF-6D multi-attribute utility scores are valued on a cardinal scale ranging between 0 and 1.0, with 0 representing dead and 1.0 representing full health [55]. For the SF-6D, only two out of six dimensions (physical functioning, role limitations) reflect physical aspects of health, while other dimensions (social functioning, pain, mental health, vitality, and emotional) relates to non-physical aspects of health. By contrast, most HUI3 attributes reflect the physical health of the individual (vision, hearing, speech, ambulation, and dexterity).

We used the following outcome variables of interest: HUI3 and SF-6D multi-attribute utility scores and the difference between HUI3 and SF-6D utility scores. The minimum clinically important difference in multi-attribute utility score is considered to be 0.03 for the HUI3 [57] and 0.04 for SF-6D [58, 59].

Empirical analyses

We combined IPD across the BLS and VICS cohorts. To identify whether our assessments of agreement between the HUI3 and SF-6D measures should be disentangled by birth status, we initially estimated the association between VP/VLBW status and HRQoL in adulthood using one-stage IPD analysis, which could be implemented either using fixed or random effects [60]. Fixed effects models were used because individuals born VP/VLBW and controls were enrolled across distinct geographical regions and time frames. This implies the presence of systematic differences across the BLS and VICS cohorts. However, we also utilized random effects as a robustness check. Models were adjusted for age and sex of the participants, mode of delivery (cesarean section vs vaginal delivery), and number of days in hospital after birth, as well as for the harmonized socio-demographic/socio-economic variables: maternal education level at birth or during childhood and maternal ethnicity.

We computed means, standard deviations, and t tests for unequal variances, medians, and Kruskal–Wallis tests to assess differences in agreement between HUI3 and SF-6D multi-attribute utility scores within VP/VLBW individuals, controls, and the combined sample. To identify statistically significant predictors that explain observed differences between HUI3 and SF-6D multi-attribute scores on covariates, we used generalized mixed models in a one-step approach. Models were estimated using multivariate linear fixed effects.

Furthermore, agreement between the HUI3 and SF-6D multi-attribute utility scores was investigated using the intra-class correlation measures and Bland–Altman plots. The analysis was performed for VP/VLBW individuals and controls separately as well as for the combined sample. An intra-class correlation coefficient less than 0.75 is indicative of moderate agreement, while an intra-class correlation coefficient greater than 0.75 indicates good agreement [60, 61]. Bland–Altman plots display the mean $\left(\frac{HUI3+SF-6D}{2}\right)$ overall scores and the difference (HUI3-SF-6D) against each other. A line of mean difference estimates systematic difference between the two instruments, with limits of agreement estimated as the mean difference plus/minus 1.96 standard deviation of the mean difference. Limits of agreement (LoA) reflect the expected range in which 95% of observed differences would lie, with wider limits of agreement indicating poorer agreement [62]. Good concordance between the

Table 2 Characteristics of VP/VLBW individuals and controls within HUI3 and SF-6D Meta-cohorts

Meta-cohort (BLS & VICS)	VP/VLBW	Controls	*p* value	Missings/*N* (Pct)
N (%)	558 (53.2)	491 (46.8)		544/1593 (34.15)
Age at QoL assessment, mean (SD)	22.65 (4.32)	23.00 (4.12)	0.24	742/1593 (46.58)
Child sex, *N* (%)				
Male	276 (49.5)	233 (47.5)		
Female	282 (50.5)	258 (52.5)	0.52	1/1593 (0.06)
Gestational age (weeks), mean (SD)	28.51 (2.84)	39.41 (1.33)	< 0.001	1/1593 (0.06)
Birth weight (grams), mean (SD)	1090 (328.43)	3373 (442)	< 0.001	1/1593 (0.06)
Maternal age at birth (years), mean (SD)	28.68 (5.31)	29.14 (4.93)	0.15	8/1593 (0.50)
Mat educ at birth or childhood, *N* (%)				
Low level (equivalent to ISCED 0 to 2)	138 (33.0)	127 (36.7)		
Medium level (equivalent to ISCED 3 to 5)	221 (52.9)	146 (42.2)		
High level (equivalent to ISCED 6 to 8)	59 (14.1)	73 (21.1)	0.01	462/1593 (29.00)
Maternal ethnicity, *N* (%)				
Caucasian	493 (92.1)	445 (92.9)		
Non-Caucasian	42 (7.9)	34 (7.1)	0.65	39/1593 (2.45)
HUI3-MAU score, mean (SD)	0.85 (0.19)	0.89 (0.15)	< 0.001	815/1593 (51.16)
SF-6D MAU score, mean (SD)	0.83 (0.12)	0.83 (0.10)	0.23	813/1593 (51.04)

Table 2 reports characteristics for VP/VLBW and controls which had non-missing HUI3 or SF-6D Multi-Attribute Utility Scores

Mat Educ maternal education, *ISCED* International Standard Classification of Education. *SD* Standard deviation. When proportions are reported *p* value is based on Fisher's exact test for equality of proportions. When means are reported the *p* value is based on a *t* test for unequal variances. Pct reports the percent of missing values.

HUI3 and SF-6D would show a mean difference close to zero with ≤ 5% of scatter points lying outside the limits of agreement.

Analyses were performed using STATA version 17 and p-values of 0.05 or less were considered statistically significant.

Results

Baseline characteristics of prospective cohort studies

Table 1 displays baseline characteristics of the participants of the BLS and VICS cohorts. Years of birth ranged from 1985 to 1986 for BLS and 1991–1992 for VICS. Pooled data consisted of 778 HUI3 assessments (417 VP/VLBW individuals and 361 controls) and 780 SF-6D assessments (411 VP/VLBW and 369 controls). The mean age at assessment was 18 years for VICS participants and 26.3 years for BLS participants. Table 2 shows the characteristics of VP/VLBW individuals and controls with non-missing HUI3 and SF-6D multi-attribute utility scores. Within the meta-cohort no statistically significant differences were found by birth status across the following characteristics: age, sex of the participants, maternal education level at birth or during childhood, and maternal ethnicity.

