Appropriateness of total hip joint replacement

JOSE M. QUINTANA1, ANTONIO ESCOBAR2, JESUS AZKARATE3, J. IGNACIO GOENAGA4 AND AMAIA BILBAO1

1Hospital de Galdakao, Unidad de Investigación, Galdakao, Vizcaya, Spain, 2Hospital de Basurto, Unidad de Investigación, Bilbao, Vizcaya, Spain, 3Hospital de Mendaro, Servicio de Traumatología, Mendaro, Guipuzcoa, Spain, 4Hospital de Santiago, Servicio de Traumatología, Vitoria, Alava, Spain

Abstract

Objective. To evaluate the appropriateness of the use of total hip joint replacements.

Design. Observational study of consecutive patients with a diagnosis of hip osteoarthritis and who had undergone total hip arthroplasty over a 1-year period from seven hospitals.

Main measures. The appropriateness of the use of hip replacement was judged by explicit criteria developed by a panel of experts using RAND methodology. The length of hospital stay during the admission and complications were recorded 6 months post-operatively. Patients were also surveyed 6 months after discharge to determine whether they believed they had recovered or their satisfaction with the intervention. Appropriateness results of this study were compared with a previous study performed with the same criteria 4 years previously.

Results. In total, 784 patients participated in the study. Indications for surgery were considered necessary in 52.2% of cases, appropriate in 21.3%, uncertain in 21.4%, and inappropriate in 5.1%. Differences were found in the rates of appropriateness exclusively from one hospital. At 6 months after discharge, differences between centres were found for the proportion of patients that reported they had recovered from surgery (range 57.7–24.8%) and in the length of hospital stay during admission (range 10–16 days). Improvement in the appropriateness rates were found for all participant hospitals during both periods.

Conclusions. We identified a low percentage of inappropriate indications and differences in some outcomes between centres. Compared with previously, there has been improvement in the use of this technique, although both periods are not methodologically comparable.

Keywords: appropriateness, outcome measurement, total hip replacement, utilization review, variation

Total hip joint replacement is performed increasingly in developed countries [1,2]. At the same time, surgical rates continue to vary across regions of different countries [3,4], which cannot be explained solely by differences in the prevalence of hip disease.

Ideally, health care systems should function such that appropriate care increases and inappropriate care decreases. Reducing overuse of a procedure should enhance the quality of care and decrease medical costs [5]. Central to such an investigation is the determination of what is considered an appropriate indication for any given procedure [6]. Unfortunately, for most conditions something other than rigorous data on efficacy or effectiveness must be used to determine the criteria of appropriateness [7]. A method that combines expert opinion with available scientific evidence was developed by the RAND University of California at Los Angeles group [8]. This method has been used to evaluate the appropriateness of a variety of medical and surgical interventions [9,10].

The goal of the present study was to apply explicit criteria, developed using the RAND appropriateness methodology, to examine the appropriateness of the indications for patients with osteoarthritis undergoing total hip arthroplasty in various hospitals and their outcomes; and to compare the results with previous appropriateness results [11].

Methods

Development of explicit criteria

The criteria for measuring the appropriateness of hip replacement in patients with osteoarthritis were developed according to the previously described RAND appropriateness method [8] according to the following steps. Firstly, we conducted an extensive literature review to summarize the existing knowledge of efficacy, effectiveness, risks, costs, and
opinions about the use of hip replacement in patients with osteoarthritis.

Secondly, based on this review, we developed comprehensive and detailed lists of mutually exclusive and clinically specific scenarios (indications) in which hip replacement might be performed. The list contained 216 scenarios for the indication of total hip arthroplasty. Each scenario contained sufficient detail that patients within a given scenario were reasonably homogeneous. These scenarios included the following variables: age, diagnosis, surgical risk [based on the American Society of Anesthesiologists (ASA) criteria], previous non-surgical procedures performed, pain [based on the need for medication and the effect on pain; relation to rest, and sleep or night disturbance; rhythm; and intensity], and functional limitations assessment [based on the American College of Rheumatology classification, and need for a mobility aid]. The last two variables were categorized as minor, moderate, or severe. Detailed descriptions of the variables and their categories, definitions, and references were published elsewhere [12].

