ARTHRITIS & RHEUMATOLOGY
Vol. 66, No. 8, August 2014, pp 2134–2143
DOI 10.1002/art.38685

Use of a Validated Algorithm to Judge the Appropriateness of Total Knee Arthroplasty in the United States

A Multicenter Longitudinal Cohort Study

Daniel L. Riddle, William A. Jiranek, and Curtis W. Hayes

Objective. **In previous studies conducted outside the US, ~20% of total knee arthroplasty (TKA) surgeries were judged to be inappropriate. The present study was undertaken to determine the prevalence rates of TKA surgeries classified as appropriate, inconclusive, and inappropriate in a knee osteoarthritis population in the US.**

Methods. **We used a modification of a validated appropriateness classification system and applied it to patients in the Osteoarthritis Initiative data set who underwent TKA. A variety of preoperative data were used in the classification, including Western Ontario and McMaster Universities Osteoarthritis Index pain and physical function scores, radiographic features, knee motion and laxity measures, and age.**

Results. **Data on 205 patients who underwent TKA were examined. The prevalence rates for classification of the procedure as appropriate, inconclusive, and inappropriate were 44.0% (95% confidence interval [95% CI] 37–51%), 21.7% (95% CI 16–28%), and 34.3% (95% CI 27–41%), respectively.**

Conclusion. **Approximately one-third of TKA surgeries were judged to be inappropriate. Variation in the characteristics of patients undergoing TKA was extensive. These data support the need for consensus development of criteria for patient selection among US practitioners treating patients who are potential candidates for TKA. Among the important issues, consensus development needs to address variation in patient characteristics and the relative importance of preoperative status and subsequent outcome.**

Several recent high-profile reports have described the dramatic growth in the number of total knee arthroplasties (TKAs) performed in the US (1–4). Between 1991 and 2010, for example, the annual volume of TKA surgeries among Medicare beneficiaries increased by 161.5%, and per capita utilization increased by 99.2% over the same period (1). Some have suggested that TKA is overutilized (2) or that overutilization may be one factor explaining large per capita increases in the rate of TKA surgery (1). Cram and colleagues (1) contend that recent growth in the number of TKA procedures is likely due both to an increase in utilization of a highly effective procedure and to overutilization of a procedure for which determination of need is highly reliant on subjective criteria.

Any determination of the extent to which TKA surgery is appropriate or inappropriate requires the use of valid appropriateness criteria. Such criteria as applied to patients undergoing TKA have been developed in

This article was prepared using an Osteoarthritis Initiative (OAI) public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

The OAI is a public–private partnership comprising five contracts (N01-AR-2-2258, N01-AR-2-2259, N01-AR-2-2260, N01-AR-2-2261, and N01-AR-2-2262) funded by the National Institutes of Health and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories, Novartis Pharmaceuticals Corporation, GlaxoSmithKline, and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. The authors of the current article were not funded as part of the OAI.

Daniel L. Riddle, PT, PhD, FAPTA, William A. Jiranek, MD, Curtis W. Hayes, MD: Virginia Commonwealth University, Richmond.

Dr. Jiranek has received consulting fees, speaking fees, and/or honoraria from DePuy, Inc. (more than $10,000) and receives royalties from DePuy, Inc. for hip and knee prostheses. Dr. Hayes has received consulting fees from BioClinica, Inc. (more than $10,000).

Address correspondence to Daniel L. Riddle, PT, PhD, FAPTA, Virginia Commonwealth University, PO Box 980224, Richmond, VA 23298-0224. E-mail: dlriddle@vcu.edu.

Submitted for publication October 9, 2013; accepted in revised form April 22, 2014.

other countries (5–8) but have not, to our knowledge, been formally developed or studied in the US.

The most commonly recommended approach for establishing appropriateness criteria for elective surgical procedures is the RAND/University of California, Los Angeles (UCLA) method (9–11). First, a systematic literature review of risks of, benefits of, and indications for the procedure is conducted. Second, an extensive and mutually exclusive set of clinical scenarios (typically numbering in the hundreds) is written to capture the full range of potential patient scenarios reflecting all potentially important clinical indications. Third, an expert panel is formed to conduct a modified Delphi survey to classify each scenario as appropriate, inconclusive, or inappropriate for the procedure. A rating of "appropriate" indicates that the expected benefits of the procedure outweigh the expected harms to the extent that the procedure is justified. A rating of "inconclusive" indicates either that the expected benefits and harms are roughly equal or that there was a lack of consensus among panel members. A rating of "inappropriate" indicates that the expected harms outweigh the expected benefits.

The most extensively studied RAND/UCLA-based appropriateness algorithm for TKA is the approach developed in Spain by Escobar and colleagues (5,12–15). Those authors conducted a systematic review of TKA evidence related to indications, effectiveness, and risks and used this evidence to develop 624 clinical scenarios based on the following literature-based key variables: symptom behavior, functional status, extent and location of radiographically documented arthritis, age, knee joint mobility and stability, and prior history of surgical and nonsurgical treatment. A modified Delphi survey approach was used with 2 independent national panels (n = 11 each) of arthroplasty surgeons (n = 18) and physiatrists or rheumatologists (n = 4). Reliability of recommendations between the 2 panels regarding the judgment of whether TKA for each scenario was appropriate, inappropriate, or inconclusive was found to be high (weighted $\kappa = 0.75$). In a subsequent study of 792 patients who underwent TKA (14), the largest improvements on the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (16) 6 months following surgery were demonstrated in patients whose TKA was judged as appropriate based on the appropriateness criteria (5), and the smallest improvements occurred in those whose TKA was judged as inappropriate.