Relationship between VP/VLBW status and HRQoL using HUI3 vs SF-6D

Using a one-stage IPD meta-analysis, to identify whether our assessments of agreement between the HUI3 and SF-6D measures should be disentangled by VP/VLBW status, we initially estimated the association between VP/VLBW status and HRQoL in adulthood. The adjusted impact of VP/VLBW status on the HUI3 multi-attribute utility score was -0.04 (95% CI − 0.06, − 0.01) with no significant impact on the SF-6D multi-attribute utility score (Table 3). To understand the sources of identified differences we present the additional evidence in Online Appendix A (Tables A.1, A.2). We utilized random effects models and reported results in Online Appendix B. Further evidence on the association between VP/VLBW status and HRQoL in adulthood using HUI3 and SF-6D can be found in a recent study [63].

Comparison of HRQoL assessed by the HUI3 and SF-6D

Table 4 displays descriptive and inferential statistics for HUI3 and SF-6D multi-attribute utility scores for each group considered. Mean and median estimates for HUI3 multi-attribute utility scores were consistently higher compared with their respective SF-6D values. All differences were clinically [57–59] and statistically significant within the

Table 3 One-stage IPD meta-analyses: Impact of preterm birth on HUI3-MAU score and SF-6DMAU score all cohorts combined

	Health utilities index mark 3 MAU score						Short form 6D MAU score					
	Unadjusted	95%CI	*p* value/*N*	Adjusted	95%CI	*p* value/*N*	Unadjusted	95%CI	*p* value/*N*	Adjusted	95%CI	*p* value/*N*
βVPT/VLBW	− 0.04 (0.01)	[− 0.06, − 0.02]	0.01/778	− 0.04 (0.01)	[− 0.06, − 0.01]	0.01/615	− 0.01 (0.01)	[− 0.02, 0.01]	0.45/780	− 0.01 (0.01)	[− 0.03, 0.00]	0.18/617
Mode of delivery				− 0.00 (0.01)	[− 0.03, 0.03]	0.91				− 0.00 (0.01)	[− 0.03, 0.02]	0.52
Hosp days				− 0.00 (0.00)	[− 0.00, 0.00]	0.728				− 0.00 (0.00)	[− 0.00, 0.00]	0.55
Sex (female)				− 0.02 (0.01)	[− 0.05, 0.01]	0.99				− 0.03 (0.01)	[− 0.05, − 0.02]	< 0.001
Mat educ (medium vs low)				0.00 (0.01)	[− 0.03, 0.02]	0.07				0.01 (0.00)	[− 0.01, 0.03]	0.22
Mat educ (high vs low)				0.02 (0.01)	[− 0.00, 0.06]	0.09				0.01 (0.00)	[− 0.01, 0.03]	0.34
Age at assessment				− 0.01 (0.00)	[− 0.02, 0.01]	0.05				− 0.00 (0.00)	[− 0.02, 0.01]	0.52
Cohort effects included	No			Yes			No			Yes		

Method: Linear Fixed Effects Models. HUI3-MAU and SF-6D column results are based on the following cohorts: BLS at 26 and VICS at 18. Mode of Delivery is an indicator for Cesarean section. Hosp. Days—number of days in hospital after birth. All models controlled for cohorts' fixed effects. Robust standard errors in round parentheses

Table 4 HUI3 and SF-6D utility scores, differences between scores and quantification of agreement by VP/VLBW, controls, and combined sample

Cohorts	VP/VLBW	Controls	Combined	*p* value
BLS cohort				
N (%)	259 (53.1)	229 (46.9)	488 (100.0)	
HUI3-MAUI, mean (SD)	0.85 (0.18)	0.89 (0.14)	0.87 (0.16)	0.01
SF-6D MAUI, mean (SD)	0.83 (0.11)	0.84 (0.09)	0.83 (0.10)	0.49
Median Δ, (min; max)	0.05 (− 0.75; 0.40)	0.07 (− 0.61; 0.37)	0.06 (− 0.75; 0.40)	0.02
Mean Δ, (95% CI)	0.02 (− 0.00; 0.04)	0.05 (0.03; 0.07)	0.04 (0.02; 0.05)	0.02
VICS cohort				
N (%)	299 (53.3)	262 (46.7)	561 (100.0)	
HUI3-MAUI, mean (SD)	0.86 (0.20)	0.90 (0.15)	0.88 (0.18)	0.08
SF-6D MAUI, mean (SD)	0.82 (0.13)	0.83 (0.11)	0.82 (0.12)	0.35
Median Δ, (min; max)	0.04 (0.02; 0.07)	0.07 (0.05; 0.10)	0.06 (0.04; 0.07)	0.16
Mean Δ, (95% CI)	0.08 (− 0.83; 0.49)	0.08 (− 0.58; 0.36)	0.08 (− 0.83; 0.49)	0.10
BLS and VICS cohorts				
N (%)	558 (53.2)	491 (46.8)	1049 (100.0)	
HUI3-MAUI, mean (SD)	0.85 (0.19)	0.89 (0.15)	0.87 (0.17)	< 0.001
SF-6D MAUI, mean (SD)	0.83 (0.12)	0.83 (0.10)	0.83 (0.11)	0.23
Median Δ, (min; max)	0.06 (− 0.83; 0.49)	0.08 (− 0.61; 0.37)	0.07 (− 0.83; 0.49)	0.01
Mean Δ, (95% CI)	0.03 (0.01; 0.05)	0.06 (0.04; 0.07)	0.04 (0.03; 0.06)	0.01

Δ denotes the difference between HUI3 and SF-6D Multi-Attribute Utility Scores. *SD* standard deviation. When means are reported the *p* value is based on a paired *t* test. When medians are reported the *p*-value is based on a Kruskal–Wallis test

meta-cohort across all groups considered ($p < 0.01$). Table 5 shows the estimates from regressing differences between HUI3 and SF-6D multi-attribute scores on covariates. The evidence suggests that none of the variables considered was a statistically significant predictor of observed differences between multi-attribute scores.

The correlation coefficient (ρ) between HUI3 and SF-6D multi-attribute utility scores for the BLS and VICS cohorts was computed. The evidence showed that ρ between the two multi-attribute utility scores within VICS was 0.45 ($\rho = 0.51$ for VP/VLBW individuals and $\rho = 0.31$ for controls), which was higher compared with $\rho = 0.35$ within the BLS cohort ($\rho = 0.37$ for VP/VLBW individuals and $\rho = 0.33$ for controls). Within the meta-cohort, the ICC was 0.40 for the VP/VLBW sample, 0.29 for the controls, and 0.36 for the combined sample. Overall, the evidence suggests that the HUI3 and SF-6D multi-attribute scores had moderate or low correlation.