Thirdly, we convened a national panel of 12 specialists (nine orthopaedic surgeons, one rheumatologist, one rehabilitation medical specialist, and one family physician), all of whom had experience in the field. The panelists were provided with the literature review and the list of indications and asked to rate each indication for its appropriateness in performing the procedure, considering the average patient and average physician at the time the panel met in 1997. Appropriateness was defined as indicating that the expected health benefit would exceed the expected risks by a sufficiently wide margin to make each procedure worth performing after choosing the best surgical alternative available for the patient.

Ratings were scored on a nine-point scale. The use of total hip arthroplasty for a specific indication 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, or uncertain if the median rating was between 4 and 6 or if the members of the panel disagreed. Disagreement was defined as occurring when at least one-third of the panelists rated an indication from 1 to 3 and at least another third rated it from 7 to 9, agreement if no more than three panelists rated the indication outside the three-point spread (1–3, 4–6, 7–9) containing the median, and undetermined if neither agreement nor disagreement was found. This method did not attempt to force panelists to reach agreement on appropriateness.

The ratings were confidential and took place in two rounds, using a modified Delphi process. The first round was performed before the panel meeting. The results were collated and presented to the panelists at a meeting on day 1. Each panelist received the anonymous ratings of the other panelists as well as a reminder of his or her own ratings. After extensive discussion, panelists revised the indications according to the definitions. Finally, all scenarios that were considered appropriate during the second round were rated by the panel on a nine-point scale to determine those that were considered necessary. Necessity was defined as meaning that a procedure is not only appropriate but crucial and that it would be improper care not to recommend it in a given clinical situation, as defined by previous authors [13]. Scenarios with a median necessity rating from 7 to 9 and with no disagreement were considered necessary. The other appropriate scenarios were considered elective. A total of 68 scenarios were rated. It was rated only by the nine orthopedic surgeons.

The results of the work of the expert panel were reported previously [14].

Data collection

This prospective observational study was conducted in seven public teaching hospitals belonging to the Basque Health Service—Osakidetza, a local government agency in the Basque Country that is part of the Spanish National Health Service. Physicians in each hospital were blinded to the study goals.

Consecutive patients undergoing total hip arthroplasty, who were followed in any of the seven hospitals, were eligible for the study. Patients with malignant, severe organic, or psychiatric diseases were excluded. Between March 1999 and March 2000, 1495 patients were placed on waiting lists to undergo this technique. Of these, 183 were excluded. Of 1312 who fulfilled the selection criteria, 1138 agreed to participate and completed the questionnaires sent to them before the intervention. Of these, 784, which is the sample in the present study, presented with osteoarthritis and had accessible medical records. The ethics review board at our hospital approved the project.

To collect data and determine appropriateness, we developed a computerized algorithm based on the panel results. We also developed data collection questionnaires to retrieve data from the medical records that included variables before the intervention and at admission and discharge, including the intervention and complications 6 months after discharge. Besides those variables belonging to the appropriateness algorithm, other variables collected included sociodemographic data, local and general complications peri- and post-intervention, readmission, reintervention, death, and length of hospital stay. Six months after discharge, all medical records were reviewed again to determine whether the patient had been readmitted, had 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 the form. The physicians were trained to retrieve the medical record data in a similar way and were tested for agreement among them. They were provided with a manual that included the definitions of all variables and categories. An agreement study of the responses of the reviewers and the members of the research team was performed for the main variables (diagnosis, surgical risk, death, readmissions, reintervention, and length of hospital stay). 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 kappa value for dichotomous variables (diagnosis, surgical risk, death, readmission, reintervention) of >0.99 and an intraclass correlation coefficient for length of hospital stay of 0.99.
Six months after surgery, patients received a questionnaire in which they were asked to rate the status of their hip compared with 1 year previously and to answer questions related to their recovery and global satisfaction with the results of the intervention. The sources of information and time of data collection of the main variables of the appropriateness algorithm and outcomes are summarized in Table 1.

