Published in final edited form as:
Arthritis Care Res (Hoboken). 2015 March ; 67(3): 349–357. doi:10.1002/acr.22428.

Using Surgical Appropriateness Criteria to Examine Outcomes of Total Knee Arthroplasty in a United States Sample

Daniel L. Riddle, PT, PhD, FAPTA[1,2], **Robert A. Perera, PhD**[3], **William A. Jiranek, MD**[2], and **Levent Dumenci, PhD**[4]

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

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

[3]Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia, USA

[4]Department of Social and Behavioral Health, Virginia Commonwealth University, Richmond, Virginia, USA

Abstract

Objective—We determined outcomes for patients classified as appropriate, inconclusive or inappropriate for total knee arthroplasty (TKA) using a modified version of a validated appropriateness algorithm. Outcome measurement was conceptualized as short-term postoperative change attributable primarily to surgery and rehabilitation (two-months) and as longer term post-operative change and recovery (one- and two-year).

Methods—Pre-operative and yearly post-operative WOMAC Function, KOOS Symptoms and KOOS Pain scores were examined for persons undergoing primary TKA in the Osteoarthritis Initiative. Multi-group two-piece latent growth curve modeling was used to determine differences in outcome variable changes for each group from pre- to two-months post-surgery as well as over a two-year post-operative period.

Results—Data from 167 persons with primary TKA were examined. Prevalence rates of appropriate, inconclusive and inappropriate judgments were 47.9%, 20.8%, and 31.3%, respectively. The inappropriate group showed no change at two months following surgery while appropriate and inconclusive groups had substantial improvement in all outcomes. One-year and two-year post-operative recovery outcomes were not significantly different among the three groups.

Conclusion—The inappropriate group was unchanged two months after surgery and, on average, improved by 2.3 WOMAC Function points from pre-surgery to one year following surgery based on our models. Appropriate and inconclusive groups improved by an average of 19.8 WOMAC Function points at one year post-surgery. These data provide a compelling case for

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.

The authors had no financial support or other benefits from commercial sources for the work reported on in the manuscript, or any other financial interests which could create a potential conflict of interest or the appearance of a conflict of interest with regard to the work.

consensus building efforts to define eligibility criteria for TKA with the goals of reducing variation in patient selection and optimizing both change over time and final outcomes.

Keywords

Knee; arthroplasty; prognosis; pain; function

Extensive variation in patients' characteristics and in surgical rates for patients undergoing a variety of procedures has been well documented. For example, a substantial literature supports the presence of geographic(1), racial and ethnic disparities in total knee arthroplasty (TKA)(2). High rates of variation are not isolated to TKA and have been found for multiple surgical procedures(3). Lawson and colleagues(3), among others(4) have suggested that high variations in surgery rates are likely attributable, at least in part, to either under- or over-utilization.

Attempts to quantify the extent of overuse or underuse of a medical procedure require the use of appropriateness criteria to judge indications for the procedure. The RAND-UCLA Appropriateness Method has been used extensively in the US and elsewhere(3;5) to determine the extent of over- or under-utilization of a variety of procedures including bariatric surgery(6), hysterectomy(7), and TKA(8). We used appropriateness criteria for TKA developed by Escobar and colleagues(8) and described in a prior publication(9), to study utilization of TKA. To our knowledge, this will be the first study of TKA appropriateness conducted in the US.

We used variables(9) in the Osteoarthritis Initiative dataset(10) to reflect the appropriateness classification system used by Escobar and colleagues(8). We calculated prevalence rates for 175 patients undergoing isolated TKA in OAI and found a rate of 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.

The purpose of the current study was to determine if outcomes following TKA vary for the three classification groups(8). Outcome measurement in TKA can be conceptualized either as a journey, that is, change in outcome score over time, or as a destination, the outcome score the person ends up with(11). Both are important. The journey captures the magnitude of improvement (or worsening) while the destination describes the extent to which pain and functional status at a later point in time is acceptable to the patient. We hypothesized that subjects classified as inappropriate would have smaller changes in pain and function (i.e., a shorter journey) following surgery as compared to persons classified as inconclusive or appropriate for TKA(12). In regard to the destination, because the OAI data does not report Patient Acceptable Symptom State (PASS) measures(13;14), we were unable to compare groups on PASS. However, pre-operative scores are powerful predictors of post-operative status(15-17) such that persons with less pain and better function prior to surgery should have less pain and better function after surgery compared to more severely involved patients. We therefore hypothesized that the one- and two-year post-operative destination outcome scores would be better (i.e. less pain and better function) for persons classified as inappropriate as compared to the other two groups.

METHODS

A prior publication(9) summarized the eligibility criteria, radiographic measures and additional classification criteria of the Escobar et al system(8) and modifications to the system for application to Osteoarthritis Initiative (OAI) data. Briefly, variables in the OAI were matched to variables used by Escobar and colleagues (see Table 1) in an algorithm validated by Escobar et al(8) and designed to categorize surgeries as appropriate, inappropriate or inconclusive. No modification was made to algorithm structure. Figures 1a and 1b illustrate the algorithm used to classify surgeries into the three groups. Subjects eligible for the current study underwent isolated TKA surgery and no follow-up knee or hip replacement surgery during the 5-year study period. In our prior work, we classified 175 of 205 people with TKA and who had complete data for appropriateness classification. Because we were interested in exploring outcome of isolated TKA, we excluded 38 of the 205 who had an additional hip or knee arthroplasty during the study period.