The Bland–Altman plots were constructed by birth status (see Fig. 1) and showed a mean difference of 0.06 (95% CI 0.04, 0.07), i.e., HUI3 multi-attribute utility scores for controls were higher than the SF-6D multi-attribute utility scores for controls. The mean difference for VP/VLBW individuals was 0.03 (95% CI 0.01, 0.05), meaning that HUI3 multi-attribute utility scores were higher than SF-6D multi-attribute utility scores in this group. In the Bland–Altman plot (Fig. 1), the data points deviate widely from the agreement line at low levels of mean utility and the relationship between the difference in HUI3 and SF-6D utilities shifts in magnitude but not in direction. The same pattern is observed by combining the VP/VLBW sample with controls (Fig. 2), generating a mean difference between the paired observations of 0.04 (95% CI 0.03, 0.06). Notably, in all groups considered, the Bland–Altman plots showed a funneling effect with stronger agreement as the mean overall utility score approached 1.0. However, in the Bland–Altman plots, the 95% LoA ranged from − 0.30 to 0.37 within the VP/VLBW sample, − 0.22 to 0.34 within controls, and − 0.27 to 0.36 within the combined sample. Most importantly, in all three groups considered (VP/VLBW, controls, and the combined sample), the 95% agreement differences were far wider than the clinically meaningful differences postulated for the HUI3 and SF-6D.

Discussion

This study provides the first comparative evaluation of the HUI3 and SF-6D among adults born VP/VLBW and normal birthweight or term born controls. The results show a considerable degree of disagreement between the two sets of multi-attribute utility scores, consistent with previous reports for specific diseases [31, 33, 37, 40]. The patterns underlying differences vary, however, in a number of important aspects when compared with previous research. Our results identified less agreement compared with previous comparative evaluations of the HUI3 and SF-6D measures. Interestingly, our study found that agreement between the HUI3 and SF-6D

Table 5 HUI3 and SF-6D utility scores, differences between scores and quantification of agreement by VP/VLBW, and controls

	Outcome	Δ				Outcome	Δ			
	Coefficient	SE	Lower 95% CI	Upper 95% CI	*p* value	Coefficient	SE	Lower 95% CI	Upper 95% CI	*p* value
Age at assessment	− 0.00	0.00	− 0.01	0.00	0.84	− 0.00	0.01	− 0.02	0.02	0.94
Sex (female)	0.02	0.02	− 0.01	0.06	0.17	0.02	0.02	− 0.01	0.06	0.27
Mode of delivery	0.00	0.00	− 0.03	0.04	0.82	0.00	0.02	− 0.03	0.04	0.81
Hosp days	− 0.0	0.00	− 0.03	0.04	0.44	− 0.00	0.00	− 0.03	0.04	0.45
Mat educ (medium vs low)	0.01	0.02	− 0.04	0.05	0.78	0.01	0.02	− 0.03	0.05	0.87
Mat educ (high vs low)	0.05	0.03	− 0.01	0.11	0.09	0.05	0.03	− 0.01	0.10	0.14
Cohort effects included	No					Yes				

Linear Fixed Effects Models. The outcome variable is Δ which denotes the difference between HUI3 and SF-6D Multi-Attribute Utility Scores. Models were adjusted for age and sex of the participants, mode of delivery (Cesarean section vs vaginal delivery), and number of days in hospital after birth, as well as for the harmonized socio-economic variables: maternal education level at birth or during childhood and maternal ethnicity

SE Robust standard error, *CI* confidence interval

measures was weaker in term-born or normal birthweight controls compared with VP/VLBW individuals.

Overall, the HUI3 and SF-6D measures disagree substantially because VP/VLBW status was found to be associated with minimal important decrements in utility score when health status was ascertained with the HUI3 and not the SF-6D. Furthermore, results show discordance between the outputs of the HUI3 and SF-6D in VP/VLBW individuals, controls, and the combined sample. This implies that the HUI3 and SF-6D each provide unique information on different aspects of health status across the groups considered and suggests that the HUI3 better captures preterm-induced changes to HRQoL in adulthood.

The evidence consistently demonstrates that the HUI3 and SF-6D instruments are not interchangeable for use in cost-utility based decision-making for interventions that target adults born VP/VLBW [64, 65]. Because our study also investigated concordance between the HUI3 and SF-6D in term-born or normal birthweight controls, the findings imply that the measures might also not be interchangeable for use in more general population samples.

Furthermore, given the evidence provided in this study regarding level of agreement between the HUI3 and SF-6D measures overall, our findings imply that studies focused on capturing the physical and cognitive effects of interventions should employ the HUI3 as a primary instrument, with the SF-6D as a potential supplementary measure. Our study implies that the HUI3 may be preferred to the SF-6D for studies designed at quantifying physical and cognitive aspects of health particularly since for SF-6D, only two out of six dimensions (physical functioning, role limitations) reflect physical aspects of health, while other dimensions (social functioning, pain, mental health, vitality, and emotional) relates to non-physical aspects of health. However, most HUI3 attributes reflect the physical health of the individual (vision, hearing, speech, ambulation, and dexterity). Prioritization of a preffered multi-attribute utility measure might increase the value of research design and potentially reduce unnecessary research costs related to primary data collection. Our results indicate that the HUI3 and SF-6D instruments are not interchangeable for use in clinical, population research, and cost-effectiveness based decision-making that considers the long-term consequences of VP/VLBW status [64, 65].

Our overall results are consistent with the differences in the HUI3 and SF-6D descriptive systems. Specifically, given that the HUI3 explicitly asks about a person's vision, dexterity, ambulation, and cognition, while SF-6D does not, it is perhaps expected that VP/VLBW individuals, who are known to have impaired outcomes associated with these attributes [24–28], have lower levels of utility according to the HUI3 than according to the SF-6D. The evidence shows that discrepancies in the health descriptive systems of the

