Patient-reported outcome in total hip replacement

A COMPARISON OF FIVE INSTRUMENTS OF HEALTH STATUS

M. Ostendorf,
H. F. van Stel,
E. Buskens,
A. J. P. Schrijvers,
L. N. Marting,
A. J. Verbout,
W. J. A. Dhert

From the University Medical Center, Utrecht, The Netherlands

Our aim was to define the minimum set of patient-reported outcome measures which are required to assess health status after total hip replacement (THR). In 114 patients, we compared the pre-operative characteristics and sensitivity to change of the Oxford hip score (OHS), the Western Ontario and McMaster Universities osteoarthritis index (WOMAC), the SF-36, the SF-12 (derived from the SF-36), and the Euroqol questionnaire (EQ-5D).

At one year after operation, very large effect sizes were found for the disease-specific measures, the physical domains of the SF-12, SF-36 and the EQ-5D$_{index}$ (1.3 to 3.0). Patients in Charnley class A showed more change in the OHS, WOMAC pain and function, the physical domains of the SF-36 and the EQ-5D$_{vas}$ ($p < 0.05$) compared with those in the Charnley B and C group. In this group, the effect size for the OHS more than doubled the effect sizes of WOMAC pain and physical function. We found high correlations and correlations of change between the OHS, the WOMAC, the physical domains of the SF-12 and the SF-36 and EQ-5D$_{index}$. The SF-36 and EQ-5D scores at one year after operation approached those of the general population. Furthermore, we found a binomial distribution of the pre-operative EQ-5D$_{index}$ score and a pre-operative discrepancy and post-operative agreement between the EQ-5D$_{vas}$ and EQ-5D$_{index}$.

We recommend the use of the OHS and SF-12 in the assessment of THR. The SF-36 may be used in circumstances when smaller changes in health status are investigated, for example in the follow-up of THR. The EQ-5D is useful in situations in which utility values are needed in order to calculate cost-effectiveness or quality-adjusted life years (QALYs), such as in the assessment of new techniques in THR.

■ M. Ostendorf, MSc, PhD, Student
■ A. J. Verbout, MD, PhD, Professor, Orthopaedic Surgeon
■ W. J. A. Dhert, MD, PhD, Director of Orthopaedic Research
Department of Orthopaedics, University Medical Center, P O Box 85500, 3508 GA, Utrecht, The Netherlands.

■ H. F. van Stel, PhD, Clinical Epidemiologist
■ E. Buskens, MD, PhD, Clinical Epidemiologist
■ A. J. P. Schrijvers, PhD, Professor of Health Services
Julius Centre for Health Sciences and Primary Care, University Medical Center, PO Box 85500, 3508 GA, Utrecht, The Netherlands.

■ L. N. Marting, MD
Department of Orthopaedics, Sint Antonius Hospital, PO Box 2500, 3430 Nieuwegein, The Netherlands.

Correspondence should be sent to Dr M. Ostendorf.

doi:10.1302/0301-620X.86B6.14950 $2.00

J Bone Joint Surg [Br] 2004;86-B:801-8.
Received 11 September 2003; Accepted 4 November 2003

Total hip replacement (THR) is an effective treatment which improves function and relieves pain in the hip secondary to severe osteoarthritis or other diseases which affect the joint.[1] In an environment of limited resources, health-status questionnaires are of particular importance when comparing the cost-to-benefit ratio of medical interventions.[2] In the evaluation of THR, the Western Ontario and McMaster Universities osteoarthritis index (WOMAC) and the SF-36 health status questionnaire (SF-36) have been especially recommended as patient-centred questionnaires.[3] However, the 12-item Oxford hip score (OHS) and the Euroqol questionnaire have also been suggested, the first because of its brevity and site-specificity and the second because it allows comparisons to be made with the effect of other health-care interventions.[4,5] The psychometric characteristics of these four questionnaires have been reported extensively in the literature. Our aim was to define a minimum set of outcome measures to assess health status in THR. We therefore compared the baseline characterics and sensitivity to change of the OHS, the WOMAC, the SF-36, the SF-12 health status questionnaire (SF-12) which is derived from the SF-36 and the Euroqol questionnaire (EQ-5D). We also investigated the pre- and post-operative distribution of the Euroqol scores. Finally, the SF-36 and Euroqol scores were compared with published data on age and gender-matched norms (personal communication).[6,7]

Patients and Methods

Between April 1999 and September 2000, 147 patients were recruited from three different hospitals, one university hospital and two regional hospitals. They were asked to participate in the study after their orthopaedic surgeon had put them on the waiting list for primary total hip replacement. Patients were excluded from the study if they were under the

age of 18 years, suffering from rheumatoid arthritis, were unable to fill out the questionnaires or if they had previously undergone a THR on the contralateral hip. Before and at three and 12 months after operation four questionnaires were mailed to consenting patients together with a letter with additional information about the study and an informed consent form. If necessary, missing questions were answered by telephone. To allay anxieties about discussing their care and experiences with surgery, patients were assured that their surgeons would not receive any information on patients' answers to any questions. Approval was obtained from the Ethical Committees of all the participating hospitals.