Ghomrawi and colleagues contend that TKA appropriateness criteria like those developed by Escobar and colleagues are among the most powerful tools for improving quality of care and controlling costs (2). Because in studies in other countries only 60–80% of arthroplasty procedures were found to be appropriate, Ghomrawi et al suggested that similar overutilization of TKA might be occurring in the US. Given that no US appropriateness criteria for TKA have been developed, we used a modified version of the criteria described by Escobar et al to make an initial approximation of the proportion of knee arthroplasties performed in the US that may be inappropriate. While Escobar and colleagues' system was not designed for US patients, we believe the key criteria used in the system (i.e., pain and functional status, extent of radiographically evident arthritis, age, and knee joint impairment) are likely among the most important criteria for US patients as well (17). Our purpose was to use a modified version of the Escobar appropriateness criteria (5) to estimate the proportion of TKA procedures in the US that would be classified as appropriate, inconclusive, and inappropriate. We hypothesized that the prevalence of TKAs judged to be inappropriate would be similar to that demonstrated in earlier studies (11,13,14), i.e., ~20%.

PATIENTS AND METHODS

Patients. We used a subset of 4,796 persons enrolled in the Osteoarthritis Initiative (OAI) from which to derive the study population for the present study. The OAI is a privately and National Institutes of Health–funded multicenter longitudinal (5-year) prospective natural history study of persons with or at high risk of knee OA. The data collection was approved by the Institutional Review Boards of each of the following participating sites: the University of Maryland (Baltimore, MD), The Ohio State University (Columbus, OH), the University of Pittsburgh (Pittsburgh, PA), and Memorial Hospital of Rhode Island (Pawtucket, RI).

Criteria for exclusion from the study were as follows: 1) rheumatoid arthritis, 2) having undergone bilateral knee arthroplasty or already having established plans for bilateral knee arthroplasty in the next 3 years, 3) bilateral end-stage radiographic knee OA, and 4) use of ambulatory aids other than a single straight cane for >50% of the time. In addition, men weighing >130 kg and women weighing >114 kg were excluded for technical reasons, because these patients were unlikely to successfully undergo yearly magnetic resonance imaging (MRI) examinations required in the OAI protocol.

Over the study period, 216 patients in the OAI underwent knee replacement surgery. For persons who had TKA on both knees in the same year (n = 18), we randomly selected either the right or the left knee for study. A total of 11 persons with unicompartmental knee replacement (1 lateral, 1 patellofemoral, 9 medial) during the study period were excluded because of our focus on TKA (Figure 1).

Radiographic measures. Knee radiography was performed by highly trained technicians. The OAI investigators used a standardized radiographic technique (standing, semi-

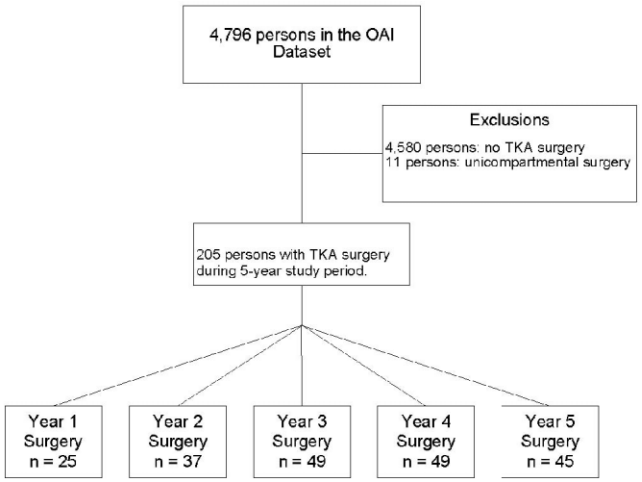

Figure 1. Disposition of the patients in the Osteoarthritis Initiative (OAI) assessed for inclusion in the study. TKA = total knee arthroplasty.

flexed, posteroanterior projection) (18); assessment of joint space width is more accurate and reproducible with this approach as compared to knee-extended views (17–21).

The Kellgren/Lawrence (K/L) scale (22) and the Osteoarthritis Research Society International (OARSI) scales (23,24) were used to quantify the pattern and severity of tibiofemoral arthritis of both knees. K/L grades range from 0 to 4. A grade of 0 is normal, 1 indicates doubtful narrowing of joint space and possible osteophyte(s), 2 indicates definite osteophytes and possible narrowing of joint space, 3 indicates the presence of definite joint space narrowing with some sclerosis and possible deformity of bone ends, and 4 indicates large osteophytes, marked narrowing of joint space, severe sclerosis, and definite deformity of the distal tibia or femur. The OARSI scale ranges from 0 to 3 and is used to grade the extent of joint space narrowing in both the medial and the lateral tibiofemoral compartments. A grade of 0 is normal, 1 indicates mild narrowing (1–33%), 2 indicates moderate narrowing (34–66%), and 3 indicates severe narrowing (67–100%). The lack of lateral or sunrise projections in the OAI precludes radiographic K/L grading of the patellofemoral compartment in these patients. Therefore, we developed a K/L-based surrogate measure using OAI MRIs for the subset of patients (n = 34) for whom a patellofemoral OA grade was required in order to apply the algorithm, due to the location of knee OA involvement. All radiographs and MRIs were obtained yearly, and we used the images obtained at the most recent visit prior to TKA surgery.

For assessment of the tibiofemoral joints, we used the highly reliable radiographic scoring data provided by OAI investigators. Test–retest reliability was substantial to almost perfect (25), with weighted kappa coefficients for both K/L and OARSI grades ranging from 0.70 to 0.87 for repeated independent readings of 300 randomly selected knee radiographs obtained 3–9 months apart. For assessment of the patellofemoral joints, an experienced musculoskeletal radiologist (CWH), who was blinded with regard to clinical and radiographic data, used a modified K/L system based on MRIs, as noted above, to determine the extent of patellofemoral OA (Table 1). The weighted kappa for measures repeated by this radiologist over a 6-month interval was 0.80 (95% confidence interval [95% CI] 0.61–0.99).