We used the Western Ontario and McMasters Universities Arthritis Index (WOMAC) Function Scale, the KOOS Pain Scale and the KOOS Symptoms Scale to comprehensively examine two-year outcomes following TKA surgery. The WOMAC Function scale asks the patient to rate the difficulty associated with 17 common activities such as walking, standing and stair climbing. The WOMAC Function scale has been extensively validated(18;19) and ranges from 0 (normal function) to 68 (severely affected function). The Knee Injury and Osteoarthritis Outcome Score (KOOS) Symptoms scale is a reliable and valid measure of seven-day knee level symptoms pertaining to knee swelling, grinding, catching, stiffness and motion(20;21). The KOOS Pain scale is a reliable and valid measure of function-related pain and contains nine items(20;21). We chose the KOOS Pain scale over the WOMAC Pain scale because KOOS's 9 items provide a more comprehensive and therefore likely a more internally consistent estimate of function-related pain than the 5-item WOMAC Pain scale. Both KOOS scales are scored from 0 to 100 with 100 indicating no pain or symptoms. Subjects completed each outcome measure during the visit prior to knee replacement surgery and annually for two years following surgery. We used the data collected prior to surgery to classify patients into one of three categories of appropriate, inappropriate or inconclusive(22).

Data Analysis

We were interested in determining whether either the journey (i.e., changes over time) or the destination (i.e., one- or two-year outcome) was different among the three classification groups. In addition, we needed an analytic approach that accounted for the fact that in the OAI, time from surgery to the pre- and postoperative study visits varied for each subject. Two-piece latent growth curve (LGC) modeling with individually varying times of observations and its multi-group extension(23) were used to estimate the trajectories (i.e. the journeys) of outcome scores for each of the appropriateness groups. LGC modeling was chosen for its ability to model individually varying times of outcome measurement, to model data that is missing at random using the full information maximum likelihood estimation method, to estimate non-linear trends, and to test for differences in trajectories between groups.

A two-piece LGC model was used to account for the trends expected to appear in the data. Three model-based intercept points (i.e., specific time points following surgery) were considered: (a) two-months post-surgery, (b) one-year post-surgery, and (c) two-year post-surgery. The model with the intercept set at two-months post-surgery allowed us to compare three appropriateness groups after post-surgical pain has mostly subsided and post-acute rehabilitation approaches completion (24). Similarly, setting the intercept at one- and two-year post-surgery allowed us to compare model predicted outcomes (i.e. destinations) between the three appropriateness groups after a more complete recovery. With this model specification, two slopes were estimated simultaneously with an inflection point being specified at two-months post-surgery. The first slope (i.e., Slope 1) represented the estimated change in the outcome due to surgery and post-surgical rehabilitation (i.e., the difference from pre-surgery to two-months post-surgery), and the second slope (i.e., Slope 2) represented the longer-term post-surgery recovery period (i.e., the slope from two-month post-surgery until two-years post-surgery).

The models were first fit in an exploratory manner to each of the three outcomes separately for each group. Similar to the mixed model and random coefficient regression, LGC modeling allows intercepts and slopes to be random variables. Once the initial two-piece LGC model was fit, non-significant variance and covariance parameters were fixed to zero to obtain the most parsimonious model. Once the model for each appropriateness category was determined for a particular outcome, multi-group LGC analysis was used to test mean differences in the three intercepts and two slopes for the three classification groups. We combined groups that showed no statistical difference in the estimated growth trajectories in the multi-group LGC analysis. The estimated trajectories for each outcome measure are presented graphically for each appropriateness group along with 90% confidence intervals using parameter estimates from the multi-group LGC analysis. All fixed and random growth parameters reported in this study were statistically significant when $p < 0.05$. Mplus (version 6.11) software was used to fit the models.

RESULTS

A total of 167 persons underwent primary unilateral TKA surgery and no other arthroplasty surgery during the five year period. Of these, 144 persons (86.2%) had complete data for classifying appropriateness. The mean ages for the three classification groups along with other demographic variables and outcome scores over the perioperative period are reported in Table 2. A total of 19, 30, 35, 39 and 44 TKA surgeries were conducted in years one through five, respectively. The mean number of days from surgery to the pre-operative study visit and each subsequent post-operative visit over the course of two years of recovery are summarized in Table 2. The prevalence rates of appropriate, inconclusive and inappropriate classifications were 47.9% (95%CI, 39.7, 56.1, n=69), 20.8% (95%CI, 14.2, 27.4, n=30) and 31.3% (95%CI, 23.7, 38.9, n=45), respectively.

Results for the two-piece latent growth curve models for each outcome and appropriateness group are presented in the supplementary appendix. When comparing the appropriate and inconclusive groups, no significant differences were found in KOOS symptoms

$\left(\chi^2_{(3)}=7.7 \quad p=0.052\right)$, WOMAC function $\left(\chi^2_{(3)}=0.0 \quad p=0.998\right)$, or KOOS pain $\left(\chi^2_{(3)}=1.7 \quad p=0.634\right)$. Given these results, we combined data from appropriate and inconclusive groups for subsequent analyses.

