

Health-Related Quality of Life and Appropriateness of Knee or Hip Joint Replacement

José M. Quintana, MD, PhD; Antonio Escobar, MD; Inmaculada Arostegui, MHSc; Amaia Bilbao, MSc; Jesús Azkarate, MD, PhD; J. Ignacio Goenaga, MD; Juan C. Arenaza, MD

Background: We studied the association between explicit appropriateness criteria for total hip joint replacement (THR) and total knee replacement (TKR) with changes in health-related quality of life of patients undergoing these procedures.

Methods: Prospective observational study of 1576 consecutive patients with diagnoses of osteoarthritis on waiting lists to undergo THR or TKR. Explicit appropriateness criteria using the RAND appropriateness method were applied. Patients completed 2 questionnaires that measured health-related quality of life, the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) and the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), before the procedure and 6 months afterward.

Results: Patients who were considered appropriate candidates for these procedures had greater improvements than those who were considered inappropriate candidates in all 3 WOMAC domains (pain, functional limitation, and stiffness; THR: 43.0, 40.6, and 40.4 vs 14.7,

19.1, and 15.9; TKR: 34.9, 32.5, and 30.2 vs 23.2, 18.9, and 17.1; $P < .001$ for all comparisons). Patients who underwent THR and were judged to be appropriate candidates had greater improvements in the physical function, role-physical, bodily pain, and social function domains of the SF-36 than those judged to be inappropriate candidates (34.4, 35.1, 33.1, and 26.6 vs 19.6, 9.2, 5.7, and 7.0; $P = .04$, $P = .03$, $P < .001$, and $P < .001$, respectively). Appropriate candidates for TKR demonstrated greater improvement in the social function domain of the SF-36 after the procedure than those deemed inappropriate candidates (19.9 vs 7.9; $P = .004$) but not in the other domains of functional status.

Conclusions: These results suggest a direct relationship between explicit appropriateness criteria and better health-related quality-of-life outcomes after THR and TKR surgery. Our results support the use of these criteria for clinical guidelines or evaluation purposes.

Arch Intern Med. 2006;166:220-226

Author Affiliations: Unidad de Investigación, Hospital de Galdakao, Galdakao (Dr Quintana and Ms Bilbao), and Unidad de Investigación (Dr Escobar) and Servicio de Traumatología (Dr Arenaza), Hospital de Basurto, Bilbao, Spain; Departamento de Matemática Aplicada, Estadística e Investigación Operativa, Universidad del País Vasco, Lejona, Spain (Ms Arostegui); Servicio de Traumatología, Hospital de Mendara, Mendara, Spain (Dr Azkarate); and Servicio de Traumatología, Hospital de Santiago, Vitoria, Spain (Dr Goenaga).

AS LIFE EXPECTANCIES INCREASE, the rates of hip joint and knee replacement are expected to increase.¹⁻³ Although these procedures are expensive,⁴⁻⁶ they also are among the most effective in terms of patient benefits. Substantial variations in the indications for a variety of surgical procedures, including hip joint and knee replacement, have been reported during the last 20 years.^{7,8}

The RAND appropriateness method⁹ combines expert opinion with available scientific evidence to create explicit appropriateness criteria. Following this model, our group assembled 2 panels of experts: one to develop appropriateness criteria for total hip joint replacement (THR)¹⁰ and the other for total knee replacement (TKR).¹¹

In an effort to validate the explicit appropriateness criteria, which share similar variables, we conducted a prospective observational study to examine the relationship between appropriateness evaluation and outcomes measured by 2 validated

health-related quality of life (HRQoL) instruments: the generic Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36)¹² and the specific Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC).¹³ We hypothesized that if the appropriateness criteria did indeed offer good clinical guidance, patients considered appropriate would have higher HRQoL improvements in all relevant domains in these instruments.

METHODS

EXPLICIT CRITERIA DEVELOPMENT

First, we performed an extensive literature review to summarize the existing knowledge on the effectiveness and risks of THR and TKR for treating patients with osteoarthritis. Second, from this review, comprehensive and detailed lists of mutually exclusive and clinically specific scenarios (indications) were developed in which THR or TKR might be performed. This list contained 216 scenarios for THR and 624 for TKR. For THR, these scenarios included the

