Osteoarthritis and Cartilage Open 6 (2024) 100482

Contents lists available at ScienceDirect

Osteoarthritis and Cartilage Open

journal homepage: www.elsevier.com/journals/osteoarthritis-and-cartilage-open/2665-9131

Head-to-head comparison of appropriate use criteria for knee arthroplasty: A multicenter cohort study

Daniel L. Riddle [a,*], Levent Dumenci [b]

[a] Departments of Physical Therapy, Orthopaedic Surgery and Rheumatology, 900 East Leigh Street, Room 4:100, Virginia Commonwealth University, Richmond, VA, USA
[b] College of Public Health, Department of Epidemiology and Biostatistics, Temple University, Philadelphia, PA, USA

ARTICLE INFO

Handling Editor: Professor H Madry

Keywords:
Knee
Arthroplasty
Classification
Appropriateness

ABSTRACT

Objective: To determine, in a head-to-head comparison, which of two RAND-based knee replacement appropriateness criteria is optimal based on comparison to an externally validated method of judging good versus poor outcome.

Design: Longitudinal data from the Osteoarthritis Initiative (OAI) and the Multicenter Osteoarthritis Study (MOST) were combined to produce a dataset of 922 persons with knee arthroplasty, 602 of which had adequate data for RAND classification and had their surgery within one year prior to a study visit. Data were used to determine appropriateness classification (i.e., Appropriate, Inconclusive, Rarely Appropriate) using modified versions of the first-generation and second-generation Escobar system. Growth curve analyses and multivariable regression were used to compare the two systems.

Results: Neither system associated with the gold standard measure of good versus poor outcome. Distributions of appropriateness categories for the second-generation system were inconsistent with current evidence for knee arthroplasty outcome. For example, 16% of participants were classified as Appropriate and 64% as Rarely Appropriate for pain outcome. Distributions for the first-generation system aligned with current evidence.

Conclusion: The first-generation modified version of the Escobar appropriateness system is superior to the newer version but neither version associated with our gold standard growth curve analyses. Both systems only differentiate between patient classification groups preoperatively and up to ten months following surgery. Reliance on appropriateness criteria to inform long-term outcome is not warranted.

1. Introduction

Knee arthroplasty (KA) is the most common major surgery conducted in the US totaling approximately one million procedures per year [1]. Success of KA in relieving pain and improving function, combined with low complication rates have rendered this procedure one of the most effective and cost-effective surgical procedures. Despite the remarkable success and growth in utilization of KA, approximately 20% of patients [2] (i.e., approximately 200,000 US patients per year) [1] have persistent pain and/or compromised daily life activity after surgical recovery. These data are worrisome and suggest, in part, that the decision to undergo surgery may not have been optimized for some patients. Lack of surgical optimization is supported by wide variation in disease and symptom burden that exists in patients undergoing KA [3,4]. Although KA is an elective procedure, timeliness of surgery is an important determinant of success. Utilizing KA too early may expose patients to albeit low, but serious risks of a major operation, yet result in little to no improvement in function and pain compared to preoperative status [5]. On the other hand, waiting too long to have surgery [6,7], in addition to unnecessarily prolonging suffering, may result in worse outcomes and an increased risk of obesity, diabetes, cardiovascular disease and subsequent premature death [8].

An evidence-based decision tool is needed to fully account for major prognostic factors impacting KA outcome and to inform the clinic discussion between the patient and surgeon. Appropriateness criteria (AC) were originally proposed more than three decades ago to define baseline classification criteria using a comprehensive prognostic literature synthesis, for predicting treatment outcome for interventions not conducive to randomized trial evidence (i.e., knee arthroplasty) and to develop evidence-based metrics to help determine the timeliness of poor outcome risks for surgical procedures [9,10]. Recognizing that all stakeholders (e.g., patients, patients' families, providers and payers) would likely

* Corresponding author.
E-mail addresses: dlriddle@vcu.edu (D.L. Riddle), ldumenci@temple.edu (L. Dumenci).

https://doi.org/10.1016/j.ocarto.2024.100482
Received 8 March 2024; Accepted 5 May 2024

benefit if evidence-based criteria for KA decision-making were implemented widely, we conducted a series of investigations over the last several years examining a variety of appropriate use criteria including first [5,11,12] and second-generation [13,14] Escobar systems, as well as AAOS arthroplasty appropriateness systems [15,16]. We demonstrated that a modified version of the first-generation Escobar system could be used to identify a subgroup of KA patients experiencing minimal benefit (i.e., poor-responders) up to two years following KA [5,11]. We have also developed and preliminarily validated a new second-generation Escobar system using data from Spain [13,14]. The purpose of the current study was to conduct a head-to-head comparison of the first-generation Escobar system [5,11,12] to the newly developed second-generation Escobar system [13,14] using multicenter data from two NIH-funded US-based longitudinal studies. The two RAND-based Escobar systems are arguably the strongest and most extensively studied KA appropriateness classification systems available [5,7,11–14,17]. A head-to-head comparison is urgently needed to determine which method is superior going forward. The overarching hypothesis was that the second-generation Escobar system will be superior to the first-generation system in differentiating good outcome from poor outcome using an externally validated method [18,19].

2. Methods

Data on knee arthroplasty (KA) outcome from two public-use National Institutes of Health (NIH) funded independent datasets were analyzed [20,21]. The Osteoarthritis Initiative (OAI), is a prospective longitudinal cohort study with yearly assessment and nine years of follow-up. A total of 4796 participants between the ages of 45 and 79 years were consented, beginning in 2004. The purpose of OAI was to study the natural history, risk factors, outcomes, onset and progression of knee tibiofemoral OA. Participants in OAI were recruited from four clinical sites: 1) the University of Maryland School of Medicine 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.

