



# Responsiveness of the EQ-5D-3L and EQ-5D-5L in patients following total hip or knee replacement

Xuejing Jin<sup>1</sup> · Fatima Al Sayah<sup>1</sup> · Arto Ohinmaa<sup>1</sup> · Deborah A. Marshall<sup>2,3</sup> · Jeffrey A. Johnson<sup>1</sup>

Accepted: 3 May 2019 / Published online: 14 May 2019  
© Springer Nature Switzerland AG 2019

## Abstract

**Purpose** To examine the responsiveness of the EQ-5D-3L and EQ-5D-5L among total hip/knee replacement (THR/TKR) patients.

**Methods** The EQ-5D (3L or 5L) and Western Ontario and McMaster University Osteoarthritis Index (WOMAC) have been routinely administered to all THR/TKR patients before and at 3 months after surgery in Alberta, Canada, since 2010. Patients were included in this analysis if they completed the WOMAC and the same version of EQ-5D at baseline and 3-month follow-up. The WOMAC was used as an anchor to categorize patients into 9 subgroups according to the relative change from baseline, i.e., no change, and 4 categories each for the amount of deterioration or improvement: large ( $\geq 70\%$ ), moderate ( $50\% \leq \text{change} < 70\%$ ), small but important ( $20\% \leq \text{change} < 50\%$ ), and very mild ( $0 < \text{change} \leq 20\%$ ). The responsiveness of the EQ-5D-3L and EQ-5D-5L for each subgroup was assessed using effect size, standardized response mean, and Guyatt's Responsiveness Index.

**Results** A total of 1594 patients completed the EQ-5D-3L and WOMAC (60% females, mean age 66 years,  $N = 646$  [41%] THR), and 3180 completed the EQ-5D-5L and WOMAC (60% females, mean age 66 years,  $N = 1352$  [43.2%] THR) at baseline and 3-month follow-up. For both THR and TKR patients with “small but important” improvement, the EQ-5D-5L was consistently more responsive than the EQ-5D-3L.

**Conclusion** Our study demonstrated that the EQ-5D-5L is more responsive than the EQ-5D-3L in identifying health-related quality of life changes in THR/TKR patients. We recommend using the EQ-5D-5L in longitudinal studies in this patient population.

**Keywords** EQ-5D · Responsiveness · Measurement properties · Total hip replacement · Total knee replacement

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11136-019-02200-1>) contains supplementary material, which is available to authorized users.

✉ Jeffrey A. Johnson  
jeffreyj@ualberta.ca

<sup>1</sup> School of Public Health, University of Alberta, 2-040 Li Ka Shing Centre for Health Research Innovation, Edmonton, AB T6G 2E1, Canada

<sup>2</sup> Department of Community Health Sciences, University of Calgary, HRIC Building Room 3C58, 3280 Hospital Drive NW, Calgary, AB T2N 4Z6, Canada

<sup>3</sup> Alberta Bone and Joint Health Institute, Calgary, Canada

## Introduction

For patients suffering from pain or function limitations caused by end-stage of osteoarthritis (OA, hip and/or knee) or joint injuries, joint replacement, including total hip replacement (THR) and total knee replacement (TKR), is one of the most effective ways to relieve the symptoms and improve patients' functioning and health-related quality of life (HRQL) [1]. In 2015, there were 166 cases of hip replacements and 126 knee replacements per 100,000 people in the Organization for Economic Co-operation and Development (OECD) countries [2]. On average, the rate of knee replacements almost doubled and the rate of hip replacements increased by 30% in OECD countries from 2000 to 2015 [2]. In Canada, the numbers of THR and TKR in 2016–2017 were 55,981 and 67,169 (increased by 18% and 15% since 2011–2012), respectively, which placed a burden

of more than 1 billion Canadian dollars on the Canada health system [3].

Due to the continued growth of the number of joint replacements and emphasizing patient-centered care in this population, HRQL of patients following THR and TKR has been considered an important outcome of these procedures, that is, the change in HRQL before to after the surgery. Therefore, instruments used to measure HRQL (and change in HRQL) of patients following the THR/TKR need to have the ability to pick up such change (i.e., responsiveness) [4].

HRQL of the THR/TKR patients can be measured using both OA-specific measures and generic measures [5–7]. Generic preference-based measures can provide utility scores that can be used in calculating quality-adjusted life years (QALYs) for use in cost-utility analyses and also allow comparisons among different interventions and health conditions. Developed by the EuroQol Group [8], the EQ-5D [9] is one of the most commonly used generic preference-based measures in clinical trials and cost-utility analyses [10–13]. It also has been routinely applied in THR/TKR programs in the UK [14], Sweden [15], and Canada (Alberta) [16]. There are two versions of the EQ-5D: the 3L and the 5L. For routine outcome measurement among the THR/TKR patients, the 3L version is used in the UK and Sweden. In Alberta, the 3L version was used initially, but has recently, and gradually, been replaced by the 5L version.