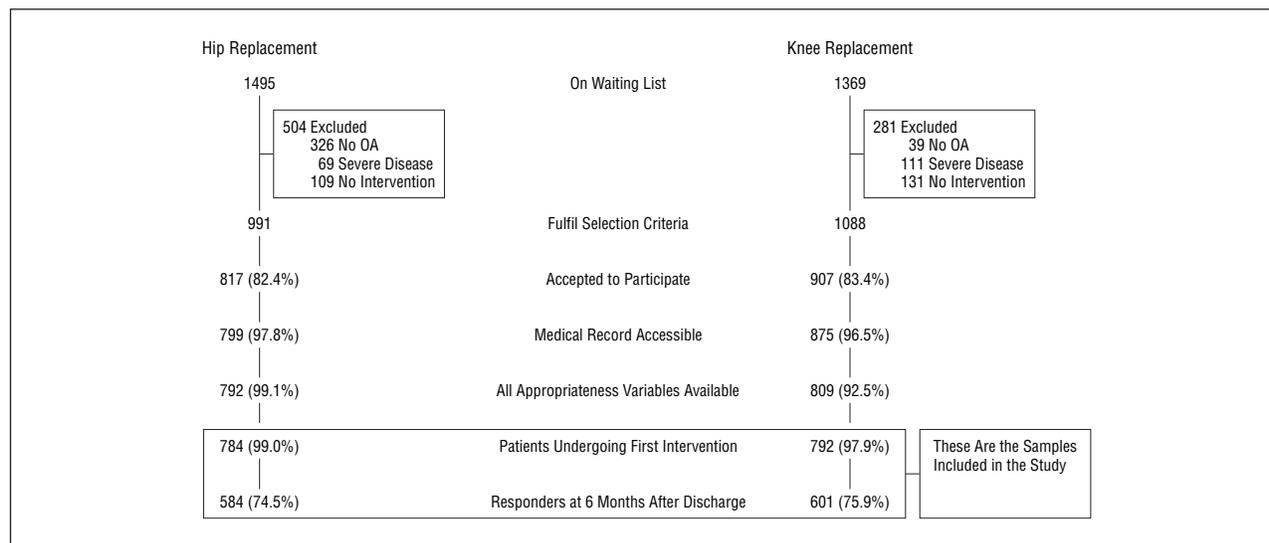


Figure. Patient recruitment and losses. Exclusion criteria included a main diagnosis of not hip or knee osteoarthritis (OA); malignant, severe organic, or psychiatric diseases; and failure to undergo the surgical intervention for any reason (death, intervention at another hospital, or refusal to undergo the intervention) 1 year after inclusion in the study. Each percentage is estimated based on the previous frequency.

following variables: age, bone quality measured by x-ray examination according to the classification of Singh et al,¹⁴ surgical risk (based on the American Society of Anesthesiologists criteria¹⁵), previous nonsurgical procedures performed, pain, and functional limitations assessment (based on the American College of Rheumatology classification¹⁶ and need for a mobility aid). For TKR, the scenarios included the following variables: age, previous surgical interventions, anatomical location, symptoms and functional impairment, joint mobility and stability, and radiology of the lesion (based on the Ahlbäck classification¹⁷).

An appropriate procedure is one in which “the expected health benefit exceeds the expected negative consequences by a sufficiently wide margin that the procedure is worth doing, exclusive of cost.”^{9(p54)} Ratings were based on a 9-point scale. The use of each procedure for a specific scenario was considered appropriate if the panel’s median rating was between 7 and 9 without disagreement, inappropriate if the value was between 1 and 3 without disagreement, and uncertain if the median rating was between 4 and 6 or if panel members disagreed. Disagreement was defined as a minimum of one third of the panelists rating an indication from 1 to 3 and a minimum of one third rating it from 7 to 9.

Third, we formed 2 independent national panels. The panelists were provided with the literature review and the list of indications and asked to rate each one for the appropriateness of performing each procedure. The ratings were confidential and took place in 2 rounds, using a modified Delphi process. The results of both panels were reported previously.^{10,11}

DATA COLLECTION

The prospective observational study took place in 5 large and 2 medium-sized public teaching hospitals with similar human and technical resources located in the Basque Country, serving a total population of 2 million inhabitants. These medical institutions belong to the network of public hospitals of the Basque Health Care Service—Osakidetza, a local government that is part of the Spanish National Health Service, which provides free unrestricted care to nearly 100% of the population. Physicians in each hospital were blinded to the study goals. The hospitals’ ethics review boards approved both projects.

Consecutive patients with osteoarthritis scheduled to undergo THR or TKR in any of the 7 hospitals were eligible for

the study. Between March 1999 and March 2000, 1495 patients were placed on waiting lists to undergo THR and 1369 to undergo TKR. Patients with severe comorbidities, such as cancer, terminal disease, or psychiatric conditions, and those whose main diagnosis was not hip or knee osteoarthritis or who failed to undergo the surgical intervention 1 year after inclusion in the study were excluded from analysis. All patients were assessed before the procedure and 6 months afterward. The **Figure** shows the recruitment process.

All patients on the waiting list for THR or TKR were sent a letter that described the study and asked for their voluntary participation. This mailing included the SF-36 and WOMAC questionnaires and sociodemographic information. A reminder letter was sent to patients who had not replied after 15 days. We sent the questionnaires again and contacted by telephone those who still had not replied after another 15 days. Six months after the intervention, patients were sent another letter with the questionnaires and additional questions on the clinical aspects of their disease and satisfaction with the intervention. The satisfaction question was dichotomized as being satisfied or not. At this time, patients answered a transitional question about their joint improvement after the intervention. The possible responses included “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” or “a great deal worse.” Those who had not replied were followed up as described previously.