The Multicenter Osteoarthritis study (MOST) is a prospective longitudinal cohort study with seven years of follow-up [21]. MOST investigated knee osteoarthritis by evaluating potentially modifiable risk factors for disease and poor pain and physical function outcomes. A total of 3026 participants were aged 50–79 years, with recruitment beginning in 2003. Participants were assessed at baseline, and at follow-up months 15, 30, 60, 72 and 84. Participants were recruited from two communities at the following clinical sites: 1) University of Iowa in Iowa City, Iowa, 2) University of Alabama, Birmingham in Birmingham, Alabama. MOST data are publicly available at https://agingresearchbiobank.nia.nih.gov/studies/most/. OAI data are available at https://nda.nih.gov/oai/. Both studies required all participants to read and sign IRB approved consent forms from each site prior to participation.

2.1. Knee arthroplasty samples from OAI and MOST

Participants reported if KA was conducted and if so, the date of surgery was confirmed with medical record review. A total of 922 participants underwent KA on at least one knee in the combined dataset, with 427 in OAI and 495 in MOST. KA was time varying in both OAI and MOST, meaning that KA could have occurred at any time between repeated study visits. We included only those participants who had their study visit within one year of their KA. Systematic review evidence indicates that self-reported preoperative pain measures are stable if taken a year or less prior to KA surgery [22]. From the total of 922 participants, n = 698 (76%) had complete data to allow for appropriateness classification [5,11]. Of these, a total of 602 of 698 participants (86.2%), 354 in OAI and 248 in MOST, had their preoperative study data collected within a year of KA. The combined sample of 602 participants were used in all primary analyses. A total of 20 of 354 (5.6%) participants in OAI had a partial KA. MOST did not report whether participants had a total or partial KA.

2.2. First-generation modified Escobar appropriateness variables

We used a previously validated set of measures [5,11], to classify patients as Appropriate, Inconclusive or Rarely Appropriate for KA. Appropriateness classification variables were Kellgren and Lawrence knee OA grade [23], and number of osteoarthritic knee compartments, as determined by validated anterior and lateral knee radiographs in MOST and knee radiographs and MRI in OAI [24]. Age was classified as < 55 years or 55–65 years or > 65 years. The first-generation Escobar system was modified by using Kellgren and Lawrence [23] radiographic ratings in place of Ahlback [25] classification and using combined WOMAC Pain and Disability scores to quantify knee symptoms in place of patient ratings of pain behavior and medication usage [12].

We substituted one variable to allow for use of the MOST data. The one exception was the knee joint mobility and stability variable. Because knee mobility and stability examinations were not conducted in MOST, we relied on the Knee Osteoarthritis Outcome Survey Sports and Recreational Activities self-report item #4, which was included in both datasets. If the preoperative item #4 "Difficulty pivoting and twisting on the injured knee" was rated as moderate or worse, the participant was coded as abnormal on the knee mobility and stability item while a score of none or minor was coded as normal.

Both the first-generation modified Escobar system and the second-generation system relied on the RAND Appropriateness system for classification [9]. The possible categories are Appropriate, Inconclusive, and Inappropriate. Consistent with other evidence [26], we elected to use the term Rarely Appropriate instead of Inappropriate to describe this RAND category. The algorithms used to define the Appropriate, Inconclusive and Rarely Appropriate classifications from both the first-generation and second-generation Escobar appropriateness systems [5,11] appear in Supplemental file 1.

2.3. Second-generation Escobar appropriateness variables

The second-generation system used the same methods described for the first-generation system including radiographic OA status [23], WOMAC Pain and Disability scores, included separately, and age, quantified using the following categories, <55 years, 55–65 years, >65–85 years and >85 years. Additionally, the second-generation system included more contemporary prognostic measures of psychological distress [27] and comorbidity [18,19]. In both the OAI and MOST, depressive symptoms were quantified using the Center for Epidemiologic Studies depression scale [28] and general psychological distress was quantified using the SF-12 Mental Component Summary score [29]. Comorbidity was quantified using the modified Charlson comorbidity index [30].

2.4. Preoperative and postoperative self-reported outcome variables

Both datasets included the outcomes of interest. Outcome variables were the Western Ontario and McMaster Universities Arthritis Index (WOMAC) Pain and Disability scales. The WOMAC Pain scale ranges from 0 to 20 with higher scores equating to worse pain with activity. The WOMAC Disability scale ranges from 0 to 68 with higher scores equating to more difficulty with daily activity. Both scales have been extensively validated in persons with KA [31,32]. The SF-12 Physical Component Summary (SF-12 PCS) [33], a validated generic health related quality of life scale ranging from 13 to 69 with higher scores equating to better physical health related quality of life was used in a sensitivity analysis.

The WOMAC Pain, WOMAC Disability and SF-12 PCS measures were obtained during a presurgical and three postsurgical follow-up visits in both datasets. Average number of months from the date of surgery to the preoperative visit was 5.9 months, the first postsurgical follow-up

occurred at an average of 10.3 months, the second follow-up visit mean was 23.4 months and the third follow-up mean was 40.0 months following surgery. Table 1 provides descriptive statistics for demographic variables.

2.5. Preoperative variables and the first- and second-generation criteria

In addition to the first and second generation appropriateness criteria variables, we included several additional preoperative variables based on prior evidence supporting prognostic utility for outcome association following KA [18,19,27,34–39]. Contralateral knee preoperative WOMAC Pain scores were used. Depressive symptoms were quantified with the CES-D, a validated depressive symptom measure [28]. The SF-12 Mental Component Summary score was completed and ranged from 0 (severe mental health dysfunction) to 100 (excellent mental health) [40]. Age in years, sex, self-reported race (African American or other), and BMI in kg/m^2 were quantified. Comorbidity was measured with the validated modified Charlson comorbidity index [41]. Preoperative opioid use, yes or no, was recorded. Bodily pain areas (n = 16) were measured in the following way: a study participant indicated on a body diagram all major joint regions, including both, shoulders elbows, wrists, hands, ankles and feet and including the cervical and lumbar spine. Pain had to be present on most days in the past 30 days. For hip pain questions, symptoms on most days for at least 1 month during the past 12 months was reported. Educational attainment was determined by asking the participant to indicate one of the following: less than high school degree, high school degree, some college, college degree, some graduate school, graduate school degree. Table 2 lists the variables in each system.