The 5L version of the EQ-5D was developed to enhance the measurement properties of the instrument, since the 3L version has important limitations including high ceiling effect and limited sensitivity to change [17–20]. The available evidence on measurement properties of the 3L versus the 5L among patients following THR/TKR is mainly based on cross-sectional comparisons; however, longitudinal comparisons between the 3L and 5L versions are lacking. Greene et al. [21] and Conner-Spady et al. [17] have shown the 5L version is superior to the 3L in terms of construct validity in patients undergoing THR/TKR. Two recent assessments of longitudinal measurement properties have focused on the 5L version. Conner-Spady et al. [22] reported that the EQ-5D-5L has appropriate responsiveness in THR/TKR patients; however, they only evaluated the responsiveness for the 5L version among a relatively small sample ( $n = 269$  THR,  $n = 268$  TKR) over a 12-month period. Since other factors (e.g., rehabilitation, development of other symptoms/conditions) may occur in such a relatively long term, the change in HRQL may not be the real surgery-related change, affecting the responsiveness evaluation results. Bilbao and colleagues' study [23] supported the reliability, validity, and responsiveness (over a 6-month period) of the EQ-5D-5L in patients with hip or knee osteoarthritis; however, since only 20.2% patients received joint replacement during the follow-up, their findings may not be appropriate to generalize to the THR/TKR patient population.

Our objective was to examine and compare the responsiveness of the EQ-5D-3L and EQ-5D-5L over a 3-month time period in patients following THR/TKR.

## Methods

### Sample and materials

This is a retrospective cohort analysis using data from the Alberta Bone and Joint Health Data Repository. This data repository is operated by the Alberta Bone and Joint Health Institute and was launched in 2010 [24]. The EQ-5D (3L or 5L) and the Western Ontario and McMaster University Osteoarthritis Index (WOMAC) have been routinely administered to all THR/TKR patients before surgery, 3 months, and 1 year after surgery at 13 hospitals in Alberta, Canada. Typically, patients complete the EQ-5D and WOMAC during the first clinical assessment; otherwise, patients could complete the measures before the first clinic visit through a hyperlink included in the appointment confirmation email, or after the clinic visit through a link in a follow-up email. For those who do not complete the two measures during, before, or after the first clinical assessment, their pre-surgery measurement could be done during, before, or after one of the subsequent clinic visits. Since patients' HRQL might change over time between the first clinical assessment and the surgery, and given our intention to examine whether the EQ-5D can capture pre- and post-surgery change in HRQL, a clear definition of baseline measurement was necessary. Therefore, we defined the baseline measurement as those pre-surgery measurements completed between 14 days before, and 30 days after, the first clinical assessment. Patients were included in this analysis if they: (1) had a primary THR or TKR; (2) completed the EQ-5D and WOMAC in the time period from 14 days before to 30 days after the first assessment; and (3) completed the WOMAC and the same version of EQ-5D at baseline and 3-month follow-up. All clinics used the EQ-5D-3L between 2010 and 2012, and gradually shifted to using the EQ-5D-5L between 2013 and 2016, after 2016, all clinics have used the EQ-5D-5L.

In the present analysis, we only employed data from baseline and 3-month follow-up after surgery and did not use the 1-year follow data, because changes in HRQL caused by THR/TKR usually happen during the first three months after the surgery, after that, HRQL change will be mainly impacted by many other factors (e.g., aging, comorbidities, rehabilitation, and so on). Therefore, 3-month is an appropriate duration to assess the responsiveness of EQ-5D in patients following THR/TKR.

## Measures

The EQ-5D has 5 dimensions: mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD). In the 3L version (the initial version was introduced in 1990 by the EuroQol group), each dimension has three severity levels: “no problems” (level 1), “some problems” (level 2), and “unable to” (level 3, “confined to bed” for MO) describing 243 distinct health states [25]. The 5L version has the same descriptive system including the five dimensions but has five severity levels for each dimension: “no problems” (level 1), “slight problems” (level 2), “moderate problems” (level 3), “severe problems” (level 4), and “extreme problems” (level 5) describing 3125 distinct health states [26]. We used the EQ-5D-3L and EQ-5D-5L Canadian value sets [27, 28] to calculate the index scores. The index scores range from -0.340 (health state “33333”) to 1 (health state “11111”) for the EQ-5D-3L [27] and range from -0.148 (health state “55555”) to 0.949 (health state “11111”) for the EQ-5D-5L [28].

The WOMAC is a commonly used, self-administered instrument to evaluate hip and knee osteoarthritis-related symptoms and functional limitations. It includes 24 items measuring pain (5 items), stiffness (2 items), and physical functioning limitations (17 items) [29]. The 5-point Likert version of the WOMAC was employed in this study. For each item, the score ranges from 0 for no symptoms or functional limitations to 4 for extreme symptoms or functional limitations. The ranges of the raw scores of the aforementioned subscales are 0–20, 0–8, and 0–68, successively. All three raw subscale scores can be linearly converted to transformed scores ranging from 0 (worst) to 100 (best). For each patient, the transformed overall score is the average of the three subscales’ transformed scores [29]. In a sensitivity analysis, the WOMAC overall score was calculated using a simpler method: converted the sum of the 24 items’ score to a 0 (worst) to 100 (best) scale [29].

Previous studies have reported that the WOMAC is responsive to change in HRQL among patients following joint replacement (or with OA) [30–32].

## Statistical analysis

Patients’ baseline characteristics including age, gender, pre-surgery risk factors, body mass index (BMI), number of weeks between the baseline and surgery, and patients’ responses to the WOMAC and EQ-5D were compared using rank-sum test (continuous variables) and Chi-square test (categorical variables) between the 3L and 5L samples. Ceiling and floor effects were explored at both health state level and dimension level for both the 3L and 5L samples. If more than 15% of the patients

reported the best or worst possible levels or health states, then ceiling or floor effects were considered presented [33].

