Published in final edited form as:
Arthritis Rheumatol. 2014 August ; 66(8): 2134–2143. doi:10.1002/art.38685.

Using a Validated Algorithm to Judge the Appropriateness of Total Knee Arthroplasty in the United States: A Multi-Center Longitudinal Cohort Study

Daniel L. Riddle, PT, PhD, FAPTA[1,2], **William A. Jiranek, MD**[2], and **Curtis W. Hayes, MD**[3]

[1]Department of Physical Therapy, Virginia Commonwealth University, Richmond, Virginia

[2]Department of Orthopaedic Surgery, Virginia Commonwealth University, Richmond, Virginia

[3]Department of Radiology, Virginia Commonwealth University, Richmond, Virginia

Abstract

Objective—We used a modified version of validated appropriateness criteria to determine the prevalence rates of total knee arthroplasty (TKA) surgeries that were classified as appropriate, inconclusive or inappropriate. Based on prior evidence, we hypothesized that the prevalence of TKA surgeries classified as inappropriate would approximate 20%.

Methods—The appropriateness classification system was adapted for use on persons undergoing TKA in the Osteoarthritis Initiative dataset. A variety of pre-operative data were used including WOMAC Pain and Physical Function scores, radiographic and knee motion and laxity measures and age. Prevalence rates for classifications of appropriate, inconclusive and inappropriate were calculated.

Results—Data from 205 persons with TKA were examined. The prevalence rate was 44.0% (95%CI= 37, 51) for classifications of appropriate, 21.7% (95%CI = 16, 28) for inconclusive classifications and 34.3% (95%CI =27, 41) for inappropriate classifications.

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

Keywords

Knee; arthroplasty; prognosis

Corresponding Author Address: Daniel L. Riddle, Department of Physical Therapy, Basement, West Hospital, Room B-100, Virginia Commonwealth University, Richmond, Virginia 23298-0224, Phone: 804-828-0234, Fax: 804-828-8111, dlriddle@vcu.edu.

Competing Interest Statement: The authors had no competing interests that that could potentially and inappropriately influence (bias) their work and conclusions.

The authors received no financial support and have no financial conflict of interest associated with the work reported in this paper.

Several recent high-profile publications have described the dramatic growth in utilization of total knee arthroplasty (TKA) in the United States (1–4). Between 1991 and 2010, for example, the annual volume of TKA surgeries among Medicare beneficiaries increased 161.5% and per capita utilization increased 99.2% over the same period (1). Some have suggested that TKA is over-utilized (2) or that over-utilization may be one factor explaining large per capita increases in TKA surgery (1). Cram and colleagues (1) contend that recent growth in TKA utilization is likely due both to an increase in utilization of a highly effective procedure and over-utilization of a procedure that 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. These criteria, as applied to patients undergoing TKA, to our knowledge, have not been formally developed or studied in the US but have been developed in other countries (5–8).

The most commonly recommended approach for establishing appropriateness criteria for elective surgical procedures is the RAND/UCLA method (9–11). First, a systematic literature review of risks, benefits and indications for the procedure is conducted. Second, an extensive and mutually exclusive set of clinical scenarios (typically numbering in the hundreds) are written to capture the gamut 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 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 a lack of consensus among panel members was found. An “inappropriate” rating indicates 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). The 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 radiographic arthritis, age, knee joint mobility and stability, and prior history of surgical and non-surgical treatment. A modified Delphi survey approach was used with two independent national panels (n=11 each) of arthroplasty surgeons (n=18) and physiatrists or rheumatologists (n=4). Reliability of recommendations between the two panels was found to be high (Weighted Kappa = 0.75) for judging whether TKA for each scenario was judged to be appropriate, inappropriate or inconclusive. A subsequent study of 775 TKA patients judged as appropriate based on the appropriateness criteria (5) demonstrated the largest WOMAC Scale improvements 6 months following surgery and patients judged as inappropriate had the smallest improvements (14).

Ghomrawi and colleagues contend that 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 studies in other countries reported that 60% to 80% of arthroplasty procedures were found to be appropriate, Ghomrawi et al suggested that similar over-utilization in TKA was possible in the US. Given that no appropriateness criteria for TKA have been developed in the US, we used a modified version of the Escobar et al appropriateness criteria to make an initial approximation of the proportion of knee arthroplasties that may be inappropriate in the US. While the Escobar et al system was not designed for US patients, we contend that the key criteria used in the system (i.e., pain and functional status, extent of radiographic arthritis, age and knee joint impairment) are likely among the most important criteria for US patients as well (16). Our purpose was to use a modified version of the Escobar et al (5) appropriateness criteria to estimate the proportion of TKA procedures classified as appropriate, inconclusive and inappropriate. We hypothesized that the prevalence rate of TKAs judged to be inappropriate would be similar to prior reports (11;13;14) and approximate 20%.

