Comparing the validity and responsiveness of the EQ-5D-5L to the Oxford hip and knee scores and SF-12 in osteoarthritis patients 1 year following total joint replacement

Barbara L. Conner-Spady[1] · Deborah A. Marshall[1] · Eric Bohm[2] · Michael J. Dunbar[3] · Tom W. Noseworthy[1]

Abstract

Purpose (1) To assess responsiveness of the EQ-5D-5L compared to Oxford hip and knee scores and the SF-12 in osteoarthritis patients undergoing total hip (THR) or knee (TKR) replacement surgery; (2) to compare distribution and anchor-based methods of assessing responsiveness.

Methods Questionnaires were mailed to consecutive patients following surgeon referral for primary THR or TKR and 1 year post-surgery. We assessed effect size (ES), standardized response mean (SRM), and standard error of measurement (SEM). Minimum important difference (MID) was the mean change in patients reporting somewhat better in hip or knee, health in general, and those who were satisfied with surgery (5-point scales). Responders were compared using MID versus 1 and 2SEM.

Results The sample of 537 (50% TKR) was composed of 56% female with a mean age of 64 years (SD 10). EQ-5D-5L ES was 1.86 (THR) and 1.19 (TKR) compared to 3.00 and 2.05 for Oxford scores, respectively. MID for the EQ-5D-5L was 0.22 (THR) and 0.20 (TKR) for patients who rated their hip or knee as somewhat better. There was a wide variation in the MID and the percentage of responders, depending on the joint, method of assessment, and the outcome measure. The percent agreement of responder classification using 2SEM vs. MID ranged from 79.6 to 99.6% for the EQ-5D-5L and from 69.4 to 94.8% for the Oxford scores.

Conclusions Responsiveness of the EQ-5D-5L was acceptable in TKR and THR. Caution should be taken in interpreting responder to TJR based on only one method of assessment.

Keywords EQ-5D-5L · Validity · Responsiveness · Total joint replacement · Minimum important difference · Oxford hip and knee Scores

Introduction

Osteoarthritis (OA) is a chronic, generally progressive, disease characterized by symptoms of joint pain and functional disability [1]. As the disease progresses, joint pain worsens and performing the activities of daily life becomes difficult. Health-related quality of life (HRQL) worsens, commonly affecting ones mood, the ability to have a comfortable sleep, and to enjoy social and recreational activities. When conservative treatment is no longer effective in providing pain relief and functional improvement, a decision may be made for total joint replacement (TJR).

Increasing emphasis is being placed on patient-reported outcome measures (PROMs) to evaluate the impact of TJR on HRQL [2, 3]. On average, patients undergoing TJR achieve significant improvement during the first 6–12 months following surgery [2, 4]; unfortunately, 7–30% of patients report little or no improvement or are not satisfied with the surgical results [1, 5–8] with poorer results for knee as compared to hip surgery.

✉ Barbara L. Conner-Spady
barbara.spady@ualberta.net

1   Department of Community Health Sciences, University of Calgary, 3rd Floor, TRW Building, 3280 Hospital Drive NW, Calgary, AB T2N 4Z6, Canada

2   Concordia Hip and Knee Institute, 310-1155 Concordia Avenue, Winnipeg, MB R2K 2M9, Canada

3   Department of Orthopaedic Surgery, Dalhousie University, 1796 Summer Street, Suite 4822, Halifax, NS B3H 4R2, Canada

Condition-specific and generic PROMs are often used in combination in arthroplasty registries to assess the HRQL of patients with TJR [9]. PROMs must be tested in the intended population to ensure that they meet acceptable standards of reliability and validity. Validity is the degree to which empirical evidence and theoretical rationales support the interpretability, appropriateness, and usefulness of a score [10]. One approach to assessing the validity of a new HRQL tool is to compare it with generic and condition-specific tools recognized as being reliable and valid in the population of interest. Convergent validity is the degree to which the scores correlate with other measures that assess a similar construct. An important aspect of validity in any outcome measure is the instrument's responsiveness, its ability to detect change in health status when it has occurred. While condition-specific PROMs usually have higher responsiveness, it is important to include generic PROMs to compare outcomes across different populations [11].

The EQ-5D-3L is a generic preference-based HRQL instrument for describing and valuing health [12]. It is one of the most common generic instruments used in arthroplasty registries [9] and is useful in economic evaluations where utility values are used to calculate quality-adjusted-life-years (QALYs) [13, 14]. It is based on a descriptive system that defines health in terms of 5 dimensions, each with three response options: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Responsiveness of the EQ-5D-3L compared to condition-specific measures in TJR shows that the EQ-5D-3L is less responsive than Oxford hip (OHS) and knee (OKS) scores [3, 13] and other joint-specific measures [13, 15]. The inability of the EQ-5D-3L to discriminate small to moderate differences in health status and its poor responsiveness were attributed to the wording of mobility [13] and a strong ceiling effect [15]. The three response options limit the range of responses for individuals with moderate to severe disability [16, 17]. For example, the three response options for mobility lack the flexibility of responses necessary to describe the mobility limitations of OA patients who may need a walking aid but are not confined to bed.

The recently developed 5-level EQ-5D, the EQ-5D-5L, addresses these issues by expanding the range of responses for each dimension from three to five levels and modifies the labels for some of the options. Thus, the wording for the most severe level of mobility was changed from "I am confined to bed" to "I am unable to walk about" [18]. Respondents also rate their overall health on a 0–100 visual analogue scale (EQ-VAS). The EQ-5D-5L has been compared to the EQ-5D-3L in population surveys [19–21] and clinical groups [22–26]. Its responsiveness has not been reported in longitudinal studies of patients having TJR.

The purpose of this study is twofold: (1) to assess the validity of the generic EQ-5D-5L and compare it to two commonly used PROMs (condition-specific and generic) in patients with OA with total hip (THR) or knee (TKR) replacement surgery; validity evidence included ceiling effects, convergent validity, and responsiveness; (2) to compare distribution-based and anchor-based methods of assessing responsiveness one year following TJR.

Methods

A questionnaire was administered to consecutive patients referred to orthopaedic practices in Halifax, Nova Scotia (QE11 Health Sciences Centre) and Winnipeg, Manitoba (Concordia Joint Replacement Group) from March 2012 to February 2014. In Halifax, the orthopaedic surgeons at the QE11 work loosely as a group. They receive about 95% of referrals from the Halifax area (Capital District) and approximately 50% of the referrals from the rest of the Province. The Concordia Joint Replacement Group consists of four orthopaedic surgeons, who practice as a group. They receive approximately 50–70% of all referrals in Winnipeg.