We used the WOMAC as an anchor and categorized patients into the following nine groups according to their relative change in the transformed overall WOMAC score from baseline: large (decrease from baseline  $\geq 70\%$ ), moderate ( $70\% >$  decrease from baseline  $\geq 50\%$ ), small but important ( $50\% >$  decrease from baseline  $\geq 20\%$ ), and very mild deterioration ( $20\% >$  decrease from baseline  $> 0$ ), no change (no change in WOMAC overall score), very mild ( $0 <$  increase from baseline  $< 20\%$ ), small but important ( $20\% \leq$  increase from baseline  $< 50\%$ ), moderate ( $50\% \leq$  increase from baseline  $< 70\%$ ), and large improvement ( $70\% \leq$  increase from baseline) [29]. For those who had baseline WOMAC overall score as 0, we replaced their baseline score with 0.5. Patients who achieved at least a small but important improvement were defined as responders, and those who failed to achieve a small but important improvement were defined as non-responders. For both the 3L and 5L samples, the difference in baseline characteristics between the responder and non-responders was tested using Wilcoxon rank sum and Chi-square tests where applicable.

Since the anchor per se should measure very similar construct as the target measure under investigation [34], we assessed the correlation between the anchor, i.e., change in overall WOMAC score and change in the EQ-5D-3L and EQ-5D-5L index scores using a priori criteria as recommended by Guyatt and colleagues [35]: the correlation of the anchor with the change of the EQ-5D (3L and 5L) index scores from baseline to 3 months should be larger than 0.5, i.e., at least a moderate correlation [36]. Spearman correlation coefficients between the anchor and the two versions of EQ-5D were calculated by treating the change in WOMAC overall score as a continuous variable and a categorical variable (based on the above-defined 9 groups).

The responsiveness of the EQ-5D-3L and EQ-5D-5L for patients in each of the above-defined groups was assessed using effect size (ES, change in scores divided by standard deviation of the baseline score), standardized response mean (SRM, change in scores divided by standard deviation of the change), and Guyatt’s Responsiveness Index (GRI, change in scores divided by the standard deviation of the change for the stable patients). In our analysis, the stable patients for GRI calculation were defined as patients who had “very mild improvement” or “very mild deterioration” or remained the same (“no change”) on the WOMAC overall score. The magnitudes of these coefficients were interpreted as small (0.2–0.49), moderate (0.5–0.79), or large ( $\geq 0.8$ ) [37].

## Results

### General characteristics of participants

Between April 2010 and March 2017, a total of 1998 patients had THR and completed the WOMAC and the same version of EQ-5D at baseline and 3-month follow-up. Among these patients, 646 completed the 3L version (59.4% females, mean age 66.1 years) and 1352 completed the 5L version (58.1% females, mean age 65.3 years). During the same time period, a total of 2776 patients had TKR and completed the WOMAC and the same version of EQ-5D at baseline and 3-month follow-up. Among these patients, 948 completed the EQ-5D-3L (60.7% females, mean age 66.1 years) and 1828 patients completed the EQ-5D-5L (61.8% females, mean age 66.5 years). These patients were included in our analysis (Fig. 1). In both the THR and TKR samples, there were no statistically significant differences between the 3L and 5L samples in terms of age, sex, or pre-surgery body mass index (BMI) (Table 1). Patients who completed the EQ-5D-5L had longer time interval between the baseline and the surgery than those who completed the EQ-5D-3L. The mean baseline WOMAC overall scores and EQ-5D index scores for THR patients (WOMAC 36.4, EQ-5D-3L 0.52, EQ-5D-5L 0.38) were lower than that for the TKR patients (WOMAC 41.5, EQ-5D-3L 0.57, EQ-5D-5L 0.47) (Table 1). Ceiling and floor effects were not considered presented (all < 0.3%) for both versions at the health state level at baseline, while the EQ-5D-3L consistently presented a larger ceiling effect on SC and A/D dimensions and larger floor effect on UA and P/D dimensions than the EQ-5D-5L. Ceiling and floor effects for the MO dimension

were not presented for both versions (all < 5.0%); however, for the EQ-5D-3L, more than 94% patients selected level 2 (Table 1). For patients following TKR in both the 5L and 3L samples, there was no significant difference in terms of age, gender, pre-surgery BMI, and number of pre-surgery risk factors between the responders and non-responders, except for the non-responders were younger than the responders in the 5L sample. Similarly, for the patients following THR, there was no significant difference between responders and non-responders, except for the non-responders had more pre-surgery risk factors than the responders in the 3L sample, and had lower BMI in the 5L sample. The responders consistently had lower EQ-5D index scores at baseline than non-responders across all samples (see Supplementary Table 1).

### Correlation between anchor and EQ-5D

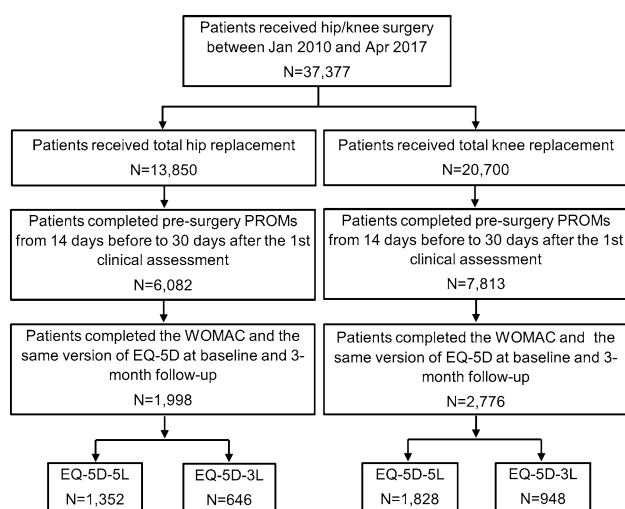
For both versions of the EQ-5D in both the THR and TKR samples, the correlations between the change in EQ-5D index score and change in the WOMAC overall score (as either continuous or categorical variables) met the a priori criteria (Table 2).