Methods

Subjects

Subjects were derived from a subset of 4,796 persons who were enrolled in the Osteoarthritis Initiative (OAI), an NIH and privately funded natural history multicenter prospective 5-year longitudinal study of persons with or at high risk for knee osteoarthritis (OA). The data collection was approved by the Institutional Review Boards of each of the following participating sites: (1) the University of Maryland in Baltimore, Maryland, (2) the Ohio State University in Columbus, Ohio, (3) the University of Pittsburgh in Pittsburgh, Pennsylvania, and (4) Memorial Hospital of Rhode Island, in Pawtucket, Rhode Island.

Subjects were excluded if they had; 1) rheumatoid arthritis, 2) bilateral knee arthroplasty or pre-existing plans to undergo bilateral knee arthroplasty in the next 3 years, 3) bilateral end-stage radiographic knee OA, 4) used ambulatory aids other than a single straight cane for more than 50% of the time. In addition, men weighing more than 130 kgs and women weighing more than 114kg were excluded for technical reasons because these subjects were unlikely to successfully undergo yearly MRI examinations required in the OAI protocol.

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

Radiographic Measures

The OAI investigators used a standardized radiographic technique (standing semi-flexed posteroanterior (PA) projection) and extensively trained technologists to obtain knee radiographs each year over the study period (17). More accurate and reproducible assessment of joint space width is obtained with this approach as compared to knee extended views (17–20).

The Kellgren and Lawrence (KL) scale and the Osteoarthritis Research Society International (OARSI) scale was used to quantify the pattern and severity of tibiofemoral arthritis of both knees. KL grades range from 0 to 4 (21–23). A grade of 0 was normal, 1 indicated doubtful narrowing of joint space and possible osteophyte(s), 2 indicated definite osteophytes and possible narrowing of joint space, 3 indicated the presence of definite joint space narrowing with some sclerosis and possible deformity of bone ends, and 4 indicated large osteophytes, marked narrowing of joint space, severe sclerosis and definite deformity of the distal tibia or femur (23). The OARSI scale ranges from 0 to 3 and is used to grade the extent of joint space narrowing for both the medial and lateral tibiofemoral compartments. A grade of 0 was normal, 1 indicated mild (1 to 33%) narrowing, 2 indicated moderate (34–66%) narrowing and 3 indicated severe (67 to 100%) narrowing (21;22). The lack of lateral or sunrise projections in the OAI precludes radiographic KL grading of the patellofemoral compartment. Therefore, the authors developed a KL-based surrogate measure using OAI MR images for the subset of subjects (n=34) who required a patellofemoral OA grade in the algorithm. All radiographs and MRIs were obtained yearly and we used the images obtained at the visit prior to TKA surgery.

For the tibiofemoral joints, we used the highly reliable radiographic scoring data provided by OAI investigators. Test-retest reliability was substantial to almost perfect (24) with weighted Kappa (K_w) coefficients for both KL and OARSI grades ranging from 0.70 to 0.87 for repeated independent readings of 300 randomly selected knee films separated by 3 to 9 months (25). For the patellofemoral joints, an experienced musculoskeletal radiologist (CWH) who was blinded to clinical and radiographic data, used a modified KL system based on MR images (see Table 1). The extent of patellofemoral OA of 34 subjects who required grading for Escobar et al classification (See Figures 2 and 3) was determined. The weighted Kappa for measures repeated in a blinded fashion by CWH over a 6-month interval was $K_w = 0.80$ (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 (grades 2 and 3) and severe (grades 4 and 5). The Ahlbäck grade of slight is approximately equivalent to a Kellgren and Lawrence (KL) grade of 3 while Ahlbäck grades of moderate and severe approximate a KL grade of 4(26). Reliability among Ahlbäck and KL scores has been shown to be substantial (Kappa ranging from 0.63 to 0.78) (26;27).

Additional Classification Criteria

Age was classified using the categories defined as <55 years, 55 to 65 years, and > 65 years by Escobar and colleagues (5). To quantify the extent of pain and functional loss, or what Escobar and colleagues refer to as symptomatology, we used combined scores from the highly reliable and valid (28;29) WOMAC Pain and WOMAC Physical Function scales (n=22 items) for the surgical knees obtained at the visit prior to surgery. Each item in WOMAC is scored from 0 to 4 (0 = none, 1=mild, 2= moderate, 3=severe, 4=extreme) for a total score range of 0 to 88. We split combined WOMAC Pain and Function scores into four categories to reflect the slight, moderate, intense and severe symptomatology groupings defined by Escobar et al. We reasoned that if patients' scores on combined WOMAC Pain and Physical Function scores were 0 to 11, this score was equivalent to up to half of the items marked as mild. Cut scores of 12 to 22, 23 to 33 and 34 and higher on combined WOMAC scores 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 marked moderate (i.e., 12 to 22), then the patient could be classified as having moderate symptomatology. This approach allowed us to have a mutually exclusive and exhaustive scoring system for the WOMAC which, in our view, approximates the symptomatology criterion defined by Escobar and colleagues (5).

