Rheumatology 2009;48:1402–1409
Advance Access publication 2 September 2009
doi:10.1093/rheumatology/kep264

Decision trees for indication of total hip replacement on patients with osteoarthritis

José M. Quintana[1], Amaia Bilbao[2], Antonio Escobar[3], Jesus Azkarate[4] and Jose I. Goenaga[5]

Objective. To develop a decision tree based on health-related quality of life outcomes rather than expert consensus for determining the appropriateness of total hip replacement (THR) among patients with hip OA.
Methods. This is a prospective observational study of two independent cohorts. The derivation cohort included 590 patients recruited from seven hospitals between March 1999 and March 2000. The validation cohort included 339 patients recruited from six hospitals between September 2003 and September 2004. Socio-demographic and clinical data were collected for the participants, all of whom completed the WOMAC before hip replacement and 6 months later. Univariate and Regression Trees, by classification and regression trees (CART), analyses were performed in the derivation cohort. The decision trees derived in the derivation cohort were validated in the validation cohort.
Results. Main variables that predicted change in the WOMAC pain and functional limitation domains were pre-intervention pain or functional limitation and the application of non-surgical treatments. CART analysis showed that when pre-intervention pain was classified as minor, or WOMAC pain or functional limitation scores were $\leqslant 40$, there was an odds ratio of 0.076 (95% CI 0.031, 0.185) of having an expected gain after THR in the WOMAC pain domain of >30 or >25 in the WOMAC functional limitation domain.
Conclusions. A simple decision tree based on WOMAC outcomes can help to determine the appropriate application of THR. It could also be used to evaluate clinical practice or for quality control.

Key words: Hip osteoarthritis, Hip replacement, Quality of life, Appropriateness, Decision trees.

Introduction

OA of the hip is one of the most prevalent chronic conditions, affecting as many as 7 in 100 older adults [1–3]. Total hip replacement (THR) is frequently performed to alleviate the pain and disability associated with hip OA [4]. Substantial variability exists in the performance of this procedure [5–7]. In an effort to reduce this variation and improve health care quality, various medical organizations have proposed guidelines for the indication of THR [8, 9], and diverse research teams have developed explicit criteria for determining the appropriateness of THR. These efforts are generally based on the work of expert panels using different methodologies [5, 10, 11].

Various clinical parameters have been used to evaluate the effectiveness of THR. There is growing interest in the use of self-administered questionnaires, usually measured before and after the intervention. Although early studies used generic instruments such as the Medical Outcomes Study 36-item Short-Form Health Survey (SF-36), the use of disease-specific instruments such as the WOMAC may be more accurate [12, 13].

The goal of this study was to develop, and validate, explicit criteria for the indication of THR in patients with hip OA based on changes measured by a disease-specific instrument before and after THR in two independent, prospective cohorts.

Methods

Two prospective cohorts were recruited from several public teaching hospitals affiliated with the Basque Health Service-Osakidetza, a local government agency in the Basque Country, which is part of the Spanish National Health Service. The Basque Health Service provides free unrestricted care to nearly 100% of the population. These hospitals serve a population of 2 million inhabitants. The ethics review board of each hospital approved the study. All patients gave their informed consent before participating in the study.

One cohort (the derivation cohort) was used to develop the appropriateness criteria. Consecutive patients scheduled to undergo THR in any of seven hospitals between March 1999 and March 2000 were eligible for the study.

The other cohort (the validation cohort) was used to validate the appropriateness criteria. Consecutive patients scheduled to undergo THR in any of six hospitals between September 2003 and September 2004 were eligible for the study. In both cohorts, patients with a malignant pathology or other organic or psychiatric condition that rendered them unable to participate or to complete the questionnaires were excluded. Data collection and methodology for both cohorts were the same.

We collected data from the hospital medical records and directly from the patients. To retrieve data from the medical records, we developed standardized data collection questionnaires that included socio-demographic data, the primary patient complaint, American Surgical Association (ASA) surgical risk [14], bone quality [15], previous non-surgical treatments (correctly done if aspirin, acetaminophen, NSAIDs or narcotic analgesics at regular doses during 6 months with no pain relief; weight control treatment, if overweight; and physical therapies done) [16], presence of contralateral hip, knee or ankle OA, weight, height and comorbidities, including all those on the Charlson Comorbidity Index [17].

In each cohort, all patients on the waiting list for THR were sent a letter informing them of the study and asking for their voluntary participation. Also included in the mailing were the WOMAC questionnaire [18] and additional questions comprising a short form regarding level of pain, pain control and functional limitation. A reminder letter was sent to patients who had not replied after 15 days. Those who had still not responded after an additional 15 days were sent the questionnaires again, and were contacted by telephone. This procedure was repeated 6 months after the intervention.

[1]Unidad de Investigación, Hospital de Galdakao – CIBER Epidemiología y Salud Pública (CIBERESP), Galdakao, [2]Fundación Vasca de Innovación e Investigación Sanitarias (BIOEF) – CIBER Epidemiología y Salud Pública (CIBERESP), Sondika, [3]Unidad de Investigación, Hospital de Basurto – CIBER Epidemiología y Salud Pública (CIBERESP), Bilbao, [4]Servicio de Traumatología, Hospital de Mendaro, Mendaro, Guipuzkoa and [5]Servicio de Traumatología, Hospital de Santiago, Vitoria, Alava, Spain.