2.6. Data analysis

Analyses were conducted in three steps. First, separately for WOMAC Pain and WOMAC Disability outcomes, a two-piece latent class growth curve model (LCA) with individually varying times of observations was used to estimate poor and good outcomes using one preoperative and three postoperative measurements. The first piece represents the short-term change in outcome from pre-surgery to the knot, set at 10-months post-surgery, and the second piece from the knot to the last measurement occasion modeled the long-term changes in outcome. Data collection took place during study visits, which, relative to the timing of KA surgery, varied by person. Estimating curves with individually varying times of observations allows for the time of outcome measurement relative to the surgery to differ between persons. A two-piece LCA growth model used to define good and poor outcomes was based on our prior work [18,19] showing that the two-class solution (i.e., good versus poor outcome) was both optimal and externally validated. The 95% CI, entropy, and the mean most likely class membership a posteriori latent class probabilities were used to determine the quality of the latent class solution. Non overlapping CIs, entropy $\geq$0.70, and a posteriori latent class probabilities of $\geq$0.80 indicate good separation [42,43].

The second step was applied to first-and second-generation appropriateness outcomes. Two-piece latent growth curve modeling with individually varying times of observations was used to estimate curves for appropriate, inconclusive, and rarely appropriate categories of both appropriateness criteria.

For the third step, association models including a set of covariates were added to the LCA using a logistic link function. For each outcome,

Table 1
Preoperative sample characteristics (n = 602).

Variable	OAI (n = 354)	Missing OAI	MOST (n = 248)	Missing MOST
Age in yrs, mean (sd)	67.9 (8.5)	0	68.9 (7.4)	0
Female sex, n (%)	208 (58.8)	0	162 (65.3)	0
Race/Ethnicity, n (%)		1		0
American Indian or Alaska Native	–		–	
Asian	4 (1.1)		–	
Black or African American	41 (11.6)		27 (10.9)	
Hispanic or Latino/a	6 (1.7)		3 (1.2)	
Native Hawaiian or Other Pacific Islander	–		–	
White	298 (84.4)		221 (89.1)	
Other/not reported	10 (2.8)		1 (0.4)	
Education highest grade, n (%)		1		0
Less than high school	7 (2.0)		10 (4.0)	
High school degree	59 (16.7)		71 (28.6)	
Some college	90 (25.4)		72 (29.0)	
College degree	59 (16.7)		39 (15.7)	
Some graduate school	27 (7.6)		18 (7.3)	
Graduate degree	111 (31.4)		38 (15.3)	
Body Mass Index, mean (%)	30.2 (4.9)	4	32.3 (6.3)	1
Preop WOMAC Pain Score, mean (sd)	7.2 (3.8)	0	7.9 (4.0)	0
Preop WOMAC Function, mean (sd)	23.4 (11.8)	0	27.3 (11.9)	0
Preop SF-12 PCS Score, mean (sd)	37.7 (9.4)	59	36.7 (9.1)	3
Additional prognostic indicators				
Depressive symptoms, mean (sd)	6.9 (6.8)	7	8.0 (7.6)	0
Mental Health Summary, mean (sd)	56.0 (8.5)	9	55.0 (9.0)	18
Comorbidity, mean (sd)	0.4 (0.7)	7	0.7 (1.0)	0
Opioid use - yes (%)	37 (10.5)	0	44 (17.7)	0
Widespread pain - yes (%)	139 (39.3)	13	129 (52.0)	0
Pre-operative data timing in days				
Preoperative visit, mean (sd)	−174.7 (94.4)	0	−192.3 (95.4)	0
First postoperative visit, mean (sd)	197.3 (96.1)	8	453.2 (359.1)	1
Second postoperative visit, mean (sd)	555.5 (104.6)	48	925.7 (379.2)	45
Third postoperative visit, mean (sd)	916.7 (116.8)	94	1742.4 (138.8)	108
First-generation Classification		34		6
Appropriate, n (%)	169 (52.8)		175 (72.3)	
Inconclusive, n (%)	61 (19.1)		29 (12.0)	
Rarely Appropriate, n (%)	90 (28.1)		38 (15.7)	
Second-generation Classification		0		0
Appropriate, n (%)	48 (13.6)		49 (19.8)	
Inconclusive, n (%)	56 (15.8)		64 (25.8)	
Rarely Appropriate, n (%)	250 (70.6)		135 (54.4)	

Table 2
Baseline indication criteria variables used in the 1st and 2nd generation Escobar systems.

Escobar System	Indication Criteria	Measurement Scale
1st generation modified system	Age	<55 years, 55–65 years, >65 years
	Radiology: Kellgren and Lawrence (KL) grade	KL ≤ 2, KL = 3, KL = 4
	Knee osteoarthritis location	Unicompartmental, bicompartmental, tricompartmental
	Symptomatology: Combined WOMAC Pain and Disability Scale	Slight (0–11)
		Moderate (12–22)
		Intense (23–33)
		Severe (≥34)
	Mobility and stability:	moderate or worse = positive, none or minor = neg
	KOOS Sports and Recreation item #4	
	Appropriateness Rating	Rarely appropriate
		Uncertain
		Appropriate
2nd generation system	Age	<55 years, 55–65 years, >66–85 years, >85 years
	Radiology: Kellgren and Lawrence (KL) grade	KL ≤ 2, KL ≥ 3
	Knee Osteoarthritis Location	Unicompartmental,
		More than one compartment
		Either uni or multiple compartment*
	Pain: WOMAC Pain Scale scored 0 (best) to 100 (worst)	Slight (<3 S)
		Moderate (35–50)
		Severe (>50)
	Disability: WOMAC Disability Scale scored 0 (best) to 100 (worst)	Slight (<35)
		Moderate (35–54)
		Severe (>54)
	Anxiety or Depression: Hospital Anxiety and Depression Scale	≤10 in both anxiety and depression
	score 0 (best) to 21 (worst)	>10 on either anxiety or depression
	PAIN CATASTROPHIZING: Pain	≤30, >30
	Catastrophizing Scale scored 0 (best) to 52 (worst)	
	Comorbidities	None, At least one
	Appropriateness Rating	Inappropriate
		Uncertain
		Appropriate

two sets of models of association were estimated: (a) only one covariate was included, and (b) all covariates were included (i.e., multivariable). We included codes from first- and second-generation Escobar appropriateness variables (i.e., Inconclusive, and Rarely Appropriate for both systems with Appropriate used as the referent) along with the preoperative prognostic variables. Multivariable association models included all variables in the univariate association model, regardless of the level of significance, to avoid chance reporting [44]. WOMAC Pain scores were not used in WOMAC Disability association models and vice versa because of high multicollinearity between the two measures [45]. Twenty multiply imputed datasets were used to handle missing covariate data. Models were determined separately for the primary outcomes, WOMAC Pain, and WOMAC Disability.