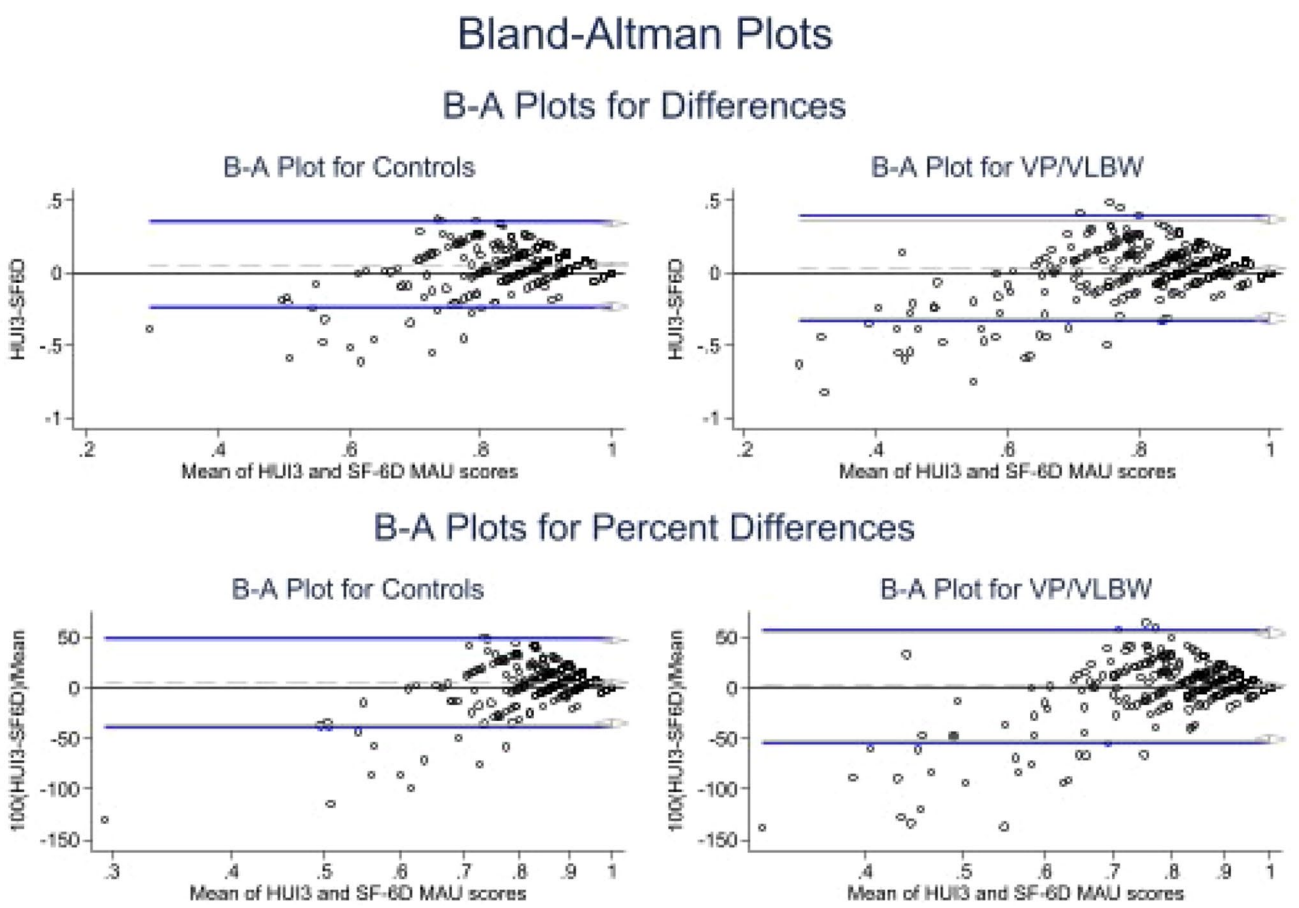

Fig. 1 The Bland–Altman plots by VP/VLBW status

HUI3 and SF-6D instruments may drive the differences in multi-attribute utility scores of VP/VLBW individuals and controls in adulthood. Our study demonstrates that variation in the descriptive systems of the measures is likely to be a major contributory factor to variation in the utility scores. Results of this research corroborate the conclusions of a study that analyzed patients in several disease areas and found that the EQ-5D, SF-6D, HUI3, 15D, QWB, and AQoL-8D instruments measure related but different constructs [44]. Also, the study concluded that the instruments differ in their relationship to different health dimensions, and the differences are primarily the result of the instruments' descriptive systems.

Our study advances the literature because we provide clear evidence that differences in descriptive systems explain, at least in part, disagreement found between the outputs of the HUI3 and SF-6D measures. The evidence shows that the discordance between the outputs is observed within both adults born VP/VLBW and controls. However, differences related to HUI3 and SF-6D valuation protocols and utility ranges may also partly contribute to the differences in multi-attribute utility scores we document in this study. Furthermore, the study is the first in the literature to use a meta-analysis in this context combining data from two longitudinal prospective cohort studies.

This study does not infer that the HUI3 measure is generally preferable to SF-6D when health outcomes associated with clinical or public health interventions are ascertained. Rather, it provides insights for future research related to agreement between the HUI3 and SF-6D measures and suggests that the HUI3 classification system, unlike the SF-6D, is able to capture consequences of VP/VLBW status in adulthood, which is consistent with prior documented patterns reported in the disability literature [24–28]. We are not arguing against the use of the SF-6D or other preference-based HRQoL measures to investigate consequences of VP/VLBW status.

However, this study provides insight for stakeholders seeking to understand what instruments to use for comparative effectiveness research related to preterm birth and low birthweight. Further investigation is needed to understand the between-measure discrepancies attributable to descriptive classification systems for other measures, including the EQ-5D which is widely recommended in