Final models were tested by fitting the three groups simultaneously with the intercept parameterized at two months post-surgery. Since no differences were found between the appropriate and inconclusive groups, the growth parameters for these models were set to be equal. Based on the estimates from the individual group models, slope 1 for the inappropriate group was allowed to differ from the appropriate/inconclusive group and the intercept and slope 2 were constrained to be equal. Model tests of these constraints show no significant differences for KOOS Symptoms $\left(\chi^2_{(5)}=10.7 \quad p=.06\right)$, KOOS Pain $\left(\chi^2_{(5)}=3.6 \quad p=.60\right)$, or WOMAC 8 Function $\left(\chi^2_{(5)}=1.0 \quad p=.96\right)$. Given that the intercept was specified at the inflection point of the model, and that the intercept and slope 2 were not found to significantly differ from each other, intercepts at one- and two-year post-surgery were constrained to be equal. Statistical tests comparing the alternatively parameterized models are equivalent. The final estimated growth trajectories for the three groups and three outcomes are illustrated in Figure 2. Differences are seen in trajectories for the preoperative to two-month postoperative period across groups and the shared trajectory beginning at 2 months post-surgery.

DISCUSSION

We expected the inappropriate group to have smaller improvements (shorter journeys) following surgery because they had less preoperative pain and better function, but to also have better two-month, one-year and two-year outcomes as compared to inconclusive and appropriate groups. A substantial literature supported both hypotheses(15-17). We found that patients classified as inappropriate had no significant change in any of the three outcomes measures over the early perioperative period in contrast to substantial changes for the other two groups. WOMAC Function scores, for example, were unchanged in the inappropriate group at two months post-surgery while the appropriate and inconclusive groups improved by an average of 15.8 points, a substantial and clinically important improvement(25). We consider this difference to be important because evidence supports substantial improvements in pain and self-reported function after the first few months following surgery(26;27). Longer term changes in self-reported function for the inappropriate group were statistically significant and improved by 6.4 WOMAC Function points relative to pre-surgical scores two years following surgery. Appropriate and inconclusive groups, in contrast, improved by 23.8 points based on the models over the same interval. Because patients classified as inappropriate demonstrated no change over the short term and small improvements in symptoms and pain up to two years after surgery, we contend that patients classified as inappropriate received substantially less function and pain-related benefit (their journey resulted in small changes) in contrast to patients in the appropriate and inconclusive groups.

For destination outcomes, our study was limited because we did not have measures of PASS to determine whether patients found their outcomes to be acceptable. As an alternative, we determined whether the inappropriate group had higher self-reported function, less symptoms and less pain at two months, one and two years post-surgery than the appropriate and inconclusive groups. Because preoperative status is among the most powerful predictors of postoperative status(17;27;28) we expected the inappropriate group to have better follow-up function and less symptoms and pain than the other two groups. Contrary to our hypothesis, we found that all three groups had essentially the same short and longer term destination outcomes. Our 2-year outcomes generally indicated continued improvement relative to the 1-year and earlier time points which are consistent with recent outcomes literature(29-31).

The patterns of change across the three outcome measures for the three groups also differed. For the appropriate and inconclusive groups, the greatest improvement occurred between the pre-operative and two-months post-operative assessments. These early improvements primarily reflect therapeutic effects attributable to surgery and post-operative rehabilitation. Continued slight improvements, reflected by the significant slopes in the statistical models occurred over the extended post-operative period which is consistent with the literature (17;32). For the inappropriate group, the most substantial improvements occurred only during the longer term post-operative period and only for KOOS Pain and Symptom measures and these improvements were relatively small. Patterns of change over the course of study for the inappropriate group have not been described in the literature, to our knowledge. We speculate that because these patients had mild pain and functional loss prior to surgery, their journey toward recovery was both delayed and substantially less overall than the other two groups. The immediate surgical trauma effects on pain and function in a group of patients with mild pre-operative pain and functional deficits apparently resulted in no early post-operative improvement. Only after recovery from the acute surgical effects did these subjects experience mild improvement in pain and knee symptoms.

Quintana and colleagues(12) reported 6-month outcomes for TKA patients using the Escobar et al system(8). Persons classified as inappropriate had a 40% to 50% improvement in their WOMAC Pain and Physical Function scores which contrasts to our findings of an approximate 10% improvement six months post-surgery. The substantial differences in pre-operative scores for the patients classified as inappropriate in the two samples may, in part, explain these seemingly contradictory findings. In our study, patients in the inappropriate group had average pre-operative WOMAC Physical Function scores of 12.8 (sd=9.7) as compared to average scores of 31.4 (sd not reported) in the study by Quintana and colleagues. Lower WOMAC scores indicate better functioning. We speculate that because our patients in the inappropriate group had less pain and disability prior to surgery compared to the Quintana et al sample, changes following surgery were substantially less. Replication studies with larger samples are clearly needed to clarify this issue.

Quintana and colleagues reported a prevalence of 12.4% of patients classified as inappropriate for TKA(12). This was substantially lower than the 31.3% rate in our study. While our modifications to the algorithm are a potential contributor to this difference, we attribute most of these differences to the greater variation in patient characteristics in the US sample including the higher proportion of US patients with milder levels of pain, knee OA and functional loss.