Additionally, Weighted Kappa and observed % agreement were conducted for head-to-head comparisons. Inferential statistics could not be applied for direct comparison because the data for both systems came from the same sample and were therefore not independent.

Missing post-surgery outcome measurements were handled using full information maximum likelihood method. The individually varying times of observations feature of the model was necessary to avoid bias in growth parameters by incorrectly assuming that all observations were obtained at a fixed timepoint relative to surgery at each measurement occasion [46,47] Two sensitivity analyses were conducted. The first used PCS-12 as the outcome (n = 499) and the second excluded persons with unicompartmental arthroplasty (n = 24). MPlus was used for all analyses [48].

3. Results

The combined sample consisted of 602 participants with KA, of which 354 were recruited to OAI and 248 participated in MOST. The average age was 68.4 (SD = 8.0) years and a total of 68 participants (11.3%) self-reported as being Black or African American. The average preoperative WOMAC Pain score was 7.5 (SD = 3.9). Sample characteristics for the two studies are summarized in Table 1. Weighted Kappa comparing the two systems was fair [49] at 0.23 (se = 0.02) and observed % agreement was 40.2% (See Table 3).

The latent class two-piece growth model with individually-varying times of observations for good and poor classes demonstrated clear separation for WOMAC Pain, and WOMAC Disability (see Fig. 1, panels A and B). Entropy was 0.87 for WOMAC Pain and 0.74 for WOMAC Disability. The most likely class membership a posteriori latent class probabilities for the WOMAC Pain latent class were 0.88 for poor outcome and 0.98 for good outcome. For WOMAC Disability latent class probabilities were 0.87 for poor outcome and 0.94 for good outcome indicating precise assignment of outcome classes. The WOMAC Pain two-piece latent class growth curves (LCA) appear in Fig. 1 WOMAC Pain curves in panel A WOMAC Disability in panel B.

WOMAC Pain two-piece LGCs for first- (panel A) and second-generation (panel B) systems appear in Fig. 2 and WOMAC Disability curves appear in Fig. 3. See supplemental file 2 for 95% CIs of all curve estimates. The most notable difference between them is that the second-generation system classified only 16.11% of participants as Appropriate for KA while 63.96% were classified as Rarely Appropriate. This contrasts with the first-generation system that classified 61.21% as Appropriate for KA and 22.78% as Rarely Appropriate. Additionally, LGCs for both systems were similar in that differences in magnitude of change between the three classifications was largest from the preoperative to the first post-operative visit. After the first postsurgical follow-up visit, changes

Table 3
Comparisons of ratings between 1st and 2nd generation appropriateness systems.

	2nd Generation System			
1st Generation System	Appropriate	Inconclusive	Rarely Appropriate	Totals
Appropriate	91	95	158	344
Inconclusive	6	16	68	90
Rarely Appropriate	0	9	119	128
Totals	97	120	345	562