**Statistical analysis**

The unit of study was the patient. In cases in which two interventions were carried out during the study period (eight cases, 1.0%), we included the first intervention. Descriptive statistics included frequency tables and mean and standard deviations. Chi-squared and Fisher’s exact tests were used to test for statistical significance among proportions. For continuous variables (e.g., age), ANOVA and Scheffé’s test for multiple comparisons were performed in the univariate analysis.

Logistic regression models were used to compare hospitals by appropriateness (appropriate versus uncertain and inappropriate). Hospital 2 served as the comparison centre. Logistic regression models also were used to analyse differences between hospitals for the outcomes studied. For those cases, Hospital 3 served as the comparison group. In both cases, the hospital with the worst results was chosen as the reference group. Finally, adjustment by appropriateness was performed in all models.

$P < 0.05$ was considered significant unless otherwise noted. All statistical analyses were performed using SAS for Windows statistical software, version 8.0.

**Results**

During the 1-year recruitment period, 784 patients fulfilled the selection criteria and had complete information on relevant variables. No statistically significant differences were found among the seven hospitals for patient age, sex, percentage of adequate previous treatments performed, pain, or functional limitations. Differences were detected only in surgical risk, which was due exclusively to Hospital 2 with a significantly higher percentage of patients with high (ASA IV) surgical risk (8.9%) compared with Hospitals 4, 5, and 7 (Table 2).

Indications for surgery were considered necessary in 52.2% of cases, appropriate in 21.3%, uncertain in 21.4%, and...
inappropriate in 5.1%. Hospital 2 had 10% of cases classified as inappropriate, which was significantly different from the other centres \( (P = 0.04) \). No other statistically significant differences were found between the hospitals. The inappropriateness rates from five hospitals were below 5%. In 52.2% of cases, the indication was necessary, with only Hospitals 2 and 6 below 50% \( (P < 0.001) \) (Table 3). Compared with the results of a previous study performed at five of the seven hospitals, we found statistically significant changes in the appropriateness rates in hospitals between the two study periods \( (P < 0.001) \) for all five comparisons except in Hospital 2 and 5 where differences in the inappropriate categories were not found between the two periods.

The most frequently encountered scenarios in the field study were patients with severe pain and moderate-to-severe functional limitations, all with low surgical risk, and classified as appropriate as defined by our panelists. Uncertain cases were those patients classified as having moderate pain and functional limitations. On the other hand, the inappropriate scenarios were those with minor-to-moderate pain and functional limitations (Table 4). Of the 216 theoretical scenarios, 81 \( (37.50\%) \) were used in the field study.

No differences were found between hospitals in local or general complications peri- or post-operatively until 6 months after discharge. In addition, no differences were found between hospitals regarding death, readmission, or reintervention rates.

Regarding other outcomes, we found statistically significant differences between centres (Table 5). Differences were found in the length of hospital stay, which ranged from 10 (Hospital 7) to 16 (Hospital 3) days. Statistically significant differences by Scheffé’s test were found between Hospital 1 and 2 and 5 where differences in the inappropriate categories were not found between the two periods.

### Table 3

Frequencies and percentages (in parentheses) of total hip arthroplasty appropriateness by hospital

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Necessary</th>
<th>Appropriate</th>
<th>Uncertain</th>
<th>Inappropriate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (n = 130)</td>
<td>68 (52.3)</td>
<td>24 (18.5)</td>
<td>30 (23.1)</td>
<td>8 (6.2)</td>
</tr>
<tr>
<td>2 (n = 90)</td>
<td>43 (47.8)</td>
<td>19 (21.1)</td>
<td>19 (21.1)</td>
<td>9 (10.0)</td>
</tr>
<tr>
<td>3 (n = 41)</td>
<td>22 (53.7)</td>
<td>10 (24.4)</td>
<td>7 (17.1)</td>
<td>2 (4.9)</td>
</tr>
<tr>
<td>4 (n = 152)</td>
<td>78 (51.3)</td>
<td>31 (20.4)</td>
<td>37 (24.3)</td>
<td>6 (4.0)</td>
</tr>
<tr>
<td>5 (n = 149)</td>
<td>77 (51.7)</td>
<td>34 (22.8)</td>
<td>32 (21.5)</td>
<td>6 (4.0)</td>
</tr>
<tr>
<td>6 (n = 56)</td>
<td>26 (46.4)</td>
<td>15 (26.8)</td>
<td>14 (3)</td>
<td>1 (1.8)</td>
</tr>
<tr>
<td>7 (n = 166)</td>
<td>95 (57.3)</td>
<td>34 (20.5)</td>
<td>29 (17.5)</td>
<td>8 (4.8)</td>
</tr>
</tbody>
</table>

\( P\) value (differences between studies) \( <0.001 \)

Chi-squared test for the comparison within hospitals between the two periods \( (1997 \text{ versus } 1999–2000) \).