Escobar and colleagues (5) used the Ahlbäck radiographic grading system to classify the extent of OA as slight (Ahlbäck grade 1), moderate (grade 2 or 3), or severe (grade 4 or 5) (26). An Ahlbäck grade of slight is approximately equivalent to a K/L grade of 3, while Ahlbäck grades of moderate and severe approximate a K/L grade of 4. Reliability among Ahlbäck and K/L scores has been shown to be substantial (κ = 0.63–0.78) (26,27).

Additional classification criteria. Age was classified using the categories defined by Escobar and colleagues (5), i.e., <55 years, 55–65 years, and >65 years. To quantify the extent of pain and functional loss (referred to by Escobar and colleagues as symptomatology), we used combined scores from the highly reliable and valid (16,28) WOMAC pain and WOMAC physical function scales (22 items) obtained at the most recent visit prior to surgery. Each item in the WOMAC is scored from 0 to 4 (0 = none, 1 = mild, 2 = moderate, 3 = severe, 4 = extreme), for a total score range of 0–88. We divided combined WOMAC pain and function scores into 4 categories to reflect the slight, moderate, intense, and severe symptomatology groupings defined by Escobar et al. We reasoned that if a patient had a combined WOMAC pain and physical function score of 0–11, this score was equivalent to up to half of the items being marked as mild. Combined WOMAC scores of 12–22, 23–33, and ≥34 were used to demarcate moderate, intense, and severe symptomatology respectively, as defined by Escobar and colleagues. For example, if the patient's score was equivalent to that obtained when up to half of the WOMAC items were graded as moderate (i.e., 12–22), then the patient could be classified as having moderate symptomatology. This approach allowed us to derive a mutually exclusive and exhaustive scoring system from the WOMAC which, in our view, approximates the symptomatology criterion defined by Escobar et al.

In the knee joint mobility and stability criterion defined by Escobar and colleagues (5), patients are categorized as having limited mobility/stability if they have either ≤90 degrees of knee motion or >5 mm of medial or lateral gapping during stress testing of an extended knee. To adapt the OAI data, we classified patients as having limited mobility when

Table 1. Magnetic resonance imaging–based grading of patellofemoral osteoarthritis

Grade	Definition
0	Normal
1	No definite osteophyte (may have other limited cartilage/bone/periarticular changes, but no joint space narrowing)
2	Definite osteophyte. Focal cartilage loss without extensive involvement (i.e., no joint space narrowing)
3	Osteophyte plus significant cartilage loss involving at least 1 facet and/or trochlear surface (i.e., some joint space narrowing)
4	Osteophyte plus complete cartilage loss involving >50% of the medial and/or lateral patellofemoral compartment (i.e., at least 1 surface of bone-on-bone joint space narrowing)

Table 2. Criteria used to assess the appropriateness of total knee arthroplasty in patients with knee OA, in the study by Escobar et al (5) and as modified in the present study*

Criterion	Escobar et al	Present study
Age	<55 years *or* 55–65 years *or* >65 years	<55 years *or* 55–65 years *or* >65 years
Radiologic findings	Slight (Ahlbäck grade I) *or* moderate (Ahlbäck grades II and III) *or* severe (Ahlbäck grades IV and V)	Slight (K/L grade ≤3) *or* moderate (K/L grade 4) *or* severe (K/L grade 4)
Localization	Unicompartmental tibiofemoral *or* unicompartmental plus patellofemoral *or* tricompartmental	Unicompartmental tibiofemoral *or* unicompartmental plus patellofemoral *or* tricompartmental
Knee joint mobility and stability	Preserved mobility and stable joint (≥0–90° range of motion and absence of medial or lateral gapping of >5 mm in the extended knee) *or* limited mobility and/or unstable joint (<90° range of motion and/or medial or lateral gapping of >5 mm in the extended knee)	Preserved mobility and stable joint (<5° flexion contracture and normal or minor medial or lateral gapping in the 20° flexed knee) *or* limited mobility and/or unstable joint (≥5° flexion contracture and/or moderate or severe medial or lateral gapping in the 20° flexed knee)
Symptomatology	Slight (sporadic pain [e.g., when climbing stairs], daily activities typically carried out, NSAID use for pain control) *or* moderate (occasional pain [e.g., when walking on level surfaces], some limitation of daily activities, NSAID use to relieve pain) *or* intense (pain almost continuous [e.g., when walking short distances or standing for <30 minutes], frequent use of NSAIDs, may require crutch or cane) *or* severe (pain at rest, daily activities always significantly limited, frequent use of analgesics-narcotics/NSAIDs, frequent use of walking aids)	Slight (mild overall functional loss and function-related pain [e.g., up to half of WOMAC pain and physical function scale items scored from 0 to 11]) *or* moderate (moderate overall functional loss and function-related pain [e.g., up to half of WOMAC pain and physical function scale items scored from 12 to 22]) *or* intense (intense overall functional loss and function-related pain [e.g., up to half of WOMAC pain and physical function scale items scored from 23 to 33]) *or* severe (severe overall functional loss and function-related pain [e.g., more than half of WOMAC pain and physical function scale items scored ≥34])

* OA = osteoarthritis; K/L = Kellgren/Lawrence; NSAID = nonsteroidal antiinflammatory drug; WOMAC = Western Ontario and McMaster Universities OA Index.