Assessment of quality of life. The health status of the patients was measured before and at three and 12 months after operation. We used two disease-specific questionnaires, the OHS and the WOMAC, and two general-health questionnaires, the SF-36 and the EQ-5D. We calculated the SF-12 scores from the answers to the SF-36 questionnaire as described by Kosinski.[8] In order to assess the effects of comorbidity on outcomes, the index of co-existent disease (ICED) was scored using data from the medical records of the patients.[9]

Oxford hip score.[10] The OHS is a disease-specific measure consisting of 12 questions which assess pain and function of the hip in relation to different activities of daily life. Each question is answered by ticking a position on a five-point ordinal scale. Responses are then totalled in order to obtain a score between 12 and 60. A low score indicates less difficulty and a high score greater difficulty. The OHS was developed specifically to assess the outcomes of hip replacement surgery and has been shown to be consistent, reproducible, valid and sensitive to clinical change.[10,11] The Dutch OHS has shown to be valid and reliable in measuring outcome in patients who have had THR.[12]

Western Ontario and McMaster Universities (WOMAC) osteoarthritis index.[13] The WOMAC is a well-known, disease-specific measure which is widely used for measuring outcome after THR.[3] Using a Likert scale, patients rate themselves on multiple items which are grouped in three domains, pain, stiffness and difficulty in function. The maximum score is 20 points for pain, 8 for stiffness and 68 points for clinical function. A low score indicates less difficulty and a high score greater difficulty.

SF-36. This is a widely used measure of general-health status.[14] The 36 items are grouped on eight scales which are designed to represent fully the WHO definition of health: physical function, role limitations due to physical problems, role limitation due to emotional problems, social function, bodily pain, vitality, mental health and perceptions of general health. Aaronson et al[6] translated the test into Dutch and tested its validity and reliability.

Euroqol. This is a standardised generic instrument for describing and assessing health-related quality of life, and identifies 243 possible health states.[15] Patients describe their own health state on five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. One of three levels of severity is chosen for each dimension, no problems, some/moderate problems or extreme problems for the dimension in question. A single, weighted utility score, the EQ-$5D_{index}$, is calculated from the five dimensions.[7] Perfect health and death have utility scores of one and zero respectively and states worse than death (< 0) are possible. Patients are also asked to value their current state of health on a thermometer scale from 0 ('worst imaginable') to 100 ('best imaginable'). We have rescaled this score, the Q-$5D_{vas}$ to, one and zero for comparison with the EQ-$5D_{index}$. The Euroqol therefore generates two overall values (utilities) for the quality of life (one from a social perspective, the EQ-$5D_{index}$ and the other from the patient's perspective, the EQ-$5D_{vas}$). It is recommended for use in combination with other more detailed generic measures such as the SF-36.[15]

SF-12. The SF-12 health survey, a short form with 12 items, is a generic measure of health status which was developed to provide an alternative to the SF-36 for the purpose of monitoring large samples from general and specific populations. The SF-12 is attractive because it is short, easy to administer, has proven reliability and validity, and can be printed on one page. It has norm-based scores, which means it has a mean of 50 and an SD of 10 in the general population.[16] We did not use the actual questionnaire in our study, but we calculated the SF-12 physical (PCS-12) and mental summary scores (MCS-12) from the responses to the SF-36 in order to compare the SF-12 scores and changes in scores with the other questionnaires.[8]

Analysis of data. Missing values for the WOMAC were approximated as long as 40% of the items had values, using the mean values of the remaining items of the subscale. Missing values for the SF-36 subscales were also approximated according to the SF-36 guidelines.[17] Because the OHS, the EQ-5D and the SF-12 are short questionnaires with 12, five and 12 items, respectively, missing values were not approximated and questionnaires with missing values were not included in the analysis. We assessed the pre- and post-operative floor and ceiling effects of all the questionnaires as a measure of the validity of their content. Patients who achieve the best possible score on a questionnaire cannot demonstrate improvement on a subsequent application of the same questionnaire even if they have improved clinically.[18] This is referred to as the ceiling effect. The floor effect is the opposite.

Statistical analysis. To assess construct validity, we calculated Spearman correlation coefficients in order to examine the relationship between instruments at the pre-operative measurement. We calculated changes in scores and effect sizes for all the measures of outcome (one-year post-operative score minus the pre-operative score). Effect sizes show the relative magnitude of a change as measured by different instruments.[19] For this study we calculated the effect sizes as the difference between the pre-operative and one-year post-operative scores divided by the SD of the pre-operative

Table I. Details of the patients before THR

	Number	Percentage
Females	71	*62.3*
Males	43	*37.7*
Diagnosis		
Primary osteoarthritis	95	*83.3*
Avascular necrosis	9	*7.9*
Post-traumatic	5	*4.4*
Dysplasia	3	*2.6*
Legg-Calvé-Perthes' disease	2	*1.8*
Charnley class		
A	97	*85.1*
B	13	*11.4*
C	4	*3.5*
Comorbidity (ICED)		
None	39	*34.2*
Mild	53	*46.5*
Moderate	19	*16.7*
Severe	3	*2.6*
Previous surgery		
None	100	*87.7*
Internal fixation	7	*6.1*
Femoral osteotomy	4	*3.5*
Femoral osteotomy and shelf procedure	2	*1.8*
Core decompression of femoral head	1	*0.9*
Prosthesis		
Cemented	103	*90.4*
Uncemented	11	*9.6*
Mean age (SD; range)	67.6 years (10.1; 35.9 to 88.8)	
Mean waiting time (SD; range)	6 months (3.6; 3 weeks to 18 months)	