The SF-36¹² covers 8 domains and 2 summary scales, physical and mental. The scores for the SF-36 scales range from 0 to 100, with a higher score indicating better health status. The SF-36 has been translated into Spanish and validated in Spanish populations.¹⁸

The WOMAC¹³ covers 3 dimensions: pain, stiffness, and physical function. We used the categorical version with 5 response levels for each item. The data were standardized to a range of values from 0 to 100, with 0 representing the best health status and 100 the worst possible status. The original and Spanish questionnaire versions are reliable, valid, and sensitive to the changes in the health status of patients with hip or knee osteoarthritis.^{19,20}

We retrieved data from the hospital and physician medical records that included variables before the intervention, at admission, and at discharge. Besides those variables that belonged to the appropriateness algorithm, other variables collected included sociodemographic data, all comorbidities included on the Charlson Comorbidity Index,²¹ local and gen-

eral complications perioperatively and postoperatively, re-intervention, death, and length of hospital stay. Six months after discharge, all medical records were reviewed to determine if the patient had been readmitted, had any complication resulting from the intervention, or had died.

Three physicians blinded to the specific study goals extracted the data from the patients' medical records and recorded them on a prepared form. The physicians were trained to retrieve the medical record data in a particular way and were tested for agreement. Members of the research team also reviewed those variables in a sample of 121 records to test the accuracy of the data retrieved by the reviewers. This review resulted in a κ value for dichotomous variables (diagnosis, surgical risk, death, readmissions, and re-intervention) above 0.99 and an intraclass correlation coefficient for length of hospital stay of 0.99.

STATISTICAL ANALYSIS

The unit of study was the patient. In cases in which 2 interventions were performed in the same patient, we selected the first one. Longitudinal data analysis was performed in all available cases. A general linear model was executed with HRQoL as the dependent variable over time for the 8 domains, the 2 summary measures of the SF-36, and the 3 WOMAC domains. Appropriateness was the independent variable, adjusted by sex. Multilevel analysis with mixed models was performed to test differences among hospitals in HRQoL improvement for the 3 appropriateness categories.

For all WOMAC and SF-36 domains, we estimated, by procedure, the standard error of measurement by the following formulae:

$$\text{SEM} = \text{SD} \times \sqrt{1 - R},$$

where SD is the standard deviation of the sample at baseline and R is the reliability coefficient.²² We used the Cronbach α as a reliability measure.²³ From the standard error of measurement, we derived the minimal detectable change (MDC), which is the result of the following multiplication:

$$\text{SEM} \times z \text{ score} \times \sqrt{2}.$$

We established a 95% confidence interval, which corresponds to a z score of 1.96. The MDC represents the smallest change in score that likely reflects true change more than measurement error alone.²⁴ The MDC proportion is the proportion of patients with changes in scores that exceeded the MDC. The minimal clinically important difference (MCID) has been defined as the smallest difference between the scores in a questionnaire that the patient perceives to be beneficial. It is an anchor-based method that fixes a threshold that demarcates trivial from small but important differences. The MCID was calculated for those patients who, at the 6-month visit, answered that their articulation was "somewhat better than before the intervention" to a transitional question.²⁵ The MCID proportion reflects the proportion of the sample with change scores exceeding the MCID.

All effects were considered statistically significant at $P < .05$, unless otherwise noted. All statistical analyses were performed using SAS for Windows statistical software, version 8.2 (SAS Institute Inc, Cary, NC).

RESULTS

A total of 1576 patients were included in this study, with 784 undergoing THR and 792 undergoing TKR (Figure).

No statistically significant differences occurred among responders and nonresponders at 6 months regarding socio-demographic variables, main clinical characteristics (including pain and functional limitation), or appropriateness evaluation.

The mean age of the patients undergoing THR was 69.1 years, and 48.3% were women. For patients undergoing TKR, the mean age was 71.9 years, and 73% were women. The inappropriateness rate was higher for TKR than THR (12.4% vs 5.2%).

Of the patients who underwent THR, compared with patients who underwent procedures that were deemed inappropriate, those undergoing appropriate procedures had significantly higher improvements in the physical function, role-physical, bodily pain, and social function domains and the physical component summary scale of the SF-36, as well as the 3 domains of the WOMAC. Differences were also observed in the improvement between those judged as uncertain compared with those judged inappropriate in the same domains, except for the social function of the SF-36 (**Table 1**). Among patients who underwent TKR, significant differences occurred by category of appropriateness for the social function of the SF-36, as well as for the 3 domains of the WOMAC (**Table 2**).

Multilevel analysis showed no differences among hospitals in HRQoL improvement by appropriateness categories for either surgical intervention. Therefore, we did not include the hospital as an interaction variable in the models.

The MDC values ranged from 12 to 29 for the WOMAC domains and from 19 to 42 for the SF-36 domains. For the whole sample, the proportion of patients who surpassed the MDC was higher than 50% for the SF-36 physical function and the WOMAC pain and functional limitation in both procedures (**Table 3**). The MCID values varied by procedure and questionnaire. More than half of patients undergoing THR surpassed the MCID in all SF-36 domains; approximately 75% did so on all the WOMAC domains. The percentages were lower for patients undergoing TKR.

A significantly higher proportion of patients undergoing THR judged as appropriate candidates surpassed the MDC or MCID than those judged as inappropriate candidates on all the WOMAC domains and the physical function, role-physical, bodily pain, social function, and mental health domains of the SF-36 (**Table 4**). Among patients undergoing TKR, a similar association was observed with all the WOMAC domains and the vitality and social function domains of the SF-36 (Table 4).