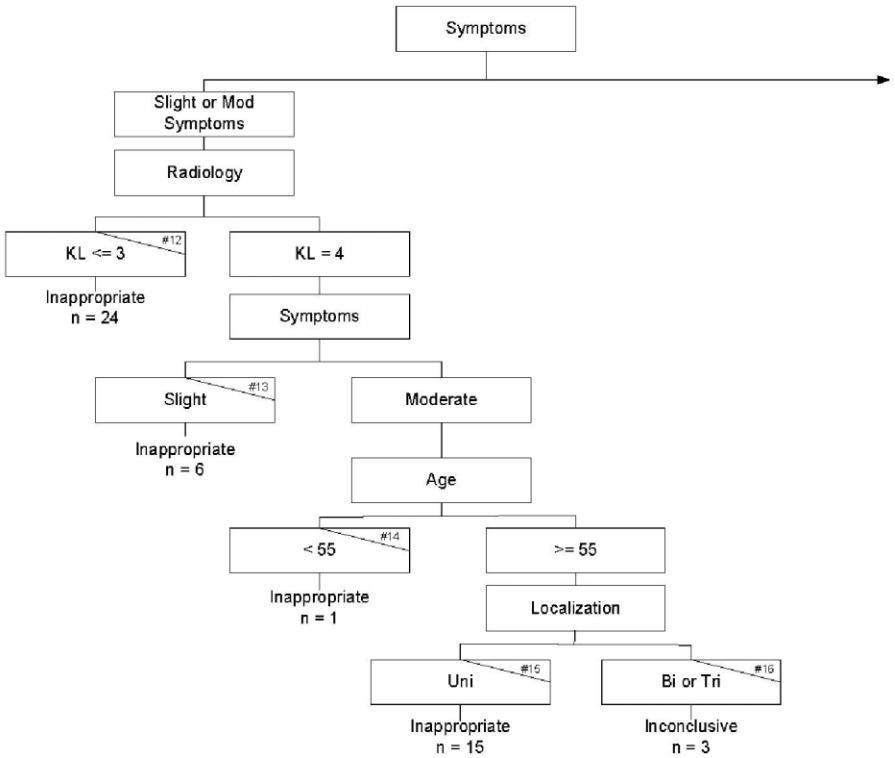

Figure 2. Left side of the algorithm, modified from that developed by Escobar and colleagues (5), for use in classifying total knee arthroplasty (TKA) procedures as appropriate, inappropriate, or inconclusive. Shown below the terminal node of each branch of the algorithm is the number of patients who met the criteria for classifying the TKA as inappropriate (nodes 12, 13, 14, and 15) or classifying the appropriateness/inappropriateness as inconclusive (node 16). Mod = moderate; K/L – Kellgren/Lawrence; Uni = 1 compartment; Bi = 2 compartments; Tri = 3 compartments.

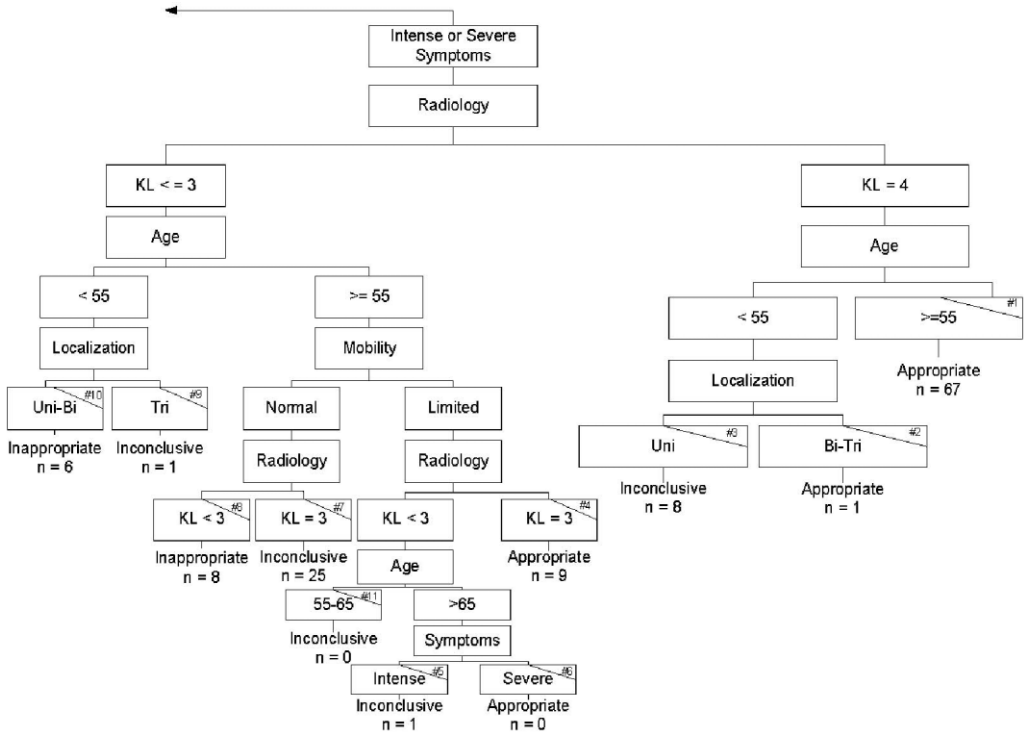

Figure 3. Right side of the algorithm, modified from that developed by Escobar and colleagues (5), for use in classifying TKA procedures as appropriate, inappropriate, or inconclusive. Shown below the terminal node of each branch of the algorithm is the number of patients who met the criteria for classifying the TKA as appropriate (nodes 1, 2, 4, and 6) or inappropriate (nodes 8 and 10) or classifying the appropriateness/inappropriateness as inconclusive (nodes 3, 5, 7, 9, and 11). See Figure 2 for definitions.

they either had a flexion contracture of ≥5 degrees or were graded as having moderate or severe medial or lateral gapping during valgus or varus stress testing with the knee flexed to 20 degrees. We judge these criteria as being reasonably close approximations of those used by Escobar and colleagues. The complete list of classification criteria recommended by Escobar and colleagues and the modifications made for the current study are listed in Table 2. Escobar and colleagues used a classification and regression tree approach (29) to confirm the classification criteria. These classification algorithms, adapted for use with OAI data, are illustrated in Figures 2 and 3.