Patients meeting the inclusion criteria were adults (> 18 years), referred to an orthopaedic surgeon as likely candidates for a primary TJR due to OA. Ethics approval was obtained from the University of Calgary Conjoint Health Research Ethics Board, the University of Manitoba Health Research Ethics Board, and the Nova Scotia Health Authority Research Ethics Board.

An information sheet and the questionnaire with a postage-paid return envelope were mailed shortly after the clinic received the referral form. Non-responders received a 2-week follow-up reminder and a 4-week repeat survey package. Patients were assessed following referral to the orthopaedic surgeon to obtain baseline measures and at 12 months following surgery. A questionnaire was also sent following the orthopaedic consultation to assess test–retest reliability of the outcome measures.

Outcome measures

Patients completed HRQL questionnaires in the following order: the OHS or OKS, the EQ-5D-5L, EQ-VAS, and the Short Form 12 (SF-12). The OHS and OKS are 12-item questionnaires designed to assess the pain and symptoms of patients having THR [27] or TKR [28, 29]. The 12 items, each measured on a 5-point scale, assess hip or knee pain, mobility (walking, climbing, changing position), self-care (washing, drying, putting on socks), and activities such as shopping. If 2 or less items were left unanswered, we replaced the missing data with the mean value of the other responses, as per the scoring protocol for the OHS and OKS [29]. Summary scores range from 0 (most severe symptoms) to 48 (least symptoms). Because the OHS and OKS were

designed specifically for patients having a THR or TKR, they were used as the 'gold standard' by which to assess responsiveness of the EQ-5D-5L. To calculate the EQ-5D-5L index, we used the value set described by van Hout et al. [30] to convert the descriptive system to a summary index score. The SF-12 is a generic HRQL questionnaire which produces two standardized summary scores, the Physical (PCS-12) and Mental Component Summary (MCS-12) [31], that range from 0 (worst) to 100 (best). The PCS-12 includes items related to general health, daily activities, and the effect of pain on work, both outside and inside the home. The MCS-12 items assess emotional health and social activities. The SF-12 has been used to assess HRQL in TJR [11, 13, 32, 33]. Missing values for the EQ-5D-5L and SF-12 were not imputed. Demographic variables included age, sex, marital status, education, the problem joint (hip or knee), and previous TJR in another joint.

Data analysis

The distribution of the EQ-5D-5L responses across the five levels at baseline and 12 months post-surgery was assessed. This assessment is important to ensure that the scale captures the full range of potential responses and that changes over time can be detected. Significant ceiling or floor effects weaken the ability of a scale to detect deterioration or improvement in health status. Ceiling and floor effects were calculated as the percentage of patients showing the best and worst health status for each outcome measure at baseline and 12 months post-surgery. Floor or ceiling effects were defined as small (1–15%) and moderate (> 15%) [34].

Convergent validity

We assessed the convergent validity between the EQ-5D-5L and the other measures with Pearson correlation coefficients. Based on the literature [3, 13, 35, 36] and item content, we hypothesized that the EQ-5D-5L and the Oxford scores would correlate strongly (> 0.5), as both tools measure similar constructs relevant to OA patients— pain, mobility, self-care, and activities. We hypothesized that correlations of the EQ-5D-5L with the PCS-12 and EQ-VAS would be moderate (0.3–0.5) as both the PCS-12 and EQ-VAS assess general health, and that correlations with the MCS-12 would be moderate, as only one EQ-5D-5L item assesses mental health.

Responsiveness

Because responsiveness coefficients vary by population and context, multiple approaches to assessing responsiveness are recommended [37]. We used both distribution-based and anchor-based approaches. Distribution-based approaches

include the effect size (ES), the standardized response mean (SRM), and the standard error of measurement (SEM). The ES was calculated by dividing the difference in baseline and 1-year post-surgery scores by the standard deviation of the baseline score. The SRM was calculated by dividing the difference score by the standard deviation of the difference. Effect sizes were interpreted as small (0.2), medium (0.5), and large (0.8) using Cohen's conventions [38].

All PROM assessments have some measurement error [39]. The SEM is the standard deviation of errors of measurement associated with scores from a particular group of respondents. The SEM is a function of the reliability of the score and the standard deviation of scores. As reliability coefficients are sample dependent, we used intraclass correlation coefficients (ICCs) to assess the test–retest reliability of the outcome measures at 2 timepoints: at baseline and following surgeon consultation. The SEM was calculated by taking the square root of one minus the reliability and multiplying the results by the standard deviation of each outcome measure at baseline. A 95% confidence band (2SEM) was used around individual scores to provide 'reasonable limits' for estimating the true score [40]. The 2SEM criterion was applied to the change scores to classify patients as improved (> 2SEM) versus the same or worse. The procedure was repeated using a 68% confidence band (1SEM). Both 2SEM [41, 42] and 1SEM [43] have been used in the HRQL literature to classify patients into those who improved following surgery vs. those who did not.

The anchor-based method examines the relationship between a PROM and an external criterion or anchor, either clinical or patient-based, to assign people into several groups reflecting their degree of change [39, 44–46]. The minimum important difference (MID) is defined as the smallest difference in a PROM that is perceived by patients as beneficial or harmful [37, 39, 47, 48]. MIDs vary by population, context, method, the choice of anchor, the length of follow-up, and on the strength of the relationship between the HRQL measure and the anchor [37, 39]. A fundamental aspect of estimating the MID is to define the subgroup of people who have changed by a minimal amount [37, 39, 47]. The most commonly used method to identify this subgroup is to ask patients at follow-up how much they have changed since baseline; this is also referred to as a global transition question [37, 46]. Typically the MID is assessed by asking patients to provide a retrospective judgment about whether they have improved, stayed the same, or worsened over a period of time; the mean change of the group that reports getting a little better or a little worse is used to estimate the MID [37, 46, 47]. This is the method that we used. As recommended [37, 39], we calculated the MID using three different anchor questions, each with five response options, 1 year following surgery. (1) How would you rate your hip or knee now compared to when you first saw the orthopaedic

surgeon? (2) How would you rate your health in general now compared to when you first saw the orthopaedic surgeon? (3) How satisfied are you with your hip or knee replacement? The response options for the first two questions were: much better, somewhat better, about the same, somewhat worse, and much worse. The response options for the satisfaction question were: very satisfied, satisfied, neither satisfied nor dissatisfied, dissatisfied, and very dissatisfied. We calculated the MID as the mean change score between baseline and one year post-surgery in patients who rated (1) their hip or knee as "somewhat better", (2) their health in general as "somewhat better", and (3) their satisfaction as "satisfied".