For the knee joint mobility and stability criterion defined by Escobar and colleagues (5), patients were categorized as limited when they had either less than 0° to 90° of knee motion or greater than 5 millimeters of medial or lateral gapping during stress testing of an extended knee. For the OAI data, we classified patients as having limited mobility when they either had a ≥ 5° flexion contracture 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. These criteria are, in our view, reasonably close approximations of the criteria used by Escobar and colleagues. The complete list of classification criteria recommended by Escobar and colleagues (5) and the modifications made in the current study are listed in Table 2. Escobar and colleagues used a Classification and Regression Tree approach (30) to confirm the classification criteria. These classification algorithms, adapted for OAI data, are illustrated in Figures 2 and 3.

Data Analysis

Patients were classified as appropriate, inappropriate or inconclusive for TKA surgery based on the 16 terminal nodes of the algorithms developed by Escobar et al. (Figure 2 and 3). We combined totals for the 6 uncertain, 4 appropriate, and 6 inappropriate nodes and report prevalence rates along with 95% confidence intervals (CIs) for each of these combined nodes.

Results

We had a total of 205 persons with TKA surgery during the 5-year period. A total of 175 (85.4%) patients had complete data for all classification variables. The most common reason for incomplete data was missing pre-operative radiographs (see Table 3). Patients had a mean age of 66.9 years and 59.5% were female (see Table 3). Age (t=0.46, p=0.65), combined WOMAC (t=1.1, p=0.29), sex (χ^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, 45 TKR surgeries were conducted in years 1 through 5 respectively. The mean number of days from the pre-operative study visit to the surgery day was 177.9 days (sd= 99; range: 2–464 days).

Appropriate classifications

Of the 175 subjects with complete data, 77 (44.0%, 95%CI= 37, 51) were classified as appropriate. Terminal nodes classified as appropriate appear as #1, #2, #4, and #6 in Figure 3. The great majority of TKAs classified as appropriate (n = 67, 87.0%) had intense or severe symptoms, KL scores of 4, and were at least 55 years of age. All but one of the remaining appropriate TKAs (n=9, 11.7%) had intense or severe symptoms, KL scores of 3, limited mobility and were 55 years or older.

Inconclusive classifications

A total of 38 of 175 subjects (21.7%, 95%CI = 16, 28) were classified as inconclusive. Terminal nodes classified as inconclusive are labeled #16 in Figure 2 and #3, #5, #7, #9, #11, in Figure 3. The most common combination of findings for TKAs classified as inconclusive was the presence of intense or severe symptoms, a KL grade of 3, aged at least 55 years and normal mobility (n=25, 65.8%). Persons younger than 55 years with intense or severe symptoms and a KL grade of 4 in only one compartment were in the second most common node (n=8, 21.1%) classified as inconclusive.

Inappropriate classifications

There were 60 of 175 subjects (34.3%, 95%CI =27, 41) classified as inappropriate for TKA (see Figures 2 and 3). Terminal nodes indicating inappropriate classifications are labeled #12–15 in Figure 2 and #8 and #10 in Figure 3. Most TKAs classified as inappropriate were either in a group that had slight or moderate symptoms and KL grades of 3 or less (n=24, 40.0%) or a group that was 55 years or older with moderate symptoms and a KL grade of 4 in only one compartment (n=15, 25.0%).

Discussion

Ours is the first study in the U.S., to our knowledge, that compares previously validated appropriateness criteria (5) with actual TKA surgery cases in an extremely well documented sample. Importantly, the approach by Escobar and colleagues was intended for estimations of TKA appropriateness in groups of patients but not for individual patients (14) and we strongly endorse this approach. For example, The Escobar et al system does not account for medical comorbidities or BMI, factors known to influence outcome and risk of complications (31;32).

Many patients struggle with the decision to undergo TKA surgery (34). Patients must consider their symptomatic severity and psychological readiness among other issues, as well as surgical risk along with recommendations of the surgeon and other members of the healthcare team. In addition to the variables examined in this study, surgeons consider a host of other patient-specific variables when recommending for or against a primary TKA surgery. Ultimately, surgical decisions likely include many other issues 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 in our study was the percentage of patients (34.3%, 95%CI=27, 41) classified as inappropriate for TKA surgery. As seen in Figures 2 and 3, classifications of inappropriate are driven firstly by the presence of slight or moderate symptoms and secondly, by pre-surgical KL scores, usually <=3 but sometimes a grade of 4. Symptoms and KL scores were the two strongest predictors of appropriateness judgments 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 inappropriate classifications are either younger age (<55 years) or knee mobility impairments.