### Responsiveness

For the THR and TKR patients in the “small but important improvement” groups, the EQ-5D-5L (ES, THR 0.90, TKR 0.95; SRM, THR 1.20, TKR 1.13; GRI, THR 0.95, TKR 1.06) was consistently more responsive than the EQ-5D-3L (ES, THR 0.64, TKR 0.69; SRM, THR 0.69, TKR 0.73; GRI, THR 0.75, TKR 0.73). According to the pre-defined classification of magnitude of coefficients, for both versions, the three coefficients were “small” for the “very mild improvement” groups in both the THR and TKR samples (Table 3). The EQ-5D-5L consistently had larger ES, SRM, and GRI than the EQ-5D-3L for “moderate improvement” and “large improvement” groups in both the THR and TKR samples (Table 3).

### Sensitivity analysis

We explored defining the WOMAC overall score using a simple approach in the sensitivity analysis. More than 98% of patients were categorized into the same subgroup defined by the relative change in WOMAC overall score in the base-case. Correlations between the EQ-5D and the anchor, and value of the responsiveness statistics were extremely similar to the results from the base-case analysis (Supplementary Tables 2 and 3).



**Fig. 1** Flowchart for including patients for analysis. WOMAC Western Ontario & McMaster Universities Osteoarthritis (Index)

**Table 1** Patient baseline characteristics

	Hip			Knee		
	EQ-5D-5L N=1352	EQ-5D-3L N=646	<i>p</i> value*	EQ-5D-5L N=1828	EQ-5D-3L N=948	<i>p</i> value*
Age (years)	65.3 (11.6)	66.1 (11.1)	0.269	66.5 (9.0)	66.1 (9.4)	0.237
Female [ <i>n</i> (%)]	785 (58.1)	384 (59.4)	0.561	1129 (61.8)	575 (60.7)	0.563
Number of pre-surgery risk factors	1.8 (1.5)	2.1 (1.7)	0.004	2.0 (1.6)	2.5 (1.9)	< 0.001
Pre-surgery BMI (kg/m <sup>2</sup> )	29.9 (5.6)	30.0 (6.4)	0.332	32.9 (6.4)	33.1 (6.9)	0.946
Baseline to surgery (weeks)	14.6 (9.9)	11.7 (8.4)	< 0.001	17.7 (11.6)	13.6 (9.3)	< 0.001
<b>WOMAC</b>						
Overall	35.4 (16.5)	37.1 (15.9)	0.024	40.0 (15.7)	42.9 (16.4)	< 0.001
Pain	37.3 (18.0)	38.7 (16.9)	0.053	40.8 (17.1)	43.2 (18.0)	0.003
Stiffness	35.3 (19.9)	36.8 (19.2)	0.039	37.3 (19.3)	41.2 (20.3)	< 0.001
Function limitation	34.9 (17.0)	36.6 (16.7)	0.026	40.1 (16.5)	43.0 (16.9)	< 0.001
<b>EQ-5D</b>						
Index score	0.38 (0.24)	0.52 (0.19)	< 0.001	0.47 (0.23)	0.57 (0.18)	< 0.001
Health state “11111” [ <i>n</i> (%)]	1 (0.1)	0 (0)		3 (0.2)	3 (0.3)	
Health state “55555/33333” [ <i>n</i> (%)]	3 (0.2)	1 (0.1)		1 (0.1)	0 (0)	
MO level 1 [ <i>n</i> (%)]	18 (1.3)	27 (4.2)		43 (2.4)	47 (5.0)	
MO level 5/3 [ <i>n</i> (%)]	59 (4.4)	4 (0.6)		25 (1.4)	8 (0.8)	
SC level 1 [ <i>n</i> (%)]	265 (19.6)	327 (50.6)		798 (43.7)	710 (74.9)	
SC level 5/3 [ <i>n</i> (%)]	12 (0.9)	6 (0.9)		9 (0.5)	11 (1.2)	
UA level 1 [ <i>n</i> (%)]	11 (0.8)	35 (5.4)		68 (3.7)	81 (8.5)	
UA level 5/3 [ <i>n</i> (%)]	190 (14.1)	158 (24.5)		99 (5.4)	115 (12.1)	
P/D level 1 [ <i>n</i> (%)]	2 (0.2)	1 (0.1)		9 (0.5)	6 (0.6)	
P/D level 5/3 [ <i>n</i> (%)]	251 (18.6)	310 (48.0)		206 (11.3)	378 (39.9)	
A/D level 1 [ <i>n</i> (%)]	408 (30.2)	293 (45.4)		663 (36.3)	540 (57.0)	
A/D level 5/3 [ <i>n</i> (%)]	47 (3.5)	32 (5.0)		37 (2.0)	45 (4.8)	

Number and brackets present mean (standard deviation) unless otherwise stated

*BMI* body mass index, *WOMAC* Western Ontario and McMaster Universities Osteoarthritis Index, *MO* mobility, *SC* self-care, *UA* usual activities, *P/D* pain/discomfort, *A/D* anxiety/depression

\*Rank-sum test for continuous variables and Chi-square test for categorical variables

**Table 2** Spearman correlation coefficients between the change in EQ-5D and WOMAC (anchor)

	Hip		Knee	
	EQ-5D-5L N=1352	EQ-5D-3L N=646	EQ-5D-5L N=1828	EQ-5D-3L N=948
Treating the change in WOMAC overall score as a categorical variable	0.56	0.56	0.55	0.54
Treating the change in WOMAC overall score as a continuous variable	0.71	0.68	0.60	0.57

*WOMAC* Western Ontario and McMaster Universities Osteoarthritis Index

## Discussion

Our study demonstrated that both versions of the EQ-5D were responsive among patients following the THR or TKR; however, the EQ-5D-5L had better capabilities in

picking-up changes in HRQL than the EQ-5D-3L.