Any evaluation of the effectiveness of TJR depends on the definition of a "successful" treatment [49]. In TJR, the MID has been used as a threshold to determine the number of patients who are classified as responders to surgery [8, 50–54] or as having a successful outcome [55]. We used the

MID as a cutpoint to dichotomize patients into responders and non-responders to TJR [8, 51]. Patients with a change score greater than the MID were classified as "responders". Finally, we cross-tabulated the responder categories defined by the MID with those defined by patients who improved using the 1SEM and 2SEM criteria. We calculated the percent agreement for each comparison.

Results

Questionnaires were mailed to 1738 patients who were referred to an orthopaedic surgeon; 975 patients returned the baseline questionnaire (Fig. 1). There was no significant difference in age or sex in responders vs. non-responders. Of the responders, 592 were offered and had a TJR; of these, 541 completed both the baseline and 12-month post-surgery
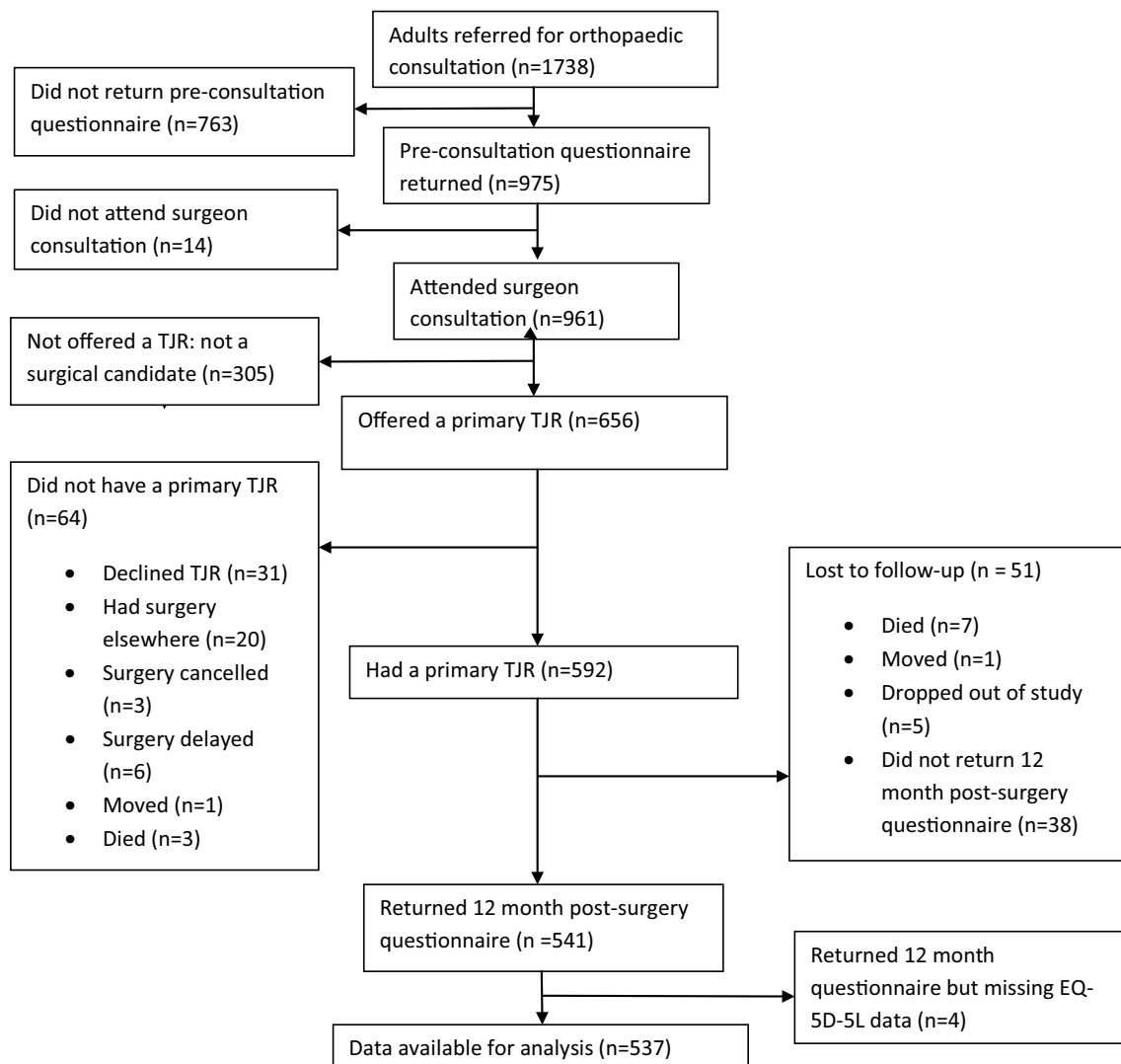

Fig. 1 Flow diagram showing recruitment of patients

questionnaire. Our sample included 537 patients with baseline and one-year post-surgery EQ-5D-5L data. Eighty-two percent of patients had their surgery in Winnipeg and 18% in Halifax. Fifty percent ($n = 268$) had a knee replacement, 19% ($n = 100$) had had a previous TJR on another joint, 56% ($n = 302$) were female, and the mean age was 64.1 years (SD 10.3). Table 1 provides baseline descriptive statistics for the demographics and outcome measures.

Table 2 shows the EQ-5D-5L responses for hip and knee replacement at baseline and one year post-surgery. All response options were used for mobility and self-care at both time points. The majority of patients reported "moderate" or "severe" problems at baseline for mobility, usual activities, and pain, and "slight" or "no" problems one year post-surgery for all five dimensions. Ceiling effects for the EQ-5D-5L post-surgery ranged from 18 (TKR) to 36% (THR) compared to 5% and 23%, respectively, for the Oxford scores (Table 3). There were no floor effects for the EQ-5D-5L or the SF-12.

As hypothesized, correlation coefficients of the EQ-5D-5L with the Oxford scores were strong; correlation of change scores were 0.72 for THR and 0.74 for TKR (Table 4). Correlations of EQ-5D-5L change scores with the other measures ranged from 0.32 (EQ-5D-5L and MCS-12) to 0.53 (EQ-5D-5L and PCS-12).

The effect sizes for the EQ-5D-5L were 1.86 (hip) and 1.19 (knee) (Table 5). These were smaller than effect sizes for the Oxford scores (3.00 and 2.05) and the PCS-12 (2.05 and 1.62), respectively.

The percentage improved with the EQ-5D-5L at one year post-surgery based on 2SEM was 77% for THR and 55% for TKR. This percentage was similar to the PCS-12 for THR and lower than the Oxford scores (Table 6).