Our definitions of slight and moderate symptoms were based on quartile splits of combined WOMAC Pain and Function scales. Persons with slight or moderate symptoms had combined WOMAC scores of 22 or less out of a total of 88 points. Combined WOMAC Pain and Function scores prior to TKA surgery typically average in the high 40s to low 50s (33;34). Our TKA patients with mild or moderate symptoms who were classified as inappropriate (n=46) had an average combined WOMAC score of 18 (sd=11.4) indicating these subjects had pain and functional loss that was less than half that of the average patient undergoing TKA.

One factor that may have influenced pain and functional status severity was the number of days from WOMAC assessment to surgery. The 46 subjects classified with mild or moderate symptoms in the inappropriate category completed the WOMAC scale a mean of 195 days (sd=96, range= 7, 378 days) prior to surgery. We correlated the WOMAC combined score with days from surgery and found a Pearson $r = 0.07$ ($p = 0.64$) indicating that time from surgery was not associated with combined WOMAC score severity. Either no worsening (35) or very slight worsening (36) on the order of 1 or 2 WOMAC scale points occurs in patient samples during the 6 to 12 month period prior to TKA surgery. Given that approximately half of our patients’ data in the inappropriate group were collected less than 6 months from 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 only a few highly symptomatic activities versus a more global functional complaint, use of the highest (worse) scoring WOMAC item may be better suited to classification than the total WOMAC score. In an a posteriori sensitivity analysis we identified the highest (worse) single item from the pre-operative 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 the original analysis and we suspect the differences are likely attributable to the greater error associated with single items as compared to the multi- item combined WOMAC (37).

An age of <55 years was another criterion that was combined with symptom and KL scores to classify 7 patients as inappropriate for TKA. This age threshold is an arbitrary standard though a US-based consensus document (16) and a population-based survey of Canadian surgeons suggests an age of <55 years as reason to question TKA candidacy (38). TKA utilization is, however, increasing in younger patients not only in the US (39) but also in Europe (40) which indicates that consensus is lacking. If we reconsidered this admittedly arbitrary age criterion and reclassified those in the inappropriate category who were < 55 years as inconclusive, the inappropriate cases sample would total 53 persons (30% of the study sample).

The most common inappropriate classification for subjects with intense or severe symptoms was attributed to having pre-operative KL scores of 2 or less. A KL score of 2 or less indicates no joint space narrowing. Most commonly, recommendations for TKA require the presence of moderate or severe arthritis (38) or joint failure (16) implying at least some degree of joint space narrowing.

Our subjects classified as inappropriate generally had either mild or moderate symptoms or KL scores of 2 or lower. Patients seek TKA primarily because of their knee pain and the associated impact of pain on daily life (41). Given that most of these subjects either had pain and functional loss profiles that were less than half that of 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. For patients classified as appropriate for TKA, 67 (87%) reported intense or severe symptoms (mean combined WOMAC scores of 39.9 (sd=11.3), had KL grades of 4 and were 55 years of age or older (mean age=69 years (sd=6.1)). This age, symptom and disease status profile more closely approximates the typical patient who undergoes TKA surgery (33;34).

Not surprisingly, subjects classified as inconclusive had the most heterogeneous sets of findings. The most common category of inconclusive ratings consisted of subjects (n=25) with intense or severe symptoms, aged 55 or more years, normal mobility and KL grades of 3. The Escobar et al 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 demonstrated the greatest disagreement so it is not surprising that the profiles of these subjects are the most varied.

Escobar and colleagues included OA pain/anti-inflammatory medication use in the symptomatology assessment (see 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/anti-inflammatory 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 three classification categories differed in the proportion of subjects who reported using non-prescription or prescription pain/anti-inflammatory medications for more than half the days over the past 30 days, as reported during the OAI visit prior to TKA surgery. A total of 72% of TKA subjects used these medications and there were no differences among the classification categories (χ^2=0.84, p=0.66).