No differences were observed among the 3 appropriateness categories for any of the clinical indicators analyzed up to 6 months after discharge, except for the fatality rate among patients undergoing THR. Patients reported higher satisfaction with the intervention in the appropriate group than in the inappropriate group for both procedures. Among THR patients, those who underwent appropriate procedures reported better perception of their general health status compared with the previous year, greater relief of symptoms, and greater recovery than those who underwent inappropriate procedures (**Table 5**).

Table 1. Preintervention Health-Related Quality-of-Life and Improvement Scores 6 Months After Total Hip Joint Replacement by Appropriateness in 784 Patients*

	Preintervention			P Value†	Improvement at 6 Months			P Value†
	Appropriate (n = 575)	Uncertain (n = 168)	Inappropriate (n = 41)		Appropriate (n = 434)	Uncertain (n = 120)	Inappropriate (n = 30)	
SF-36								
Physical functioning	17.98	29.05	40.74	<.001	34.44	32.79	19.64	.04
Role-physical	6.52	22.71	40.63	<.001	35.13	39.27	9.24	.03
Bodily pain	26.13	43.41	60.07	<.001	33.09	28.64	5.74	<.001
General health	57.31	63.19	68.43	<.001	4.79	4.26	-2.49	.21
Vitality	39.08	52.37	54.75	<.001	21.28	17.61	10.96	.09
Social functioning	50.87	66.91	77.03	<.001	26.57	17.36	7.01	<.001
Role-emotional	66.75	75.33	85.51	<.001	13.45	13.59	1.73	.30
Mental health	57.66	66.03	69.43	<.001	14.67	11.36	7.45	.26
PCSS	26.09	30.96	35.70	<.001	12.02	11.89	3.87	.006
MCSS	46.16	50.36	50.78	<.001	5.27	3.74	2.73	.52
WOMAC								
Pain	59.61	42.23	27.08	<.001	-42.98	-30.23	-14.68	<.001
Functional limitation	69.62	55.58	42.92	<.001	-40.61	-34.85	-19.08	<.001
Stiffness	63.01	50.07	33.72	<.001	-40.41	-33.84	-15.87	<.001

Abbreviations: MCSS, mental component summary scale; PCSS, physical component summary scale; SF-36, Medical Outcomes Study 36-Item Short-Form Health Survey; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index.

*Data are adjusted for sex. Preintervention: higher SF-36 scores indicate better health-related quality of life; higher WOMAC scores indicate worse health-related quality of life.

†P values are for the global comparisons among the 3 groups (F test of fixed effects).

Table 2. Preintervention Health-Related Quality-of-Life and Improvement Scores 6 Months After Total Knee Replacement by Appropriateness in 792 Patients*

	Preintervention			P Value†	Improvement at 6 Months			P Value†
	Appropriate (n = 557)	Uncertain (n = 137)	Inappropriate (n = 98)		Appropriate (n = 414)	Uncertain (n = 108)	Inappropriate (n = 79)	
SF-36								
Physical functioning	19.49	33.11	38.66	<.001	25.79	20.90	19.91	.12
Role-physical	11.26	20.42	30.35	<.001	30.50	26.05	30.84	.75
Bodily pain	31.09	43.37	53.95	<.001	19.74	16.11	15.45	.48
General health	56.56	61.89	62.07	<.001	2.42	4.93	3.98	.65
Vitality	39.67	50.26	56.59	<.001	15.47	10.77	7.54	.09
Social functioning	52.04	65.62	77.63	<.001	19.89	13.13	7.88	.004
Role-emotional	61.77	70.86	84.28	<.001	11.38	5.88	-0.88	.12
Mental health	57.81	62.25	67.64	<.001	9.73	8.17	5.96	.60
PCSS	27.63	32.22	33.75	<.001	8.97	7.99	9.25	.74
MCSS	45.15	48.62	52.39	<.001	3.98	1.80	-0.14	.11
WOMAC								
Pain	57.94	50.26	42.58	<.001	-34.90	-26.64	-23.02	<.001
Functional limitation	65.41	54.36	46.24	<.001	-32.47	-23.71	-18.86	<.001
Stiffness	58.93	53.22	46.47	.009	-30.25	-20.34	-17.09	<.001

Abbreviations: PCSS, physical component summary scale; MCSS, mental component summary scale; SF-36, Medical Outcomes Study 36-Item Short-Form Health Survey; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index.

*Data are adjusted for sex. Preintervention: higher SF-36 scores indicate better health-related quality of life; higher WOMAC scores indicate worse health-related quality of life.

†P values are for the global comparisons among the 3 groups (F test of fixed effects).

COMMENT

This prospective observational study of more than 1500 patients who underwent THR or TKR supports the validity of the criteria developed by the RAND appropriateness method for 2 procedures that share many characteristics, such as the role of symptoms in the clinical decision process or the outcome measures. In general, we found that patients who were deemed appropriate candidates by

the criteria were more likely to have had better outcomes and greater improvements in HRQoL following THR or TKR than patients deemed inappropriate candidates.