The MID for the EQ-5D-5L was 0.22 (hip) and 0.20 (knee) for patients who rated their hip or knee, respectively, as somewhat better (Table 7). We did not assess the MID for patients who reported their outcomes as somewhat worse, as there were insufficient numbers in this category (5 or less). The percentage of patients classified as "responders" one year following surgery varied with the outcome measure, the anchor, and the surgical joint. The percentage of patients who improved using the EQ-5D-5L ranged from 57 (health now) to 81% (hip knee now) for THR and 54–68%, respectively, for TKR. The percentage improved was higher for hip patients than knee patients with both the EQ-5D-5L and Oxford scores. Similar to the percentage of patients who improved based on exceeding 1 or 2SEM, the percentage of patients categorized as "responders" with the MID was generally the highest using the Oxford scores.

Table 1 Baseline statistics for outcome measures and demographics for hip and knee replacement patients

	Hip					Knee				
	Mean	SD	Min	Max	n	Mean	SD	Min	Max	n
Outcome measure										
EQ-5D-5L	0.35	0.25	− 0.51	0.80	269	0.39	0.27	− 0.36	0.84	268
EQ-VAS	65.18	19.39	10.00	100.00	268	65.74	18.57	10.00	100.00	265
Oxford	17.42	7.76	0.00	40.00	269	17.57	8.08	1.00	42.00	268
PCS-12	28.54	9.02	10.00	59.91	267	28.59	8.25	6.56	56.71	267
MCS-12	49.59	12.05	13.45	77.20	267	51.52	12.25	15.64	74.65	267
Demographics										
Age	63.98	10.87	25.18	95.00	269	64.21	9.67	33.52	88.55	268
Sex F n (%)	139 (52)				269	163 (61)				268
Married/partner n (%)	211 (79)				267	189 (72)				264
Post high school n (%)	174 (65)				269	158 (59)				268

EQ-5D-5L index: range − 0.59 (health state worse than death) to 1 (full health)

EQ-VAS (EQ-5D-5L visual analogue scale): range 0 (worst health) to 100 (best health)

Oxford Hip (OHS) or Knee (OKS) score: range 0 (worst) to 48 (best)

PCS-12 Short Form-12 Physical Component Summary: range 0 (worst) to 100 (best)

MCS-12 Short Form-12 Mental Component Summary: range 0 (worst) to 100 (best)

Percent missing data: n missing/total (%):

Age, sex education 0/269 (0.0%); Marital status 2/269 (0.7%)

Age, sex, education 0/268 (0.0%); Marital status 4/268 (1.3%)

Hip EQ-5D-5L 0/269 (0.0%); EQ-VAS 1/269 (0.4%); OHS 0/269 (0.0%); PCS-12 2/269 (0.7%); MCS-12 2/269 (0.7%)

Knee EQ-5D-5L 0/268 (0.0%); EQ-VAS 3/268 (1.1%); OKS 0/268 (0.0%); PCS-12 1/268 (0.4%); MCS-12 1/268 (0.4%)

Table 2 Distribution of EQ-5D-5L responses at baseline and 1 year post-surgery

Dimension	Level of severity								
	No problems		Slight problems		Moderate problems		Severe problems		Extreme/unable	
Baseline hip										
Mobility	8	3.0%	22	8.2%	112	41.6%	121	45.0%	6	2.2%
Self-care	39	14.5%	72	26.8%	123	45.7%	34	12.6%	1	0.4%
Usual activities	6	2.2%	29	10.8%	109	40.5%	100	37.2%	25	9.3%
Pain/discomfort	0	0.0%	12	4.5%	114	42.4%	113	42.0%	30	11.2%
Anxiety/depression	84	31.2%	89	33.1%	77	28.6%	13	4.8%	6	2.2%
Post-surgery hip										
Mobility	167	62.1%	64	23.8%	25	9.3%	11	4.1%	2	0.7%
Self-care	209	77.7%	43	16.0%	13	4.8%	2	0.7%	2	0.7%
Usual activities	167	62.1%	67	24.9%	26	9.7%	9	3.3%	0	0.0%
Pain/discomfort	131	48.7%	94	34.9%	33	12.3%	10	3.7%	1	0.4%
Anxiety/depression	212	78.8%	37	13.8%	20	7.4%	0	0.0%	0	0.0%
Baseline Knee										
Mobility	2	0.7%	25	9.3%	134	50.0%	103	38.4%	4	1.5%
Self-care	91	34.0%	75	28.0%	77	28.7%	24	9.0%	1	0.4%
Usual activities	6	2.2%	45	16.8%	117	43.7%	78	29.1%	22	8.2%
Pain/discomfort	2	0.7%	16	6.0%	109	40.7%	109	40.7%	32	11.9%
Anxiety/depression	113	42.2%	70	26.1%	65	24.3%	17	6.3%	3	1.1%
Post-surgery knee										
Mobility	116	43.3%	78	29.1%	55	20.5%	17	6.3%	2	0.7%
Self-care	197	73.5%	47	17.5%	19	7.1%	4	1.5%	1	0.4%
Usual activities	120	44.8%	82	30.6%	48	17.9%	12	4.5%	6	2.2%
Pain/discomfort	68	25.4%	110	41.0%	66	24.6%	19	7.1%	5	1.9%
Anxiety/depression	174	64.9%	63	23.5%	20	7.5%	5	1.9%	6	2.2%

hip ($n=269$); knee ($n=268$)

Table 3 Floor and ceiling effects for outcome measures for hip and knee (percent)

	EQ-5D-5L	EQ-VAS	Oxford	PCS-12	MCS-12
Hip					
Ceiling					
Baseline	0.0	0.7	0.0	0.0	0.0
Post-surgery	35.7	4.9	23.2	0.0	0.0
Floor					
Baseline	0.0	0.0	1.1	0.0	0.0
Post-surgery	0.0	0.0	0.0	0.0	0.0
Knee					
Ceiling					
Baseline	0.0	1.1	0.0	0.0	0.0
Post-surgery	18.3	4.9	4.5	0.0	0.0
Floor					
Baseline	0.0	0.0	0.0	0.0	0.0
Post-surgery	0.0	0.0	0.0	0.0	0.0

On the suggestion of a reviewer, using a method described by Lee et al. [56], we compared our MID results to those using simple linear regression. Lee et al. calculated a minimal clinically important difference (MCID), using the slope of the line for the change in outcome measure for levels of patient satisfaction. Using this method, we used the slope of the line for the change in EQ-5D-5L and Oxford scores for the change according to level of perceived improvement

Table 4 Correlation coefficients of EQ-5D-5L with other outcome measures

	Baseline		Change scores	
	Hip	Knee	Hip	Knee
Oxford	0.78	0.79	0.72	0.74
PCS-12	0.49	0.45	0.48	0.53
MCS-12	0.37	0.45	0.33	0.32
EQ-VAS	0.39	0.31	0.37	0.43

Pearson correlation coefficients all sig ($p < 0.001$)

in the hip or knee. We collapsed the levels of improvement into three groups: "much better", "somewhat better", and "about the same/worse" due to small numbers in the "about the same" or "worse" groups. The MCID for the EQ-5D-5L, identified by simple linear regression, was 0.24 (95% CI 0.11–0.37) and 0.20 (95% CI 0.12–0.27) for THR and TKR, respectively. The analysis was repeated for the OHS and OKS; results were 12.26 (95% CI 7.98–16.53) for the OHS and 8.78 (95% CI 6.36–11.21) for the OKS.