A history of prior surgical management of the knee undergoing TKA also was included in the study by Escobar and colleagues. We examined whether classifications of appropriate, inappropriate or inconclusive demonstrated different proportions of persons who had knee surgery prior to TKA. We found that 37% of TKA patients reported prior surgery on the involved knee and there were no differences among the 3 classification categories (χ^2=3.7, p=0.16). In the Escobar et al study, a history of prior surgical management explained only 3% of the variability in classification (as compared, for example, to 62% of variability explained by symptomotology). Our data also suggest that prior surgical history 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). While the system is, in our view, the most sound and validated of available appropriateness criteria, it may not be generalizable to current US patients. While the three most heavily weighted criteria, pain, functional loss and extent of knee OA have been frequently cited as key factors driving TKA candidacy (16;42) the Escobar et al criteria were based on evidence published prior to 1999. Medical comorbidities and BMI, for example, were not accounted for in the system and recent studies have reported effects of comorbidities and extreme obesity on complication rate and outcome (31;32;43). The OAI sites are located in the Midwest, East and Northeast US and participants lived in the surrounding communities. It is unclear whether these data represent TKA appropriateness rates in the entire US. Future research should better account for area variation in TKA use (44) when estimating appropriateness rates.

Importantly, the Escobar et al system is conceptually grounded in the assumption that TKA should be conducted on persons with severe pain and functional loss who are late in the OA disease process. While these are the persons who demonstrate the greatest improvements following TKA (34), the literature lacks consensus on this issue (45;46). The lack of inclusion of comorbidity and obesity data in guiding classification further reinforces the importance of not applying the Escobar et al system to individual patients for decision making but rather for applying to patient groups. In addition, our measures of knee mobility and stability were not obtained at all yearly visits and therefore may have underestimated the proportion of persons who scored positive on this criterion. A total of 8 persons who were classified as inappropriate scored a negative on the Mobility and Stability criterion and some of these may have been false negatives.

We used the Escobar et al system (5) to obtain an initial estimate of the proportion of appropriate, inappropriate and inconclusive TKAs in the US. In our view, work should now focus on developing a consensus-based appropriateness classification system for US patients.

In conclusion, the rate of TKAs determined to be inappropriate was higher than we expected in that we found approximately a third of TKAs conducted in OAI to be inappropriate when applying a modified version of the Escobar et al (5) appropriateness criteria. This finding is driven primarily by the bias grounding the Escobar et al criteria; that persons who are the best candidates for TKA are 55 years and older with severe levels of pain, functional loss and knee OA. Because there is no consensus in the US on TKA candidacy, extensive variation among TKA patients' characteristics exists, particularly with regard to knee pain, OA severity and extent of functional loss. It is likely this variation will continue until consensus is reached on the key criteria that drive decisions to recommend TKA to patients.

Acknowledgments

The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the

Department of Health and Human Services, 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. This manuscript was prepared using an 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 authors of the current paper were not funded as part of the OAI.

Role of the funding source: The OAI is a public-private partnership comprised of five contracts funded by the National Institutes of Health. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc." The authors of the current paper are not part of the OAI investigative team. The funding source played no role in the conduct or reporting of this study.

Reference List

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(12):1227–36. [PubMed: 23011713]
2. Ghomrawi HM, Schackman BR, Mushlin AI. Appropriateness criteria and elective procedures--total joint arthroplasty. N Engl J Med. 2012; 367(26):2467–9. [PubMed: 23268663]
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(3):201–7. [PubMed: 22298051]
4. Slover J, Zuckerman JD. Increasing use of total knee replacement and revision surgery. JAMA. 2012; 308(12):1266–8. [PubMed: 23011717]
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(1):57–70. [PubMed: 12701939]
6. Lofvendahl S, Bizjajeva S, Ranstam J, Lidgren L. Indications for hip and knee replacement in Sweden. J Eval Clin Pract. 2011; 17(2):251–60. [PubMed: 20860582]
7. Naylor CD, Williams JI. Primary hip and knee replacement surgery: Ontario criteria for case selection and surgical priority. Qual Health Care. 1996; 5(1):20–30. [PubMed: 10157268]
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. [PubMed: 18062051]
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(11):1133–43. [PubMed: 23017632]
10. Lee CN, Ko CY. Beyond outcomes--the appropriateness of surgical care. JAMA. 2009; 302(14): 1580–1. [PubMed: 19826028]
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(12):1433–40. [PubMed: 22184308]
12. Ang DC, James G, Stump TE. Clinical appropriateness and not race predicted referral for joint arthroplasty. Arthritis Rheum. 2009; 61(12):1677–85. [PubMed: 19950319]
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. [PubMed: 20977745]
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(2):220–6. [PubMed: 16432092]
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(14):1576–84. [PubMed: 18663171]
16. NIH Consensus Panel. NIH Consensus Statement on total knee replacement December 8–10, 2003. J Bone Joint Surg Am. 2004; 86-A(6):1328–35. [PubMed: 15173310]

17. Kothari M, Guermazi A, von IG, 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(9):1568–73. [PubMed: 15150666]

18. 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(3):344–6. [PubMed: 20450420]

19. 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(6):1308–20. [PubMed: 12064851]