Table II. Mean scores (SD), effect sizes (%) and floor and ceiling effects for OHS, WOMAC, SF-12, SF-36 subscales, EQ-$5D_{vas}$ and EQ-$5D_{index}$ before operation and one year after THR. Higher OHS and WOMAC scores indicate poorer results while higher SF-36 and EQ-5D scores show better results

	Mean score pre-operatively	Mean score at one year	Effect size (%)	Floor/ceiling effect pre-operatively (%)		Floor/ceiling effect post-operatively (%)	
OHS	42.5 (7.9)	19.0 (7.7)	*3.0*	*0.0*	*0*	*0*	*13.5*
WOMAC							
Pain	11.7 (3.5)	3.6 (4.3)	*2.3*	*0.9*	*0.0*	*0*	*36.2*
Stiffness	4.9 (1.7)	2.1 (1.7)	*1.7*	*5.6*	*0.0*	*1.0*	*16.2*
Function	42.7 (11.4)	15.6 (15.0)	*2.4*	*0.0*	*0.0*	*0*	*7.8*
SF-12 PCS	30.5 (8.3)	45.6 (9.6)	*1.8*	*0.0*	*0.0*	*0*	*0.0*
SF-12 MCS	41.4 (12.5)	49.7 (12.2)	*0.7*	*0.0*	*0.0*	*0*	*0.0*
SF-36							
Bodily pain	28.5 (16.5)	72.1 (26.3)	*2.6*	*11.2*	*0.0*	*1.9*	*32.4*
Physical function	22.0 (16.9)	60.7 (26.0)	*2.3*	*9.5*	*0.0*	*1.0*	*1.9*
Role physical	11.0 (24.0)	55.3 (47.1)	*1.9*	*78.3*	*2.8*	*33.7*	*41.3*
Role emotional	51.7 (43.5)	68.7 (43.0)	*0.4*	*37.7*	*36.8*	*24.0*	*60.6*
Mental health	73.5 (19.4)	79.4 (19.8)	*0.3*	*0.0*	*5.7*	*0*	*12.4*
Vitality	55.1 (21.3)	68.4 (21.0)	*0.6*	*0.0*	*0*	*0*	*4.8*
Social function	55.7 (29.7)	79.2 (25.7)	*0.8*	*10.3*	*12.1*	*2.9*	*46.7*
General health	63.8 (19.8)	66.7 (20.4)	*0.1*	*0.6*	*0.9*	*0*	*4.8*
EQ-$5D_{vas}$	0.59 (0.22)	0.75 (0.18)	*0.7*	*0.6*	*0*	*0*	*4.9*
EQ-$5D_{index}$	0.35 (0.31)	0.76 (0.27)	*1.3*	*0.0*	*0*	*0*	*39.8*

scores. Effect sizes of 0.2, 0.5 and 0.8 were regarded as indicating small, medium and large degrees of change respectively.[20] The Wilcoxon signed-rank and the Mann-Whitney U test were used to compare the change in scores among the different groups of patients. Spearman correlation coefficients were calculated to examine whether change in one measure corresponded with a change in other measures. Each patient was matched to his or her expected population norm for the SF-36 and Euroqol based on a five-year age interval and gender (personal communication).[6,7]

Results

Of the 147 patients, 114 had completed questionnaires at the one-year measurement. Three had died, two had a revision for infection, one had a femoral fracture, two had undergone total knee replacement and 12 had under-

Table III. Changes in the scores (SD) for the OHS, WOMAC, SF-12, SF-36 subscales, EQ-5D$_{vas}$ and EQ-5D$_{index}$ before and at 12 months after THR Charnley class A (n = 97) and Charnley class B and C patients (n = 17). The change scores and effect sizes were calculated for all the questionnaires by comparing the pre-operative scores and the scores after one year. Higher OHS and WOMAC scores indicate poorer results while higher SF-12, SF-36 and EQ-5D scores show better results