Data analysis. For each patient, TKA surgery was classified as appropriate, inappropriate, or inconclusive based on the 16 terminal nodes of the algorithms developed by Escobar et al (Figures 2 and 3). We combined totals for the 6 nodes representing inconclusive, 4 representing appropriate, and 6 representing inappropriate and report prevalence rates along with 95% CIs for each of these combined nodes.

RESULTS

A total of 205 patients (mean age 66.9 years; 122 [59.5%] female, 83 [40.5%] male) underwent TKA surgery during the 5-year study period. Complete data on all classification variables were available for 175 of these patients (85.4%). The data that were most frequently missing were preoperative radiographs (Table 3). Age ($t = 0.46$, $P = 0.65$), combined WOMAC score ($t = 1.1$, $P = 0.29$), sex ($\chi^2 = 2.13$, $P = 0.13$), and body mass index ($t = 0.38$, $P = 0.70$) were not significantly different among those with and those without missing classification data. A total of 25, 37, 49, 49, and 45 TKA surgeries were performed in years 1 through 5, respectively. The mean ± SD number of days from the preoperative study visit to the surgery day was 177.9 ± 99 (range 2–464).

TKAs classified as appropriate. In the 175 subjects with complete data, TKA was classified as appropriate in 77 (44.0%) (95% CI 37–51%). The great majority of patients in whom TKA was classified as appropriate (n = 67 [87.0%]) had intense or severe symptoms and a K/L score of 4 and were at least 55 years of age (Figure 3). All but 1 of the remaining patients whose TKA was classified as appropriate (n = 9

Table 3. Characteristics of the 205 patients who underwent total knee arthroplasty*

Female	122 (59.5)
Age, mean ± SD (range) years	66.9 ± 46 (8.5–83)
Race†	
White or Caucasian	170 (83.3)
Black or African American	26 (13.0)
Other	8 (3.7)
Baseline body mass index, mean ± SD (range) kg/m^2	29.8 ± 19.8 (4.8–43.5)
Presurgery WOMAC score, mean ± SD (range)‡	32.1 ± 16.0 (0–86)
K/L grade§	
0	2 (1.1)
1	2 (1.1)
2	16 (9.0)
3	56 (31.3)
4	103 (57.5)
OARSI score medial compartment/OARSI score lateral compartment§	
0	47 (26.3)/132 (73.7)
1	14 (7.8)/5 (2.8)
2	44 (24.6)/13 (7.3)
3	74 (41.3)/29 (16.2)
≥5° knee flexion contracture or moderate or severe laxity¶	84 (41.0)
Patellofemoral score#	
1	1 (2.9)
2	10 (29.4)
3	17 (50)
4	6 (17.7)

* Except where indicated otherwise, values are the number (%). WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index; K/L = Kellgren/Lawrence; OARSI = Osteoarthritis Research Society International.
† Data missing on 1 patient.
‡ Data missing on 12 patients.
§ Data missing on 26 patients.
¶ Flexion contracture measures were obtained only at baseline; knee laxity measures were available only during years 2 and 3.
# In the patients for whom patellofemoral grades were required for classification using the system of Escobar et al (5).

[11.7%]) had intense or severe symptoms, a K/L score of 3, and limited mobility and were at least 55 years of age (Figure 3).

TKAs whose appropriateness was classified as inconclusive. The classification of TKA was inconclusive in 38 of the 175 patients (21.7%) (95% CI 16–28%) (Figures 2 and 3). The most common combination of findings for TKAs classified as inconclusive was the presence of intense or severe symptoms, a K/L grade of 3, age ≥55 years, and normal mobility (n = 25 [65.8%] [Figure 3]). The second most common combination of findings in this group was the presence of intense or severe symptoms, age <55 years, and a K/L grade of 4 in only 1 compartment (n = 8 [21.1%]).

TKAs classified as inappropriate. In 60 of the 175 patients (34.3% [95% CI 27–41%), TKA was classified as inappropriate (Figures 2 and 3). Most TKAs classified as inappropriate were either in a group of patients who had slight or moderate symptoms and a K/L grade of ≤3 (n = 24 [40.0%]) or a group who were age ≥55 years and had moderate symptoms and a K/L grade of 4 in only 1 compartment (n = 15 [25.0%]) (Figure 2).

DISCUSSION

This is, to our knowledge, the first study to compare previously validated appropriateness criteria (5) with actual TKA surgery cases in an extremely well-documented US sample. It is important to note that the approach described by Escobar and colleagues (5) was intended for estimation of TKA appropriateness in groups of patients but not for individual patients (14), and we strongly endorse this approach. For example, their system does not take into account medical comorbidities or body mass index, factors known to influence outcome and risk of complications (30,31).

Many patients struggle with the decision of whether to undergo TKA surgery (32). Patients must consider issues such as their symptom severity and psychological readiness, as well as surgical risk and recommendations of the surgeon and other members of the health care team. In addition to the variables examined in this study, surgeons consider numerous other patient-specific variables when recommending for or against primary TKA surgery. Ultimately, surgical decisions likely entail many factors beyond those included in any single set of appropriateness criteria and as a result, we suspect that any appropriateness criteria will have limitations that may restrict application for some individuals.

The most important and likely most controversial finding of our study was the percentage of patients for whom TKA was classified as inappropriate (34.3% [95% CI 27–41%]). As seen in Figures 2 and 3, classifications of inappropriate are driven first by the presence of slight or moderate symptoms and second by presurgical K/L grade, (usually ≤3 but sometimes applying as well to patients with a grade of 4). Symptoms and K/L grades were the two strongest predictors of judgments with regard to TKA appropriateness in the regression models tested by Escobar and colleagues and are therefore weighted heaviest in the models. The third and fourth criteria most commonly contributing to classification of TKA as inappropriate are younger age (<55 years) or knee mobility impairment.