Our finding that the EQ-5D-5L was responsive to change in HRQL of patients following the THR or TKR is in line with a recently published study [22]. The responsiveness of the EQ-5D-5L had also been examined and confirmed in patients with stroke [38, 39], adolescent idiopathic scoliosis

**Table 3** Change in EQ-5D index scores and coefficients of effect size

	No change or deterioration			Improvement			No change or deterioration			Improvement		
	Very mild	Small but important	Moderate	Large	Very mild	Small but important	Moderate	Large	Very mild	Small but important	Moderate	Large
<b>EQ-5D-5L (hip)</b>												
<i>N</i> (%)	27 (2.0)	131 (9.7)	125 (9.3)	1018 (75.3)	78 (4.3)	281 (15.4)	243 (13.3)	1100 (60.2)	126 (6.9)	0.58 (0.22)	0.56 (0.19)	0.39 (0.22)
Baseline	0.48 (0.26)	0.61 (0.20)	0.55 (0.20)	0.32 (0.22)	0.58 (0.22)	0.58 (0.19)	0.56 (0.19)	0.39 (0.22)	0.62 (0.18)	0.70 (0.18)	0.80 (0.13)	0.80 (0.15)
3-month	0.43 (0.23)	0.79 (0.15)	0.80 (0.15)	0.82 (0.14)	0.53 (0.24)	0.76 (0.15)	0.80 (0.13)	0.80 (0.15)	0.70 (0.18)	0.76 (0.15)	0.24 (0.18)	0.41 (0.22)
Δ	-0.05 (0.25)	0.18 (0.15)	0.25 (0.17)	0.50 (0.22)	-0.05 (0.21)	0.19 (0.16)	0.24 (0.18)	0.41 (0.22)	0.08 (0.16)	0.19 (0.16)	1.26	1.86
ES	0.27	0.90	1.25	2.27		0.95	1.26	1.86	0.44	0.95	1.07, 1.46	1.76, 1.96
95% CI	-0.12, 0.66	0.64, 1.15	0.98, 1.52	2.16, 2.38		0.77, 1.12	1.07, 1.46	1.76, 1.96	0.19, 0.69	0.77, 1.12	1.33	1.86
SRM	0.33	1.20	1.47	2.27		1.13	1.33	1.86	0.50	1.13	1.13, 1.53	1.76, 1.96
95% CI	-0.06, 0.72	0.93, 1.46	1.19, 1.75	2.16, 2.38		0.95, 1.30	1.13, 1.53	1.76, 1.96	0.25, 0.75	0.95, 1.30	1.41	2.41
GRI	0.32	0.95	1.32	2.63		1.06	1.41	2.41	0.47	1.06	1.21, 1.61	2.30, 2.52
95% CI	-0.08, 0.70	0.69, 1.20	1.04, 1.58	2.51, 2.75		0.88, 1.23	1.21, 1.61	2.30, 2.52	0.22, 0.72	0.88, 1.23		
<b>EQ-5D-3L (knee)</b>												
<i>N</i> (%)	19 (2.9)	77 (11.9)	75 (11.6)	454 (70.3)	56 (5.9)	205 (21.6)	133 (14.0)	492 (51.9)	62 (6.5)	0.65 (0.18)	0.63 (0.15)	0.51 (0.18)
Baseline	0.61 (0.14)	0.66 (0.14)	0.62 (0.15)	0.47 (0.19)	0.65 (0.18)	0.64 (0.16)	0.63 (0.15)	0.51 (0.18)	0.68 (0.13)	0.74 (0.10)	0.77 (0.13)	0.80 (0.13)
3-month	0.55 (0.20)	0.75 (0.16)	0.77 (0.14)	0.80 (0.14)	0.62 (0.19)	0.75 (0.17)	0.77 (0.13)	0.80 (0.13)	0.74 (0.10)	0.75 (0.17)	0.14 (0.15)	0.29 (0.19)
Δ	-0.06 (0.11)	0.09 (0.13)	0.15 (0.13)	0.34 (0.20)	-0.03 (0.17)	0.11 (0.15)	0.14 (0.15)	0.29 (0.19)	0.06 (0.13)	0.11 (0.15)	0.93	1.61
ES	0.27	0.64	1.00	1.74		0.69	0.93	1.61	0.46	0.69	0.68, 1.18	1.47, 1.75
95% CI	-0.34, 0.88	0.32, 0.96	0.66, 1.33	1.58, 1.89		0.49, 0.89	0.68, 1.18	1.47, 1.75	0.10, 0.82	0.49, 0.89	0.93	1.53
SRM	0.23	0.69	1.15	1.65		0.73	0.93	1.53	0.46	0.73	0.68, 1.18	1.38, 1.67
95% CI	-0.38, 0.83	0.36, 1.01	0.80, 1.49	1.50, 1.80		0.53, 0.93	0.68, 1.18	1.38, 1.67	0.10, 0.82	0.53, 0.93	0.93	1.93
GRI	0.25	0.75	1.25	2.75		0.73	0.93	1.93	0.40	0.73	0.68, 1.18	1.38, 1.67
95% CI	-0.36, 0.85	0.42, 1.07	0.89, 1.59	2.57, 2.93		0.53, 0.93	0.68, 1.18	1.38, 1.67	0.04, 0.75	0.53, 0.93	0.68, 1.18	1.78, 2.08

The change categories were defined according to the change in WOMAC overall score, deterioration and no change were grouped together due to small sample size in each category

Number and brackets present mean (standard deviation), unless other statements

CI confidence interval, ES effect size, GRI Guyatt's responsiveness index, SRM standardized response mean, WOMAC Western Ontario and McMaster Universities Osteoarthritis (Index), Δ changes score (3-month minus baseline)

[40], chronic obstructive pulmonary disease [41], and meningitis [42]. The small ceiling and floor effects of the EQ-5D-5L at the health state level were in line with findings from studies in similar populations [17, 23].