One factor that may have influenced our findings was the number of days between data collection sessions and the day of surgery. There were no significant differences in the number of days from surgery to the data collection sessions for the three groups (see Table 2) but there were subtle differences. Persons in the inappropriate group had their pre-operative visit, on average, approximately three weeks earlier than persons in the other two groups. We suspect it is unlikely this difference played a role in our findings. Pain or functional status worsening prior to surgery is very gradual and likely only affects pain and function scores by a few points if at all, after waiting times of six months to a year(33;34). We also correlated our pre-operative KOOS Pain and WOMAC Physical Function scores with days prior to surgery and found no association (r = 0.01 and r = -0.04, p>0.5). Our analytic approach combined with prior evidence supports the notion that number of days between pre-operative assessment and surgery was unlikely to influence our findings.

The group classified as inconclusive (20.8% of our sample) was found to respond similarly to TKA surgery as those classified as appropriate. However, the inconclusive category is, by definition a group that could not be classified as either appropriate or inappropriate in Escobar et al's original work(8). This group had the greatest variation among the characteristics used for classification and, in our view, likely represents one of the more formidable challenges to refining classification in future research.

Our study has some substantial limitations. Our sample size is small for an outcome study particularly when considering that we stratified the sample into three groups. Because of this small sample, we were unable to adjust for factors that may have influenced longer term outcome such as the number of other symptomatic joints(35). Our small sample size also likely influenced statistical power particularly for KOOS symptoms score comparisons for the appropriate and inconclusive groups. For example, the p value for appropriate and inconclusive group comparisons was p= 0.052 for the two-piece LGC models. We chose to interpret this according to our preset significance level of p< 0.05. We would need a larger study to detect significant differences between appropriate and inconclusive for KOOS symptoms, assuming approximately the same differences as that reported in this study. The p values for pain and function outcome measures were large (i.e., p ≥ 0.6) which suggests that the clinical importance of the KOOS symptoms findings are likely to be minimal given that pain and function outcomes were very similar for the appropriate and inconclusive groups. A larger sample replication study also would provide important information about the extent to which our findings might generalize to other groups. We also were unable to compare PASS among the groups(36). With this said, the magnitude of improvement in persons classified as inappropriate one year following surgery was 2.3 WOMAC Function points as compared to an average 19.8 point WOMAC improvement for persons classified as appropriate or inconclusive. A change of two WOMAC Function points, while not directly comparable to PASS estimates for patients with TKA(13;36) is unlikely, in our view, to be substantial enough to achieve a PASS for most patients.

We cannot rule out the possibility of regression to the mean for patients with high levels of pain and disability like those in the appropriate and inconclusive groups and floor effects because of mild pain and disability such as those in the inappropriate group. Knee arthritis pain fluctuates(37) and persons in the inappropriate group may have reported mild pain and disability pre-operatively because they were adept at avoiding pain related activity or they may have completed preoperative forms when their symptoms were relatively mild. Alternatively, persons in the other two groups may have more commonly completed pre-operative forms on a more symptomatic day which may have artificially inflated scores. A strength of the OAI is that it is not designed as a knee arthroplasty study and no care was provided as part of OAI so it is unlikely that social desirability(38), the tendency to please a health care practitioner, played a role in the study. We modified the original appropriateness classification system proposed by Escobar and colleagues. These modifications may have resulted in some misclassification.

Losina and Katz contended that it was time for health care providers, funders and policy makers to determine whether the journey, the destination or both should drive determinations of TKA appropriateness(11). Our data provides some evidence to support this endeavor. Persons who were classified as inappropriate had no significant change in pain, function and knee symptoms over the two-month postoperative period. This contrasts to the substantial improvements over the same period in appropriate and inconclusive groups. In addition, despite having higher functioning and substantially less pain pre-operatively, the destination outcomes of patients classified as inappropriate were not statistically different as compared to persons classified as appropriate or inconclusive. Overall, improvements in the group classified as inappropriate were small and in some cases did not reach clinical significance one year following surgery.

The small sample size available does provide limited power for detecting differences between groups. Further comparison of appropriate and inconclusive groups would benefit from studies with higher power. Despite this, our findings of differences between groups are statistically robust with clear clinically important differences among groups in the magnitude and pattern of change. Our study design precludes us from concluding either that the appropriateness criteria we studied are valid for US patients or that patients receive TKA only when they need it. Our study does, however, provide evidence to suggest that one or the other (or both) of these inferences is likely to be false. To advance the science, future work needs to examine the validity of appropriateness criteria and the extent to which TKA decisions are medically defensible. Our findings support consensus building efforts by orthopedic surgeons, other members of the healthcare team, and additional key stakeholders including patients to define eligibility criteria for TKA with the goal of reducing variation in patient characteristics and maximizing outcome.

Acknowledgments

Funding Source: 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.