Submitted 6 April 2009; revised version accepted 23 July 2009.

Correspondence to: José M. Quintana, Unidad de Investigación, Hospital Galdakao-Usansolo, Barrio Labeaga s/n, 48960 Galdakao, Vizcaya, Spain.
E-mail: josemaria.quintanalopez@osakidetza.net

The WOMAC is a disease-specific, self-administered questionnaire developed to evaluate patients with hip or knee OA [18]. It uses a multi-dimensional scale composed of 24 items grouped into three dimensions: pain (5 items), stiffness (2 items) and physical function (17 items). We used the categorical version with five response levels representing different degrees of intensity (none, mild, moderate, severe or extreme) for each item. Responses were scored from 0 (none) to 4 (extreme). The higher the score, the poorer the quality of life. The data were standardized to a range of values from 0 to 100, where 0 represents the best health status possible and 100 the worst. Missing data were dealt with according to the 'half-scale' method. Over time, a reduction in the overall score represents an improvement. The original questionnaire is reliable, valid and sensitive to changes in the health status of patients with hip or knee OA [18]. The WOMAC has been translated into Spanish and validated in Spain [19].

The mailing that each patient received also included questions regarding level of pain, pain control and functional limitation, which we will refer to as the short scales. The structure of those variables has been described previously [16] but, in short, the pain short scale included four questions (about the need for medication and the effect on pain; relation to rest and sleep or night disturbance; rhythm; and intensity), and the functional limitations another three (based on the ACR classification [20], and need for a mobility aid).

Statistical analyses

The unit of study was the patient. In cases where a patient received two interventions during the recruitment period, we selected the first one performed.

To describe the samples, we used means and S.D., as well as frequencies and percentages. Within each cohort, we compared socio-demographic data, clinical data and domains of the WOMAC questionnaire at baseline between patients who responded to the questionnaires 6 months after THR and those who did not. We also compared those characteristics between the derivation and validation cohorts. Chi-square or Fisher's exact tests were performed to compare categorical variables and the *t*-test or Wilcoxon non-parametric test was used to compare continuous variables.

Data from the derivation cohort were used to identify variables that predicted the appropriateness of THR based on changes in the WOMAC. We performed univariate general linear models, setting two dependent variables: change from baseline to 6 months in the WOMAC pain domain and change from baseline to 6 months in the WOMAC functional limitation domain. Independent variables were age, gender, bone quality (normal or deficient), surgical risk (ASA I–III or IV), Charlson Comorbidity Index (0, 1 or >1), previous non-surgical treatment (performed adequately or inadequately), contralateral hip or knee arthritis and pre-intervention pain and functional limitation. We used two alternative measures for pre-intervention pain and functional limitation: (i) pain and functional limitation short scales [16], or (ii) WOMAC pain and functional limitation domains. The β-coefficient estimated and R^2 are provided for each independent variable.

The derivation cohort was used to compile algorithms in decision tree form by means of classification and regression trees (CART) analysis. We performed different regression trees depending on the dependent variable employed: the change in WOMAC pain domain or the change in WOMAC functional limitation domain. In both cases, the independent variables were those identified as significant in the general linear models. Pain and functional limitation at baseline were measured in two different ways: using short scales, or using the WOMAC pain and functional limitation domains at baseline. Thus, finally four regression tress were derived: (i) regression tree for change in WOMAC pain domain considering pain and functional limitation short scales at baseline as independent variables; (ii) regression tree for change in WOMAC pain domain considering pain and functional limitation WOMAC scales at baseline as independent variables; (iii) regression tree for change in WOMAC functional limitation domain considering pain and functional limitation short scales at baseline as independent variables; and (iv) regression tree for change in WOMAC functional limitation domain considering pain and functional limitation WOMAC scales at baseline as independent variables. For each node of the decision trees, we provided the estimated mean change and corresponding 95% CI, as well as the sample size in each node with the percentage in relation to the entire corresponding sample. The regression trees were validated in the validation cohort by estimating the same information as before for each node of the trees, and comparing the results of each node with those obtained in the derivation cohort by means of the *t*-test or the Wilcoxon non-parametric test.

From the various trees, we constructed a summary tree presented as a flow chart for determining the appropriateness of THR. In this context, we defined appropriateness as the gain in terms of functionality and reduction of pain that exceeds the cut-off points of the minimal clinically important difference (MCID) established for this type of procedure [21, 22]. We compared the appropriateness indication tree with the actual results, setting cut-points of appropriate gain based on a prior study of ours establishing MCID for THR [21], and on MCID by appropriateness categories as developed by a panel of experts [22]. These cut-points were used as references for considering as appropriate a final node tree if the gain was $\geqslant 40$ for WOMAC pain or functional limitation domains. We compared those cases categorized as inappropriate in the proposed tree *vs* the other categories with the actual gain in WOMAC pain (>30) and functional limitations (>25) domains at 6 months after the intervention, as cut-offs for uncertain or appropriate interventions. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and odds ratio (OR), with their 95% CIs, as well as area under the receiver operator characteristic curve (AUC) were estimated in both cohorts.

All effects were considered as statistically significant at $P < 0.05$. All statistical analyses were performed using SAS for Windows statistical software, version 9.1 (SAS Institute, Carey, NC, USA), except the CART analysis where we employed the S-Plus 2000 (MathSoft, 1999).