We cross-classified the responder category (responder or non-responder) of the change scores with the MID (calculated with 3 different anchors) versus 1 and 2 SEM. Table 8 shows the percent agreement for responder category for each comparison. For the EQ-5D-5L, there was 96.3% agreement in the classification of responder status for THR and 86.9% for TKR using 2 SEM vs. the MID with hip or knee improvement as the anchor. A comparison of the percentage of patients who were classified in the same category (responder or non-responder) using 2SEM versus the MID showed agreement between 79.6 and 99.6% with the EQ-5D-5L and between 69.4% and 94.8% with the Oxford scores. The percentage agreement using the two methods was generally higher with 2SEM compared to 1SEM for the Oxford scores and EQ-5D-5L.

Table 5 Effect size (ES) and standardized response mean (SRM) for outcome measures for hip and knee

Joint	EQ-5D-5L		EQ-VAS		Oxford		PCS-12		MCS-12	
	ES	SRM	ES	SRM	ES	SRM	ES	SRM	ES	SRM
Hip	1.86	1.53	0.88	0.90	3.00	2.29	2.05	1.54	0.47	0.50
Knee	1.19	1.04	0.63	0.55	2.05	1.54	1.62	1.15	0.15	0.17

Sample size (% missing data) for ES and SRM (Baseline and 12-month post-surgery data) for each outcome measure: Hip: EQ-5D-5L $n = 269$ (0.0%); EQ-VAS $n = 268$ (0.4%); OHS $n = 267$ (0.7%); PCS-12 $n = 259$ (3.7%); MCS-12 $n = 259$ (3.7%). Knee: EQ-5D-5L $n = 268$ (0.0%); EQ-VAS $n = 264$ (1.5%); OKS $n = 268$ (0.0%); PCS-12 $n = 257$ (4.1%); MCS-12 $n = 257$ (4.1%)

ES effect size, *SRM* standardized response mean

Table 6 Percent of patients improved > 1 and > 2 standard errors of measurement (SEM)

	ICC[a]	Baseline SD	1SEM[b]	2SEM[b]	% improved 1SEM	% improved 2SEM
Hip						
EQ-5D-5L	0.75	0.25	0.13	0.25	87.4%	77.3%
EQ-VAS	0.59	19.39	12.42	24.83	54.1%	31.3%
Oxford hip score	0.82	7.76	3.29	6.59	96.3%	94.0%
PCS-12	0.66	9.02	5.26	10.52	84.6%	73.4%
MCS-12	0.72	12.05	6.38	12.75	42.5%	27.0%
Knee						
EQ-5D-5L	0.75	0.27	0.13	0.27	74.5%	54.7%
EQ-VAS	0.59	18.54	11.87	23.75	43.3%	25.5%
Oxford knee score	0.82	8.10	3.43	6.87	88.8%	83.9%
PCS-12	0.66	8.25	4.81	9.62	79.0%	64.2%
MCS-12	0.72	12.25	6.48	12.97	32.7%	14.4%

[a] *ICC* Intraclass correlation coefficient based on test–retest reliability. Sample size for ICCs: EQ-5D-5L ($n = 591$); EQ-VAS ($n = 586$); Oxford scores ($n = 585$); PCS-12 ($n = 548$); MCS-12 ($n = 548$)

[b] *SEM* standard error of measurement

Table 7 Minimum important difference (MID) and percent improved for each outcome measure for hip and knee

Anchor	Response	EQ-5D-5L		EQ-VAS		Oxford Score		PCS-12		MCS-12	
		MID	%	MID	%	MID	%	MID	%	MID	%
Hip											
Hip/knee[a]	Somewhat better	0.22	81	1.35	79	10.68	89	4.15	88	3.60	52
Health[b]	Somewhat better	0.41	57	9.34	67	19.93	68	13.42	64	5.91	44
Satisfaction[c]	Satisfied	0.30	72	13.09	54	17.56	75	14.11	63	2.88	54
Knee											
Hip/knee[a]	Somewhat better	0.20	68	0.42	72	11.13	69	3.47	81	3.84	40
Health[b]	Somewhat better	0.28	54	7.75	57	16.49	53	11.96	53	2.03	49
Satisfaction[c]	Satisfied	0.26	56	9.90	55	13.45	64	10.20	63	1.78	49

Minimum important difference (MID) based on mean change score between baseline and 1-year post-surgery for each response

Percent improved based on change score > MID

[a]How would you rate your hip or knee now compared to when you first saw the orthopaedic surgeon?

[b]How would you rate your health in general now compared to when you first saw the orthopaedic surgeon?

[c]How satisfied are you with your hip or knee replacement?

Table 8 Percent agreement of responder category (responder or non-responder) with minimum important difference (MID) vs. 1 and 2 standard errors of measurement (SEM)

Anchor	MID versus 2SEM		MID versus 1SEM	
	Hip (%)	Knee (%)	Hip (%)	Knee (%)
EQ-5D-5L				
Hip or knee now	96.3	86.9	93.7	92.9
Health now	79.6	99.6	69.5	79.5
Satisfaction	94.4	98.9	84.4	81.0
EQ-VAS				
Hip or knee now	52.2	53.8	75.0	71.6
Health now	64.6	68.2	87.3	86.0
Satisfaction	77.6	70.1	99.6	87.9
Oxford				
Hip or knee now	94.8	85.8	92.5	81.0
Health now	73.8	69.4	71.5	64.6
Satisfaction	80.9	80.2	78.7	75.4
PCS-12				
Hip or knee now	85.7	83.3	96.9	98.1
Health now	91.1	89.1	79.9	74.3
Satisfaction	90.0	98.8	78.8	84.0
MCS-12				
Hip or knee now	75.3	73.9	90.7	92.2
Health now	83.0	65.8	98.5	84.0
Satisfaction	72.6	65.8	88.0	84.0

Discussion

In this study, we assessed the convergent validity and responsiveness of the EQ-5D-5L and compared it to condition-specific and generic PROMs in patients who had either THR or TKR. We also compared both distribution-based and anchor-based methods of assessing responsiveness. Baseline Oxford scores and PCS-12 were similar to pre-TJR scores in other studies [3, 57, 58], while EQ-5D-5L index scores were similar to pre-surgery EQ-5D-3L index scores in THR [13] and TJR [3] patients. The baseline MCS-12 was similar to other pre-TJR scores [11, 57] and healthy population norms [31]. Compared to EQ-5D-3L rheumatology studies where the majority of patients used level 2 (some to moderate) for mobility, usual activities, and pain/discomfort [16, 35, 59], responses to the EQ-5D-5L were more evenly distributed across the five levels.