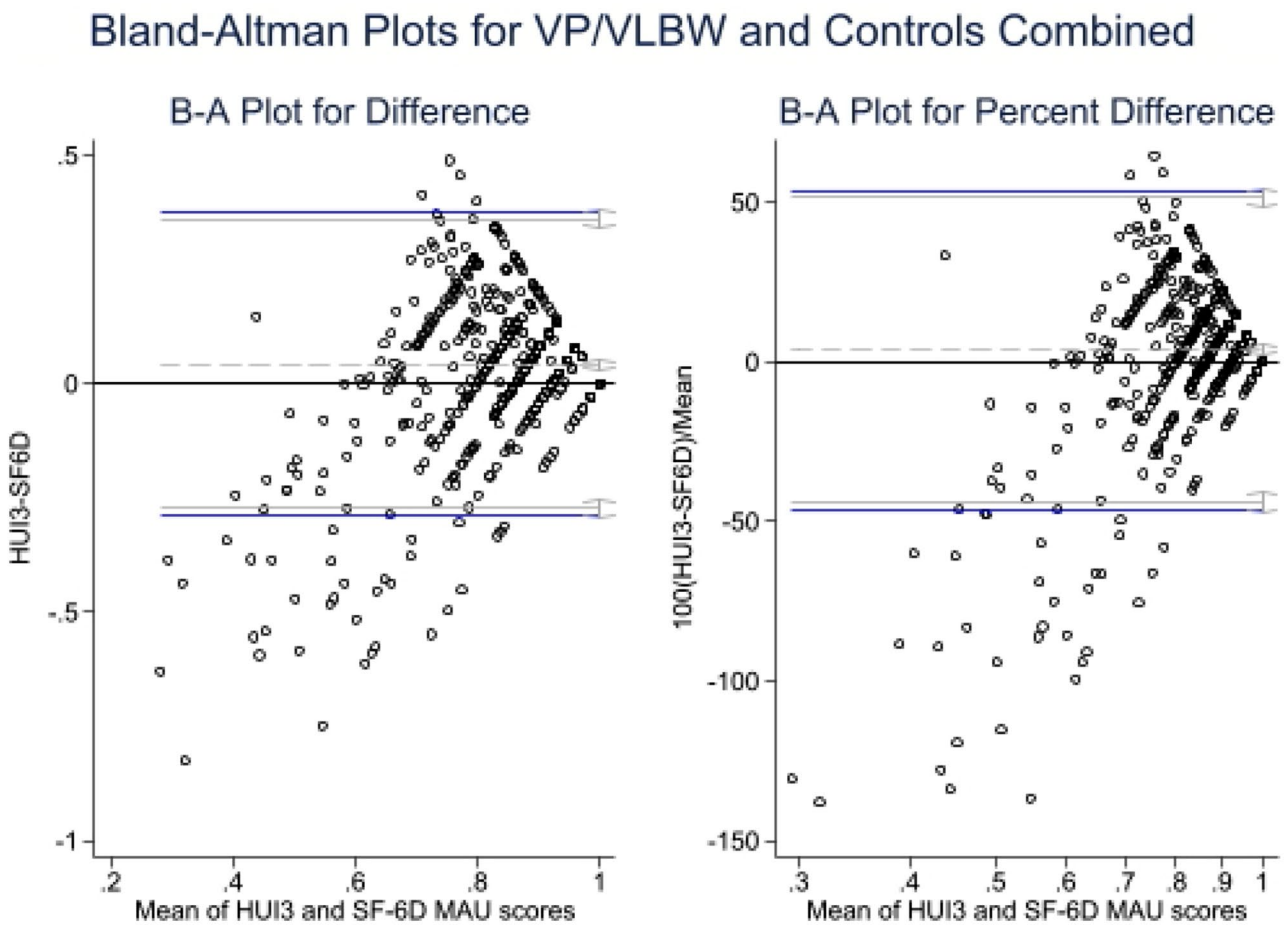

Fig. 2 The Bland–Altman plots for VP/VLBW and controls combined

HTA guidelines [8, 17, 66–69] and other measures to inform the methodological debate and guide the selection of the most appropriate HRQoL instruments. Overall, the current study highlights the need to carefully consider the outcomes of interest and the characteristics being studied of the condition for an appropriate selection of HRQoL instrument.

Strengths and limitations

The data structure made it possible to examine the agreement between measures within VP/VLBW individuals, normal birthweight or term born controls, and within the combined sample. We were able to assess the validity of the results by replicating the main finding across different populations, which strengthens the study's conclusions and which had not been studied previously as far as we are aware. This is the major strength of this study. Furthermore, our study ascertained agreement between the outputs of the HUI3 and SF-6D measures using controls selected from the general populations in Germany and Australia. This implies that results of this study may be generalizable to populations from Germany and Australia.

Another strength of this study is that we were able to confirm VP/VLBW status in each participant due to the rigorous recruitment, data collection, and follow-up methods utilized by the participating cohorts, which also harmonized relevant socio-demographic factors. Furthermore, our study employed socioeconomically diverse samples of VP/VLBW individuals and controls. Finally, results of this study are not affected by biases associated with proxy parental reporting [70] because participating cohorts used self-reported HRQoL data.

It is important to note that the scoring algorithms for the HUI3 and SF-6D differ in certain respects. Thus, while our study shows that the utility differences we found are driven by the underlying concepts of health being measured, the methods employed are not able to measure the contributory effects of valuation protocols, i.e., differences in scoring algorithms. A further limitation is that our study included

cohorts from only two countries. Thus, replication of this study with data from other countries, particularly low- or middle-income countries, would be a valuable contribution to the literature.

Our report did not include the EQ-5D in this comparative evaluation because no individual study that contributed to the RECAP platform assessed HRQoL using the EQ-5D. This is a limitation because a recent review identified that the EQ-5D is the most frequently recommended multi-attribute utility instrument in HTA guidelines [17]. Thus, our study is not able to provide comprehensive evidence regarding the most appropriate preference-based HRQoL measure to ascertain utility scores in adulthood for VP/VLBW individuals or for normal birthweight or term born controls. Comparing agreement of the EQ-5D, HUI3, and SF-6D for VP/VLBW individuals and normal birthweight or term born controls offers a fruitful direction for further investigation.

Conclusion

The evidence from two longitudinal cohort studies conducted in Australia and Germany demonstrates poor agreement between the HUI3 and SF-6D in VP/VLBW individuals and normal birthweight or term born controls. It may be beneficial to use both the HUI3 and SF-6D instruments when evaluating health outcomes of interventions related to gestational age at birth and/or birthweight. However, studies focused on measuring physical or cognitive aspects of health will likely benefit from prioritizing the use of the HUI3 in order to better detect and quantify the effects of health interventions or assess outcomes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11136-023-03344-x.

Acknowledgements We would like to thank Sylvia van der Pal, Nicole Baumann, and Marina Mendonca for contributing to data harmonization. We thank the participants in the BLS and VICS cohorts and their families that contributed data to this study.

Author contributions Study Conception: CB. Study Design: CB and SP. Design of empirical strategy: CB. Implementation of data analysis: CB. Acquisition of data: CB, DW, LWD, PB, PJA, JLYC, and SP. Interpretation of data: CB and SP. Drafting of the manuscript: CB. Critical revision of the paper for important intellectual content: CB, DW, LWD, PB, PJA, JLYC, and SP.

Funding This work was supported by Grant 733280 from the European Commission as part of the Research on European Children and Adults Born Preterm (RECAP) Consortium Preterm Project, an EU Horizon 2020 study to construct a platform combining data from cohort studies of very preterm birth in Europe. The Victorian Infant Collaborative Study was supported by Grants 491246 and 546519 from the National Health and Medical Research Council of Australia. SP receives support from the UK National Institute for Health Research (NIHR) Senior Investigator (NF-SI-0616-10103) and from the UK NIHR Applied Research Collaboration Oxford and Thames Valley.