Our definitions of slight and moderate symptoms

were based on quartiles of combined WOMAC pain and function scores. Patients with slight or moderate symptoms had combined WOMAC scores of ≤22 of a possible total of 88. Combined WOMAC pain and function scores prior to TKA surgery typically average in the high 40s to low 50s (32,33). Patients with mild or moderate symptoms in the present study for whom TKA was classified as inappropriate (n = 46) had a mean ± SD combined WOMAC score of 18 ± 11.4), indicating that their pain and functional loss was less than half that of the average patient undergoing TKA.

One factor that may have influenced the severity of pain and functional loss was the number of days from WOMAC assessment to surgery. The 46 patients with mild or moderate symptoms in whom TKA was classified as inappropriate completed the WOMAC scale a mean ± SD of 195 ± 96 days (range 7–378) prior to surgery. However, we tested the correlation between combined WOMAC score and number of days from surgery and found a Pearson's r of 0.07 ($P = 0.64$), indicating that time from surgery was not associated with the combined WOMAC score. Either no worsening (34) or very slight worsening (35) (on the order of 1 or 2 points on the WOMAC scale) occurs in patient populations during the 6–12-month period prior to TKA. Given that data on approximately half of our patients in the inappropriate TKA group were collected <6 months before surgery, we suspect that any undetected mild worsening over longer waiting periods had minimal effect on our findings, though this is a limitation of the study design.

If patients elected to undergo TKA because of severe symptoms experienced in conjunction with only a few activities as opposed to a more global functional disability, use of the highest (worst) scoring WOMAC item may be better suited to classification than the total WOMAC score. In an a posteriori sensitivity analysis we identified the highest (worst) single item from the presurgery WOMAC for each patient and used this single item score to classify symptomatology. We applied this new symptomatology rating to appropriateness classifications as applied in the main analysis and found rates of 37.7%, 19.4%, and 42.9% for ratings of appropriate, inconclusive, and inappropriate, respectively. These ratings were reasonably similar to those obtained in the original analysis, and we believe the differences are likely attributable to the greater error associated with single items (36) as compared to the multi-item combined WOMAC.

An age of <55 years was another criterion that was combined with symptom and K/L scores to classify TKA as inappropriate in 7 patients. This age threshold is an arbitrary standard, although a US-based consensus document (17) and a population-based survey of Canadian surgeons suggest that an age of <55 years is reason to question TKA candidacy (37). TKA utilization is, however, increasing in younger patients not only in the US (38) but also in Europe (39), which indicates that consensus is lacking. If we reconsidered this admittedly arbitrary age criterion and reclassified those in the inappropriate TKA category who were <55 years old as instead being in the category "inconclusive," the number of cases of inappropriate TKA would total 53 (30% of the study sample).

The most frequent reason TKA was classified as inappropriate in patients with intense or severe symptoms was a preoperative K/L grade of ≤2. A K/L score of ≤2 indicates that there is no joint space narrowing. Most commonly, recommendations for TKA require the presence of moderate or severe arthritis (37) or joint failure (17), implying the presence of at least some degree of joint space narrowing.

Patients in this study in whom TKA was classified as inappropriate generally had either mild or moderate symptoms or K/L grades of ≤2. Patients seek knee replacement primarily because of knee pain and the associated impact of pain on daily life (40). Given that most of these patients either had pain and functional loss scores that were less than half those in typical patients undergoing TKA or they had no joint space narrowing, it seems reasonable to question whether TKA was the most appropriate intervention for this subgroup. Among patients in whom TKA was classified as appropriate, 67 (87%) reported intense or severe symptoms (mean ± SD combined WOMAC score 39.9 ± 11.3), had a K/L grade of 4, and were age ≥55 years (mean ± SD 69 ± 6.1). This symptom, disease status, and age profile more closely approximates the typical patient who undergoes TKA (32,33).

Not surprisingly, the group of patients in whom the appropriateness of TKA was classified as inconclusive had the most heterogeneous sets of findings. The most common category of inconclusive ratings consisted of patients (n = 25) with intense or severe symptoms, age ≥55 years, normal mobility, and a K/L grade of 3. Escobar and colleagues' system is a consensus-based classification system built via a series of Delphi surveys (5). It is the inconclusive category in which the Delphi participants in their study demonstrated the greatest disagreement so it is not surprising that the profiles of these patients are the most varied.

Escobar et al included OA pain/antiinflammatory

medication use in their symptomatology assessment (Table 2). We chose to use only WOMAC pain and function scores to rate symptomatology. Our rationale was that medication usage and pain and functional status may not be strongly associated for a variety of reasons, and therefore may not allow for clear classification decisions. A patient may, for example, report severe pain and functional loss yet not use pain/antiinflammatory medication because of intestinal bleeding or cardiovascular risks. In lieu of including medication data in the classification, we used OAI data to determine whether the 3 classification categories differed in the proportion of patients who reported using nonprescription or prescription pain/antiinflammatory medications for more than half the days over the previous 30 days, as reported during the OAI visit prior to TKA surgery. A total of 72% of the patients had used these medications, and there were no differences among the classification categories ($\chi^2 = 0.84$, $P = 0.66$).

History of prior surgery on the knee undergoing TKA was also assessed in the study by Escobar et al. We examined whether the classification categories of appropriate, inappropriate, or inconclusive included different proportions of subjects who had undergone knee surgery prior to TKA. We found that 37% of the patients reported prior surgery on the involved knee, with no differences among the 3 classification categories ($\chi^2 = 3.7$, $P = 0.16$). In Escobar's study, a history of prior surgery explained only 3% of the variability in classification (as compared, for example, to 62% of variability explained by symptomatology). Our data also suggest that prior surgery is not a key variable associated with appropriateness ratings.