Results

For the derivation cohort, 1495 patients were placed on waiting lists to undergo THR while 634 were for the validation cohort. The selection process and losses on follow-up is displayed as supplementary data (supplementary Figure 1) available at *Rheumatology* Online. Differences in socio-demographic, clinical and health-related quality of life variables between responders and non-responders were not observed in both cohorts, except for the Charlson Comorbidity Index and presence of contralateral hip OA in the derivation cohort, and for the Charlson Comorbidity Index and the BMI in the validation cohort. Statistically significant differences were observed between the derivation and validation cohorts in bone quality, WOMAC scales of pain and functional limitation, and the functional limitation short scale, with generally poorer results observed in the validation cohort (see supplementary Table 1 available as supplementary data at *Rheumatology* Online).

In the univariate analysis in the derivation cohort, variables that significantly predicted change in the pain domain of the WOMAC included the ASA surgical risk, previous non-surgical treatments performed adequately, pre-intervention pain or functional limitation scores of the WOMAC domains, pain short scale and functional limitation short scale (only between severe and

TABLE 1. Univariate analysis ($n = 590$) of derivation cohort

	Changes in WOMAC pain domain at 6 months			Changes in WOMAC functional limitation domain at 6 months		
	β-Parameter	*P*-value	R^2, %	β-Parameter	*P*-value	R^2, %
Age	0.08	0.4855	0.08	0.18	0.0934	0.49
Age			0.15			0.62
50–70 *vs* <50	3.86	0.5116		9.04	0.1133	
>70 *vs* <50	4.92	0.4032		10.41	0.0684	
Gender			0.03			0.003
Men *vs* women	−0.77	0.6784		−0.22	0.9046	
Bone quality			0.001			0.04
Normal *vs* deficient	0.18	0.9437		1.17	0.6359	
Surgical risk			0.67			0.35
ASA IV *vs* ASA I–III	−13.85	0.0500		−9.24	0.1586	
Charlson Comorbidity Index			0.17			0.28
1 *vs* 0	−0.45	0.8306		−1.92	0.3504	
>1 *vs* 0	2.41	0.3794		1.50	0.5688	
Previous non-surgical treatment			2.06			0.45
Adequate *vs* inadequate	6.45	0.0005		2.92	0.1096	
Contralateral hip OA	−1.30	0.4918	0.08	−3.86	0.0346	0.79
Contralateral knee OA	−6.19	0.0567	0.64	−6.35	0.0472	0.70
WOMAC pre-intervention pain			38.76			14.32
40–60 *vs* 0–40	14.77	<0.0001		10.91	<0.0001	
60–80 *vs* 0–40	29.85	<0.0001		18.84	<0.0001	
80–100 *vs* 0–40	47.10	<0.0001		26.36	<0.0001	
WOMAC pre-intervention functional limitation			25.59			26.27
40–60 *vs* 0–40	13.85	<0.0001		14.79	0.0052	
60–80 *vs* 0–40	25.20	<0.0001		26.53	0.0001	
80–100 *vs* 0–40	39.53	<0.0001		39.02	<0.0001	
Pain short scale			12.53			5.91
Moderate *vs* minor	16.29	0.0006		14.00	0.0034	
Severe *vs* minor	30.49	<0.0001		22.24	<0.0001	
Functional limitation short scale			5.55			2.92
Moderate *vs* minor	4.43	0.1467		6.64	0.0263	
Severe *vs* minor	14.23	<0.0001		12.02	<0.0001	

Models performed by means of general linear models.

minor categories). The same variables were significantly related to change in the WOMAC functional limitation domain with the exception of surgical risk and previous non-surgical treatments performed adequately (Table 1). Other variables were studied such as having contralateral hip or knee OA, BMI, age or gender. Relation with changes in the two areas of the WOMAC was not observed for them, except for patients with contralateral hip or knee OA who improved significantly less (4 and 6 points less) in the WOMAC functional limitation domain than those without OA on those joints.

The CART analysis included only those variables that had a statistically significant impact in the univariate analysis. An improvement higher than 30 points in WOMAC pain scores after THR surgery was observed for patients with pre-intervention pain classified as severe or moderate but with functional limitation between moderate to severe according to the short scales (Fig. 1a) or pre-intervention WOMAC pain scores >40 (Fig. 1b). Improvement higher than 25 points in the WOMAC functional limitation domain after THR were found in patients with moderate pain and moderate to severe functional limitation (Fig. 2a). The gains were clinically relevant for pre-intervention WOMAC functional limitation domain values >60 or between 40 and 60 accompanied by a level of pain >40 (Fig. 2b). The results generated in the validation cohort were similar to those of the derivation cohort (Figs 1 and 2). Statistically significant differences were not observed in any node for mean WOMAC changes between the derivation and validation cohorts except for pain changes based on WOMAC pre-intervention scales for those with pre-intervention pain $\leqslant 40$ and functional limitation between 60 and 100. In this case, the mean score was 28.85 in the derivation cohort and 21.67 in the validation cohort ($P = 0.011$).

Also, though ASA classes did not provide with relevant information about the decision-making process, some differences in ASA results for those with severe levels of functional limitation were found in trees to predict changes on functional limitation based on pre-intervention WOMAC scores. Those with ASA IV got always poorer gains (12.35, $n = 5$) than those classified as ASA I–III (54.77, $n = 107$) but just in the derivation cohort. Such results were not replicated in the validation cohort (55.83, $n = 6$ *vs* 54.09, $n = 87$). ASA IV results were based on very small sample sizes, with high variability.