However, there are very few studies that have compared the longitudinal measurement properties of the two versions of EQ-5D. The finding that the 5L version was more responsive than the 3L in our study was inconsistent with results from a similar comparison in patients with stroke that found that the EQ-5D-3L was more responsive than the EQ-5D-5L [38]. This difference may be due to different anchors used to evaluate the responsiveness, and the different patient populations and clinical context of these health conditions. Moreover, Golicki et al. used an interim Polish EQ-5D-5L value set which was estimated using the cross-walk approach [38]; however, in our study, we used the Canadian 5L value set estimated using the time trade-off method. The differences in the value sets and the approaches used to generate the value set might impact the assessment of responsiveness of the measures.

For patients following THR/TKR, the MO dimension is highly impacted and considerable changes are typically expected in this dimension after surgery. One of the likely reasons that the 5L was more responsive than the 3L among this population would be the difference in the wording of the most severe level of the MO dimension between the two versions. The most severe level of the 3L MO dimension is “confined to bed (level 3)”, and the most severe level of the 5L MO dimension is “extreme problems (level 5)”. Previous cross-sectional studies demonstrated that for the 3L version, most patients report levels 1 and 2, with almost none in level 3 [43, 44]; for the 5L, the distribution of responses was more even across the levels. This more even distribution allows more room for change. More specifically, for the 3L MO dimension, most of the improvements after surgery can only happen from level 2 to level 1, since the level 3 is barely used; but for the 5L MO dimension, improvements can start from any level (except level 1). So, the modification from 3 levels to 5 levels in the EQ-5D, and also changing the wording of the most severe level from “confined to bed” to “extreme problems” impacted the measurement of mobility, which is highly relevant in this THR/TKR patient population. In addition, compared to the EQ-5D-3L, the EQ-5D-5L consistently presented lower ceiling/floor effect at SC, UA, P/D and A/D dimensions, which suggested that the 5L dimensions had better discriminate ability. This can also explain the superiority of the EQ-5D-5L in responsiveness.

The WOMAC user’s guide suggested a meaningful change in HRQL using the following criteria [29]: 1) 20/50/70% improvement in the WOMAC pain subscale, or 2) 20/50/70% improvement in WOMAC pain and 20/50/70% improvement in WOMAC stiffness or WOMAC physical function subscales. In order to maximize the conceptual

overlap between the anchor and the EQ-5D, we used the 20/50/70% (increase or decrease) of baseline WOMAC overall score to categorize patients’ HRQL change from the baseline. Our approach captured the HRQL change in all three aspects (pain, stiffness and physical function). However, since we used relatively strict criteria compared to the WOMAC user’s guide, our study may have overestimated the responsiveness of the EQ-5D relative to the usual interpretation of the WOMAC. In addition, we only used WOMAC as the anchor in this study. Benson et al. [45] reported that the EQ-5D-3L has moderate (Pearson’s  $r > 0.5$ ) correlation with howRu and Oxford hip and knee scores in patients undergoing THR/TKR, those measures can be explored as anchors in further study.

Our study added evidence to the comparison of longitudinal measurement properties between the EQ-5D-3L and the EQ-5D-5L. This is important for orthopedic researchers and end-users when making decisions about which version to use to measure change in HRQL over time in this patient population. This evidence can also support switching the routine application of the EQ-5D-3L in the THR and TKR fields to using the EQ-5D-5L. Moving from the 3L version to the 5L version in this particular setting (Alberta, Canada) would potentially lead to an increase in QALY gained from the surgery in cost-utility analysis.

Our study had several limitations to be considered. In the Alberta Hip and Knee Replacement Project, patients only completed either the 3L or the 5L version of EQ-5D, so we were not able to conduct a head-to-head comparison. The 3L and 5L samples used in our analysis were comparable at main demographics; however, the 5L sample consistently had lower WOMAC scores than the EQ-5D-3L. Even though the mean differences between the 3L and 5L were small (1.4–3.9 on a 0–100 scale), the bias from the potential heterogeneity in baseline quality of life may still affect the results. In addition, due to the shifting from the 3L to the 5L, patients who completed the EQ-5D-3L received the surgery earlier than those completed the EQ-5D-5L. Since waiting time in Alberta for THR/TKR has been getting longer in the recent years [46], the time interval between the baseline and surgery was longer for the EQ-5D-5L samples, which may cause bias. Further head-to-head comparisons on the longitudinal measurement properties between the 3L and 5L versions of EQ-5D might add more valuable information to this field. In our data set, for those who had completely WOMAC and EQ-5D records at baseline, only about 1/3 completed the WOMAC and the same version of the EQ-5D at 3-month follow-up. We did not apply any imputations and compare those who had or did not have 3-month records in the present study. This was because of the complicated reasons for losing samples, which includes the switch from the 3L to 5L, logistic issues at each clinic, and patients’ readmissions and joint repairs (end of follow-ups of the initial THR/TKR). Losing such an amount of patients may

bring bias; however, we were not able to estimate the direction and magnitude to inform the generalizability of this study by using our current data set.