Ceiling effects for the EQ-5D-5L one year following surgery were moderate (> 15%) but were less than those reported for the EQ-5D-3L at 6 months [60] and one year post-surgery [13, 15], which was reported as high as 84% [15]. The post-surgery ceiling effect for the OHS was also moderate and has been reported in other studies for the OHS and OKS [33, 61].

Convergent validity

Our results supported the convergent validity of the EQ-5D-5L in TJR. Both baseline and change score correlations between the EQ-5D-5L and Oxford scores were strong and similar to those in a cross-sectional study of TJR [35]. Correlations were stronger than those between the EQ-5D-3L and Oxford scores reported in other studies [3, 13, 35, 36].

Responsiveness

Although effect sizes for the EQ-5D-5L were large (> 0.8), they were smaller than effect sizes for the Oxford scores and the PCS-12. Compared to the few studies that have reported ES for the EQ-5D-3L in TJR, effect sizes for the EQ-5D-5L in our study were higher than those reported for the EQ-5D-3L in THR patients [13] and similar in TKR patients [15] one year following surgery. ESs for Oxford scores were similar to those in other studies of patients with TKR [3] and THR [13, 33, 58]. In a comparison of responsiveness of the OHS and EQ-5D-3L, Dawson et al. [62] also found a larger ES for the OHS one year following revision hip replacement. ICCs were similar or lower than those in the few studies that have reported test–retest reliability in similar populations [28, 35, 63, 64]. The percentage improved using 2SEM was higher for hip patients than knee patients for all outcome measures. This is consistent with other studies that report a higher percentage of patients showing improvement with hip vs. knee surgery [4, 11].

We assessed MIDs using three patient-rated questions as anchors: perceived change in health status, change in hip or knee, and satisfaction with surgery 12 months following surgery. The MID for each PROM varied depending on the joint (hip or knee) and the anchor. For the EQ-5D-5L, the MID ranged from 0.20 (hip or knee somewhat better for TKR) to 0.41 (health somewhat better for THR). In comparison, Walters and Brazier reported a MID of 0.12 for the EQ-5D-3L for TKR patients who reported some change in health from baseline to 6 months [65]. Although there is some concern regarding the application of the MID for preference-based HRQL instruments in cost-effectiveness analysis [66], the EQ-5D-5L is also used in clinical evaluation, where the MID is well accepted. Our results for the MID of 11 points for OHS and OKS for patients who reported their hip/knee problem as "somewhat better" were similar to the mean values reported by Beard et al. [58] of 11 and 9 points for the OHS and OKS, respectively, for patients whose response was "a little better" to a question that compared their hip/knee problems 6 months after surgery to before surgery. Our MID results for the EQ-5D-5L and Oxford scores were similar to those using the linear regression method described by Lee et al. [56].

The percentage of responders with the EQ-5D-5L varied from 54 to 81%, and with the Oxford scores, from 53 to 89%. In comparison, other longitudinal studies using the MID to determine responders to TJR have reported the percentage of responders ranging from 54 [53] to 70% [8, 11]. However, these studies varied in method of assessment, hip or knee joint, and follow-up time. Similar to other studies [11, 53, 67], a higher percentage of hip patients compared to knee patients were classified as responders. Of the three anchors we used to calculate the MID, the anchor most specific to

the hip and knee outcome resulted in the highest proportion of responders for all outcome measures except the MCS-12.

Comparison of distribution and anchor-based methods

Both distribution-based and anchor-based methods have been used to assess responsiveness but few studies have done concurrent comparisons of both approaches [43, 46, 68]. In our study, the percent agreement (responders or non-responders) using 2SEM and MID ranged from 79.6 to 99.6% for the EQ-5D-5L and from 69.4 to 94.8% for the Oxford scores. The percent agreement varied with the measures but was generally higher with 2SEM compared to 1SEM with the EQ-5D-5L and Oxford scores. In contrast, Wyrwich et al. [43] reported that 1SEM closely approximated the MCID in patients with chronic obstructive pulmonary disease. However, they did not report a comparison of the 2SEM with the MCID. Based on a systematic literature review of the MID, Norman et al. [46] concluded that the MID estimates were close to one-half a SD, which is equivalent to 1SEM when the test–retest reliability is 0.75. However, in one study, this method of defining responders resulted in only 54% of patients classified as having a "good TJR outcome" 6 months following surgery [53].

Limitations of this study were a poor response rate for patients following referral but there was no significant difference in demographic variables between responders and non-responders. Based on the percentage of responders who were offered surgery, not all of the non-responders would have been surgical patients. This could explain some of the non-response, as the survey questions may not have been relevant for patients who were not appropriate candidates for TJR. However, the majority of surgical patients who returned the baseline survey also completed the one year follow-up survey. The analysis for test–retest reliability may be biased by the orthopaedic consultation between the pre- and post-consultation assessments.

Conclusion

This study is the first to report on the responsiveness of the EQ-5D-5L in a longitudinal study of patients undergoing TJR. Our results support its use as a generic HRQL measure to be used alongside a condition-specific measure in assessing PROMs in primary TJR. Compared to the few studies that have assessed the responsiveness of the EQ-5D-3L in primary TJR [3, 13, 15, 60], the EQ-5D-5L had a more even distribution of responses across the five dimensions, less ceiling effect, and a larger ES for THR. Although the EQ-5D-5L was less responsive than the Oxford scores, this finding is consistent with other comparative studies of

condition-specific and generic HRQL measures [15, 41]. We showed that there is wide variation in the MID and the percentage of responders, depending on the joint assessed, the method of assessment, and the outcome measure. Therefore, interpretation of a successful outcome of TJR should take these factors into consideration. Caution should be taken in the interpretation of the percentage of "responders" to TJR based on only one method of assessing the responder classification.

Compliance with ethical standards

Conflict of interest The first author is a member of the EuroQol Group, a not-for-profit group that develops and distributes instruments that assess and value health.

Informed consent All persons gave their informed consent prior to inclusion in the study.