A flow chart for determining the appropriateness of THR is presented in Fig. 3. We estimated the precision parameters of this tree for interventions classified as inappropriate *vs* the other categories compared with the actual gain in WOMAC pain (>30) and functional limitation (>25) domains at 6 months after the intervention (Table 2). For the derivation cohort, the sensitivity and specificity were 91.94 and 45.80%, respectively. In the validation cohort, sensitivity and specificity were 94.96 and 41.03%, respectively.

Discussion

Data from two independent prospective cohorts were used to develop and validate decision trees for the appropriateness of THR among patients with hip OA based on clinical parameters and changes in pain and functional limitations 6 months after the intervention. Data from CART analyses were synthesized into a flow chart (Fig. 3) that represents a proposed decision tree for determining the appropriateness of THR that would be easy to use in the clinic and in research.

To the best of our knowledge, this is the first time that clinical and objective data generated from prospective studies have been used to develop appropriateness criteria for THR. Prior efforts to develop such criteria have generally relied on expert opinion rather than outcomes. Some investigators used a modified Delphi process [5, 11]. Others, including our group, used the well-known Research and Development-University of California Los Angeles appropriateness methodology [23]. Some clinical guidelines, such as those established by the US National

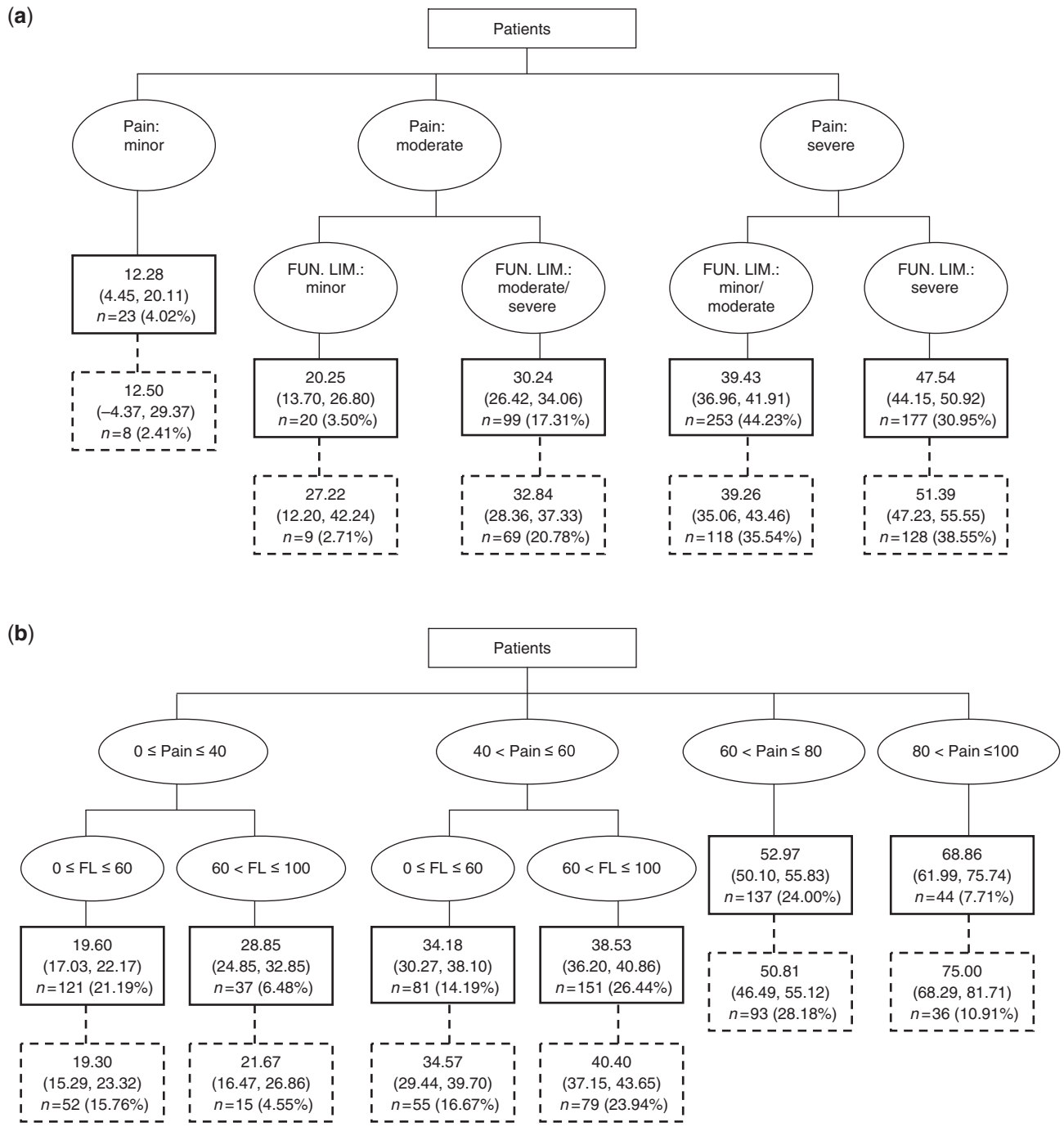

FIG. 1. Regression trees for changes in the WOMAC pain domain after THR in the derivation (solid-line rectangles) and validation (dotted-line rectangles) cohorts based on pre-intervention pain and functional limitation short scales (**a**) and pre-intervention WOMAC scores (**b**). Information included in final squares, from above to below: mean change in WOMAC pain domain after THR for that node; 95% CI for mean change in WOMAC pain domain after THR for that node; sample size in that node and, in brackets, percentage of the entire corresponding sample. FUN. LIM.: functional limitation; FL: WOMAC functional limitation.