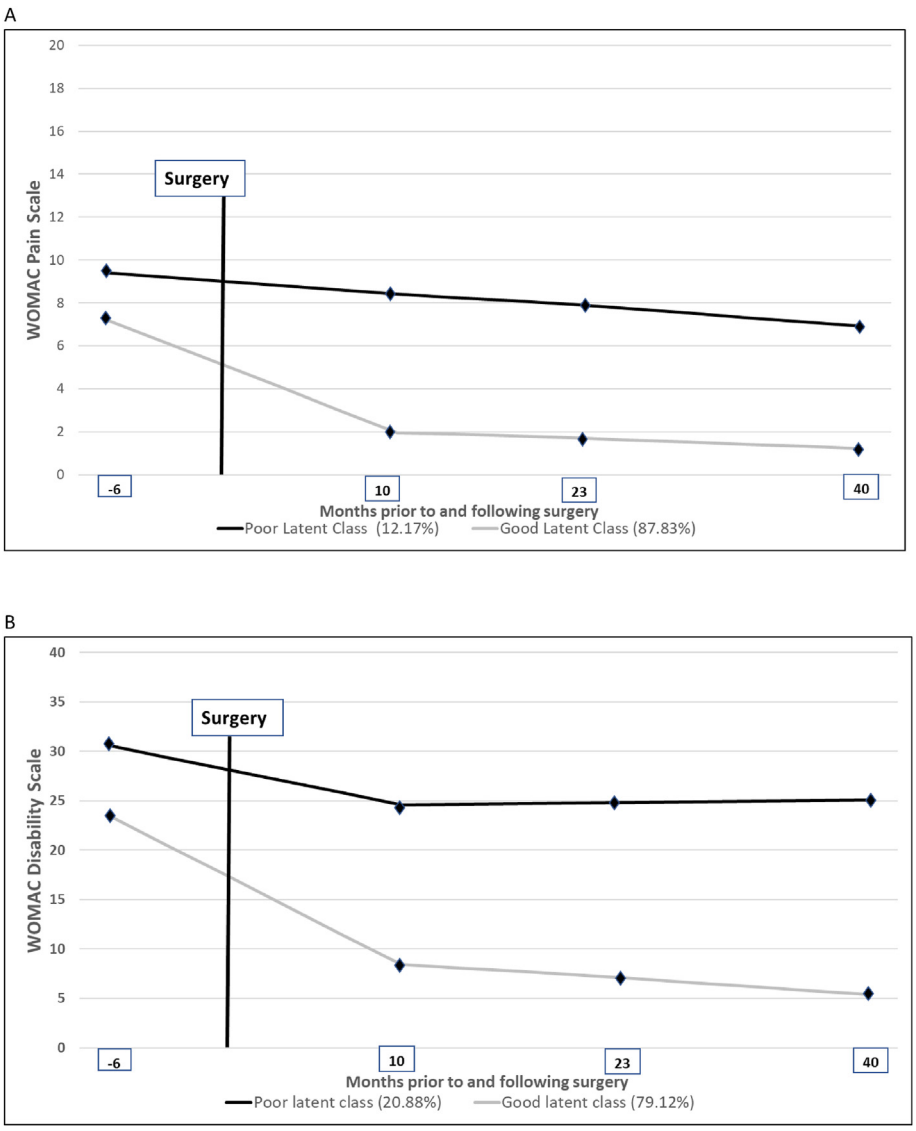

Fig. 1. Trajectories for WOMAC Pain latent classes in panel A and WOMAC Disability latent classes in panel B.

between the three classifications were very similar.

The multivariable association models indicated that first- and second-generation classifications did not associate with good versus poor WOMAC Pain or Disability outcomes (see Table 4) in multivariable models in the hypothesized manner. Only the Rarely Appropriate classification (using Appropriate classification as the referent) associated with WOMAC Disability poor outcome in the univariate analysis for first- and second-generation systems but the odds ratio was less than one. For example, the odds ratio of 0.245, 95%CI = 0.124, 0.487 indicated that participants classified as Rarely Appropriate, relative to Appropriate, were less likely to be in the poor WOMAC Disability outcome second-generation latent class. Sensitivity analyses using the SF-12 PCS as the outcome measure and a second sensitivity analysis that excluded participants with partial KA were consistent with the main analyses (see Supplemental file 3). Table 5 illustrates the proportions of participants with good versus poor WOMAC outcomes stratified by 1st or 2nd generation appropriateness system.

4. Discussion

This study found that the first-generation Escobar KA appropriateness system is the preferred system for clinical use over the second-generation system. The primary reason for this preference is the distribution of classification ratings. Approximately 62% of patients were classified as Appropriate and 23% as Rarely Appropriate using the first-generation system, estimates that align with prior work [5,11]. The second-generation system, in contrast, classified only 16% as Appropriate and 64% as Rarely Appropriate for KA pain outcome, estimates that are unrealistic given the high rate of success of KA [50]. Neither system associated with our gold standard method of classifying outcome as good or poor.

The second generation system requires ratings of more severe pain and functional loss as compared to the first-generation system, to classify patients as Appropriate for KA. A greater emphasis on more severe symptoms led to an unacceptably low rate of classifications of Appropriate and an unacceptably high rate of Rarely Appropriate classifications relative to current evidence. For example, to be rated as Appropriate, the second-generation system required patients to have severe knee pain with activity (i.e., WOMAC Pain scores of 10 or greater) and aged older than 65 years, or both severe pain and severe functional loss (i.e., WOMAC Disability scores of $\geq$37).

The prognostic association models (see Table 4) allowed us to indirectly compare the first- and second-generation systems to determine if either system associated with previously validated good versus poor latent class outcome. Neither system associated with participant membership in either the good or poor outcome class in multivariable

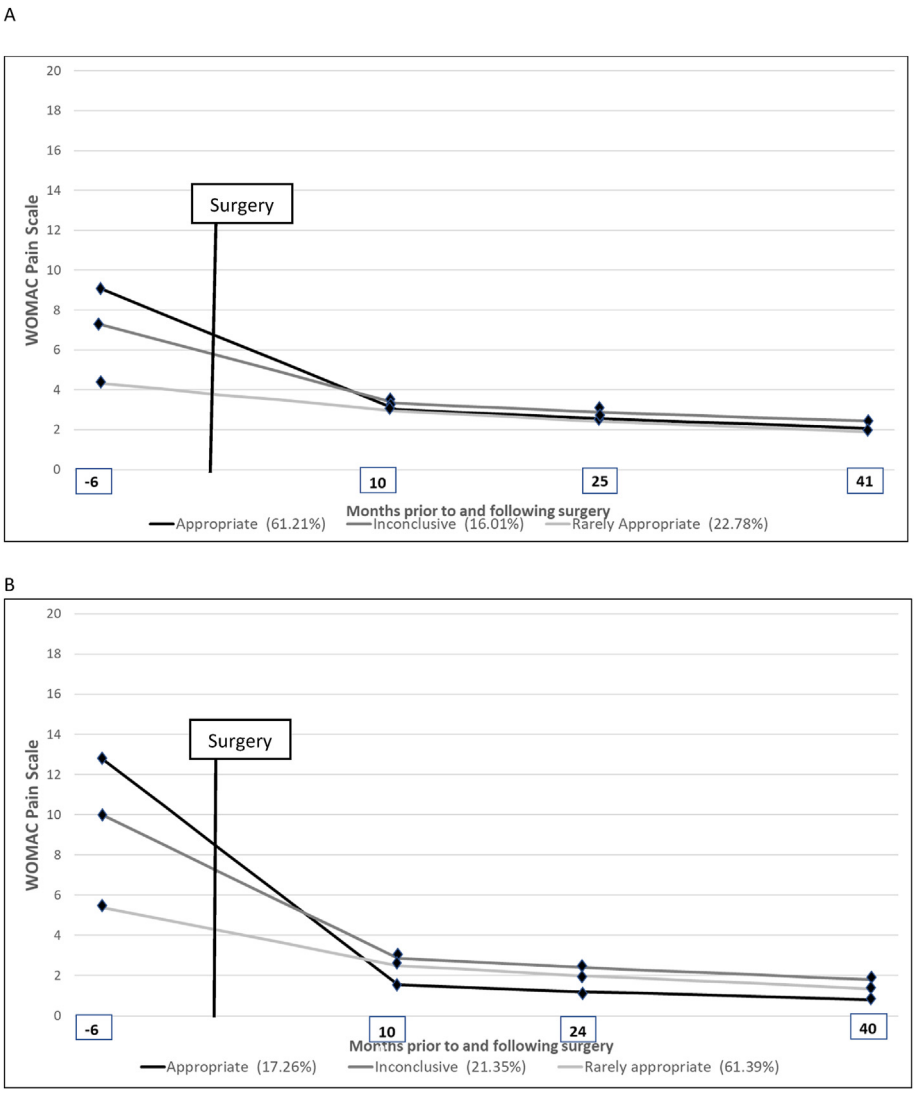

Fig. 2. Trajectories for WOMAC Pain Appropriateness classification with panel A illustrating the original first generation Escobar system and panel B illustrating the second generation system.