20. Peterfy C, Li J, Zaim S, Duryea J, Lynch J, Miaux Y, et al. Comparison of fixed-flexion positioning with fluoroscopic semi-flexed positioning for quantifying radiographic joint-space width in the knee: test-retest reproducibility. Skeletal Radiol. 2003; 32(3):128–32. [PubMed: 12605275]

21. 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. [PubMed: 8581752]

22. Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis Cartilage. 2007; 15 (Suppl A):A1–56. [PubMed: 17320422]

23. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957; 16(4):494–502. [PubMed: 13498604]

24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–74. [PubMed: 843571]

25. Felson, DT. Central Reading of Knee X-rays for K-L Grade and Individual Radiographic Features of Knee OA. 2011.

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(8):493–6. [PubMed: 9306873]

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(1):58–63. [PubMed: 17454937]

28. 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(12):1833–40. [PubMed: 3068365]

29. 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(5 Suppl 39):S148–S153. [PubMed: 16273799]

30. Breiman, L.; Frieman, JH.; Olshen, RA.; Stone, CJ. Classification and regression trees. Belmont, California: Wadsworth; 1984.

31. 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(5):1243–52. [PubMed: 23459843]

32. 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(20):1839–44. [PubMed: 23079875]

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(9):907–18. [PubMed: 18687289]

34. Lingard EA, Katz JN, Wright EA, Sledge CB. Predicting the outcome of total knee arthroplasty. J Bone Joint Surg Am. 2004; 86-A(10):2179–86. [PubMed: 15466726]

35. 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. [PubMed: 21605398]

36. 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(5):945–54. [PubMed: 20144931]

37. Streiner, DS.; Norman, GR. Health Measurement Scales A Practical Guide to the Development and Use. 4. New York: Oxford University Press; 2008.

38. 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(4):569–75. [PubMed: 20580197]

39. Losina E, Katz JN. Total knee arthroplasty on the rise in younger patients: Are we sure that past performance will guarantee future success? Arthritis Rheum. 2012

40. Leskinen J, Eskelinen A, Huhtala H, Paavolainen P, Remes V. The incidence of knee arthrolasty for primary osteoarthritis grows rapidly among baby-boomers - a population-based study. Arthritis Rheum. 2012

41. 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(9):967–73. [PubMed: 22659599]

42. Gossec L, Paternotte S, Bingham CO III, Clegg DO, Coste P, Conaghan PG, et al. An OMERACT 10 Special Interest Group. OARSI/OMERACT Initiative to Define States of Severity and Indication for Joint Replacement in Hip and Knee Osteoarthritis. J Rheumatol. 2011; 38(8):1765–9. [PubMed: 21807799]

43. Vasarhelyi EM, MacDonald SJ. The influence of obesity on total joint arthroplasty. J Bone Joint Surg Br. 2012; 94(11 Suppl A):100–2. [PubMed: 23118394]

44. 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(12):1657–66. [PubMed: 19950326]

45. Dieppe P, Lim K, Lohmander S. Who should have knee joint replacement surgery for osteoarthritis? Int J Rheum Dis. 2011; 14(2):175–80. [PubMed: 21518317]

46. Losina E, Katz JN. Total knee replacement: pursuit of the paramount result. Rheumatology (Oxford). 2012; 51(10):1735–6. [PubMed: 22843792]

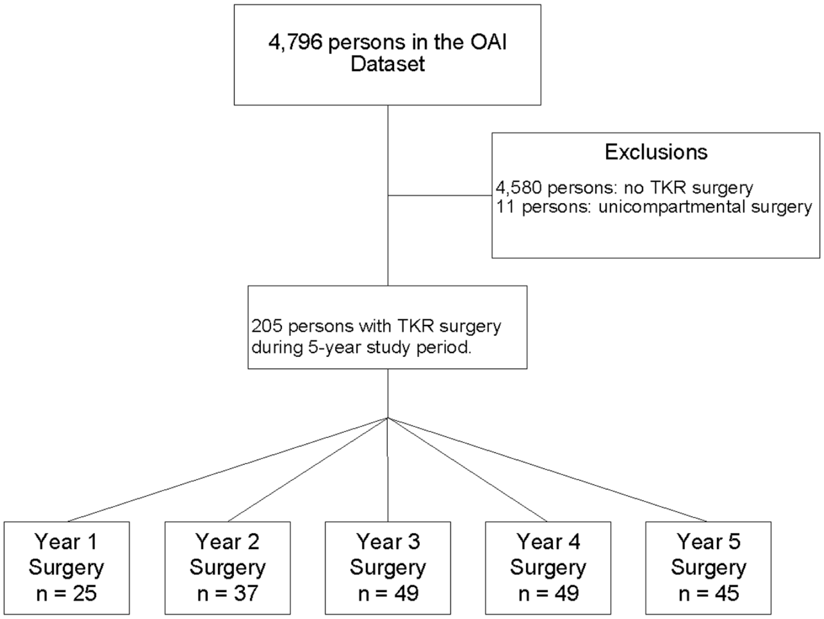
4,796 persons in the OAI Dataset
Exclusions
4,580 persons: no TKR surgery
11 persons: unicompartmental surgery
205 persons with TKR surgery during 5-year study period.
Year 1 Surgery n = 25
Year 2 Surgery n = 37
Year 3 Surgery n = 49
Year 4 Surgery n = 49
Year 5 Surgery n = 45

Figure 1.