Institutes of Health [9], are based on the opinion of experts. Although this methodology is typically classified as low quality evidence, it has frequently been the only source of criteria available.

Clinical trials, which are regarded as the gold standard for generating high-quality evidence, are not always practical for developing appropriateness criteria due to the homogeneous selection of participants and the rigorous methodology, both of which can make the results difficult to generalize to usual populations and usual clinical care. Cohort studies with appropriate follow-up and control of missing data can also yield high-quality evidence, which is why we chose prospective cohorts for this study.

Another methodological difficulty in deriving appropriateness criteria is choosing the most appropriate statistical techniques to maximize the potential of the data generated by a prospective cohort design and presenting the results in a practical way that makes clinical sense. Predictive models were classically developed through logistic regression models and, later, through CART analysis [24]. CART analysis is useful mainly because it allows for the presentation of results in the form of understandable decision

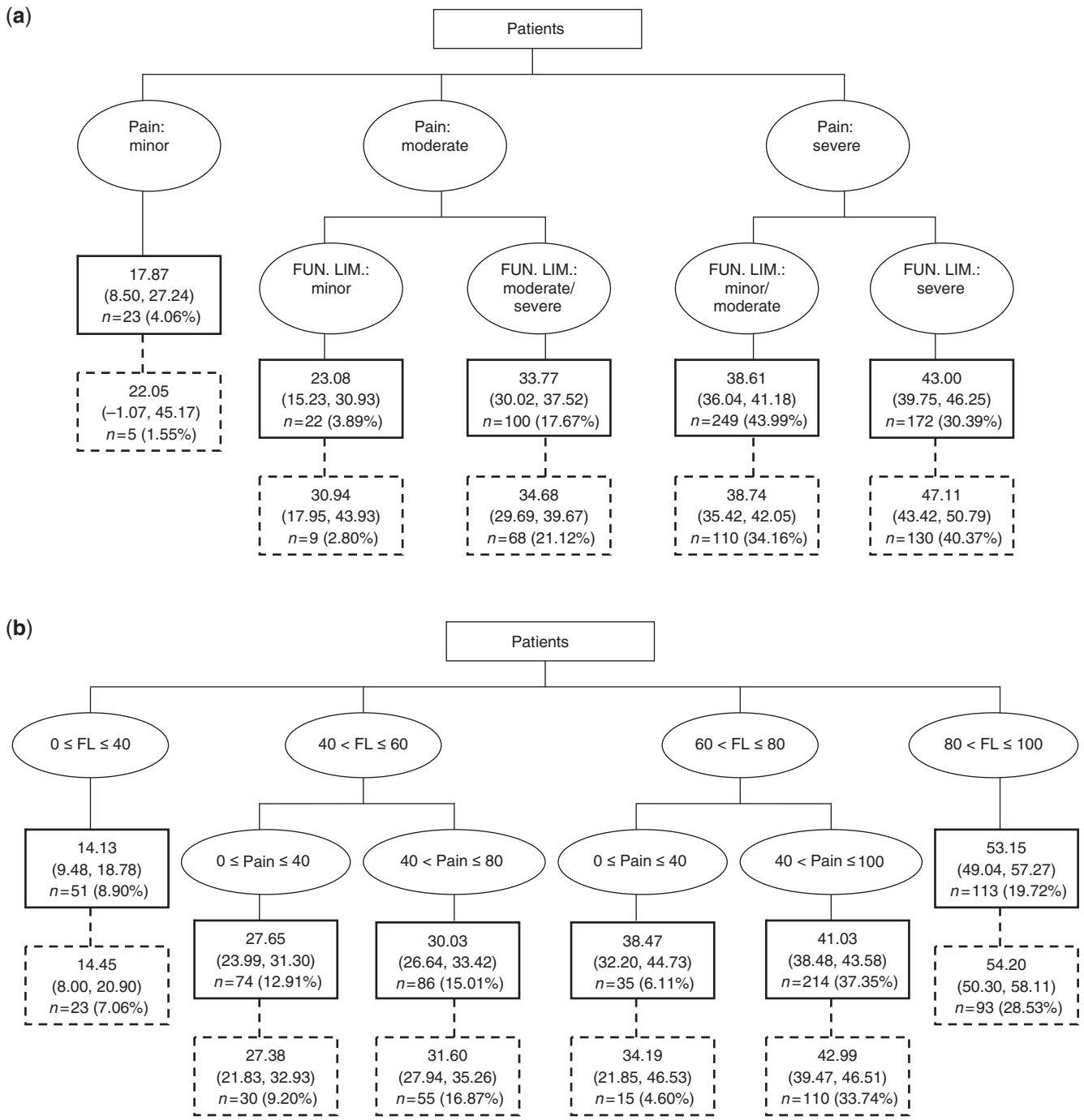

FIG. 2. Regression trees for changes in the WOMAC functional limitation domain after THR in the derivation (solid-line rectangles) and validation (dotted-line rectangles) cohorts on pre-intervention pain and functional limitation short scales (**a**) and on pre-intervention WOMAC scores (**b**). Information included on final squares, from above to below: mean change in WOMAC functional limitation domain after THR for that node; 95% CI for mean change in WOMAC functional limitation domain after THR for that node; and sample size in that node and, in brackets, percentage of the entire corresponding sample. FUN. LIM.: functional limitation; FL: WOMAC functional limitation.

trees. In some cases, the combination of both statistical techniques can yield the most useful results [25].