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. Losina E, Kessler CL, Wright EA, Creel AH, Barrett JA, Fossel AH, et al. Geographic diversity of low-volume hospitals in total knee replacement: implications for regionalization policies. Med Care. 2006; 44(7):637–45. [PubMed: 16799358]
2. Morgan RC Jr. Slover J. Breakout session: Ethnic and racial disparities in joint arthroplasty. Clin Orthop Relat Res. 2011; 469(7):1886–90. [PubMed: 21503786]
3. 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]
4. 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]
5. 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]
6. Yermilov I, McGory ML, Shekelle PW, Ko CY, Maggard MA. Appropriateness criteria for bariatric surgery: beyond the NIH guidelines. Obesity (Silver Spring). 2009; 17(8):1521–7. [PubMed: 19343019]
7. Shekelle PG, Park RE, Kahan JP, Leape LL, Kamberg CJ, Bernstein SJ. Sensitivity and specificity of the RAND/UCLA Appropriateness Method to identify the overuse and underuse of coronary revascularization and hysterectomy. J Clin Epidemiol. 2001; 54(10):1004–10. [PubMed: 11576811]
8. 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]
9. Riddle DL, Jiranek WA, Hayes CW. Using a validated algorithm to judge the appropriateness of total knee arthroplasty in the United States: A multi-center longitudinal cohort study. Arthritis Rheumatol. 2014 In press.
10. Lester G. Clinical research in OA--the NIH Osteoarthritis Initiative. J Musculoskelet Neuronal Interact. 2008; 8(4):313–4. [PubMed: 19147953]
11. Losina E, Katz JN. Total knee replacement: pursuit of the paramount result. Rheumatology (Oxford). 2012; 51(10):1735–6. [PubMed: 22843792]
12. 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]
13. Escobar A, Gonzalez M, Quintana JM, Vrotsou K, Bilbao A, Herrera-Espineira C, et al. Patient acceptable symptom state and OMERACT-OARSI set of responder criteria in joint replacement. Identification of cut-off values. Osteoarthritis Cartilage. 2012; 20(2):87–92. [PubMed: 22155074]
14. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Ann Rheum Dis. 2005; 64(1):34–7. [PubMed: 15130902]
15. Fortin PR, Clarke AE, Joseph L, Liang MH, Tanzer M, Ferland D, et al. Outcomes of total hip and knee replacement: preoperative functional status predicts outcomes at six months after surgery. Arthritis Rheum. 1999; 42(8):1722–8. [PubMed: 10446873]
16. Judge A, Arden NK, Cooper C, Kassim JM, Carr AJ, Field RE, et al. Predictors of outcomes of total knee replacement surgery. Rheumatology (Oxford). 2012; 51(10):1804–13. [PubMed: 22532699]

17. 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]

18. Bellamy N. Pain assessment in osteoarthritis: experience with the WOMAC osteoarthritis index. Semin Arthritis Rheum. 1989; 18(4 Suppl 2):14–7. [PubMed: 2786253]

19. Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, et al. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. J Rheumatol. 1997; 24(4):799–802. [PubMed: 9101522]

20. Roos EM, Lohmander LS. The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis. Health Qual Life Outcomes. 2003; 1(1):64. [PubMed: 14613558]

21. Roos EM, Toksvig-Larsen S. Knee injury and Osteoarthritis Outcome Score (KOOS) -validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcomes. 2003; 1(1):17. [PubMed: 12801417]

22. Riddle DL, Perera RA, Stratford PW, Jiranek WA, Dumenci L. Progressing toward, and recovering from, knee replacement surgery: a five-year cohort study. Arthritis Rheum. 2013; 65(12):3304–13. [PubMed: 23983118]

23. Muthen BO. Beyond SEM: General latent variable modeling. Behaviormetrika. 2002; 29:81–117.

24. Westby MD, Brittain A, Backman CL. Expert consensus on best practices for post-acute rehabilitation after total hip and knee arthroplasty: A Canada-US Delphi study. Arthritis Care Res (Hoboken). 2013

25. Escobar A, Quintana JM, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. Osteoarthritis Cartilage. 2007; 15(3):273–80. [PubMed: 17052924]

26. Davis AM, Perruccio AV, Ibrahim S, Hogg-Johnson S, Wong R, Badley EM. Understanding recovery: changes in the relationships of the International Classification of Functioning (ICF) components over time. Soc Sci Med. 2012; 75(11):1999–2006. [PubMed: 22940011]

27. Kennedy DM, Stratford PW, Riddle DL, Hanna SE, Gollish JD. Assessing recovery and establishing prognosis following total knee arthroplasty. Phys Ther. 2008; 88(1):22–32. [PubMed: 17986495]

28. Escobar A, Quintana JM, Bilbao A, Azkarate J, Guenaga JI, Arenaza JC, et al. Effect of patient characteristics on reported outcomes after total knee replacement. Rheumatology (Oxford). 2007; 46(1):112–9. [PubMed: 16735451]

29. Johnston L, Maclennan G, McCormack K, Ramsay C, Walker A. The Knee Arthroplasty Trial (KAT) design features, baseline characteristics, and two-year functional outcomes after alternative approaches to knee replacement. J Bone Joint Surg Am. 2009; 91(1):134–41. [PubMed: 19122088]

30. Williams DP, Price AJ, Beard DJ, Hadfield SG, Arden NK, Murray DW, et al. The effects of age on patient-reported outcome measures in total knee replacements. Bone Joint J. 2013; 95-B(1):38–44. [PubMed: 23307671]