References

1. Jones, C. A., Beaupre, L. A., Johnston, D. W., & Suarez-Almazor, M. E. (2005). Total joint arthroplasties: Current concepts of patient outcomes after surgery. *Clinics in Geriatric Medicine*, 21, 527–41.
2. Franklin, P. D., Lewallen, D., Bozic, K., Hallstrom, B., Jiranek, W., & Ayers, D. C. (2014). Implementation of patient-reported outcome measures in U.S. Total joint replacement registries: rationale, status, and plans. *Journal of Bone and Joint Surgery, 96*(Suppl 1), 104–109.
3. Lin, F. J., Samp, J., Munoz, A., Wong, P. S., & Pickard, A. S. (2014). Evaluating change using patient-reported outcome measures in knee replacement: The complementary nature of the EQ-5D index and VAS scores. *The European Journal of Health Economics, 15*, 489–496.
4. Ethgen, O., Bruyere, O., Richy, F., Dardennes, C., & Reginster, J. Y. (2004). Health-related quality of life in total hip and total knee arthroplasty. A qualitative and systematic review of the literature. *Journal of Bone and Joint Surgery, 86-A*, 963–974.
5. Anakwe, R. E., Jenkins, P. J., & Moran, M. (2011). Predicting dissatisfaction after total hip arthroplasty: A study of 850 patients. *The Journal of Arthroplasty, 26*, 209–213.
6. Gandhi, R., Davey, J. R., & Mahomed, N. N. (2008). Predicting patient dissatisfaction following joint replacement surgery. *The Journal of Rheumatology, 35*, 2415–2418.
7. Scott, C. E., Howie, C. R., MacDonald, D., & Biant, L. C. (2010). Predicting dissatisfaction following total knee replacement: a prospective study of 1217 patients. *The Journal of Bone and Joint Surgery, 92*, 1253–1258.
8. Judge, A., Cooper, C., Williams, S., Dreinhoefer, K., & Dieppe, P. (2010). Patient-reported outcomes one year after primary hip replacement in a European Collaborative Cohort. *Arthritis Care & Research, 62*, 480–488.
9. Rolfson, O., Eresian, C. K., Bohm, E., Lubbeke, A., Denissen, G., Dunn, J., et al. (2016). Patient-reported outcome measures in arthroplasty registries. *Acta Orthopaedica, 87*(Suppl 1), 3–8.
10. Messick, S. (1989). Validity. In (In R. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan Publishing Company.
11. SooHoo, N. F., Li, Z., Chenok, K. E., & Bozic, K. J. (2015). Responsiveness of patient reported outcome measures in total joint arthroplasty patients. *The Journal of Arthroplasty, 30*, 176–191.
12. Brooks, R. (1996). EuroQol: the current state of play. *Health Policy, 37*, 53–72.
13. Ostendorf, M., van Stel, H. F., Buskens, E., Schrijvers, A. J., Marting, L. N., Verbout, A. J., et al. (2004). Patient-reported outcome in total hip replacement. A comparison of five instruments of health status. *The Journal of Bone and Joint Surgery, 86*, 801–808.
14. Jenkins, P. J., Clement, N. D., Hamilton, D. F., Gaston, P., Patton, J. T., & Howie, C. R. (2013). Predicting the cost-effectiveness of total hip and knee replacement: a health economic analysis. *The Bone & Joint Journal, 95-B*, 115–121.
15. Giesinger, K., Hamilton, D. F., Jost, B., Holzner, B., & Giesinger, J. M. (2014). Comparative responsiveness of outcome measures for total knee arthroplasty. *Osteoarthritis Cartilage, 22*, 184–189.
16. Fransen, M., & Edmonds, J. (1999). Reliability and validity of the EuroQol in patients with osteoarthritis of the knee. *Rheumatology, 38*, 807–813.
17. Wolfe, F., & Hawley, D. J. (1997). Measurement of the quality of life in rheumatic disorders using the EuroQol. *British Journal of Rheumatology, 36*, 786–793.
18. Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research* 20, 1727–1736
19. Luo, N., Li, M., Chevalier, J., Lloyd, A., & Herdman, M. (2013). A comparison of the scaling properties of the English, Spanish, French, and Chinese EQ-5D descriptive systems. *Quality of Life Research, 22*, 2237–2243.
20. Kim, T. H., Jo, M. W., Lee, S. I., Kim, S. H., & Chung, S. M. (2013). Psychometric properties of the EQ-5D-5L in the general population of South Korea. *Quality of Life Research, 22*, 2245–2253.
21. Hinz, A., Kohlmann, T., Stobel-Richter, Y., Zenger, M., & Brahler, E. (2014). The quality of life questionnaire EQ-5D-5L: psychometric properties and normative values for the general German population. *Quality of Life Research, 23*, 443–447.
22. Kim, S. H., Kim, H. J., Lee, S. I., & Jo, M. W. (2012). Comparing the psychometric properties of the EQ-5D-3L and EQ-5D-5L in cancer patients in Korea. *Quality of Life Research, 21*, 1065–1073.
23. Jia, Y. X., Cui, F. Q., Li, L., Zhang, D. L., Zhang, G. M., Wang, F. Z., et al. (2014). Comparison between the EQ-5D-5L and the EQ-5D-3L in patients with hepatitis B. *Quality of Life Research, 23*, 2355–2363.
24. Pickard, A. S., De Leon, M. C., Kohlmann, T., Cella, D., & Rosenbloom, S. (2007). Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Medical Care, 45*, 259–263.
25. Scalone, L., Ciampichini, R., Fagiuoli, S., Gardini, I., Fusco, F., Gaeta, L., et al. (2013). Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. *Quality of Life Research, 22*, 1707–1716.
26. Janssen, M. F., Pickard, A. S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., et al. (2013). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Quality of Life Research, 22*, 1717–1727.

27. Dawson, J., Fitzpatrick, R., Carr, A., & Murray, D. (1996). Questionnaire on the perceptions of patients about total hip replacement. *The Bone & Joint Journal, 78*, 185–190.

28. Dawson, J., Fitzpatrick, R., Murray, D., & Carr, A. (1998). Questionnaire on the perceptions of patients about total knee replacement. *The Bone & Joint Journal, 80*, 63–69.

29. Murray, D. W., Fitzpatrick, R., Rogers, K., Pandit, H., Beard, D. J., Carr, A. J., et al. (2007). The use of the Oxford hip and knee scores. *The Bone & Joint Journal, 89*, 1010–1014.

30. van Hout, B., Janssen, M. F., Feng, Y. S., Kohlmann, T., Busschbach, J., Golicki, D., et al. (2012). Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health, 15*, 708–715.

31. Ware, J. Jr., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*, 220–233.

32. Gonzalez Saenz, dT., Escobar, A., Herrera, C., Garcia, L., Aizpuru, F., & Sarasqueta, C. (2010). Patient expectations and health-related quality of life outcomes following total joint replacement. *Value Health, 13*, 447–454.