To construct our decision trees, we first evaluated variables that might explain changes in two domains of the WOMAC—pain and functional limitation—before THR and 6 months later. Based on statistical significance, R^2 values, and clinical coherence, changes in pain and functional limitation, measured either via the WOMAC or with more simple questions, were the main variables we took into account. When constructing our decision trees we also took into account other variables such as age, BMI and comorbidities (using the Charlson Comorbidity Index). None of these variables, however, improved the construction of decision trees using CART analysis. An additional variable that did improve the performance of the decision trees was the performance of previous non-surgical treatments (rehabilitation, obesity control or medical therapy). However, we chose to exclude this variable from the decision trees since it seems that it should have its place just before the THR decision making process starts and as prerequisite. As a system to weight the comorbidities we used the Charlson Comorbidity Index and the ASA. Both are systems

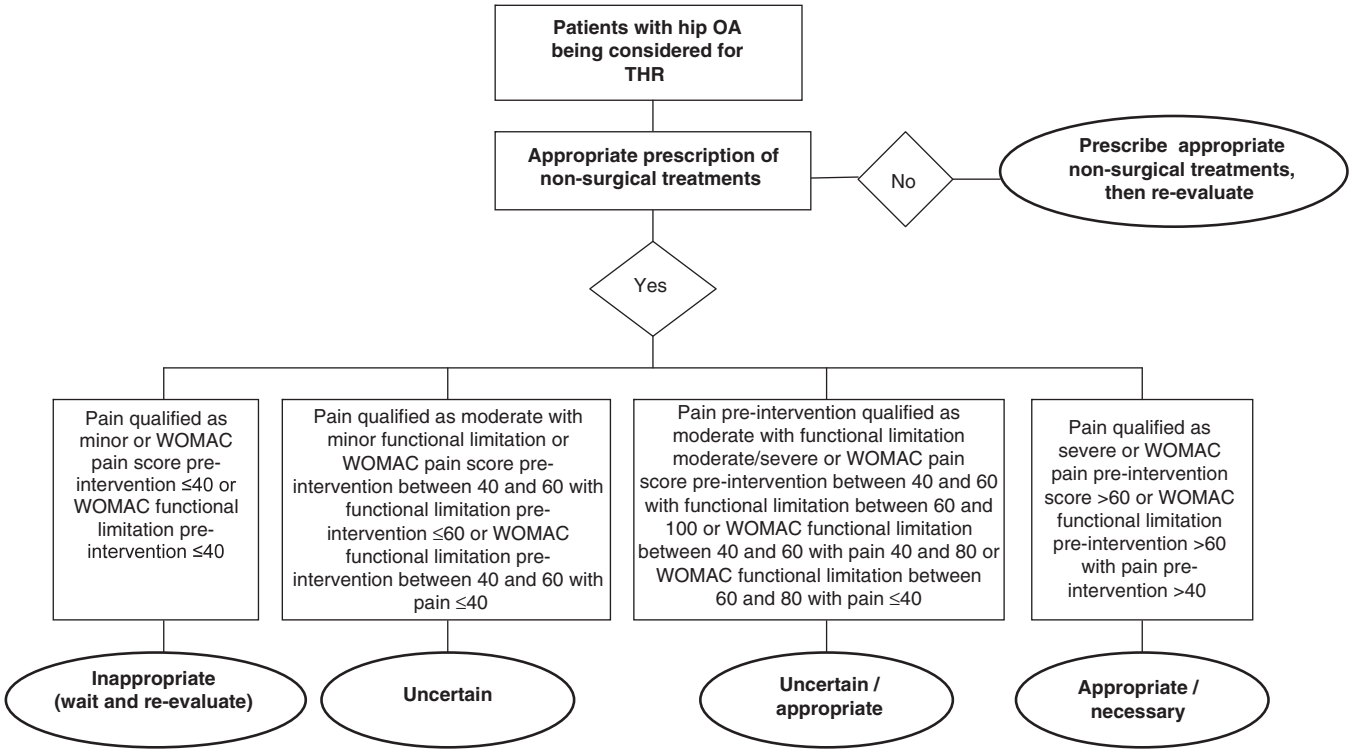

FIG. 3. Proposed decision tree for evaluating the appropriateness of THR for patients with hip OA. The recommendations of this decision tree should not be taken as a definite guide but in context and depending on other patient circumstances.

TABLE 2. Sensitivity and specificity of the final proposed appropriateness algorithm comparing with the appropriateness based on the MCID values

	Is the gain in WOMAC domains at 6 months after the intervention >30 for pain or >25 for functional limitation?					
	Derivation cohort			Validation cohort		
	Yes	No	Total	Yes	No	Total
Proposed decision tree for appropriateness						
Appropriate or uncertain	194	71	265	132	46	178
Inappropriate	17	60	77	7	32	39
Total	211	131	342	139	78	217

Values in cells represent frequencies. Derivation cohort: sensitivity 91.94%; specificity 45.80%; PPV 73.21%; NPV 77.92%; OR 9.64; 95% CI 5.28, 17.63; AUC 0.69. Validation cohort: sensitivity 94.96%; specificity 41.03%; PPV 74.16%; NPV 82.05%; OR 13.12; 95% CI 5.42, 31.75; AUC 0.68.

focused on mortality risk. Maybe the use of other comorbidities indexes could have given different results. In any case, we showed certain influence of the ASA in the results. Patients with ASA IV tend to experience smaller improvements after the intervention, although they can be extremely relevant for their quality of life.