analyses. We suspect this was the case because both the first- and second-generation systems differentiate classification subgroups based only on preoperative WOMAC scores, whereas the LCA analyses differentiates good versus poor outcome based on their entire perioperative trajectories. As shown in Fig. 1 panel A, for example, preoperative WOMAC Pain scores were substantively different for the Appropriate subgroup (9 points) as compared to the Rarely Appropriate subgroup (4.3 points). However, when considering the three follow-up visits, mean scores were approximately the same. These data indicate that Escobar appropriateness systems can differentiate among patient appropriateness subgroups only prior to surgery and for changes up to approximately 10 months post-surgery. Because scores at the first postoperative visit and beyond were approximately the same for the three classification subgroups, only the changes from the presurgical visit to the first postsurgical visit were substantially less for the Rarely Appropriate subgroup (average of 1.5 WOMAC Pain points) compared to the Appropriate subgroup (i.e., approximately 6 WOMAC Pain points). We believe these differences are clinically important. Differences between the appropriateness subgroups from the first postoperative visit onward are extremely small or non-existent.

Only one study was found that examined the validity of the second-generation Escobar system [14]. Escobar and colleagues prospectively recruited 282 patients in Spain, scheduled for KA, and determined whether preoperative differences would be found and whether changes from the preoperative visit to 6-month surgery differed among the three classification subgroups. As expected, patients classified as Appropriate had worse preoperative WOMAC scores and larger improvements six-months post-surgery relative to the other two subgroups, much like the current study. Notably, the investigators found that 142 (50.4%) were classified as Appropriate, 90 (31.9%) as Uncertain, and 50 (17.7%) classified as Rarely Appropriate. There were substantial differences between the Spanish sample in the Escobar et al. study [14] and the sample in our study. For example, the mean preoperative WOMAC Pain score in the Escobar et al. study was approximately 11 (SD = 3.8) while in our study it was 7.5 (sd = 3.9). Similar differences were seen for WOMAC Disability scores. It is likely that the differences in classification proportions between our study and the Escobar et al. study are explained mostly by baseline symptom severity.

Importantly, contemporary prognostic indicators of poor outcome risk; namely psychological distress and comorbidity were actually greater (i.e., worse) in participants classified as Appropriate as compared to participants classified as Rarely Appropriate and this was true in our study as well as the Escobar et al. validation study [14]. Contemporary prognostic indicators of poor outcome actually decrease the likelihood of classifications of Rarely Appropriate, the subgroup with minimal early improvement.

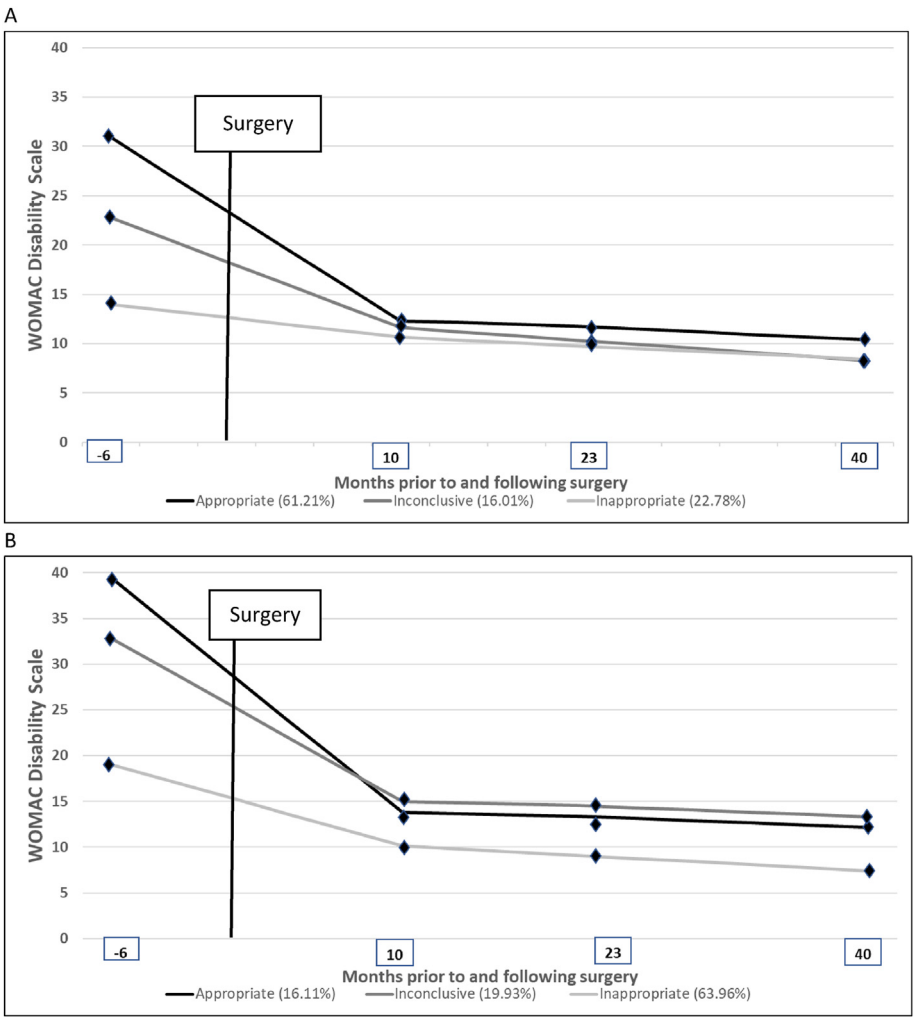

Fig. 3. Trajectories for WOMAC Disability Appropriateness classification with panel A illustrating the original first generation Escobar system and panel B illustrating the second generation system.