Determining the appropriateness of a surgical intervention is important. However, it is difficult to develop evidence-based criteria because, in most cases, high-quality evidence from clinical trials is not available, either because clinical trials are rarely performed for surgical procedures or they do not cover an important range of

Table 3. Responsiveness Indexes for WOMAC and SF-36 for Patients Undergoing Total Hip Joint or Knee Replacement

	Total Hip Joint Replacement				Total Knee Replacement			
	Cronbach α	SEM	MDC (MDC%)	MCID (MDIC%)	Cronbach α	SEM	MDC (MDC%)	MCID (MDIC%)
SF-36								
Physical functioning	.89	6.84	18.96 (72.68)	19.62 (72.0)	.88	7.29	20.19 (50.52)	10.04 (65.64)
Role-physical	.90	8.87	24.58 (53.81)	10.20 (53.81)	.88	10.33	28.64 (35.80)	7.81 (47.40)
Bodily pain	.67	14.47	40.11 (34.76)	14.74 (63.81)	.69	15.27	42.32 (19.82)	12.83 (50.88)
General health	.74	9.96	27.62 (8.62)	-0.65 (64.27)	.76	9.75	27.04 (7.27)	0.11 (52.08)
Vitality	.77	11.19	31.01 (28.26)	7.34 (68.97)	.81	10.70	29.65 (25.95)	5.42 (57.20)
Social functioning	.76	15.37	42.60 (25.71)	8.75 (64.54)	.76	15.19	42.11 (20.61)	8.77 (57.24)
Role-emotional	.94	10.69	29.63 (22.22)	-9.29 (90.22)	.95	10.15	28.13 (23.23)	2.43 (23.23)
Mental health	.87	8.43	23.37 (30.28)	6.42 (57.77)	.87	8.60	23.85 (25.43)	0.76 (59.27)
WOMAC								
Pain	.83	7.71	21.37 (78.01)	24.55 (78.01)	.79	8.24	22.85 (66.89)	22.60 (66.89)
Functional limitation	.93	4.39	12.18 (87.87)	20.80 (78.21)	.93	4.59	12.72 (76.21)	17.67 (67.84)
Stiffness	.80	10.35	28.70 (58.45)	21.79 (75.00)	.81	10.51	29.14 (42.74)	12.94 (57.60)

Abbreviations: MCID, minimal clinically important difference; MCID%, minimal clinically important difference proportion, or percentage of patients who surpassed the MCID; MDC, minimal detectable change; MDC%, minimal detectable proportion, or percentage of patients who surpassed the MDC; SEM, standard error of measurement; SF-36, Medical Outcomes Study 36-Item Short-Form Health Survey; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index.

Table 4. Relevant Changes on Health-Related Quality-of-Life Outcomes for Patients Undergoing Total Hip Joint or Knee Replacement by Appropriateness Categories*

	MDC, %			MCID, %		
	Appropriate	Uncertain	Inappropriate	Appropriate	Uncertain	Inappropriate
Total Hip Joint Replacement						
SF-36						
Physical functioning	73.9 ⁱ	72.4	56.7 ^a	73.9 ⁱ	71.6	56.7 ^a
Role-physical	53.4 ⁱ	60.6 ⁱ	32.0 ^{au}	53.4 ⁱ	60.6 ⁱ	32.0 ^{au}
Bodily pain	36.5 ⁱ	34.2 ⁱ	13.3 ^{au}	67.9 ^{ui}	56.1 ^a	36.7 ^a
General health	8.9	8.8	3.5	63.8	68.4	55.2
Vitality	31.5 ^{ui}	21.3 ^a	11.5 ^a	71.8 ^u	62.0 ^a	57.7
Social functioning	28.4 ^u	19.1 ^a	13.3	68.5 ^{ui}	54.8 ^a	46.7 ^a
Role-emotional	24.6	17.7	8.3	89.7	93.8	83.3
Mental health	33.1	24.3	15.4	61.3 ⁱ	51.4	34.6 ^a
WOMAC						
Pain	83.2 ^{ui}	68.4 ^{ai}	40.0 ^{au}	83.2 ^{ui}	68.4 ^{ai}	40.0 ^{au}
Functional limitation	89.3 ⁱ	88.9 ⁱ	63.3 ^{au}	82.5 ^{ui}	70.9 ^{ai}	46.7 ^{au}
Stiffness	61.6 ⁱ	53.5 ⁱ	33.3 ^{au}	79.0 ^{ui}	67.5 ^{ai}	46.7 ^{au}
Total Knee Replacement						
SF-36						
Physical functioning	51.0	50.5	48.1	66.3	64.8	63.6
Role-physical	34.5	35.2	43.8	48.6	39.8	51.6
Bodily pain	20.8	16.4	19.4	53.1	46.2	45.8
General health	7.0	9.5	5.6	51.4	56.2	50.0
Vitality	29.0 ^u	22.3 ^a	13.9	61.8 ^{ui}	45.7 ^a	47.7 ⁱ
Social functioning	24.9 ^{ui}	12.4 ^a	9.2 ⁱ	62.1 ⁱ	52.4	38.2 ^a
Role-emotional	25.2	21.4	15.6	25.2	21.4	15.6
Mental health	27.1	23.7	18.8	62.8 ^u	49.5 ⁱ	53.2
WOMAC						
Pain	72.8 ^{ui}	52.8 ^a	55.7 ^a	72.8 ^{ui}	52.8 ^a	55.7 ^a
Functional limitation	81.8 ^{ui}	66.4 ^a	60.8 ^a	74.2 ^{ui}	56.1 ^a	50.6 ^a
Stiffness	48.9 ^{ui}	29.3 ^a	29.1 ^a	63.9 ^{ui}	42.5 ^a	45.6 ^a

Abbreviations: MCID, minimal clinically important difference; MDC, minimal detectable change.

