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
following variables: age, bone quality measured by x-ray examination according to the classification of Singh et al., surgical risk (based on the American Society of Anesthesiologists criteria), previous nonsurgical procedures performed, pain, and functional limitations assessment (based on the American College of Rheumatology classification 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 classification). 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.” 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.

### 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-36 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.

The WOMAC 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.

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, local and gen-

---

**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.
eral complications perioperatively and postoperatively, rein-
tervention, 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 ex-
tracted the data from the patients' medical records and re-
corded 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 re-
viewed those variables in a sample of 121 records to test the
accuracy of the data retrieved by the reviewers. This review re-
resulted in a k value for dichotomous variables (diagnosis, sur-
gical risk, death, readmissions, and reintervention) 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 in-
terventions were performed in the same patient, we selected the
first one. Longitudinal data analysis was performed in all avail-
able 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 follow-
ing formulae:

\[
\text{SEM} = \frac{\text{SD} \times \sqrt{1 - R}}{	ext{SEM} \times z \text{ score} \times \sqrt{2}}
\]

where SD is the standard deviation of the sample at baseline
and R is the reliability coefficient.22 We used the Cronbach \( \alpha \)
as a reliability measure.23 From the standard error of measure-
ment, 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 corre-
sponds to a z score of 1.96. The MDC represents the smallest
change in score that likely reflects true change more than mea-
surement error alone.24 The MDC proportion is the propor-
tion of patients with changes in scores that exceeded the MDC. The
minimal clinically important difference (MCID) has been de-
efined as the smallest difference between the scores in a ques-
tionnaire 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 ar-
ticulation 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 \( P < 0.05 \),
unless otherwise noted. All statistical analyses were per-
formed using SAS for Windows statistical software, version 8.2

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 re-
ponders and nonresponders at 6 months regarding socio-
demographic variables, main clinical characteristics (in-
cluding 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 proce-
dures had significantly higher improvements in the physical
function, role–physical, bodily pain, and social func-
tion 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 be-
tween 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 hos-
pitals in HRQoL improvement by appropriateness cat-
egories 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 sur-
passed the MDC was higher than 50% for the SF-36 physi-
cal function and the WOMAC pain and functional limi-
tation in both procedures (Table 3). Among patients
who underwent TKR, more than half of patients undergo-
ing THR surpassed the MCID in all SF-36 domains; approxi-
ately 75% did so on all the WOMAC domains. The percentages
were lower for pa-

patients undergoing TKR.

A significantly higher proportion of patients under-
going THR judged as appropriate candidates surpassed
the MDC or MCID than those judged as inappropriate
for all the WOMAC domains and the physi-
cal 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 appropri-
ateness categories for any of the clinical indicators ana-
yzed up to 6 months after discharge, except for the fa-
tality 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 un-
derwent appropriate procedures reported better percep-
tion of their general health status compared with the pre-
vious year, greater relief of symptoms, and greater recovery
than those who underwent inappropriate procedures
(Table 5).
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 outcomes.

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

<table>
<thead>
<tr>
<th></th>
<th>Preintervention</th>
<th>Improvement at 6 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Appropriate (n = 575)</td>
<td>Uncertain (n = 168)</td>
</tr>
<tr>
<td>SF-36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical functioning</td>
<td>17.98</td>
<td>29.05</td>
</tr>
<tr>
<td>Role—physical</td>
<td>6.52</td>
<td>22.71</td>
</tr>
<tr>
<td>Bodily pain</td>
<td>26.13</td>
<td>43.41</td>
</tr>
<tr>
<td>General health</td>
<td>57.31</td>
<td>63.19</td>
</tr>
<tr>
<td>Vitality</td>
<td>39.08</td>
<td>52.37</td>
</tr>
<tr>
<td>Social functioning</td>
<td>50.87</td>
<td>66.91</td>
</tr>
<tr>
<td>Role—emotional</td>
<td>66.75</td>
<td>73.33</td>
</tr>
<tr>
<td>Mental health</td>
<td>57.66</td>
<td>66.03</td>
</tr>
<tr>
<td>PCSS</td>
<td>26.09</td>
<td>30.96</td>
</tr>
<tr>
<td>MCSS</td>
<td>46.16</td>
<td>50.36</td>
</tr>
<tr>
<td>WOMAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>59.61</td>
<td>42.23</td>
</tr>
<tr>
<td>Functional limitation</td>
<td>69.62</td>
<td>55.58</td>
</tr>
<tr>
<td>Stiffness</td>
<td>63.01</td>
<td>50.07</td>
</tr>
</tbody>
</table>