Data availability Information regarding the data availability can be found at https://recap-preterm.eu/for-scientists/the-recap-preterm-cohort-platform/.

Code availability The code used in this study is available from the authors upon reasonable request.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

Ethical approval Not applicable since this study describes secondary data analyses.

Consent to participate Not applicable.

Consent for publication The authors, jointly and severally, give the publisher the permission to publish the work.

References

1. Drummond Michael, F., Sculpher Mark, J., Karl, Claxton, Stoddart Greg, L., & Torrance George, W. (2015). *Methods for the economic evaluation of health care programmes*. Oxford University Press.
2. Drugs Canadian Agency, Health Technologies, others. (2006). Guidelines for the economic evaluation of health technologies: Canada.
3. Excellence Care. Guide to the Methods of Technology Appraisal, Retrieved April 4, 2013, from https://www.ncbi.nlm.nih.gov/books/NBK395867/pdf/Bookshelf_NBK395867.pdf
4. Stenman, U., Hakama, M., Knekt, P., et al. (2010). Measurement and modeling of health-related quality of life. *Epidemiology Demography Public Health., 195*, 130–135.
5. Human Services Australia. Dept., Health. (1995). *Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee: Including Major Submissions Involving Economic Analyses, November 1995*. Australian Government Pub. Service.
6. James, R. (2001). NICE: Faster access to modern treatments? Analysis of guidance on health technologies. *Bmj, 323*, 1300–1303. https://doi.org/10.1136/bmj.323.7324.1300
7. Bae, S., Lee, S., Bae, E. Y., & Jang, S. (2013). Korean guidelines for pharmacoeconomic evaluation (Second and updated version). *Pharmacoeconomics., 31*, 257–267. https://doi.org/10.1007/s40273-012-0021-6
8. Rowen, D., Azzabi Zouraq, I., Chevrou-Severac, H., & van Hout, B. (2017). International regulations and recommendations for utility data for health technology assessment. *Pharmacoeconomics., 35*, 11–19. https://doi.org/10.1007/s40273-017-0544-y

9. Mouillet, G., Falcoz, A., Fritzsch, J., et al. (2021). Feasibility of health-related quality of life (HRQoL) assessment for cancer patients using electronic patient-reported outcome (ePRO) in daily clinical practice. *Quality of Life Research., 30*, 3255–3266. https://doi.org/10.1007/s11136-020-02721-0
10. Nguyen Matthew, H., Huang Frank, F., & O'Neill, S. G. (2021). Patient-reported outcomes for quality of life in SLE: Essential in clinical trials and ready for routine care. *Journal of Clinical Medicine., 10*, 3754. https://doi.org/10.3390/jcm10163754
11. Tian-hui, C., Lu, L., & Michael, M. K. (2005). A systematic review: How to choose appropriate health-related quality of life (HRQOL) measures in routine general practice? *Journal of Zhejiang University Science B., 6*, 936–940. https://doi.org/10.1631/jzus.2005.B0936
12. Pais-Ribeiro, J. L. (2004). Quality of life is a primary end-point in clinical settings. *Clinical Nutrition., 23*, 121–130. https://doi.org/10.1016/s0261-5614(03)00109-2
13. Petrou, S., Yiu, H. H., & Kwon, J. (2019). Economic consequences of preterm birth: A systematic review of the recent literature (2009–2017). *Archives of Disease in Childhood., 104*, 456–465. https://doi.org/10.1136/archdischild-2018-315778
14. Bolbocean, C., Andújar, F. N., McCormack, M., Suter, B., & Holder, J. L. (2021). Health-related quality of life in pediatric patients with syndromic autism and their caregivers. *Journal of Autism and Developmental Disorders.* https://doi.org/10.1007/s10803-021-05030-8
15. Bolbocean, C., Rhidenour, K. B., McCormack, M., Suter, B., & Holder, J. L. (2022). COVID-19 induced environments, health-related quality of life outcomes and problematic behaviors: Evidence from children with syndromic Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders.* https://doi.org/10.1007/s10803-022-05619-7
16. Chim, L., Kelly, P. J., Salkeld, G., & Stockler, M. R. (2010). Are cancer drugs less likely to be recommended for listing by the Pharmaceutical Benefits Advisory Committee in Australia? *Pharmacoeconomics, 28*, 463–475. https://doi.org/10.2165/11533000-000000000-00000
17. Kennedy-Martin, M., Slaap, B., Herdman, M., et al. (2020). Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. *The European Journal of Health Economics., 21*, 1245–1257. https://doi.org/10.1007/s10198-020-01195-8
18. John, B., Mark, D., & Colin, G. (1999). A review of the use of health status measures in economic evaluation. *Journal of Health Services Research & Policy., 4*, 174–184. https://doi.org/10.1177/135581969900400310
19. Harrison, M. J., Davies, L. M., Bansback, N. J., et al. (2009). The comparative responsiveness of the EQ5D and SF-6D to change in patients with inflammatory arthritis. *Quality of Life Research., 18*, 1195–1205. https://doi.org/10.1007/s11136-009-9539-2
20. Cunillera, O., Tresserras, R., Rajmil, L., et al. (2010). Discriminative capacity of the EQ-5D, SF-6D, and SF-12 as measures of health status in population health survey. *Quality of Life Research., 19*, 853–864. https://doi.org/10.1007/s11136-010-9639-z
21. Rick, N., John, W., & Woodliff, D. (2003). SME survey methodology: Response rates, data quality, and cost effectiveness. *Entrepreneurship Theory and Practice., 28*, 163–172.
22. Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., et al. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet., 383*, 166–175. https://doi.org/10.1016/S0140-6736(13)62227-8
23. Kristen, O., James, W., & Raeda, A. (2021). Survey costs: Where are we and what is the way forward? *Journal of Survey Statistics and Methodology., 9*, 921–942. https://doi.org/10.1093/jssam/smaa014
24. Anderson, P. J., Miranda, D. M., Albuquerque, M. R., et al. (2021). Psychiatric disorders in individuals born very preterm/very low-birth weight: An individual participant data (IPD) meta-analysis. *EClinicalMedicine., 42*, 101216. https://doi.org/10.1016/j.eclinm.2021.101216
25. Dieter, W. (2016). Born extremely low birth weight and health related quality of life into adulthood. *The Journal of Pediatrics, 179*, 11–12. https://doi.org/10.1016/j.jpeds.2016.09.012
26. Dieter, W., Samantha, J., & Mendonca, M. (2019). The life course consequences of very preterm birth. *Annual Review of Developmental Psychology, 1*, 69–92. https://doi.org/10.1146/annurev-devpsych-121318-084804
27. Eves, R., Mendonca, M., Baumann, N., et al. (2021). Association of very preterm birth or very low birth weight with intelligence in adulthood: An individual participant data meta-analysis. *JAMA Pediatrics.* https://doi.org/10.1001/jamapediatrics.2021.1058
28. Cheong, J. L. Y., Haikerwal, A., Anderson, P. J., & Doyle, L. W. (2021). Outcomes into adulthood of infants born extremely preterm. *Seminars in Perinatology, 45*, 151483. https://doi.org/10.1016/j.semperi.2021.151483
29. Beam, A. L., Fried, I., Palmer, N., et al. (2020). Estimates of healthcare spending for preterm and low-birthweight infants in a commercially insured population: 2008–2016. *Journal of Perinatology., 40*, 1091–1099. https://doi.org/10.1038/s41372-020-0635-z
30. Horvath, H., Brindis, C. D., Reyes, E. M., Yamey, G., & Franck, L. (2017). Preterm birth: the role of knowledge transfer and exchange. *Health Research Policy and Systems., 15*, 1–14. https://doi.org/10.1186/s12961-017-0238-0
31. O'Brien, B. J., Spath, M., Blackhouse, G., Severens, J. L., Dorian, P., & Brazier, J. (2003). A view from the bridge: Agreement between the SF-6D utility algorithm and the Health Utilities Index. *Health Economics., 12*, 975–981. https://doi.org/10.1002/hec.789
32. Davison, S. N., Jhangri, G. S., & Feeny, D. H. (2009). Comparing the health utilities index mark 3 (HUI3) with the short form-36 preference-based SF-6D in chronic kidney disease. *Value in Health., 12*, 340–345. https://doi.org/10.1111/j.1524-4733.2008.00433.x
33. Abel, H., Kephart, G., Packer, T., & Warner, G. (2017). Discordance in utility measurement in persons with neurological conditions: A comparison of the SF-6D and the HUI3. *Value in Health., 20*, 1157–1165. https://doi.org/10.1016/j.jval.2017.04.008
34. Lubitz, C. C., De Gregorio, L., Fingeret, A. L., et al. (2017). Measurement and variation in estimation of quality of life effects of patients undergoing treatment for papillary thyroid carcinoma. *Thyroid., 27*, 197–206. https://doi.org/10.1089/thy.2016.0260
35. Barton, G. R., Bankart, J., Davis, A. C., & Summerfield, Q. A. (2004). Comparing utility scores before and after hearing-aid provision. *Applied Health Economics and Health Policy., 3*, 103–105. https://doi.org/10.2165/00148365-200403020-00006
36. Joern, M., & Thomas, K. (2008). Comparing preference-based quality-of-life measures: Results from rehabilitation patients with musculoskeletal, cardiovascular, or psychosomatic disorders. *Quality of Life Research., 17*, 485–495. https://doi.org/10.1007/s11136-008-9317-6
37. Abdin, E., Chong, S. A., Seow, E., et al. (2019). A comparison of the reliability and validity of SF-6D, EQ-5D and HUI3 utility measures in patients with schizophrenia and patients with depression in Singapore. *Psychiatry research., 274*, 400–408. https://doi.org/10.1016/j.psychres.2019.02.077
38. Fisk, J. D., Brown, M. G., Sketris, I. S., Metz, L. M., Murray, T. J., & Stadnyk, K. J. (2005). A comparison of health utility measures for the evaluation of multiple sclerosis treatments. *Journal of Neurology, Neurosurgery & Psychiatry., 76*, 58–63. https://doi.org/10.1136/jnnp.2003.017897