Our study has several important limitations. The most important limitation relates to use of the Escobar et al classification system (5). Although the system is, in our view, the most sound and well-validated of available appropriateness criteria, it may not be generalizable to current US patients. While the 3 most heavily weighted criteria, i.e., pain, functional loss, and extent of knee OA, have been frequently cited as key factors driving candidacy for TKA (17,41), the Escobar criteria were based on evidence published prior to 1999 and do not include features subsequently demonstrated to also have an important prognostic role. Medical comorbidities and body mass index, for example, were not accounted for in the system, and recent studies have demonstrated effects of comorbidities and extreme obesity on complication rates and outcomes (30,31,42). Additionally, the OAI sites are located in the midwestern, eastern, and northeastern US, and participants live in the communities surrounding the sites. It is unclear whether the data represent TKA appropriateness rates in the entire US. Future research should better account for area variation in TKA use (43) when estimating rates of appropriateness.

Importantly, the Escobar system is conceptually grounded in the assumption that TKA should be performed in patients with severe pain and functional loss whose OA is in a late stage. While these are the patients who generally exhibit the greatest improvements following TKA (32), the literature lacks consensus on this issue (44,45). The absence of comorbidity and obesity data in guiding classification further reinforces the importance of not applying Escobar et al's system to individual patients for decision-making, but rather applying it to patient groups. In addition, our measures of knee mobility and stability were not obtained at every yearly visit and therefore may have underestimated the proportion of patients with limitations in this regard. In 8 patients in whom TKA was classified as inappropriate, mobility/stability was scored as normal, and some of these may have been false-negatives.

In conclusion, we used a modification of the Escobar et al appropriateness criteria (5) to obtain an initial estimate of the proportions of TKAs performed in the US that are appropriate, inappropriate, and inconclusive. The rate of TKAs determined to be inappropriate was higher than we expected: we found approximately one-third of TKAs performed on OAI participants to be inappropriate based on the criteria applied. This result was driven primarily by the bias of the Escobar et al criteria toward finding persons who are age 55 years and older with high levels of pain and functional loss and severe knee OA to be the best candidates for TKA. Because there is no consensus in the US on TKA candidacy, there is extensive variation in the characteristics of patients who undergo the procedure, particularly with regard to knee pain, OA severity, and extent of functional loss. It is likely that this variation will continue until consensus is reached on the key criteria that should drive decisions to recommend TKA to patients. In our view, work should now focus on developing a consensus-based appropriateness classification system for patients in the US.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Riddle had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Riddle, Jiranek, Hayes.
Acquisition of data. Riddle, Hayes.
Analysis and interpretation of data. Riddle, Jiranek.