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*

<table>
<thead>
<tr>
<th></th>
<th>Preintervention</th>
<th>Improvement at 6 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Appropriate (n = 557)</td>
<td>Uncertain (n = 137)</td>
</tr>
<tr>
<td>SF-36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical functioning</td>
<td>19.49</td>
<td>33.11</td>
</tr>
<tr>
<td>Role—physical</td>
<td>11.26</td>
<td>20.42</td>
</tr>
<tr>
<td>Bodily pain</td>
<td>31.09</td>
<td>43.37</td>
</tr>
<tr>
<td>General health</td>
<td>56.56</td>
<td>61.89</td>
</tr>
<tr>
<td>Vitality</td>
<td>39.67</td>
<td>50.26</td>
</tr>
<tr>
<td>Social functioning</td>
<td>52.04</td>
<td>65.62</td>
</tr>
<tr>
<td>Role—emotional</td>
<td>61.77</td>
<td>70.86</td>
</tr>
<tr>
<td>Mental health</td>
<td>57.81</td>
<td>62.25</td>
</tr>
<tr>
<td>PCSS</td>
<td>27.63</td>
<td>32.22</td>
</tr>
<tr>
<td>MCSS</td>
<td>45.15</td>
<td>48.62</td>
</tr>
<tr>
<td>WOMAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>57.94</td>
<td>50.26</td>
</tr>
<tr>
<td>Functional limitation</td>
<td>65.41</td>
<td>54.36</td>
</tr>
<tr>
<td>Stiffness</td>
<td>58.93</td>
<td>53.22</td>
</tr>
</tbody>
</table>

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).
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 TKR10,11 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,
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. 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 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.

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 study 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. 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.

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

<table>
<thead>
<tr>
<th>Clinical indicator</th>
<th>Total Hip Joint Replacement</th>
<th>Total Knee Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Appropriate (n = 575)</td>
<td>Uncertain (n = 168)</td>
</tr>
<tr>
<td>Local complications</td>
<td>23 (4.0)</td>
<td>4 (2.4)</td>
</tr>
<tr>
<td>Global complications</td>
<td>25 (4.35)</td>
<td>7 (4.1)</td>
</tr>
<tr>
<td>Length of stay</td>
<td>11.82 (4.31)</td>
<td>11.1 (3.5)</td>
</tr>
<tr>
<td>Reintervention</td>
<td>27 (4.76)</td>
<td>6 (3.64)</td>
</tr>
<tr>
<td>Death</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Patient perception at 6 mo*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Better health status</td>
<td>410 (97.6)</td>
<td>114 (97.4)</td>
</tr>
<tr>
<td>Relief of symptoms</td>
<td>364 (87.29)</td>
<td>104 (88.89)</td>
</tr>
<tr>
<td>Global satisfaction</td>
<td>407 (95.54)</td>
<td>113 (95.76)</td>
</tr>
<tr>
<td>Satisfaction with intervention</td>
<td>423 (99.06)</td>
<td>118 (100)</td>
</tr>
<tr>
<td>Recovered</td>
<td>261 (62.14)</td>
<td>88 (74.58)</td>
</tr>
<tr>
<td>Better health status compared with 1 year previously</td>
<td>354 (83.7)</td>
<td>93 (79.49)</td>
</tr>
</tbody>
</table>

*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 χ² tests except for length of stay, which is from the general linear model.
In conclusion, the results of this study support the predictive validity of our explicit appropriateness criteria by showing a greater benefit in HRQoL 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.

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: José M. Ordoñez, MD, Jose M. Vilarrubias, PhD, Jordi Ballester, PhD, Carlos Barrios, PhD, Mikel Sánchez, MD, Francisco Villar, PhD, Luis Gutiérrez, 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 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.


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.