	A			B and C			
Charnley class	**Mean change between pre-operative and one year scores (SD)**	**Effect size pre-op *vs* one year**	**p value**	**Mean change pre-op *vs* 1 year (SD)**	**Effect size *vs* (SD) year**	**p value**	**p value for the difference between classes A and B/C**
OHS	24.7 (8.7)	3.1	<0.001	17.9 (11.6)	2.4	<0.001	0.026
WOMAC							
Pain	8.8 (4.3)	2.6	<0.001	4.8 (3.3)	1.1	<0.001	<0.001
Stiffness	2.9 (1.8)	1.7	<0.001	2.3 (1.5)	1.7	<0.001	0.11 (NS)*
Function	29.7 (13.3)	2.6	<0.001	15.5 (10.0)	1.4	<0.001	<0.001
SF-12 PCS	17.3 (10.6)	2.1	<0.001	8.2 (7.8)	0.9	0.002	0.002
SF-12 MCS	8.8 (11.4)	0.7	<0.001	6.4 (8.8)	0.6	0.015	0.59 (NS)
SF-36							
Bodily pain	-48.1 (23.5)	2.9	<0.001	-22.5 (18.9)	1.4	0.001	<0.001
Physical function	-42.5 (22.6)	2.5	<0.001	-18.8 (18.8)	1.1	0.003	<0.001
Role physical	-49.5 (48.7)	1.7	<0.001	-17.2 (32.6)	0.8	0.048	0.012
Role emotional	-19.1 (41.6)	0.4	<0.001	-10.4 (51.2)	0.3	0.42 (NS)	0.64 (NS)
Mental health	-5.8 (16.0)	0.3	<0.001	-8.5 (18.0)	0.5	0.08 (NS)	0.68 (NS)
Vitality	-15.0 (20.1)	0.6	<0.001	-7.3 (13.3)	0.4	0.024	0.16 (NS)
Social function	-26.2 (30.0)	0.8	<0.001	-13.3 (27.2)	0.4	0.1 (NS)	0.12 (NS)
General health	-3.8 (19.9)	0.2	0.06 (NS)	-0.3 (0.4)	0.0	0.9 (NS)	0.42 (NS)
EQ-5D$_{vas}$	-0.18 (0.23)	0.7	<0.001	-0.05 (0.23)	0.2	0.30 (NS)	0.043
EQ-5D$_{index}$	-0.43 (0.29)	1.3	<0.001	-0.30 (0.37)	1.0	0.008	0.27 (NS)

* not significant

Table IV. Spearman correlations coefficients between the pre-operative questionnaires

		WOMAC			SF-12*			
	OHS	**Pain**	**Stiffness**	**Function**	**PCS**	**MCS**	**EQ-5D$_{vas}$**	**EQ-5D$_{index}$**
WOMAC								
Pain	0.76†							
Stiffness	0.63†	0.64†						
Function	0.88†	0.75†	0.67†					
SF-12 PCS	-0.53†	-0.46†	-0.26†	-0.45†				
SF-12 MCS	-0.49†	-0.37†	-0.35†	-0.54†	0.31†			
SF-36								
Bodily pain	-0.73†	-0.70†	-0.53†	-0.71†	0.67†	0.48†	0.43†	0.68†
Physical function	-0.62†	-0.53†	-0.42†	-0.67†	0.66†	0.60†	0.32†	0.55†
Role physical	-0.39†	-0.34†	-0.20‡	-0.40†	0.53†	0.38†	0.18	0.41†
Role emotional	-0.29†	-0.20‡	-0.21‡	-0.33†	-0.01	0.79†	0.15	0.42†
Mental health	-0.31†	-0.20‡	-0.25†	-0.33†	0.06	0.82†	0.28†	0.48†
Vitality	-0.39†	-0.34†	-0.32†	-0.40†	0.27†	0.69†	0.53†	0.51†
Social function	-0.57†	-0.45†	-0.30†	-0.57†	0.48†	0.68†	0.30†	0.55†
General health	-0.23‡	-0.17	-0.16	-0.32†	0.31†	0.38†	0.44†	0.38†
EQ-5D$_{vas}$	-0.37†	-0.39†	-0.32†	-0.38†	0.38†	0.36†		
EQ-5D$_{index}$	-0.64†	-0.57†	-0.42†	-0.62†	0.42†	0.56†	0.38†	

* PCS; physical component summary score of SF-12; MCS, mental component summary score of SF-12
† $p < 0.01$ (two-tailed)
‡ $p < 0.05$ (two-tailed)

gone contralateral hip replacement. Those who had undergone a second operation after their total hip replacement were excluded from the follow-up analysis because of the effect of the second operation on the outcome scores. Thirteen patients did not return their questionnaires. We used only the results of the 114 patients with completed questionnaires in the analysis. Details of the patients are given in Table I. We compared the characteristics of patients who did not return their questionnaires with our analysed group. We found that those who did not respond at the one-year follow-up were more likely to have a post-operative complication (chi-squared test $p \leq 0.05$) and had significantly lower scores on the WOMAC pain score (3.2 *vs* 7.0, $p < 0.05$) and the SF physical functioning score (34.2 *vs* 57.0, $p < 0.05$) at three months after operation.