Those variables not included in the trees, as ASA or presence of contralateral hip or knee OA, though significantly associated to the changes in one of the WOMAC domains, were not considered appropriate to be included in the decision tree, though physicians and patients should be aware of their influence in the results.

Support to the validity of our results is also given by the fact that they match those found previously by other authors [11], as well as by our group [23], being based on panels of experts work. The variables selected and the final recommendations are quite similar in all cases. But this study contributes with a greater level of evidence because it is based on a prospective cohort design, on a robust statistical analysis and methodology, with validation of the results in a second cohort. The structure of our proposed appropriateness decision tree partially matched that of other previously developed tool by our group for prioritization [26] but their goals (appropriateness *vs* priority), development (panel of experts *vs* prospective cohort study) and variables differ.

Missing data are a key limitation of the prospective cohort design. In our study, approximately one-fourth of the patients who fulfilled the eligibility criteria and who completed the baseline questionnaires did not respond to the follow-up questionnaires 6 months after THR (72.2% response rate in the derivation cohort, and 60.8% in the validation cohort). These losses occurred despite our sending up to two mailed reminders and contacting non-respondents by telephone. No differences were observed in relevant variables when we compared responders with non-responders. Thus, although a bias may be present in our study due to missing data, it is likely to be minor and we believe the results can be generalized to the entire sample. It is possible that waiting for 6 months to assess the impact of THR may also have biased the results. Although several studies have demonstrated that improvement can be seen clearly at 6 months [21, 27–29], some investigators suggest a longer follow-up period [30].

The WOMAC provides valuable information, and has been recommended as the correct instrument for evaluating changes in studies such as this [13, 31]. Yet, it also has some inherent problems, such as ceiling or floor effect problems. When working with the WOMAC and other instruments, demonstrating clinically meaningful changes—not just statistically significant changes—is a key issue. For this purpose, we used the MCID of the WOMAC [21]. Yet, even MCID must be used carefully and not as an absolute measure, because there are several practical problems in estimating that parameter [32]. On the other hand, though it can be expected more variability of the WOMAC change after the intervention as long as the preoperative value is higher, because variability may be higher as long as the ability to improve has a wider range, in our case this is only true for patients with WOMAC pre-intervention values higher than 80 points in any of the two studied WOMAC scales. And, in terms of surpassing the MCID, those cases that do not introduce any conflict with our conclusions, since those nodes classified by us as appropriate would remain usually appropriate. Finally, as main outcomes we have considered the functional limitation and pain domains of the WOMAC as more relevant domains and excluded explicitly the stiffness dimension of WOMAC due to the fact that is composed of just two items and provide with less reliability and responsiveness than the other two domains included, as shown in some studies [19].

Our low specificity implies that we are not able to detect the inappropriate cases adequately. Nevertheless, our main objective was to be able to identify what cases were clearly necessary or appropriate to have the intervention, or, if uncertain leave the clinician the option of considering other circumstances, evidence or his/her own criteria to take a final decision.

There were differences between the derivation and the validation cohorts in four nodes of all figures of seven points in the changes in the WOMAC scores but, as reported before, just in one of those four nodes the difference was statistically significant. This could be due to the small sample sizes of some of those nodes. But, in any case, those differences do not seem to affect to our final conclusions about the appropriateness categorization of the nodes.

Finally, our decision trees should not be taken as a definite guide but in context and depending on other patient circumstances. THR clinical decision making is not just an issue about pain, stiffness and function, it is about a whole lot of other complex clinical and psycho-social issues. It is possible that the changes that may take place over time in the indication of the THR cause a broadening of the indication for this intervention. Although in countries where at the moment the waiting lists for surgery are an important problem, it is more necessary to have validated criteria to support proper clinical decision making. On this context, we hope our decision trees can work as an additional tool for the clinician when having to take a decision about a THR intervention.

In conclusion, a decision tree we created based on patient reported outcomes from two independent prospective cohorts can help determine the appropriateness of THR for patients with hip OA. The level of pain and the functional limitation caused by this condition appear to be the main drivers for the decision. The procedure is generally inappropriate or its value uncertain among patients with little pain or functional limitations or low scores on the WOMAC pain or functional limitation domains. It is deemed appropriate and necessary among patients with severe pain or functional limitations or high scores on the WOMAC pain or functional limitation domains. Before considering THR, however, non-surgical treatments for hip OA should have been instituted. These conclusions are generally in line with findings based on expert consensus, and have been formally derived and validated following recommendations for developing these kind of rules [24]. Applied to clinical decision making, the decision tree we have developed could guide the appropriate application of THR, which would reduce the undesirable variability of this procedure and increase health care quality. It can also be used to retrospectively evaluate clinical practice or quality control.

Rheumatology key messages

- Data generated from prospective studies have been used to develop appropriateness criteria for THR.
- A simple decision tree based on WOMAC outcomes can help to determine the appropriateness of THR.