*Superscript letters indicated differences among the 3 appropriateness categories (a, appropriate; u, uncertain; and i, inappropriate) at $P < .05$ by Scheffé test for multiple comparisons.

indications. For this reason, the RAND method was developed to create explicit appropriateness criteria. A major criticism of this method is the absence of studies that demonstrate the validity of such criteria.^{26,27}

Our main hypothesis was that patients classified by our explicit appropriateness criteria for THR and TKR^{10,11}

as having undergone an appropriate procedure would have larger improvements in HRQoL than patients classified as having undergone inappropriate procedures. Our results support this hypothesis. Both the SF-36 and the WOMAC demonstrated such larger improvements, but there was more evidence with the WOMAC. Therefore,

Table 5. Clinical Indicators and Patient Perception at 6 Months After Total Hip Joint or Knee Replacement by Appropriateness*

Clinical indicator	Total Hip Joint Replacement			P Value	Total Knee Replacement			P Value
	Appropriate (n = 575)	Uncertain (n = 168)	Inappropriate (n = 41)		Appropriate (n = 557)	Uncertain (n = 137)	Inappropriate (n = 98)	
Local complications	23 (4.0)	4 (2.4)	1 (2.4)	.57	37 (6.64)	11 (8.03)	5 (5.1)	.67
Global complications	25 (4.35)	7 (4.1)	0 (0)	.37	23 (4.13)	2 (1.46)	1 (1.02)	.12
Length of stay	11.82 (4.31)	11.1 (3.5)	12.5 (3.9)	.07	13.05 (5.08)	12.51 (4.42)	12.32 (4.38)	.26
Reintervention	27 (4.76)	6 (3.64)	3 (7.69)	.54	27 (4.88)	12 (8.82)	2 (2.04)	.058
Death	0 (0)	0 (0)	2 (5.0)	<.001	1 (0.18)	0 (0)	0 (0)	.81
Patient perception at 6 mo*								
Better health status	410 (97.6)	114 (97.4)	26 (92.3)	.32	374 (92.57)	99 (93.4)	64 (85.3)	.09
Relief of symptoms	364 (87.29)	104 (88.89)	21 (72.41)	.055	308 (76.24)	77 (73.3)	59 (77.63)	.76
Global satisfaction	407 (95.54)	113 (95.76)	25 (83.33)	.01	366 (89.93)	97 (91.51)	62 (80.52)	.04
Satisfaction with intervention	423 (99.06)	118 (100)	27 (90.0)	<.001	388 (95.1)	103 (95.37)	68 (87.18)	.02
Recovered	261 (62.14)	88 (74.58)	18 (60.0)	.04	219 (54.21)	54 (50.47)	39 (49.37)	.63
Better health status compared with 1 year previously	354 (83.7)	93 (79.49)	20 (66.7)	.047	267 (65.76)	72 (67.29)	44 (56.41)	.24

*Data are frequencies (percentages) except for length of stay, which is mean (SD). Percentages are for patients who responded to the questionnaire. Patients with missing data were excluded from analysis. P values are from χ^2 tests except for length of stay, which is from the general linear model.

our results support the validity of the explicit criteria because the appropriate compared with the inappropriate group had much greater benefit and similar low risks.

Our results identified several issues that deserve further comment. First, we found that the uncertain group had similar improvements to those in the appropriate group, suggesting that in some cases uncertain indications might also be appropriate.

Second, an important unresolved debate currently exists about how to determine the smallest difference between the scores in an HRQoL questionnaire that the patient perceives to be beneficial.²⁸ We tried to establish individual measures of improvement by estimating the MDC and the MCID. In our study, more patients in the appropriate group than in the inappropriate group had relevant gains on those 2 parameters on the 3 WOMAC domains and, among patients undergoing THR, on those domains of the SF-36 most related to the physical component of HRQoL. However, these scores varied a lot. As some authors have pointed out, establishing a definite MCID seems to be an almost impossible task.^{29,30}

Third, patients who underwent TKR had fewer differences among the appropriate compared with the inappropriate groups in HRQoL domains compared with those who underwent THR. Knee articulation is more complex than hip articulation, and among patients with osteoarthritis, the benefits they experience from TKR are usually less than those experienced after THR.³¹ However, even patients who underwent inappropriate TKR had substantial deterioration of their HRQoL before the intervention, measured by both HRQoL tools, especially when compared with the inappropriate cases in the THR sample, who had better HRQoL scores.