All the outcome scores showed a significant improvement at one year after operation ($p < 0.001$) apart from the SF-36 ($p = 0.09$; Table II). Very large effect sizes were found for the disease-specific measures and the physical domains

Table V. Spearman correlations of change (pre-operatively to one year after operation)

		WOMAC			SF-12*			
	OHS	Pain	Stiffness	Physical	PCS	MCS	EQ-5D$_{vas}$	EQ-5D$_{index}$
WOMAC								
Pain	0.68†							
Stiffness	0.48†	0.64†						
Function	0.66†	0.75†	0.61†					
SF-12 PCS	0.38†	0.44†	0.30†	0.57†				
SF-12 MCS	0.50†	0.33†	0.33†	0.51†	0.29†			
SF-36								
Bodily pain	-0.48†	-0.61†	-0.43†	-0.66†	0.62†	0.35†	0.38†	0.37†
Physical function	-0.41†	-0.53†	-0.39†	-0.72†	0.68†	0.42†	0.42†	0.41†
Role physical	-0.33†	-0.41†	-0.36†	-0.51†	0.75†	0.35†	0.35†	0.27†
Role emotional	-0.26†	-0.15	-0.20‡	-0.25†	0.04	0.71†	0.27†	0.31†
Mental health	-0.23‡	-0.10	-0.16	-0.22‡	0.05	0.72†	0.28†	0.36†
Vitality	-0.45†	-0.41†	-0.44†	-0.50†	0.30†	0.64†	0.45†	0.50†
Social function	-0.41†	-0.39†	-0.21‡	-0.49†	0.35†	0.57†	0.24‡	0.39†
General health	-0.08	-0.06	-0.08	-0.26†	0.22‡	0.29†	0.40†	0.22‡
EQ-5D$_{vas}$	-0.43†	-0.36†	-0.28†	-0.45†	0.37†	0.43†		
EQ-5D$_{index}$	-0.51†	-0.47†	-0.34†	-0.49†	0.31†	0.46†	0.45†	

* PCS, physical component summary score of SF-12; MCS, mental component summary score of SF-12
† $p < 0.01$ (two-tailed)
‡ $p < 0.05$ (two-tailed)

of the SF-12, SF-36 and the EQ-5D$_{index}$ (1.3 to 3.0). Effect sizes for the other domains of the SF-36 and EQ-5D$_{vas}$ were small to moderate (0.1 to 0.8).

We compared changes in scores for patients in Charnley class A (affected in one hip) and Charnley classes B and C (affected in both hips/other conditions directly impeding mobility)[21] (Table III). Patients in Charnley class A showed a greater change in their OHS, WOMAC pain and function, the physical domains of the SF-36 and the EQ-5D$_{vas}$. An interesting finding was that for the Charnley B and C group, the effect size for the OHS more than doubled the effect sizes of WOMAC pain and physical function. We did not find a significant correlation between outcome scores and the ICED co-morbidity score.

We calculated the pre-operative correlations and the correlations of change among the questionnaires in order to assess the construct validity of the questionnaires. There was a large number of significant correlations between the outcome scores at the pre-operative measurement. We found especially high correlations between the OHS, WOMAC, the physical domains of the SF-36 and the EQ-5D$_{index}$ (Table IV). We also found high correlations of change between the disease-specific scores and the disease-specific domains and a low correlation of the SF general

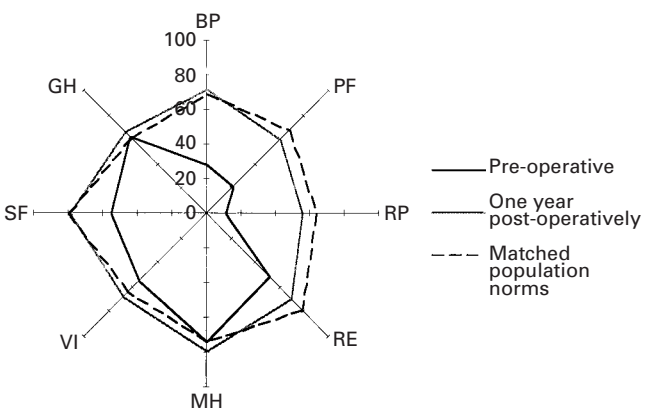

Fig. 1

Radar graph of the scores on the SF-36 health domains at the pre-operative measurement, one year after the operation and according to age- and gender-matched population norms. (BP, bodily pain; PF, physical function; RP, role physical; RE, role emotional; MH, mental health; VI, vitality; SF, social function; GH, general health).

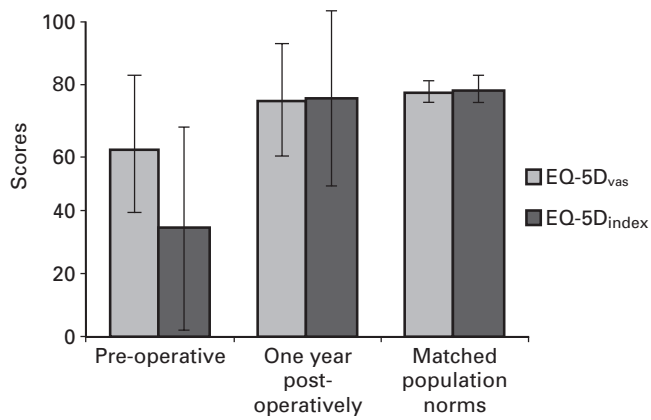

Fig. 2

EQ-5D$_{vas}$ and EQ-5D$_{index}$ scores at the pre-operative measurement, at one year after the operation and according to age- and gender-matched population norms.