31. Xie F, Lo NN, Pullenayegum EM, Tarride JE, O'Reilly DJ, Goeree R, et al. Evaluation of health outcomes in osteoarthritis patients after total knee replacement: a two-year follow-up. Health Qual Life Outcomes. 2010; 8:87. [PubMed: 20723239]

32. Fortin PR, Penrod JR, Clarke AE, St-Pierre Y, Joseph L, Belisle P, et al. Timing of total joint replacement affects clinical outcomes among patients with osteoarthritis of the hip or knee. Arthritis Rheum. 2002; 46(12):3327–30. [PubMed: 12483739]

33. 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]

34. 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]

35. Perruccio AV, Power JD, Evans HM, Mahomed SR, Gandhi R, Mahomed NN, et al. Multiple joint involvement in total knee replacement for osteoarthritis: Effects on patient-reported outcomes. Arthritis Care Res (Hoboken). 2012; 64(6):838–46. [PubMed: 22570306]

36. Judge A, Arden NK, Kiran A, Price A, Javaid MK, Beard D, et al. Interpretation of patient-reported outcomes for hip and knee replacement surgery: identification of thresholds associated with satisfaction with surgery. J Bone Joint Surg Br. 2012; 94(3):412–8. [PubMed: 22371552]

37. Hawker GA, Stewart L, French MR, Cibere J, Jordan JM, March L, et al. Understanding the pain experience in hip and knee osteoarthritis--an OARSI/OMERACT initiative. Osteoarthritis Cartilage. 2008; 16(4):415–22. [PubMed: 18296075]

38. Deshields TL, Tait RC, Gfeller JD, Chibnall JT. Relationship between social desirability and self-report in chronic pain patients. Clin J Pain. 1995; 11(3):189–93. [PubMed: 8535037]

39. Katz JN, Chang LC, Sangha O, Fossel AH, Bates DW. Can comorbidity be measured by questionnaire rather than medical record review? Med Care. 1996; 34(1):73–84. [PubMed: 8551813]

SIGNIFICANCE AND INNOVATION

- Early recovery trajectories for the subgroup of patients with total knee arthroplasty classified as inappropriate showed no change while the other two appropriateness groups showed substantial and clinically important improvement.
- Longer term trajectories were the same across the three appropriateness groups despite the inappropriate group having substantially less pain and better function prior to surgery.
- These findings provide evidence for the need to develop consensus based criteria for total knee arthroplasty.

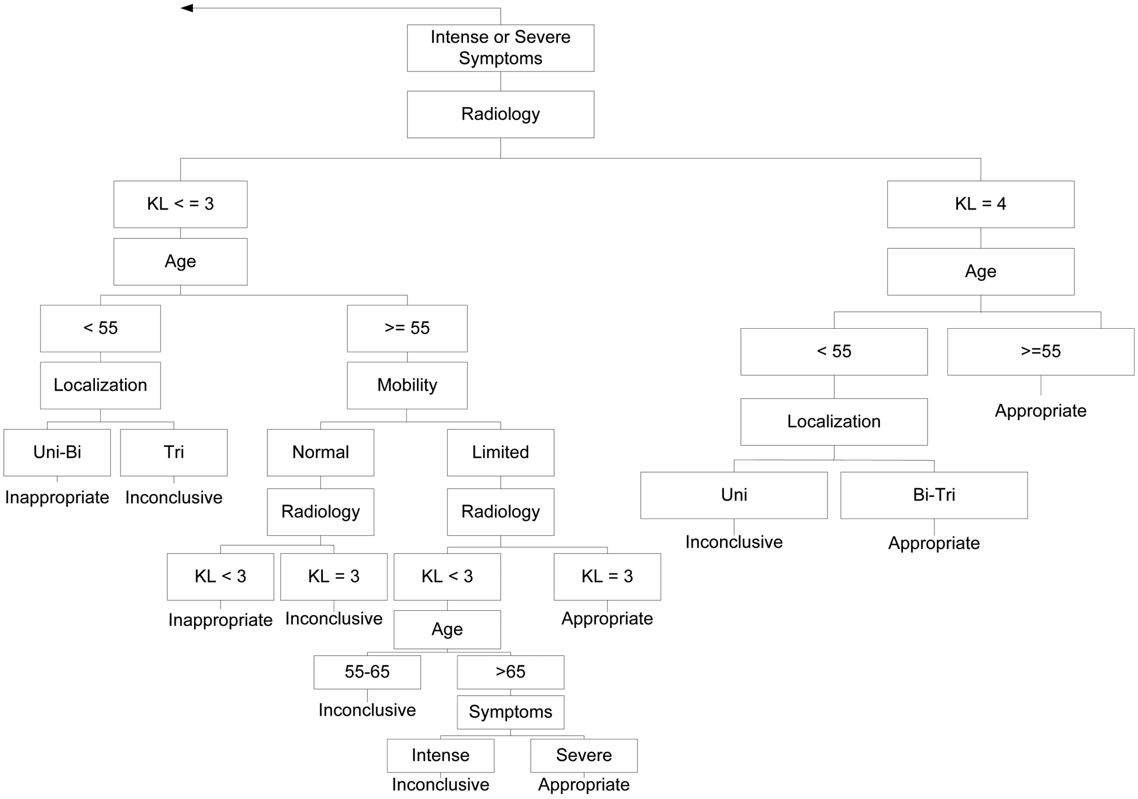

Figure 1a.