Fourth, in this study, differences between the appropriateness categories were observed in all domains of the WOMAC in both procedures but in only some of the domains of the SF-36. This finding is not unusual because, as many authors have reported, specific HRQoL questionnaires such as the WOMAC generally have greater

responsiveness than generic ones such as the SF-36.^{19,32}

Fifth, the negative consequences of the intervention (clinical complications, reinterventions, or deaths) were minor and similar among the 3 appropriateness categories. Among those undergoing THR, 2 patients in the inappropriate group died, but the deaths were unrelated to the inappropriateness of the intervention. Nevertheless, other outcomes, such as satisfaction with the intervention, differed among the appropriateness categories. Although only a small number of patients underwent an inappropriate procedure, they were slightly less satisfied with the intervention and its result, which constitutes a negative consequence from the patient's perspective.

Another recent study³³ evaluated the association between appropriateness and HRQoL among patients receiving a hip prosthesis. In that study, the surgeons were asked to complete a clinical indications form based on a set of proprietary guidelines. Then information gathered was reviewed to determine if the case profile matched the guidelines. Those investigators used the same HRQoL tools as ours. Limitations of this study were that the authors used privately published criteria, not available, to determine the appropriateness of an indication and did not attempt to test the validity of the criteria. Appropriateness was judged by a binary response: appropriate or not. The sample size of this study was smaller than ours (n=488), and the response rate was low (44%).

Our study has several strengths. We administered 2 widely used and validated HRQoL questionnaires that have been recommended by different authors for studying patients with hip or knee osteoarthritis.^{34,35} Furthermore, we collected the information prospectively in a large sample of patients, thus minimizing selection and information bias that may influence results of studies such as these. However, as a main limitation, we had a 25% rate of missing data at follow-up. Also, the fact that the appropriate group as a whole achieved a better HRQoL does not mean that each appropriate candidate procedure achieved a better HRQoL too.

In conclusion, the results of this study support the predictive validity of our explicit appropriateness criteria by showing a greater benefit in HRQoL among patients considered to be appropriate candidates for these procedures compared with those classified as inappropriate candidates. These results support the use of our criteria for clinical guidelines or to determine the degree of appropriateness and variations in the use of THR or TKR. Although variations in the indication of THR or TKR are potentially attributable to variations in clinical decision making, other issues, such as lack of human or technical resources in a specific center, could limit the generalizability of our findings in other settings. Finally, as suggested by some authors,³⁶ this method may be useful when comparing levels of appropriateness among populations but not to direct care for individual patients. When used as a utilization review tool, interventions considered inappropriate should undergo an individualized revision before being considered inappropriate.³⁷

Accepted for Publication: August 28, 2005.

Correspondence: José M. Quintana, MD, PhD, Unidad de Investigación, Hospital de Galdakao, Barrio Labeaga s/n, 48960 Galdakao, Vizcaya, Spain (jmquinta@hgda.osakidetza.net).

Financial Disclosure: None.

Funding/Support: This study was supported in part by grants from the Fondo de Investigación Sanitaria (98/001-01 to 03) and the thematic networks Red IRYSS of the Instituto de Salud Carlos III (G03/220) (Madrid, Spain). Ms Bilbao received a grant from the Department of Health of the Basque Government (Vitoria-Gasteiz, Spain).

Acknowledgment: We thank the following physicians for their contribution to this study: Jose M. Ordoñez, MD, Jose M. Vilarrubias, PhD, Jordi Ballester, PhD, Carlos Barrios, PhD, Mikel Sánchez, MD, Francisco Villar, PhD, Luis Gutierrez, PhD, Francisco Buendia, MD, Andrés Peña, MD, Félix Araluze, PhD, Joaquim Cabot, MD, Javier Vaquero, MD, Alfredo Queipo de Llano, MD, Manuel Gala, MD, Victor Alvarez, MD, Jose R. Caso, MD, Antonio Murcia, PhD, Alejandro Lizaur, MD, Jose R. Vesga, MD, Anibal Ruiz, MD, Manuel Figueroa, MD (TKR panel); Angel Alfageme, PhD, Jose M. Aranburu, PhD, Jesús Azkoaga, MD, Pedro Armandariz, PhD, Enrique Cáceres, PhD, Arsenio Diego, MD, Begoña Goicoetxea, PhD, Iñigo Guisasola, MD, Manuel Martínez-Grande, PhD, Enrique Queipo de Llano, PhD, Ramón Tobio, PhD, Jose Villar, MD (THR panel). We also thank Arantza Higuelmo, MD, Iratxe Lafuente, MSc, Alfonso Rodriguez, MD, and Ignacio Vidaurreta, MD, for their contribution to the development of the panel of experts, data retrieval, and data entry and to the Research Committee of the Galdakao Hospita. We are grateful for the support of the staff members of the different services, research, and quality units, as well as the medical records sections of the participating hospitals.