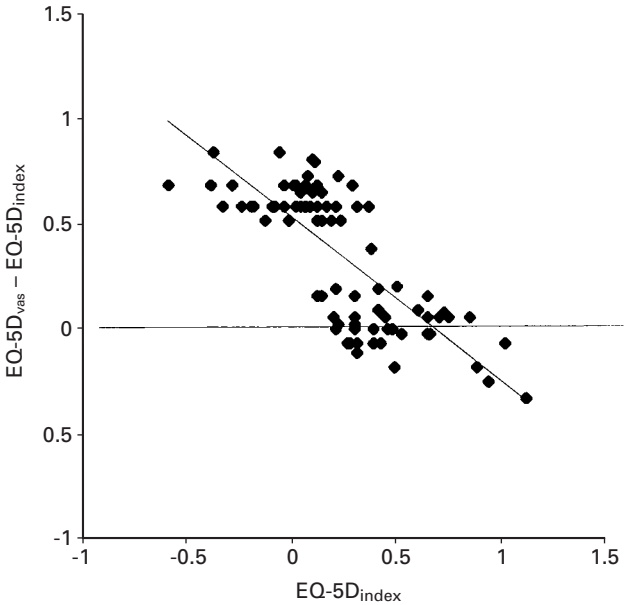

Fig. 3

Graph of the difference of the EQ-$5D_{vas}$ - EQ-$5D_{index}$ *versus* the EQ-$5D_{index}$ score. There is substantial disagreement between the scores at EQ-$5D_{index}$ levels less than 0.5. The diagonal line is a linear trend line through the data points.

health with the OHS and the WOMAC pain and stiffness (Table V). The physical and mental scores of the SF-12 (PCS and MCS) showed moderate to high correlations and correlations of change with, respectively, the physical and mental domains of the SF-36.

We compared the mean SF-36 and EQ-5D scores before and at one year after operation with the age- and gender-matched SF-36 and EQ-5D scores found in general population samples[6,7] (Figs 1 and 2). The scores at one year after operation approached the scores for the general population for both questionnaires, although we still found significant differences for the physical function, and physical and emotional role domains of the SF-36. The mental health, vitality and general health scores were even higher than the age- and gender-matched scores from the general population (Fig. 1).

We found large floor effects at the baseline for the SF-36 role physical (78%) and role emotional (38%) domains (Table II). At one year after operation most outcome scores showed a pronounced ceiling effect, mainly large (>30%) although some were moderate (>10%). The floor effects for the SF-36 role physical (34%) and role emotional (24%) remained (Table II).

We found several features in the EQ-5D scores. First, there was a binomial distribution of the pre-operative EQ-$5D_{index}$ score, which changed to a skewed normal distribution at one year after operation. Scores were especially scarce between 0.25 and 0.5. Furthermore, there were 15 patients with pre-operative Q-$5D_{index}$ scores less than zero. Particularly notable was the pre-operative discrepancy and post-operative agreement between the EQ-$5D_{vas}$ and EQ-$5D_{index}$. The pre-operative discrepancy was greatest in patients with a low EQ-$5D_{index}$ score (Fig. 3).

Discussion

For some years it has been agreed that patient-reported outcome data should define clinical success and expertise in the eyes of payers, patients and even fellow providers.[3] When limited resources must be distributed to different healthcare interventions, it is important to have information on the relative gain in health status achieved. We compared two disease-specific and three generic questionnaires in the evaluation of THR with respect to their floor and ceiling effects, content validity and sensitivity to change.

We found that patients who did not return their questionnaires at the one-year follow-up were more likely to have had a post-operative complication and had significantly lower scores on the WOMAC pain and the SF physical functioning scores at three months after operation. The fact that these patients experienced a complication after surgery may be an explanation for their not returning the questionnaires. Patients with a post-operative complication who were included in our analysed group had similar outcome scores at follow-up at one year as the patients without complications. We therefore do not expect the non-responding group to have a pronounced effect on the results at one year after operation.

All outcome scores had improved significantly at one year after operation except for the SF-36 general health domain. In our study, the OHS, together with the pain and physical function domains of the WOMAC and the SF-36, showed the largest effect sizes after THR compared with other dimensions of health status. These were thus identified as being of particular importance in the assessment of outcome after THR.[11,22] We found greater effect sizes for patients with unilateral hip involvement (Charnley A) compared with patients who had other conditions which directly impeded mobility (Charnley B and C). In this last group, the effect size of the OHS more than doubled the effect sizes of WOMAC pain and physical function. This emphasises the site-specific properties of the OHS. However, the effect size of the OHS in the Charnley B and C group is significantly smaller than that in the Charnley A group. This suggests that there is still an influence of the other conditions which impede mobility upon the OHS,[23] which is in contrast with the findings of Dawson et al[22] who did not find a significant difference in the OHS in a group with unilateral hip symptoms and a group with diffuse symptoms.