The right hand side of the algorithm modified from that originally described by Escobar and colleagues[8]. The original description of the adaptation of the algorithm by Escobar and colleagues has been published[9]. Table 1 provides a complete summary of each major category in the algorithm. Note that Figure 1a and 1b combined represent the flowchart used in the study.

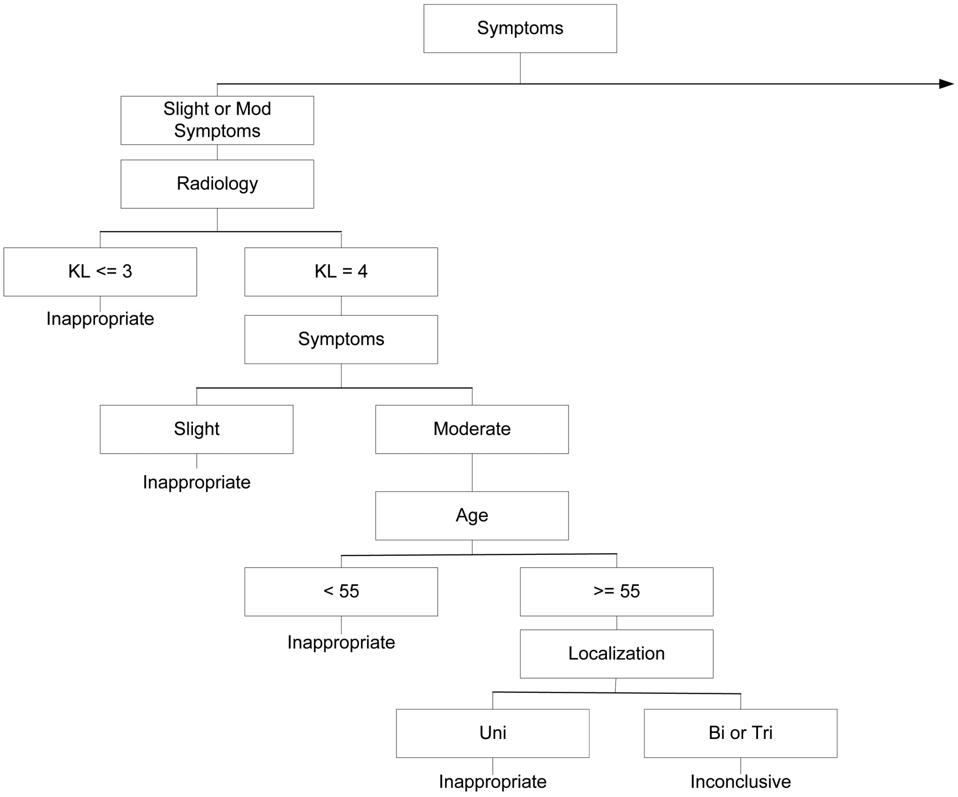

Figure 1b.

The left hand side of the algorithm modified from that originally described by Escobar and colleagues[8]. The original description of the adaptation of the algorithm by Escobar and colleagues has been published[9]. Table 1 provides a complete summary of each major category in the algorithm.