Sayah and colleagues conducted a systematic review to determine recovery trajectories following KA [50]. The authors found 21 longitudinal studies that used the WOMAC Pain scale. Mean preoperative WOMAC Pain scores ranged from approximately 8 to 12 on a 0 to 20 scale with higher scores equating to worse pain. Our mean WOMAC Pain score of 7.5 closely approaches this range while the Escobar et al. study [14] is near the upper end of this range. Our participants self-selected to participate in a longitudinal knee OA study and were not recruited from orthopaedic surgeon offices, unlike those in the Escobar et al. study and the systematic review by Sayah [50]. It is possible, therefore, that our preoperative WOMAC Pain scores either reflect normal variation in outcome scores seen in clinical practice, or that our participants were generally healthier with less psychological distressed as compared to typical patients being treated in surgeon offices and this may have led to lower self-reported pain and disability scores in our study. The severity spectrum of pain and functional loss of a sample of patients with KA will influence appropriateness ratings, particularly for the second-generation Escobar system. KA appropriateness classification systems need to account for inherent variability known to occur in preoperative self-reported pain and functional status scores [50]. It appears that the first-generation system is better at accounting for this variation whereas the second-generation system requires more severe pain and functional status scores for classifications of Appropriate and is therefore less able to account for variations in preoperative status common in many large sample studies of KA recovery [50].

Our study has several strengths including the rigorous data collection from two independent NIH funded studies, each with multiple sites. There are also important weaknesses. The datasets were not structured to specifically examine KA outcome, satisfaction was not measured, and WOMAC measures do not account for more challenging activities. Patient input was not included when developing the Escobar criteria. Time between surgery and measurement occasions, for example, varied and for some patients, their preoperative visit occurred more than one year prior to surgery. The MOST study did not report whether total or partial KA was conducted and while we found no influence of partial KA on outcome from the OAI data, the results may still have been biased. Patient satisfaction, an important outcome following KA, was not reported in either OAI or MOST. Both the first- and second-generation Escobar systems were modified to allow use of OAI and MOST data and these modifications also may have biased our results.

Table 4
Indicators of good or poor WOMAC outcome (n = 602).

Outcome	Predictor	Univariate OR	95% CI	p	Multivariable OR	95% CI	p
WOMAC Pain	First-gen RAND Class Inconclusive	1.389	0.638, 3.026	0.408	2.004	0.832, 4.875	0.121
(0 = Good; 1 = Poor)	First-gen RAND Class Rarely Appropriate	0.999	0.466, 2.143	0.998	1.446	0.515, 4.060	0.483
	Sec-gen RAND Class Inconclusive	1.339	0.565, 3.169	0.507	1.610	0.634, 4.090	0.317
	Sec-gen RAND Class Rarely Appropriate	0.600	0.272, 1.320	0.204	0.902	0.325, 2.502	0.843
	Depressive symptoms	1.008	0.969, 1.049	0.677	0.964	0.889, 1.045	0.374
	Mental Health Summary Score	0.983	0.951, 1.016	0.305	0.979	0.932, 1.028	0.399
	Age	0.973	0.939, 1.009	0.136	0.987	0.948, 1.028	0.525
	Sex (0 = male; 1 = female)	1.309	0.700, 2.445	0.399	1.033	0.500, 2.131	0.931
	Race/ethnicity (0 = nonAA; 1 = AA)	2.298	1.079, 4.897	0.031	1.961	0.861, 4.466	0.109
	BMI	1.024	0.974, 1.077	0.358	0.985	0.934, 1.039	0.581
	Comorbidity sore	0.957	0.682, 1.342	0.799	0.874	0.656, 1.164	0.357
	Opioid use	1.330	0.600, 2.947	0.483	0.706	0.244, 2.040	0.520
	Bodily pain count (no knee)	1.127	1.033, 1.230	0.007	1.074	0.970, 1.190	0.170
	Education	0.805	0.669, 0.967	0.021	0.828	0.676, 1.014	0.069
	WOMAC Pain (uninvolved)	1.202	1.113, 1.296	<0.001	1.195	1.090, 1.311	<0.001
WOMAC Disability	First-gen RAND Class Inconclusive	0.453	0.192, 1.071	0.071	0.567	0.238, 1.346	0.198
(0 = Good; 1 = Poor)	First-gen RAND Class Rarely Appropriate	0.386	0.174, 0.855	0.019	0.431	0.111, 1.669	0.223
	Sec-gen RAND Class Inconclusive	1.018	0.450, 2.118	0.961	1.963	0.792, 4.868	0.145
	Sec-gen RAND Class Rarely Appropriate	0.245	0.124, 0.487	<0.001	0.757	0.314, 1.825	0.536
	Depressive symptoms	1.074	1.039, 1.110	0.963	1.011	0.957, 1.068	0.700
	Mental Health Summary Score	0.934	0.906, 0.963	<0.001	0.951	0.907, 0.997	0.037
	Age	0.978	0.950, 1.007	0.139	1.000	0.964, 1.040	0.981
	Sex (0 = male; 1 = female)	1.407	0.812, 2.436	0.223	1.083	0.537, 2.184	0.824
	Race/ethnicity (0 = nonAA; 1 = AA)	2.258	1.105, 4.614	0.025	1.484	0.674, 3.267	0.327
	BMI	1.085	1.031, 1.141	0.002	1.055	0.986, 1.128	0.120
	Comorbidity sore	1.227	0.939, 1.065	0.134	1.262	0.824, 1.932	0.285
	Opioid use	2.186	1.119, 4.273	0.022	1.260	0.545, 2.913	0.590
	Bodily pain count (no knee)	1.155	1.062, 1.256	<0.001	1.070	0.960, 1.192	0.223
	Education	1.279	1.075, 1.521	0.005	0.925	0.741, 1.053	0.488
	WOMAC Pain (uninvolved)	1.223	1.132, 1.321	<0.001	1.164	1.053, 1.286	0.003

Table 5
Proportion of participants from 1st and 2nd generation Escobar systems with good and poor outcome stratified by appropriateness classification.