33. Garbuz, D. S., Xu, M., & Sayre, E. C. (2006). Patients' outcome after total hip arthroplasty: a comparison between the Western Ontario and McMaster Universities index and the Oxford 12-item hip score. The *Journal of Arthroplasty, 21*, 998–1004.

34. McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Quality of Life Research, 4*, 293–307.

35. Conner-Spady, B. L., Marshall, D. A., Bohm, E., Dunbar, M. J., Loucks, L., Khudairy, A., A. et al (2015). Reliability and validity of the EQ-5D-5L compared to the EQ-5D-3L in patients with osteoarthritis referred for hip and knee replacement. *Quality of Life Research, 24*, 1775–1784.

36. Benson, T., Williams, D. H., & Potts, H. W. (2016). Performance of EQ-5D, howRu and Oxford hip & knee scores in assessing the outcome of hip and knee replacements. *BMC Health Services Research, 16*, 512.

37. King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*, 171–184.

38. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates.

39. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*, 102–109.

40. Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley

41. Escobar, A., Quintana, J. M., Bilbao, A., Arostegui, I., Lafuente, I., & Vidaurreta, I. (2007). Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. *Osteoarthritis Cartilage, 15*, 273–280.

42. Quintana, J. M., Escobar, A., Bilbao, A., Arostegui, I., Lafuente, I., & Vidaurreta, I. (2005). Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. *Osteoarthritis Cartilage, 13*, 1076–1083.

43. Wyrwich, K. W., Tierney, W. M., & Wolinsky, F. D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *Journal of Clinical Epidemiology, 52*, 861–873.

44. Lydick, E., & Epstein, R. S. (1993). Interpretation of quality of life changes. *Quality of Life Research, 2*, 221–226.

45. Norman, G. R., Sridhar, F. G., Guyatt, G. H., & Walter, S. D. (2001). Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. *Medical Care, 39*, 1039–1047.

46. Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care, 41*, 582–592.

47. Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD, 2*, 63–67.

48. Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials, 10*, 407–415.

49. Ramkumar, P. N., Harris, J. D., & Noble, P. C. (2015). Patient-reported outcome measures after total knee arthroplasty: a systematic review. *Bone & Joint Research, 4*, 120–127.

50. Escobar, A., Garcia, P. L., Herrera-Espineira, C., Aizpuru, F., & Sarasqueta, C. (2013). Total knee replacement; minimal clinically important differences and responders. *Osteoarthritis Cartilage, 21*, 2006–2012

51. Terwee, C. B., Roorda, L. D., Knol, D. L., de Boer, M. R., & de Vet, H. C. (2009). Linking measurement error to minimal important change of patient-reported outcomes. *Journal of Clinical Epidemiology, 62*, 1062–1067.

52. Quintana, J. M., Aguirre, U., Barrio, I., Orive, M., Garcia, S., & Escobar, A. (2012). Outcomes after total hip replacement based on patients' baseline status: what results can be expected? *Arthritis Care & Research, 64*, 563–572.

53. Hawker, G. A., Badley, E. M., Borkhoff, C. M., Croxford, R., Davis, A. M., Dunn, S., et al. (2013). Which patients are most likely to benefit from total joint arthroplasty? *Arthritis & Rheumatism, 65*, 1243–1252.

54. Haase, E., Kopkow, C., Beyer, F., Lutzner, J., Kirschner, S., Hartmann, A., et al. (2016). Patient-reported outcomes and outcome predictors after primary total hip arthroplasty: results from the Dresden Hip Surgery Registry. *Hip International, 26*, 73–81.

55. Berliner, J. L., Brodke, D. J., Chan, V., SooHoo, N. F., & Bozic, K. J. (2016). John Charnley Award: Preoperative patient-reported outcome measures predict clinically meaningful improvement in function after THA. *Clinical Orthopaedics and Related Research, 474*, 321–329.

56. Lee, W. C., Kwan, Y. H., Chong, H. C., & Yeo, S. J. (2017). The minimal clinically important difference for Knee Society Clinical Rating System after total knee arthroplasty for primary osteoarthritis. *Knee Surgery, Sports Traumatology, Arthroscopy, 25*, 3354–3359.

57. Clement, N. D., MacDonald, D., & Simpson, A. H. (2014). The minimal clinically important difference in the Oxford knee score and Short Form 12 score after total knee arthroplasty. *Knee Surgery, Sports Traumatology, Arthroscopy, 22*, 1933–1939.

58. Beard, D. J., Harris, K., Dawson, J., Doll, H., Murray, D. W., Carr, A. J., et al. (2015). Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. *Journal of Clinical Epidemiology, 68*, 73–79.

59. Conner-Spady, B., & Suarez-Almazor, M. E. (2003). Variation in the estimation of quality-adjusted life-years by different preference-based instruments. *Medical Care, 41*, 791–801.

60. Impellizzeri, F. M., Leunig, M., Preiss, S., Guggi, T., & Mannion, A. F. (2017). The use of the Core Outcome Measures Index (COMI) in patients undergoing total knee replacement. *The Knee, 24*, 370–372.

61. Jenny, J. Y., Louis, P., & Diesinger, Y. (2014). High activity arthroplasty score has a lower ceiling effect than standard scores after knee arthroplasty. *The Journal of Arthroplasty, 29*, 719–721.

62. Dawson, J., Fitzpatrick, R., Frost, S., Gundle, R., McLardy-Smith, P., & Murray, D. (2001). Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. *The Journal of Bone and Joint Surgery, 83*, 1125–1129.

63. Naylor, J. M., Kamalasena, G., Hayen, A., Harris, I. A., & Adie, S. (2013). Can the Oxford Scores be used to monitor symptomatic

progression of patients awaiting knee or hip arthroplasty? *The Journal of Arthroplasty, 28*, 1454–1458.

64. Poitras, S., Beaule, P. E., & Dervin, G. F. (2012). Validity of a short-term quality of life questionnaire in patients undergoing joint replacement: the Quality of Recovery-40. *The Journal of Arthroplasty, 27*, 1604–1608.

65. 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.

66. Whitehurst, D. G., & Bryan, S. (2013). Trial-based clinical and economic analyses: the unhelpful quest for conformity. *Trials, 14*, 421.

67. Escobar, A., Gonzalez, M., Quintana, J. M., Vrotsou, K., Bilbao, A., Herrera-Espineira, C., et al. (2012). Patient acceptable symptom state and OMERACT-OARSI set of responder criteria in joint replacement. Identification of cut-off values. *Osteoarthritis Cartilage, 20*, 87–92.

68. Beaton, D. E., Boers, M., & Wells, G. A. (2002). Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Current Opinion in Rheumatology, 14*, 109–114.