REFERENCES

- Felson DT, Zhang Y. An update on the epidemiology of knee and hip osteoarthritis with a view to prevention. *Arthritis Rheum.* 1998;41:1343-1355.
- Fear J, Hillman M, Chamberlain MA, Tennant A. Prevalence of hip problems in the population aged 55 years and over. *Br J Rheumatol.* 1997;36:74-76.
- Williams MH, Newton JN, Frankel SJ, Braddon F, Barclay E, Gray JA. Prevalence of total hip replacement. *J Epidemiol Community Health.* 1994;48:188-191.
- Macario A, Vitez TS, Dunn B, McDonald T, Brown B. Hospital costs and severity of illness in three types of elective surgery. *Anesthesiology.* 1997;86:92-100.
- Mushinski M. Average charges for a total hip arthroplasty: geographic variations, 1995. *Stat Bull Metrop Insur Co.* 1996;77:21-28.
- Faulkner A, Kennedy LG, Baxter K, Donovan J, Wilkinson M, Bevan G. Effectiveness of hip prostheses in primary total hip replacement: a critical review of evidence and an economic model. *Health Technol Assess.* 1998;2:1-133.
- Birkmeyer JD, Sharp SM, Finlayson SR, Fisher ES, Wennberg JE. Variation profiles of common surgical procedures. *Surgery.* 1998;124:917-923.
- Coyte P, Wang PP, Hawker G, Wright JG. The relationship between variations in knee replacement utilization rates and the reported prevalence of arthritis in Ontario, Canada. *J Rheumatol.* 1997;24:2403-2412.
- Brook RH, Chassin MR, Fink A, Solomon DH, Koseoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care.* 1986;2:53-63.
- Quintana JM, Arostegui I, Azkarate J, et al. Evaluation of explicit criteria for total hip joint replacement. *J Clin Epidemiol.* 2000;53:1200-1208.
- Escobar A, Quintana JM, Arostegui I, et al. Development of explicit criteria for total knee replacement. *Int J Technol Assess Health Care.* 2003;19:57-70.
- Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36), I: conceptual framework and item selection. *Med Care.* 1992;30:473-483.
- Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC. *J Rheumatol.* 1988;15:1833-1840.
- Singh M, Nagrath AR, Maini PS. Changes in trabecular pattern of the upper end of the femur as an index of osteoporosis. *J Bone Joint Surg Am.* 1970;52:457-467.
- Schneider AJ. Assessment of risk factors and surgical outcome. *Surg Clin North Am.* 1983;63:1113-1126.
- Hochberg MC, Chang RW, Dwosh I, Lindsey S, Pincus T, Wolfe F. The American College of Rheumatology 1991 revised criteria for the classification of global functional status in rheumatoid arthritis. *Arthritis Rheum.* 1992;35:498-502.
- Ahlback S. Osteoarthritis of the knee: a radiographic investigation. *Acta Radiol Diagn (Stockh).* 1968(suppl 72):7-72.
- Alonso J, Prieto L, Anto JM. The Spanish version of the SF-36 Health Survey (the SF-36 health questionnaire). *Med Clin (Barc).* 1995;104:771-776.
- Angst F, Aeschlimann A, Steiner W, Stucki G. Responsiveness of the WOMAC osteoarthritis index as compared with the SF-36 in patients with osteoarthritis of the legs undergoing a comprehensive rehabilitation intervention. *Ann Rheum Dis.* 2001;60:834-840.
- Escobar A, Quintana JM, Bilbao A, Azkarate J, Guenaga JI. Validation of the Spanish version of the WOMAC questionnaire for patients with hip or knee osteoarthritis. *Clin Rheumatol.* 2002;21:466-471.
- Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis.* 1987;40:373-383.
- Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol.* 2004;57:1008-1018.
- Wyrwich KW, Tierney WM, Wolinsky FD. Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. *Qual Life Res.* 2002;11:1-7.
- Stratford PW, Binkley FM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. *Phys Ther.* 1996;76:1109-1123.
- Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc.* 2002;77:371-383.
- Phelps CE. The methodologic foundations of studies of the appropriateness of medical care. *N Engl J Med.* 1993;329:1241-1245.
- Shekelle PG, Chassin MR, Park RE. Assessing the predictive validity of the RAND/UCLA appropriateness method criteria for performing carotid endarterectomy. *Int J Technol Assess Health Care.* 1998;14:707-727.
- Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID). *Curr Opin Rheumatol.* 2002;14:109-114.
- Kirwan JR. Minimum clinically important difference: the crock of gold at the end of the rainbow? *J Rheumatol.* 2001;28:439-444.
- Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. *Pharmacoeconomics.* 2000;18:419-423.
- Bachmeier CJ, March LM, Cross MJ, et al. A comparison of outcomes in osteoarthritis patients undergoing total hip and knee replacement surgery. *Osteoarthritis Cartilage.* 2001;9:137-146.
- Brazier JE, Harper R, Munro J, Walters SJ, Snaith ML. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. *Rheumatology (Oxford).* 1999;38:870-877.
- Wright CJ, Chambers GK, Robens-Paradise Y. Evaluation of indications for and outcomes of elective surgery. *CMAJ.* 2002;167:461-466.
- Hawker G, Melfi C, Paul J, Green R, Bombardier C. Comparison of a generic (SF-36) and a disease specific (WOMAC) (Western Ontario and McMaster Universities Osteoarthritis Index) instrument in the measurement of outcomes after knee replacement surgery. *J Rheumatol.* 1995;22:1193-1196.
- Nilsdotter AK, Roos EM, Westerlund JP, Roos HP, Lohmander LS. Comparative responsiveness of measures of pain and function after total hip replacement. *Arthritis Rheum.* 2001;45:258-262.
- Naylor CD. What is appropriate care? *N Engl J Med.* 1998;338:1918-1920.
- Dubois RW. Appropriateness studies. *N Engl J Med.* 1994;330:433.