Our main concern was to identify the optimal set of health-outcome measures in THR. In a standard evaluation study of THR the completion of four questionnaires is a heavy burden for patients, especially when the OHS and the WOMAC measure similar symptoms as shown by their

high correlations and correlations of change. We prefer the OHS since it is a shorter, more site-specific and responsive questionnaire. It also shows fewer ceiling effects after operation and seems to be the most appropriate disease-specific questionnaire of the two tested in this study. In a study by Dunbar et al[18] on patients who underwent TKR, the Oxford knee score was also preferred, as opposed to the WOMAC and Lequesne scores, because of its higher feasibility, validity and reliability. The SF-36 captures additional, important quality-of-life domains which are influenced by a THR.[24] The SF-36 could probably be replaced by the SF-12 because of the large changes that occur between the domains which belong to the physical component summary (PCS) and mental component summary (MCS) scores. However, this is at the cost of losing detailed information about separate health-status domains. For large, cross-sectional surveys, the SF-12 is more appropriate because of its high feasibility.[18] It should be noted that we calculated the SF-12 scores from the SF-36 responses in order to reduce the number of questionnaires which patients had to complete. We did not expect the scores obtained in this way to be different from those which we would have found if the patients had completed the SF-12 questionnaire itself.

We found considerable floor and ceiling effects in the questionnaires, especially for the physical and emotional role domains of the SF-36. This can be explained by the limited answering options (yes or no) for questions about limitations in role functioning. In the new version of the SF-36 (SF-36v2) these scales are replaced by five-point Likert scales which improve precision and range for both domains.[25] The many ceiling effects post-operatively reflect the excellent outcome which can be achieved with THR in patients with severe hip disease. This is also reflected in the scores of the SF-36 and the EQ-5D, which approach general population norms at one year after operation. The OHS, WOMAC and SF-36 have shown good psychometric qualities in several studies in patients undergoing THR.[10,26] There are only a few published studies on the outcome of THR in which the EQ-5D was used.[5] The developers of the EQ-5D recommend it as a complementary instrument, not as a substitute for other instruments.[15] Hurst et al,[27] in a study of rheumatoid patients, showed that the EQ-5D is simple to use, is valid, responsive to change and reliable. The EQ-5D, especially the EQ-$5D_{index}$, showed good pre-operative correlations and correlations of change with the other questionnaires. It also showed a similar return for age- and gender-matched population norms as the SF-36.

However, our observations of the EQ-5D scores highlighted similar problems as those found by Wolfe and Hawley,[28] who studied rheumatoid patients, and Fransen and Edmonds[29] who studied patients with osteoarthritis of the knee. The binomial distribution of the EQ-$5D_{index}$ score can be explained by the crude scaling of the EQ-5D. For example, in the mobility domain, the two extreme categories are 'no problems' and 'unable/confined to bed'. Since almost no patients will be in the 'unable' category, the only movement which can be detected will be between 'no problems' and the presence of 'problems'.[28] The pre-operative difference and the post-operative agreement between the EQ-$5D_{vas}$ (patient-centred value) and the EQ-$5D_{index}$ (social value) was remarkable (Table II).[27-29] This pre-operative difference can be explained by what has been called 'response shift'. Response shift is the change in internal standards, in values, or in the perception of quality of life, which are created by changes in health state.[30] In clinical practice, this means that most patients perceive their quality of life as being better than from the perspective of the general population.

This is either because patients become accustomed to their disease or because their expectations about their health status have changed. Another explanation for the difference in pre-operative EQ-5D values is that patients with severe hip disease interpret their hip condition as a mechanical problem which has to be fixed, not as part of their general health. This is also reflected in the general health domain of the SF-36, the only domain of the SF-36 which did not improve significantly at one year after THR, and which is relatively high when compared with other domains.[31] Several patients had a pre-operative EQ-$5D_{index}$ less than zero. That is, from a social perspective their health state was worse than death. This represents the fact that normal individuals asked to consider such an existence would regard themselves as better off dead.[26] The self-assessed health of patients on the EQ-$5D_{vas}$ scale diverges from the social view in more severely disabled patients, which raises important ethical and practical questions about the use and interpretation of utility values, for example in cost-utility analyses. These peculiarities in the EQ-5D scores may justify further research into the EQ-5D, especially in the scaling of the EQ-$5D_{index}$ and the valuation of severe health states in relation to death. Further investigation may also be helpful when smaller changes in health status are being investigated, for example in the follow-up of THR.

Based on the results of our study, we recommend the OHS and SF-12 in the assessment of THR. The SF-36 may be used in smaller studies in which a more detailed description of health domains is needed and when smaller changes in health status are being investigated. The EQ-5D is useful in situations in which utility values are needed to calculate cost-effectiveness or quality-adjusted life years (QALYs), such as in the assessment of new techniques in THR.

The authors thank Dutch Healthcare Research (ZonMW) for funding this study.

No benefits in any form have been received or will be received from any commercial party related directly or indirectly to the subject of this article.

References

1. **Rissanen P, Aro S, Sintonen H, Slatis P, Paavolainen P.** Quality of life and functional ability in hip and knee replacements: a prospective study. *Qual Life Res* 1996;5:56-64.

2. **Laupacis A, Bourne R, Rorabeck C, et al.** The effect of elective total hip replacement on health-related quality of life. *J Bone Joint Surg [Am]* 1993;75-A:1619-26.