### Table 4

More frequently encountered scenarios in the field study, by appropriateness

<table>
<thead>
<tr>
<th>Number</th>
<th>Scenario</th>
<th>Appropriateness evaluation</th>
<th>No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Patient 50–70 years of age, surgical risk low, pain severe, functional limitations moderate to severe, and adequate previous treatments</td>
<td>Appropriate</td>
<td>152 (19.4)</td>
</tr>
<tr>
<td>2</td>
<td>Patient 50–70 years of age, surgical risk low, pain severe, functional limitations moderate, and inadequate previous treatments</td>
<td>Appropriate</td>
<td>49 (6.3)</td>
</tr>
<tr>
<td>3</td>
<td>Patient &gt;70 years of age, surgical risk low, pain severe, functional limitations moderate to severe and adequate previous treatments</td>
<td>Appropriate</td>
<td>114 (14.5)</td>
</tr>
<tr>
<td>4</td>
<td>Patient &gt;70 years of age, surgical risk low, pain severe, functional limitations moderate, and inadequate previous treatments</td>
<td>Appropriate</td>
<td>51 (6.5)</td>
</tr>
<tr>
<td>5</td>
<td>Patient 50–70 years of age, surgical risk low, pain and functional limitations moderate</td>
<td>Uncertain</td>
<td>39 (5.0)</td>
</tr>
<tr>
<td>6</td>
<td>Patient &gt;70 years of age, surgical risk low, pain and functional limitations moderate</td>
<td>Uncertain</td>
<td>36 (4.6)</td>
</tr>
<tr>
<td>7</td>
<td>Patient 50–70 years of age, surgical risk low, pain and functional limitations minor or moderate, and inadequate previous treatments</td>
<td>Inappropriate</td>
<td>20 (2.6)</td>
</tr>
</tbody>
</table>

Included all theoretical scenarios with a prevalence in the field study of >4%, except for those inappropriate (>2%).
Discussion

This study provides information on the appropriateness of the use of hip prosthesis among seven hospitals. We based our judgement of the appropriateness of the indication for the intervention on the explicit criteria developed by an expert panel based on RAND Appropriateness Methodology [8]. The percentage of interventions classified as inappropriate was low. Compared with previous studies in which RAND methodology had been used [15,16], the rates we found were considerably lower. Furthermore, our rates are now lower than those reported in a previous study that used similar methodology [11].

In 2000, our research group published data on variations in appropriateness of the use of hip prosthesis among five hospitals [11]. In that study, performed in 1996, the highest percentage of cases considered inappropriate was found for those with a diagnosis of osteoarthritis, which was 13.6% of the whole sample. Those classified as uncertain were also higher than in the current study (46.2% versus 21.4%). Differences between centres were also found in that study, with inappropriate rates ranging from 6.7% to 16.3% and uncertain from 42.3% to 50%. Another goal of the present study, in which we applied the same explicit criteria, was to determine whether the variations remained during the 4 years since the first study and whether important changes between the two periods could be detected. The good news is that both the rates of uncertain and inappropriate cases and the variations between hospitals decreased. However, the question of whether there were any differences between the studies remains.

We applied the same explicit criteria in both studies, and the information needed to apply the algorithm of appropriateness was similar, but the methods of data collection differed slightly. From the variables needed, diagnosis, surgical risk, and age were collected from the medical records in both studies. In the first study, a trained interviewer collected the information during direct interviews with the patients and classified the patients in one category of pain and functional limitation. However, in the present study, we surveyed the patients by mail in a standardized manner to collect information related to their pain and functional limitations level. Both methods have advantages and disadvantages. However, because there were some differences in the methodologies in the two studies and we did not evaluate the implications of...
those changes, we have to conclude that the studies are not totally comparable. Nevertheless, we understand that our results indicate a trend.