Acknowledgements

We thank I. Vidaurreta, A. Higelmo and A. Rodriguez for their contribution to the data retrieval and data entry and to the Research Committee of the Galdakao Hospital. 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. The authors also acknowledge the editorial assistance provided by Patrick J. Skerrett.

Funding: This study was supported in part by grants from the Fondo de Investigación Sanitaria (98/001-01 to 03; 01/0184), the thematic networks—Red IRYSS—of the Instituto de Salud Carlos III (G03/220) and the Department of Health of the Basque Government.

Disclosure statement: The authors have declared no conflicts of interest.

Supplementary data

Supplementary data are available at *Rheumatology* Online.

References

1 Fear J, Hillman M, Chamberlain MA, Tennant A. Prevalence of hip problems in the population aged 55 years and over: access to specialist care and future demand for hip arthroplasty. Br J Rheumatol 1997;36:74–6.
2 Corti MC, Rigon C. Epidemiology of osteoarthritis: prevalence, risk factors and functional impact. Aging Clin Exp Res 2003;15:359–63.
3 Quintana JM, Arostegui I, Escobar A, Azkarate J, Goenaga JI, Lafuente I. Prevalence of knee and hip osteoarthritis and the appropriateness of joint replacement in an older population. Arch Intern Med 2008;168:1576–84.
4 Frankel S, Eachus J, Pearson N *et al.* Population requirement for primary hip-replacement surgery: a cross-sectional study. Lancet 1999;353:1304–9.
5 van Walraven CV, Peterson JM, Kapral M *et al.* Appropriateness of primary total hip and knee replacements in regions of Ontario with high and low utilization rates. Can Med Assoc J 1996;155:697–706.
6 Merx H, Dreinhofer K, Schrader P *et al.* International variation in hip replacement rates. Ann Rheum Dis 2003;62:222–6.
7 Morris RW, Fitzpatrick R, Hajat S *et al.* Primary total hip replacement: variations in patient management in Oxford & Anglia, Trent, Yorkshire & Northern 'regions'. Ann R Coll Surg Engl 2001;83:190–6.
8 Recommendations for the medical management of osteoarthritis of the hip, & knee: 2000 update. American College of Rheumatology Subcommittee on Osteoarthritis Guidelines. Arthritis Rheum 2000;43:1905–15.
9 NIH consensus conference: Total hip replacement. NIH Consensus Development Panel on Total Hip Replacement. J Am Med Assoc 1995;273:1950–6.
10 Quintana JM, Azkarate J, Goenaga JI, Arostegui I, Beldarrain I, Villar JM. Evaluation of the appropriateness of hip joint replacement techniques. Int J Technol Assess Health Care 2000;16:165–77.
11 Naylor CD, Williams JI. Primary hip and knee replacement surgery: Ontario criteria for case selection and surgical priority. Qual Health Care 1996;5:20–30.
12 Davies GM, Watson DJ, Bellamy N. Comparison of the responsiveness and relative effect size of the Western Ontario and McMaster Universities Osteoarthritis Index and the Short-Form Medical Outcomes Study Survey in a randomized, clinical trial of osteoarthritis patients. Arthritis Care Res 1999;12:172–9.
13 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;22:1193–6.
14 Schneider AJ. Assessment of risk factors and surgical outcome. Surg Clin North Am 1983;63:1113–26.

15 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;52:457–67.
16 Quintana JM, Arostegui I, Azkarate J *et al.* Evaluation of explicit criteria for total hip joint replacement. J Clin Epidemiol 2000;53:1200–8.
17 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;40:373–83.
18 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.
19 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. Western Ontario and McMaster Universities Osteoarthritis Index. Clin Rheumatol 2002;21:466–71.
20 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.
21 Quintana JM, Escobar A, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. Osteoarthritis Cartilage 2005;13:1076–83.
22 Quintana JM, Escobar A, Arostegui I *et al.* Health-related quality of life and appropriateness of knee or hip joint replacement. Arch Intern Med 2006;166: 220–6.
23 Quintana JM, Arostegui I, Azkarate J *et al.* Evaluation by explicit criteria of the use of total hip joint replacement. Rheumatology 2000;39:1234–41.
24 Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. J Am Med Assoc 1997;277: 488–94.
25 Trujillano J, Sarria-Santamera A, Esquerda A, Badia M, Palma M, March J. Approach to the methodology of classification and regression trees. Gac Sanit 2008;22:65–72.
26 Escobar A, González M, Quintana JM, Bilbao A, Ibañez B. Validation of a prioritization tool for patients on the waiting list for total hip and knee replacements. J Eval Clin Pract 2009;15:97–102.
27 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.
28 Fortin PR, Clarke AE, Joseph L *et al.* Outcomes of total hip and knee replacement: preoperative functional status predicts outcomes at six months after surgery. Arthritis Rheum 1999;42:1722–8.
29 Davis AM, Agnidis Z, Badley E, Kiss A, Waddell JP, Gross AE. Predictors of functional outcome two years following revision hip arthroplasty. J Bone Joint Surg Am 2006;88:685–91.
30 Nilsdotter AK, Lohmander LS. Age and waiting time as predictors of outcome after total hip replacement for osteoarthritis. Rheumatology 2002;41:1261–7.
31 Salaffi F, Carotti M, Grassi W. Health-related quality of life in patients with hip or knee osteoarthritis: comparison of generic and disease-specific instruments. Clin Rheumatol 2005;24:29–37.
32 Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics 2000; 18:419–23.