39. Langfitt, J. T., Vickrey, B. G., McDermott, M. P., et al. (2006). Validity and responsiveness of generic preference-based HRQOL instruments in chronic epilepsy. *Quality of Life Research., 15*, 899–914. https://doi.org/10.1007/s11136-005-5231-3
40. Richardson, J., Khan, M. A., Iezzi, A., & Maxwell, A. (2015). Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF6D, HUI 3, 15D, QWB, and AQoL-8D multiattribute utility instruments. *Medical Decision Making., 35*, 276–291. https://doi.org/10.1177/0272989X14543107
41. Macleod, M. R., Lawson, M. A., Kyriakopoulou, A., et al. (2015). Risk of bias in reports of in vivo research: a focus for improvement. *PLoS Biology., 13*, e1002273. https://doi.org/10.1371/journal.pbio.1002273
42. Zarbin, M. (2019). Real life outcomes vs. clinical trial results. *Journal of Ophthalmic & Vision Research., 14*, 88. https://doi.org/10.4103/jovr.jovr_279_18
43. Kristofer, B., & Johnston, B. C. (2014). Recent advances in patient and proxy-reported quality of life research. *Health and Quality of Life Outcomes., 12*, 1–9. https://doi.org/10.1186/s12955-014-0110-7
44. Richardson, J., Iezzi, A., Khan, M. A., Chen, G., & Maxwell, A. (2016). Measuring the sensitivity and construct validity of 6 utility instruments in 7 disease areas. *Medical Decision Making., 36*, 147–159. https://doi.org/10.1177/0272989X15613522
45. Arnett. JJ. (2015). *The Oxford handbook of emerging adulthood*. Oxford University Press. https://books.google.md/books?hl=en&lr=&id=E7uYCgAAQBAJ&oi=fnd&pg=PP1&ots=YL2Obf_JxS&sig=iX9DyIWwTtV9jNfj3mthgTEGTAg&redir_esc=y#v=onepage&q&f=false
46. Sylvia, P., Malte, S., Manon, G., Dieter, W., & Gijsbert, V. (2020). Quality of life of adults born very preterm or very low birth weight: A systematic review. *Acta Paediatrica., 109*, 1974–1988. https://doi.org/10.1111/apa.15249
47. Stavros, P., Natnaree, K., & Kamran, K. (2020). Preference-based health-related quality of life outcomes associated with preterm birth: A systematic review and meta-analysis. *Pharmacoeconomics., 38*, 357–373. https://doi.org/10.1007/s40273-019-00865-7
48. Eryigit, M. S., Baumann, N., Jaekel, J., Bartmann, P., & Wolke, D. (2015). Neuro-cognitive performance of very preterm or very low birth weight adults at 26 years. *Journal of Child Psychology and Psychiatry., 56*, 857–864. https://doi.org/10.1111/jcpp.12358
49. Doyle, L. W., Cheong, J. L., Burnett, A., et al. (2015). Biological and social influences on outcomes of extreme-preterm/low-birth weight adolescents. *Pediatrics, 136*, e1513–e1520. https://doi.org/10.1542/peds.2015-2006
50. Darlow, B. A., Woodward, L. J., Levin, K. J., Melzer, T., & Horwood, L. J. (2020). Perinatal and childhood predictors of general cognitive outcome at 28 years in a very-low-birthweight national cohort. *Developmental Medicine & Child Neurology., 62*, 1423–1428. https://doi.org/10.1111/dmcn.14649
51. Furlong, W. J., Feeny, D. H., Torrance, G. W., & Barr, R. D. (2001). The Health Utilities Index (HUI R) system for assessing health-related quality of life in clinical studies. *Annals of Medicine., 33*, 375–384. https://doi.org/10.3109/07853890109002092
52. Furlong, W., Feeny, D., Torrance, G., et al. (1998) *Multiplicative multi-attribute utility function for the Health Utilities Index Mark 3 (HUI3) system: A technical report tech. rep*. Centre for Health Economics and Policy Analysis (CHEPA), McMaster University.
53. David, F., William, F., Michael, B., & Torrance, G. W. (1995). Multi-attribute health status classification systems. *Pharmacoeconomics, 7*, 490–502. https://doi.org/10.2165/00019053-199507060-00004
54. Torrance, G. W., Feeny, D. H., Furlong, W. J., Barr, R. D., Zhang, Y., & Wang, Q. (1996). Multiattribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2. *Medical Care*. https://doi.org/10.1097/00005650-199607000-00004
55. Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics., 21*, 271–292. https://doi.org/10.1016/s0167-6296(01)00130-8
56. Sheffield The University. SF-6D—Calculating QALYs from the SF-36 and SF-12 2021.
57. Michael, D. (2001). Introducing economic and quality of life measurements into clinical studies. *Annals of Medicine., 33*, 344–349. https://doi.org/10.3109/07853890109002088
58. Walters, S. J., & Brazier, J. E. (2005). Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Quality of Life Research., 14*, 1523–1532. https://doi.org/10.1007/s11136-004-7713-0
59. Brazier, J., Ratcliffe, J., Saloman, J., & Tsuchiya, A. (2017). *Measuring and valuing health benefits for economic evaluation*. Oxford University Press.
60. Burke, D. L., Ensor, J., & Riley, R. D. (2017). Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ. *Statistics in Medicine., 36*, 855–875. https://doi.org/10.1002/sim.7141
61. Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research., 19*, 231–240. https://doi.org/10.1519/15184.1
62. Terwee Caroline, B., Bot Sandra, D. M., Boer Michael, R., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*, 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012
63. Bolbocean, C., van der Pal, S., van Buuren, S., Anderson, P. J., Bartmann, P., Baumann, N., & Petrou, S. (2022). Health-related quality-of-life outcomes of very preterm or very low birth weight adults: Evidence from an individual participant data meta-analysis. *PharmacoEconomics*. https://doi.org/10.1007/s40273-022-01201-2
64. Frey Heather, A., & Klebanoff, M. A. (2016). The epidemiology, etiology, and costs of preterm birth in. *Seminars in Fetal and Neonatal Medicine, 21*, 68–73. https://doi.org/10.1016/j.siny.2015.12.011
65. Johnston, K. M., Gooch, K., Korol, E., et al. (2014). The economic burden of prematurity in Canada. *BMC Pediatrics., 14*, 1–10. https://doi.org/10.1186/1471-2431-14-93
66. Ye, Z., Feng Hai-ming, Qu., Ji, L. X., Wen-Juan, Ma., & Jin-hui, T. (2018). A systematic review of pharmacoeconomic guidelines. *Journal of Medical Economics, 21*, 85–96. https://doi.org/10.1080/13696998.2017.1387118
67. Devlin, N. J., & Brooks, R. (2017). EQ-5D and the EuroQol group: Past, present and future. *Applied Health Economics and Health Policy., 15*, 127–137. https://doi.org/10.1007/s40258-017-0310-5
68. Rencz, F., László, G., Drummond, M., et al. (2016). EQ-5D in central and Eastern Europe: 2000–2015. *Quality of Life Research., 25*, 2693–2710. https://doi.org/10.1007/s11136-016-1375-6
69. Kaló, Z., Adrian, G., Mirjana, H., Marcell, C., & Boerlum, K. F. (2016). HTA implementation roadmap in Central and Eastern European countries. *Health Economics, 25*, 179–192. https://doi.org/10.1002/hec.3298