We used Cohen's criteria [37] to interpret the magnitude of the SRM; however, some researchers reported that this would lead to over- or under-estimation of changes [47, 48]. We recognized the samples we studied had very little deterioration in health since hip/knee replacement surgery is generally very effective. Therefore, the estimations of effect sizes for deterioration groups were less robust because of the much smaller sample sizes for these groups. Finally, we did not estimate a minimally important difference (MID) for the EQ-5D in this analysis since more than 90% of the THR and TKR patients were categorized as having a "small but important" improvement or above in this data set. We used the recommended anchor-based statistical analysis approaches, such as receiver operating characteristic curve [35], to estimate MIDs for the EQ-5D in our analysis; however, the aforementioned unbalanced samples seemed to introduce bias. Our efforts in solving this issue are still ongoing.

## Conclusion

Our study demonstrated that the EQ-5D-5L is more responsive than the EQ-5D-3L in measuring changes in HRQL in patients following THR/TKR, relative to the WOMAC as a clinical anchor. When choosing a generic preference-based measure, we recommend using the EQ-5D-5L in longitudinal studies in this patient population. Further research comparing the responsiveness of the EQ-5D-3L and EQ-5D-5L against other anchors and in other populations is imperative to confirm these findings.

**Acknowledgments** We thank EuroQol Research Foundation for supporting this study.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** The data used in this analysis was collected as part of standard of care, and not part of a clinical study. As such, it was captured under the authority of the provincial Privacy Impact Analysis in place for quality assurance monitoring and reporting on Bone and Joint Health in Alberta (OIPC File # H2801). Participants do not provide consent.