The flow of patients through the study.

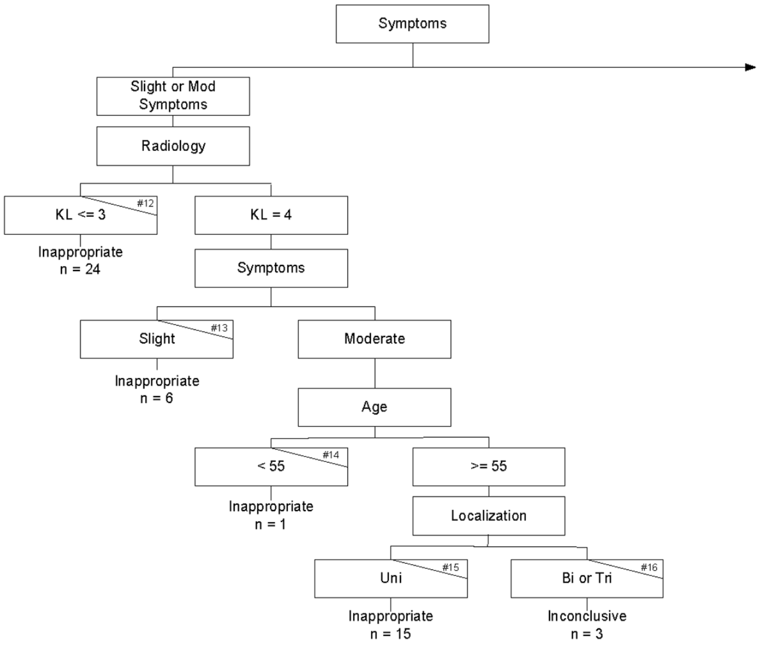

Figure 2.

The figure illustrates the left side of the algorithm modified from that developed by Escobar and colleagues for classifying total knee arthroplasty procedures as appropriate, inappropriate or inconclusive. The terminal nodes of each branch of the algorithm are labeled with the number of subjects who matched the criteria for each branch. The small number in the upper right hand corner of each terminal node indicates whether the terminal node is classified as inconclusive (# 16) or inappropriate (#s 12, 13, 14 and 15)

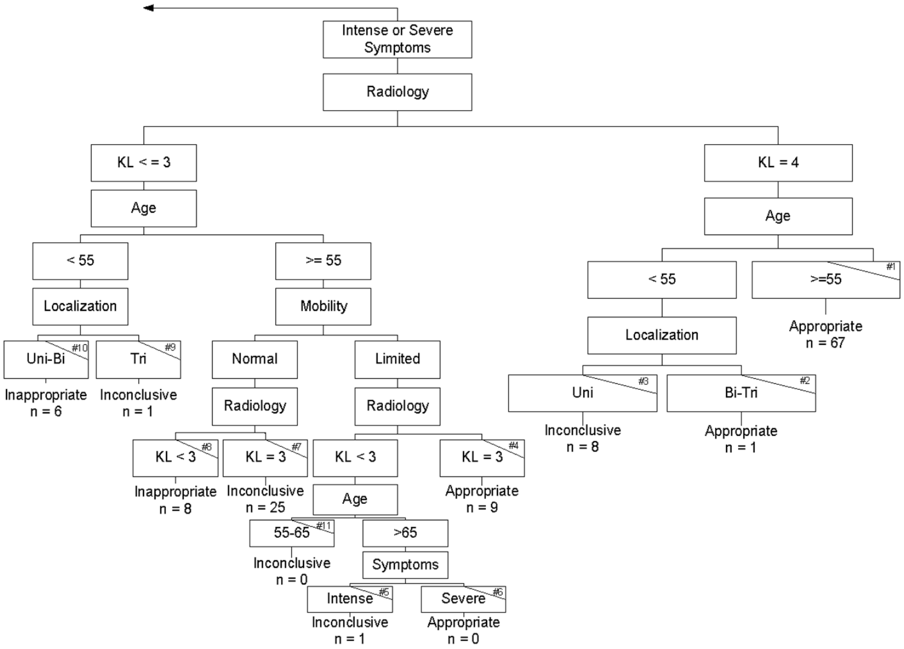

Figure 3.