70. Khadka, J., Kwon, J., Petrou, S., Lancsar, E., & Ratcliffe, J. (2019). Mind the (interrater) gap. An investigation of self-reported versus proxy-reported assessments in the derivation of childhood utility values for economic evaluation: A systematic review. *Social Science & Medicine., 240*, 112543. https://doi.org/10.1016/j.socscimed.2019.112543

Authors and Affiliations

Corneliu Bolbocean[1] · Peter J. Anderson[2,3] · Peter Bartmann[4] · Jeanie L. Y. Cheong[3,5,6] · Lex W. Doyle[3,5,6,7] · Dieter Wolke[8] · Stavros Petrou[1]

1 Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK

2 School of Psychological Sciences, Turner Institute for Brain and Mental Health, Monash University, Melbourne, VIC, Australia

3 Clinical Sciences, Murdoch Children's Research Institute, Melbourne, VIC, Australia

4 Department of Neonatology and Paediatric Intensive Care, Children's Hospital, University Hospital Bonn, Bonn, Germany

5 Newborn Services, Royal Women's Hospital, Parkville, VIC, Australia

6 Department of Obstetrics and Gynaecology, University of Melbourne, Melbourne, VIC, Australia

7 Department Paediatrics, The University of Melbourne, Melbourne, VIC, Australia

8 Department of Psychology, Warwick Medical School, University of Warwick and Division of Health Sciences, Coventry, UK