## References

- American Academy of the Orthopedic Surgeons. Total joint replacement. Accessed August 31, 2018, from <https://orthoinfo.aaos.org/en/treatment/total-joint-replacement/>.
- OECD. (2017). *Health at a glance 2017: OECD indicators*. Paris: OECD Publishing.
- Canadian Institute for Health Information. (2018). *Hip and knee replacements in Canada, 2016–2017: Canadian joint replacement registry annual report*. Ottawa: CIHI.
- Guyatt, G. H., Deyo, R. A., Charlson, M., Levine, M. N., & Mitchell, A. (1989). Responsiveness and validity in health status measurement: A clarification. *Journal of Clinical Epidemiology*, *42*(5), 403–408.
- Garbuz, D. S., Xu, M., & Sayre, E. C. (2006). Patients' outcome after total hip arthroplasty. *Journal of Arthroplasty*, *21*(7), 998–1004.
- Giesinger, J. M., Hamilton, D. F., Jost, B., Behrend, H., & Giesinger, K. (2015). WOMAC, EQ-5D and knee society score thresholds for treatment success after total knee arthroplasty. *Journal of Arthroplasty*, *30*(12), 2154–2158.
- Núñez, M., Lozano, L., Núñez, E., et al. (2009). Total knee replacement and health-related quality of life: Factors influencing long-term outcomes. *Arthritis & Rheumatology*, *61*(8), 1062–1069.
- EuroQol Group. EuroQol, About us. Accessed March 1, 2019, from <https://euroqol.org/euroqol/>.
- Rabin, R., Gudex, C., Selai, C., & Herdman, M. (2014). From translation to version management: A history and review of methods for the cultural adaptation of the EuroQol five-dimensional questionnaire. *Value in Health*, *17*(1), 70–76.
- Costa, M. L., Achten, J., Foguet, P., Parsons, N. R., & Young Adult Hip Arthroplasty Team. (2018). Comparison of hip function and quality of life of total hip arthroplasty and resurfacing arthroplasty in the treatment of young patients with arthritis of the hip joint at 5 years. *British Medical Journal Open*, *8*(3), e018849.
- Fernandes, L., Roos, E. M., Overgaard, S., Villadsen, A., & Sogaard, R. (2017). Supervised neuromuscular exercise prior to hip and knee replacement: 12-month clinical effect and cost-utility analysis alongside a randomised controlled trial. *BMC Musculoskeletal Disorders*, *18*(1), 5.
- 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. *Bone & Joint Journal*, *95*(1), 115–121.
- Rosenlund, S., Broeng, L., Holsgaard-Larsen, A., Jensen, C., & Overgaard, S. (2017). Patient-reported outcome after total hip arthroplasty: Comparison between lateral and posterior approach. *Acta Orthopaedica*, *88*(3), 239–247.
- Insight & Feedback Team, NHS England. (2017). *National patient reported outcome measures (PROMs) programme consultation report*. London: NHS.
- Kärrholm J, Lindahl H, Malchau H, et al. The Swedish hip arthroplasty register annual report 2016. Accessed August 31, 2018, from <https://registercentrum.blob.core.windows.net/shpr/t/Annual-Report-2016-B1eWEH-mHM.pdf>.
- Canadian Institute for Health Information. (2015). *PROMs forum proceedings*. Ottawa: CIHI.
- Conner-Spady, B., Marshall, D., Bohm, E., 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*(7), 1775–1784.
- Brazier, J., Roberts, J., Tsuchiya, A., & Busschbach, J. (2004). A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics*, *13*(9), 873–884.
- Johnson, J. A., & Pickard, A. S. (2000). Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada. *Medical Care*, *38*(1), 115–121.
- Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parikin, 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(10), 1727–1736.
21. Greene, M., Rader, K., Garellick, G., Malchau, H., Freiberg, A., & Rolfson, O. (2015). The EQ-5D-5L improves on the EQ-5D-3L for health-related quality-of-life assessment in patients undergoing total hip arthroplasty. *Clinical Orthopaedics and Related Research*, 473(11), 3383–3390.
  22. Conner-Spady, B., Marshall, D., Bohm, E., Dunbar, M., & Noseworthy, T. (2018). 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. *Quality of Life Research*, 27(5), 1311–1322.
  23. Bilbao, A., García-Pérez, L., Arenaza, J. C., García, I., Ariza-Cardiel, G., Trujillo-Martín, E., et al. (2018). Psychometric properties of the EQ-5D-5L in patients with hip or knee osteoarthritis: Reliability, validity and responsiveness. *Quality of Life Research*, 27(11), 2897–2908.
  24. Marshall, D. A., Christiansen, T., Smith, C., et al. (2015). Continuous quality improvement program for hip and knee replacement. *American Journal of Medical Quality*, 30(5), 425–431.
  25. EuroQol Group. EQ-5D-3L user guide. Accessed May 11, 2018, from [https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-3L\\_UserGuide\\_2015.pdf](https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-3L_UserGuide_2015.pdf).
  26. EuroQol Group. EQ-5D-5L user guide. Accessed May 11, 2018, from [https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L\\_UserGuide\\_2015.pdf](https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L_UserGuide_2015.pdf).
  27. Bansback, N., Tsuchiya, A., Brazier, J., & Anis, A. (2012). Canadian valuation of EQ-5D health states: Preliminary value set and considerations for future valuation studies. *PLoS ONE*, 7(2), e31115.
  28. Xie, F., Pullenayegum, E., Gaebel, K., et al. (2016). A time trade-off-derived value set of the EQ-5D-5L for Canada. *Medical Care*, 54(1), 98–105.
  29. Bellamy, N. (2004). *WOMAC osteoarthritis index user guide VII*. Brisbane: The University of Queensland.
  30. Impellizzeri, F. M., Mannion, A. F., Leunig, M., Bizzini, M., & Naal, F. D. (2011). Comparison of the reliability, responsiveness, and construct validity of 4 different questionnaires for evaluating outcomes after total knee arthroplasty. *Journal of Arthroplasty*, 26(6), 861–869.
  31. 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 and Cartilage*, 13(12), 1076–1083.
  32. Wolfe, F., & Kong, S. X. (1999). Rasch analysis of the Western Ontario McMaster questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. *Annals of the Rheumatic Diseases*, 58(9), 563–568.
  33. Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, 60(1), 34–42.
  34. Sonntag, M., König, H. H., & Konnopka, A. (2015). The responsiveness of the EQ-5D and time trade-off scores in schizophrenia, affective disorders, and alcohol addiction. *Health and Quality of Life Outcomes*, 13, 114.
  35. Turner, D., Schünemann, H. J., Griffith, L. E., et al. (2009). Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. *Journal of Clinical Epidemiology*, 62(4), 374–379.
  36. Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3), 69–71.
  37. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge Academic.
  38. Golicki, D., Niewada, M., Karlińska, A., et al. (2015). Comparing responsiveness of the EQ-5D-5L, EQ-5D-3L and EQ VAS in stroke patients. *Quality of Life Research*, 24(6), 1555–1563.
  39. Chen, P., Lin, K., Liang, R., Wu, C., Chen, C., & Chang, K. (2016). Validity, responsiveness, and minimal clinically important difference of EQ-5D-5L in stroke patients undergoing rehabilitation. *Quality of Life Research*, 25(6), 1585–1596.
  40. Cheung, P., Wong, C., Lau, S., & Cheung, J. (2018). Responsiveness of the EuroQoL 5-dimension (EQ-5D) in adolescent idiopathic scoliosis. *European Spine Journal*, 27(2), 278–285.
  41. Nolan, C. M., Longworth, L., Lord, J., et al. (2016). The EQ-5D-5L health status questionnaire in COPD: Validity, responsiveness and minimum important difference. *Thorax*, 71(6), 493–500.
  42. Bhadhuri, A., Jowett, S., Jolly, K., & Al-Janabi, H. (2017). A comparison of the validity and responsiveness of the EQ-5D-5L and SF-6D for measuring health spillovers: A study of the family impact of meningitis. *Medical Decision Making*, 37(8), 882–893.
  43. Oppe, M., Devlin, N., & Black, N. (2011). Comparison of the underlying constructs of the EQ-5D and Oxford Hip Score: Implications for mapping. *Value in Health*, 14(6), 884–891.
  44. Jin, X., Al Sayah, F., Ohinmaa, A., Marshall, D. A., Smith, C., & Johnson, J. A. (2019). The EQ-5D-5L is superior to the -3L version in measuring health-related quality of life in patients awaiting THA or TKA. *Clinical Orthopaedics and Related Research*. <https://doi.org/10.1097/CORR.0000000000000662>.
  45. 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(1), 512.
  46. Alberta Bone and Joint Health Institute. *Wait times for hip and knee replacement*. Accessed January 8, 2019, from <https://albertaboneandjoint.com/patients/wait-times/>.
  47. Middel, B., & van Sonderen, E. (2002). Statistical significant change versus relevant or important change in (quasi) experimental design: Some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. *International Journal of Integrated Care*, 2(4), e15.
  48. Sivan, M. (2009). Interpreting effect size to estimate responsiveness of outcome measures. *Stroke*, 40(12), e709.