REFERENCES

1. Cram P, Lu X, Kates SL, Singh JA, Li Y, Wolf BR. Total knee arthroplasty volume, utilization, and outcomes among Medicare beneficiaries, 1991-2010. JAMA 2012;308:1227–36.
2. Ghomrawi HM, Schackman BR, Mushlin AI. Appropriateness criteria and elective procedures: total joint arthroplasty. N Engl J Med 2012;367:2467–9.
3. Losina E, Thornhill TS, Rome BN, Wright J, Katz JN. The dramatic increase in total knee replacement utilization rates in the United States cannot be fully explained by growth in population size and the obesity epidemic. J Bone Joint Surg Am 2012;94: 201–7.
4. Slover J, Zuckerman JD. Increasing use of total knee replacement and revision surgery. JAMA 2012;308:1266–8.
5. Escobar A, Quintana JM, Arostegui I, Azkarate J, Guenaga JI, Arenaza JC, et al. Development of explicit criteria for total knee replacement. Int J Technol Assess Health Care 2003;19:57–70.
6. Lofvendahl S, Bizjajeva S, Ranstam J, Lidgren L. Indications for hip and knee replacement in Sweden. J Eval Clin Pract 2011;17: 251–60.
7. 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.
8. Toye F, Barlow J, Wright C, Lamb SE. A validation study of the New Zealand score for hip and knee surgery. Clin Orthop Relat Res 2007;464:190–5.
9. Lawson EH, Gibbons MM, Ko CY, Shekelle PG. The appropriateness method has acceptable reliability and validity for assessing overuse and underuse of surgical procedures. J Clin Epidemiol 2012;65:1133–43.
10. Lee CN, Ko CY. Beyond outcomes: the appropriateness of surgical care. JAMA 2009;302:1580–1.
11. Lawson EH, Gibbons MM, Ingraham AM, Shekelle PG, Ko CY. Appropriateness criteria to assess variations in surgical procedure use in the United States. Arch Surg 2011;146:1433–40.
12. Ang DC, James G, Stump TE. Clinical appropriateness and not race predicted referral for joint arthroplasty. Arthritis Rheum 2009;61:1677–85.
13. Cobos R, Latorre A, Aizpuru F, Guenaga JI, Sarasqueta C, Escobar A, et al. Variability of indication criteria in knee and hip replacement: an observational study. BMC Musculoskelet Disord 2010;11:249.
14. Quintana JM, Escobar A, Arostegui I, Bilbao A, Azkarate J, Goenaga JI, et al. Health-related quality of life and appropriateness of knee or hip joint replacement. Arch Intern Med 2006;166: 220–6.
15. 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.
16. 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.
17. NIH Consensus Panel. NIH Consensus Statement on total knee replacement December 8-10, 2003. J Bone Joint Surg Am 2004; 86-A:1328–35.
18. Kothari M, Guermazi A, von Ingersleben G, Miaux Y, Sieffert M, Block JE, et al. Fixed-flexion radiography of the knee provides reproducible joint space width measurements in osteoarthritis. Eur Radiol 2004;14:1568–73.
19. Niinimaki T, Ojala R, Niinimaki J, Leppilahti J. The standing fixed flexion view detects narrowing of the joint space better than the standing extended view in patients with moderate osteoarthritis of the knee. Acta Orthop 2010;81:344–6.
20. Brandt KD, Mazzuca SA, Conrozier T, Dacre JE, Peterfy CG, Provvedini D, et al. Which is the best radiographic protocol for a clinical trial of a structure modifying drug in patients with knee osteoarthritis? J Rheumatol 2002;29:1308–20.
21. Peterfy C, Li J, Zaim S, Duryea J, Lynch J, Miaux Y, et al. Comparison of fixed-flexion positioning with fluoroscopic semiflexed positioning for quantifying radiographic joint-space width in the knee: test-retest reproducibility. Skeletal Radiol 2003;32: 128–32.
22. Kellgren JH, Lawrence JS. Radiological assessment of osteoarthrosis. Ann Rheum Dis 1957;16:494–502.
23. Altman RD, Hochberg M, Murphy WA Jr, Wolfe F, Lequesne M. Atlas of individual radiographic features in osteoarthritis. Osteoarthritis Cartilage 1995;3 Suppl A:3–70.
24. Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis Cartilage 2007;15 Suppl A:A1–56.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
26. Petersson IF, Boegard T, Saxne T, Silman AJ, Svensson B. Radiographic osteoarthritis of the knee classified by the Ahlback and Kellgren & Lawrence systems for the tibiofemoral joint in people aged 35-54 years with chronic knee pain. Ann Rheum Dis 1997;56:493–6.
27. Toivanen AT, Arokoski JP, Manninen PS, Heliovaara M, Haara MM, Tyrvainen E, et al. Agreement between clinical and radiological methods of diagnosing knee osteoarthritis. Scand J Rheumatol 2007;36:58–63.
28. Bellamy N. The WOMAC Knee and Hip Osteoarthritis Indices: development, validation, globalization and influence on the development of the AUSCAN Hand Osteoarthritis Indices. Clin Exp Rheumatol 2005;23:S148–53.
29. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont (CA): Wadsworth; 1984.
30. Hawker GA, Badley EM, Borkhoff CM, Croxford R, Davis AM, Dunn S, et al. Which patients are most likely to benefit from total joint arthroplasty? Arthritis Rheum 2013;65:1243–52.
31. Kerkhoffs GM, Servien E, Dunn W, Dahm D, Bramer JA, Haverkamp D. The influence of obesity on the complication rate and outcome of total knee arthroplasty: a meta-analysis and systematic literature review. J Bone Joint Surg Am 2012;94: 1839–44.
32. Lingard EA, Katz JN, Wright EA, Sledge CB. Predicting the outcome of total knee arthroplasty. J Bone Joint Surg Am 2004;86-A:2179–86.
33. Chesworth BM, Mahomed NN, Bourne RB, Davis AM. Willingness to go through surgery again validated the WOMAC clinically important difference from THR/TKR surgery. J Clin Epidemiol 2008;61:907–18.
34. Ackerman IN, Bennell KL, Osborne RH. Decline in Health-Related Quality of Life reported by more than half of those waiting for joint replacement surgery: a prospective cohort study. BMC Musculoskelet Disord 2011;12:108.
35. Desmeules F, Dionne CE, Belzile E, Bourbonnais R, Fremont P. The burden of wait for knee replacement surgery: effects on pain, function and health-related quality of life at the time of surgery. Rheumatology (Oxford) 2010;49:945–54.
36. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. New York: Oxford University Press; 2008.
37. Wright JG, Hawker GA, Hudak PL, Croxford R, Glazier RH,

Mahomed NN, et al. Variability in physician opinions about the indications for knee arthroplasty. J Arthroplasty 2011;26:569–75.
38. Losina E, Katz JN. Total knee arthroplasty on the rise in younger patients: are we sure that past performance will guarantee future success? [editorial]. Arthritis Rheum 2012:64:339–41.
39. Leskinen J, Eskelinen A, Huhtala H, Paavolainen P, Remes V. The incidence of knee arthroplasty for primary osteoarthritis grows rapidly among baby boomers: a population-based study. Arthritis Rheum 2012;64:423–8.
40. Frankel L, Sanmartin C, Conner-Spady B, Marshall DA, Freeman-Collins L, Wall A, et al. Osteoarthritis patients' perceptions of "appropriateness" for total joint replacement surgery. Osteoarthritis Cartilage 2012;20:967–73.
41. Gossec L, Paternotte S, Bingham CO III, Clegg DO, Coste P, Conaghan PG, et al, for the OARSI-OMERACT Task Force Total Articular Replacement as Outcome Measure in OA. OARSI/OMERACT Initiative to define states of severity and indication for joint replacement in hip and knee osteoarthritis: an OMERACT 10 Special Interest Group. J Rheumatol 2011;38: 1765–9.
42. Vasarhelyi EM, MacDonald SJ. The influence of obesity on total joint arthroplasty. J Bone Joint Surg Br 2012;94:100–2.
43. Judge A, Welton NJ, Sandhu J, Ben-Shlomo Y. Modeling the need for hip and knee replacement surgery. Part 1. A two-stage cross-cohort approach. Arthritis Rheum 2009;61:1657–66.
44. Dieppe P, Lim K, Lohmander S. Who should have knee joint replacement surgery for osteoarthritis? Int J Rheum Dis 2011;14: 175–80.
45. Losina E, Katz JN. Total knee replacement: pursuit of the paramount result. Rheumatology (Oxford) 2012;51:1735–6.