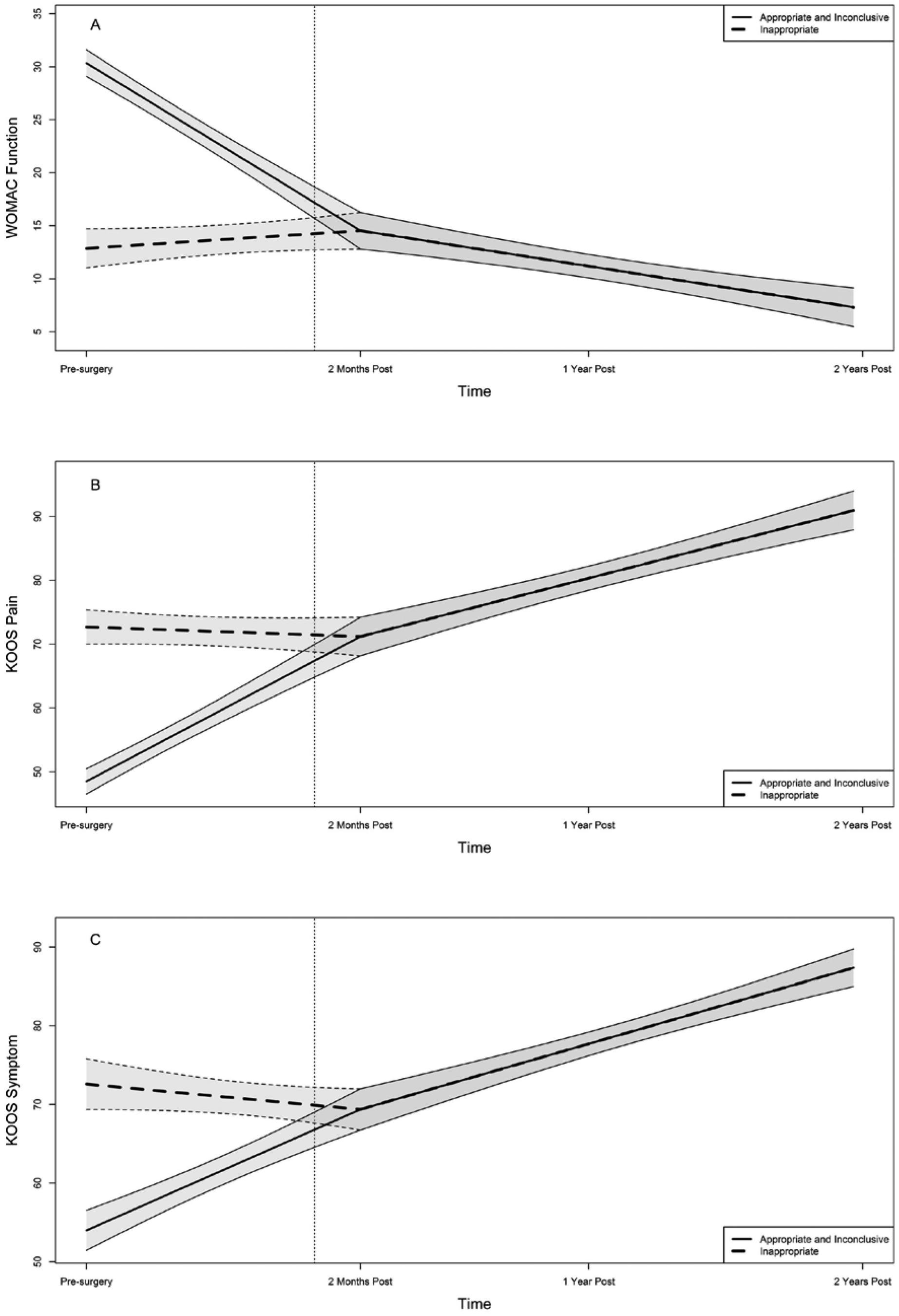

Figure 2.

The figure illustrates the early perioperative and later postoperative outcome trajectories for the WOMAC Function scale (Panel A), the KOOS Pain scale (Panel B), and the KOOS Symptoms scale (Panel C). The data were obtained pre-operatively and yearly over a two-year post-operative period. The heavy solid line represents the combined data for the appropriate and inconclusive groups while the heavy dashed line represents data from the inappropriate group. The thinner lines bounding each heavy line represent the 90% confidence intervals for the two sets of data and the dotted vertical line indicates the time of surgery. The WOMAC Physical Function score ranges from 0 to 68 with higher scores equating to worse function. The KOOS Symptoms and Pain scales range from 0 to 100 with higher scores equating to less symptoms and less pain.

Table 1

Comparison of appropriateness 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 2

Characteristics of the Three Classification Groups

	"Appropriate" Mean (sd)	"Inconclusive" Mean (sd)	"Inappropriate" Mean (sd)
Age	69.1 (7.2)	62.6 (8.6)	66.4 (10.1)
Gender (% female)	55.1	50.0	62.2
Body Mass Index	29.8 (5.0)	31.1 (5.2)	29.0 (4.1)
Pre-operative Morbidity Scale Score^			
WOMAC Physical Function+			
Preoperative (n = 69, 30, 45)*	30.3 (9.0)	30.7 (12.6)	12.8 (9.7)
One-year post-operative (n = 68, 30, 44)	12.5 (12.1)	12.8 (13.7)	13.6 (11.7)
Two-year post-operative (n = 45, 20, 32)	9.9 (9.4)	11.6 (13.6)	8.1 (10.0)
KOOS Symptoms+			
Preoperative (n = 69, 30, 45)	56.4 (17.2)	45.7 (19.6)	72.6 (17.0)
One-year post-operative (n = 68, 30, 45)	73.9 (18.8)	72.8 (20.5)	70.7 (19.9)
Two-year post-operative (n = 45, 21, 32)	82.1 (11.9)	75.2 (17.6)	85.2 (13.2)
KOOS Pain+			
Preoperative (n = 69, 30, 45)	49.4 (14.3)	44.9 (17.0)	72.7 (14.3)
One-year post-operative (n = 67, 30, 45)	75.8 (21.4)	75.2 (22.1)	75.3 (21.1)
Two-year post-operative (n = 45, 21, 32)	85.7 (17.0)	78.9 (20.5)	86.7 (17.3)
Time in days from surgery to data collection#			
Pre-operative visit	−167.6 (103.2)	−169.1 (100.1)	−198.8 (94.2)
First year post-operative visit	204.5 (98.2)	198.7 (101.1)	166.8 (84.7)
Second year post-operative visit	562.2 (110.3)	553.9 (116.6)	537.0 (95.1)

^Morbidity was quantified using the Modified Charlson Comorbidity Index (39).

*The sample sizes for each outcome variable and for each year are listed for the appropriate, inconclusive and inappropriate subgroups, respectively.

#A series of one-way analyses of variance procedures were used to compare the number of days for each classification for each time period. For the pre-operative visit, $F_{2,141}$ =1.4, *p* = 0.23, for the first post-operative visit $F_{2.140}$ = 2.3, *p* = 0.10, and for the second year visit, $F_{2.95}$ = 0.5. *p* = 0.60.

+The WOMAC Physical Function score ranges from 0 to 68 with higher scores equating to worse function. The KOOS Symptoms and Pain scales range from 0 to 100 with higher scores equating to less symptoms and less pain.