Another question remains about whether any explicit intervention was performed in the interval between the two studies. The conclusions and recommendations of our work in 1996 and the results of our field study were delivered to orthopaedic surgeons, care providers, and buyers in our area. Although they were partially followed, we have not explicitly tested this point. Additionally, we used an observational design, not a clinical trial, which would have provided better evidence. Nevertheless, we believe that a portion of the improvement that occurred over the past 4 years could be the result of our recommendations, although this was not specifically evaluated.

We also found differences between hospitals in the proportion of patients who reported that they had recovered 6 months after the intervention, in the length of the hospital stay, and in the proportion of patients who reported that they were recovered 6 months after the intervention based on the appropriateness classification of the indication. Even after adjustment by appropriateness, differences between hospitals remained. Differences in outcomes by appropriateness, favouring those that were appropriate, indicated that our appropriateness criteria have some predictive validity. The fact that differences between hospitals remained even after adjustment could be due to differences in surgical ability between orthopaedic surgeons or differences in care between centres, but neither was evaluated in this study.

A few studies performed in the 1980s and 1990s focused on the appropriateness of hip replacement but used different methodology [13,17,18]. Since then, only one Canadian study has focused on the appropriateness of hip replacement [19]. Those investigators found an appropriateness rate of 93%. To determine the appropriateness of an indication, the authors used non-publicly published criteria, from which few and poor validity data are easily available [20]. Appropriateness was judged by a binary response: appropriate or not. Their sample size was also smaller (n = 488 patients), and the response rate was low (44%). Our rate of inappropriate cases was even lower than the one in that study.

Our study had some limitations. Limitations related to the work of the panel has been described in a previous study [14]. Data collection also had some limitations. The three blinded reviewers were physicians trained to assess and record the main variables of the algorithm in a standardized manner to reduce the chances of bias. We checked their reliability with good results, and we evaluated their accuracy by checking some variables, also with excellent correlations. However, the quality of the data of some important variables necessary for the appropriateness algorithm might be questionable when depending exclusively on the medical record. In our case, the most relevant variables (age, surgical risk, diagnosis, comorbidities) were properly recorded.

Our algorithm, used as the current clinical decision-making tool in this case, is based fundamentally on pain and the functional limitations assessment of the patient. We collected that information directly from the patient using a standardized questionnaire.

This study was performed in seven public hospitals of the Spanish health care system, a public, universal-coverage system. The generalizability of our results to other health care systems is questionable, although it is likely that results from other Spanish centres may be similar.

In summary, this is an appropriateness evaluation study of patients with osteoarthritis who underwent total hip arthroplasty. Minor variations in the levels of appropriate use of hip replacement surgery still persist. Of note was an important reduction in the number of interventions that were judged inappropriate found in hospitals when we compared two time frames, which could be indirectly linked to the strategies developed by local buyers of services. Future studies should explicitly evaluate strategies implemented to reduce variation or inappropriateness, with adequate designs and measured outcomes that are relevant to clinicians and patients.

Acknowledgements

We thank the following physicians for their contribution to this study: Angel Alfageme, Jose M. Aranburu, Jesus Azkoaga, Pedro Armandariz, Enrique Cáceres, Arsenio Diego, Begoña Goicoetxea, Iñigo Guisasola, Manuel Martínez-Grande, Enrique Queipo de Llano, Ramón Tobio, and Jose Villar (total hip arthroplasty panel). We are also grateful to A. Higelmo, I. Lafuente, A. Rodriguez, and I. Vidaurreta for their contribution to the development of the panel of expert, data retrieval, and data entry. We thank the staff members of the various services, research and quality units, as well as the medical records sections of the participating hospitals. This study was supported by grants from the Fondo de Investigación Sanitaria (98/0001-01 to 03), the Department of Health of the Basque Government, Osakidetza-Servicio Vasco de Salud, and the Research Committee of Galdakao Hospital.

References


Accepted for publication 22 March 2005