The figure illustrates the right side of the algorithm modified from that proposed by Escobar and colleagues for classifying total knee arthroplasty procedures as appropriate, inappropriate or inconclusive. The terminal nodes of each branch of the algorithm are labeled with the number of subjects who matched the criteria for each branch. The small number in the upper right hand corner of each terminal node indicates whether the terminal node is classified as appropriate (#s 1, 2, 4, and 6), inconclusive (#s 3, 5, 7, 9 and 11) or inappropriate (#s 8 and 10)

Table 1

MRI 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 one facet and/or trochlear surface (i.e., some joint space narrowing)
4	Osteophyte plus complete cartilage loss involving >50% of medial and/or lateral patellofemoral compartment (i.e., at least one surface of bone-on- bone joint space narrowing)

Table 2

Comparison of criteria used by Escobar and colleagues and criteria modified for the current study

Classification Criteria: Escobar and colleagues	Classification Criteria: Current study
Age	Age
<55 years	<55 years
55 to 65 years	55 to 65 years
> 65 years	> 65 years
Radiology	Radiology
Slight (Ahlbäck grade I)	Slight (Kellgren and Lawrence grade 3 or less)
Moderate (Ahlbäck grades II and III)	Moderate (Kellgren and Lawrence grade 4)
Severe (Ahlbäck grades IV and V)	Severe (Kellgren and Lawrence grade 4)
Localization	Localization
Unicompartmental tibiofemoral	Unicompartmental tibiofemoral
Unicompartmental plus patellofemoral	Unicompartmental plus patellofemoral
Tricompartmental	Tricompartmental
Knee Joint Mobility and Stability	Knee Joint Mobility and Stability
Preserved mobility and stable joint (a minimum range of movement from 0° to 90° and absence of medial or lateral gapping of more than 5 mm. in the extended knee.)	Preserved mobility and stable joint (less than a 5° flexion contracture and normal or minor medial or lateral gapping in the 20° flexed knee.)
Limited mobility and/or unstable joint (a range of movement of less than 0° to 90° and/or medial or lateral gapping of more than 5 mm. in the extended knee.)	Limited mobility and/or unstable joint (5° or greater flexion contracture and/or moderate or severe medial or lateral gapping in the 20° flexed knee.)
Symptomatology	Symptomatology
Slight: Sporadic pain, (e.g., when climbing stairs, daily activities typically carried out) nonsteroidal anti- inflammatory (NSAID) drugs for pain control).	Slight: Mild overall functional loss and function related pain – for example, up to half of WOMAC Pain and Physical Function scale items marked as mild (scores from 0 to 11)).
Moderate: Occasional pain (e.g., when walking on level surfaces, some limitation of daily activities, NSAIDs to relieve pain.	Moderate: Moderate overall functional loss and function related pain – for example, up to half of WOMAC Pain and Physical Function scale items marked as moderate (scores from 12 to 22)).
Intense: Pain almost continuous (e.g. pain when walking short distances or standing for less than 30 minutes, limited daily activities, frequent use of NSAIDs, may require crutch or cane)	Intense: Intense overall functional loss and function related pain – for example, up to half of WOMAC Pain and Physical Function scale items marked as severe (scores from 23 to 33)
Severe: Pain at rest, daily activities always significantly limited, frequent use of analgesics- narcotics/NSAIDs, frequent use of walking aids.	Severe: Severe overall functional loss and function related pain – for example, more than half of WOMAC Pain and Physical Function scale items marked as severe (scores of 34 and higher)

Table 3

Characteristics of the Patients

	Knee Replacement Sample Mean (sd, min, max) or N (%)(n = 205)		Missing data N
Female Sex	122 (59.5)		0
Age in years*	66.9 (46, 83, 8.5)		0
Race			1
White or Caucasian	170 (83.3)		
Black or African American	26 (13.0)		
Other	8 (3.7)		
Baseline Body Mass Index	29.8 (19.8, 43.5, 4.8)		0
Pre-op WOMAC Score	32.1 (16.0, 0, 86)		12
Kellgren and Lawrence Scores			26
0	2 (1.1)		
1	2 (1.1)		
2	16 (9.0)		
3	56 (31.3)		
4	103 (57.5)		
OARSI Scores (medial compartment on left, lateral compartment on right)			26
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 degree knee flexion contracture^ or moderate or severe laxity#		84 (41.0)	0
Patellofemoral scores*			
1	1 (2.9)		
2	10 (29.4)		
3	17 (50)		
4	6 (17.7)		

^Flexion contracture measures were obtained only at baseline.

#Knee laxity measures were available only during years 2 and 3.

*Patellofemoral scores are reported for the patients who required patellofemoral grades for classification using the system by Escobar et al.