3. **Ritter MA, Albohm MJ.** Overview: maintaining outcomes for total hip arthroplasty: the past, present and future. *Clin Orthop* 1997;344:81-7.
4. **Dawson J, Fitzpatrick R, Murray D, Carr A.** Comparison of measure to assess outcomes in total hip replacement surgery. *Qual Health Care* 1996;5:81-8.
5. **Dawson J, Fitzpatrick R, Frost S, et al.** Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. *J Bone Joint Surg [Br]* 2001;83-B:1125-9.
6. **Aaronson NK, Muller M, Cohen PD, et al.** Translation, validation, and norming of the Dutch language version of the SF-36 health survey in community and chronic disease populations. *J Clin Epidemiol* 1998;51:1055-68.
7. **Dolan P.** Modeling valuations for EuroQol health states. *Med Care* 1997;35: 1095-108.
8. **Kosinski M.** Scoring the SF-12 physical and mental summary measures. *Medical Outcomes Trust Bulletin* 1997;5:3-4.
9. **Greenfield S, Apolone G, McNeil BJ, Cleary PD.** The importance of co-existent disease in the occurrence of postoperative complications and one-year recovery in patients undergoing total hip replacement. *Med Care* 1993;31:141-54.
10. **Dawson J, Fitzpatrick R, Carr A, Murray D.** Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg [Br]* 1996;78-B:185-90.
11. **Fitzpatrick R, Morris R, Hajat S, et al.** The value of short and simple measures to assess outcomes for patients of total hip replacement surgery. *Qual Health Care* 2000;9:146-50.
12. **Hoefnagels NH, Dhert WJ, Gosens T, et al.** De Oxford Heup Score, een Nederlandstalige patiëntenvragenlijst ter evaluatie van het resultaat van een totale heupoperatie. *Tijdschrift voor Orthopaedie* 2002;9:5-16.
13. **Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW.** Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 1988;15:1833-40.
14. **Ware JE Jr, Kosinski M, Bayliss MS, et al.** Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. *Med Care* 1995;33 (4 Suppl):264-79.
15. **Brooks R.** EuroQol: the current state of play. *Health Policy* 1996;37:53-72.
16. **Gandek B, Ware JE, Aaronson NK, et al.** Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project: International Quality of Life Assessment. *J Clin Epidemiol* 1998;51:1171-8.
17. How to score the SF-36 Health Survey. *Boston: Medical Outcomes Trust,* 1994.
18. **Dunbar MJ, Robertsson O, Ryd L, Lidgren L.** Appropriate questionnaires for knee arthroplasty: results of a survey of 3600 patients from The Swedish Knee Arthroplasty Registry. *J Bone Joint Surg [Br]* 2001;83-B:339-44.
19. **Kazis LE, Anderson JJ, Meenan RF.** Effect sizes for interpreting changes in health status. *Med Care* 1989;27 (3 Suppl):178-89.
20. **Cohen J.** *Statistical power analyses for the behavioural sciences.* 2nd ed. Hillsdale, Lawrence Erlbaum Associates, 1988.
21. **Charnley J.** *Low friction arthroplasty of the hip.* Berlin, etc: Springer-Verlag, 1979:23-4.
22. **Dawson J, Fitzpatrick R, Murray D, Carr A.** The problem of 'noise' in monitoring patient-based outcomes: generic, disease-specific and site-specific instruments for total hip replacement. *J Health Serv Res Policy* 1996;1:224-31.
23. **McMurray R, Heaton J, Sloper P, Nettleton S.** Measurement of patient perceptions of pain and disability in relation to total hip replacement: the place of the Oxford hip score in mixed methods. *Qual Health Care* 1999;8:228-33.
24. **Lieberman JR, Dorey F, Shekelle P, et al.** Outcome after total hip arthroplasty: comparison of a traditional disease-specific and a quality-of-life measurement of outcome. *J Arthroplasty* 1997;12:639-45.
25. **Ware JE.** SF-36 Health Survey update, 2003. http://www.sf-36.org. (accessed 21/06/04)
26. **Boardman DL, Dorey F, Thomas BJ, Lieberman JR.** The accuracy of assessing total hip arthroplasty outcomes: a prospective correlation study of walking ability and 2 validated measurement devices. *J Arthroplasty* 2000;15:200-4.
27. **Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A.** Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). *Br J Rheumatol* 1997;36:551-9.
28. **Wolfe F, Hawley DJ.** Measurement of the quality of life in rheumatic disorders using the EuroQol. *Br J Rheumatol* 1997;36:786-93.
29. **Fransen M, Edmonds J.** Reliability and validity of the EuroQol in patients with osteoarthritis of the knee. *Rheumatology (Oxford)* 1999;38:807-13.
30. **Sprangers MA, Schwartz CE.** Integrating response shift into health-related quality of life research: a theoretical model. *Soc Sci Med* 1999;48:1507-15.
31. **McGuigan FX, Hozack WJ, Moriarty L, Eng K, Rothman RH.** Predicting quality-of-life outcomes following total joint arthroplasty: limitation of the SF-36 health status questionnaire. *J Arthroplasty* 1995;10:742-7.