Escobar System	Poor Outcome (95% CI)	Good Outcome (95% CI)
1st generation System WOMAC Pain		
Appropriate (n = 344)	12.61 (9.65–15.57)	87.39 (84.43–90.35)
Inconclusive (n = 90)	14.97 (8.15–21.79)	85.03 (78.21–91.85)
Rarely Appropriate (n = 128)	10.82 (5.98–15.67)	89.18 (84.33–94.03)
2nd generation System WOMAC Pain		
Appropriate (n = 97)	15.96 (9.73–22.18)	84.04 (77.82–90.27)
Inconclusive (n = 120)	18.66 (12.76–24.57)	81.34 (75.43–87.24)
Rarely Appropriate (n = 385)	9.20 (6.64–11.73)	90.82 (88.27–93.36)
1st generation System WOMAC Disability		
Appropriate (n = 344)	28.09 (24.35–31.83)	74.91 (71.17–78.65)
Inconclusive (n = 90)	17.27 (10.81–23.73)	82.73 (76.27–89.19)
Rarely Appropriate (n = 128)	15.52 (10.38–20.65)	84.48 (79.35–89.62)
2nd generation System WOMAC Disability		
Appropriate (n = 97)	30.40 (22.47–38.84)	69.60 (61.66–77.53)
Inconclusive (n = 120)	32.98 (26.31–39.64)	67.02 (60.36–73.69)
Rarely Appropriate (n = 385)	14.70 (11.91–17.50)	85.30 (82.50–88.08)

5. Conclusion

The first-generation modified Escobar appropriateness classification system is superior to the second-generation system. The second-generation system has multiple limitations. Classification distributions do not align with current evidence and more contemporary prognostic variables for poor outcome associated more strongly with classifications of Appropriate rather than classifications of Rarely Appropriate. Both systems differentiate between groups of patients only at the preoperative visit so extent of expected improvement following KA can only be estimated within the first several months following KA. Classification categories do not differentiate between patient groups at later time periods. The first-generation system holds some promise over the second-generation system for stimulating discussions between patients and surgeons on the potential for substantial benefit or lack of meaningful benefit several months following KA, but reliance on appropriateness criteria to inform long-term outcome is not warranted.

Role of the funding source

Both studies on which this study was based were funded by the National Institutes of Health. The Osteoarthritis Initiative (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. Funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. The Multicenter Osteoarthritis Study (MOST) was supported by NIH grants from the National Institute of Aging to Drs. Lewis (U01-AG-18947), Torner (U01-AG-18832), Nevitt (U01-AG-19069), and Felson (U01-AG-18820).

Author contributions

Both authors contributed equally to the study and made substantial contributions to the conception or design of the study, drafting the paper and reviewing it critically and both approved the final paper and agree to be accountable for all aspects of the work.

We also wish to acknowledge the Orthopaedic Research and Education Foundation (OREF) grant #23-022 for providing funding for the study. We also thank Dr. Gregory Golladay for his assistance in procuring funding for the study. The granting agencies played no role in the conduct or reporting of the study.

Conflict of interest

Both authors report no competing interests or conflicts.

Acknowledgements

We wish to acknowledge the Orthopaedic Research and Education Foundation (OREF) grant #23-022 for providing funding for the study. The OREF played no role in the conduct or reporting of the study. We also thank Dr. Gregory Golladay for his assistance in procuring funding for the study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ocarto.2024.100482.

References

[1] S.M. Kurtz, K.L. Ong, E. Lau, K.J. Bozic, Impact of the economic downturn on total joint replacement demand in the United States: updated projections to 2021, J Bone Joint Surg Am 96 (8) (2014) 624–630.
[2] A.D. Beswick, V. Wylde, R. Gooberman-Hill, A. Blom, P. Dieppe, What proportion of patients report long-term pain after total hip or knee replacement for osteoarthritis? A systematic review of prospective studies in unselected patients, BMJ Open 2 (1) (2012) e000435, https://doi.org/10.1136/bmjopen-2011-000435.
[3] U.D.T. Nguyen, D.C. Ayers, W. Li, L. Harrold, P.D. Franklin, Pre-operative pain and function: profiles of patients selected for total knee replacement among surgeons in the United States, J. Arthroplasty 31 (11) (2014) 2402–2407.e2.
[4] R. Cobos, A. Latorre, F. Aizpuru, et al., Variability of indication criteria in knee and hip replacement: an observational study, BMC Muscoskel. Disord. 11 (1471–2474) (2010) 249.
[5] D.L. Riddle, R.A. Perera, W.A. Jiranek, L. Dumenci, Using surgical appropriateness criteria to examine outcomes of total knee arthroplasty in a United States sample, Arthritis Care Res. 67 (3) (2015) 349–357, https://doi.org/10.1002/acr.22428.
[6] M.M. Ward, A. Dasgupta, Regional variation in rates of total knee arthroplasty among medicare beneficiaries, JAMA Netw. Open 3 (4) (2020) e203717, https://doi.org/10.1001/JAMANETWORKOPEN.2020.3717.
[7] H.M.K. Ghomrawi, A.I. Mushlin, R. Kang, et al., Examining timeliness of total knee replacement among patients with knee osteoarthritis in the U.S, J. Bone Joint Surg. 102 (6) (2020) 468–476, https://doi.org/10.2106/jbjs.19.00432.
[8] B. Ravi, R. Croxford, P.C. Austin, et al., The relation between total joint arthroplasty and risk for serious cardiovascular events in patients with moderate-severe osteoarthritis: propensity score matched landmark analysis, BMJ 347 (2013) 1756–1833 (Electronic)):f6187.
[9] K. Fitch, S.J. Bernstein, M.D. Aguilar, et al., The RAND/UCLA Appropriateness Method User's Manual, 2001. Published online April 10, http://www.rand.org/pubs/monograph_reports/MR1269.html.
[10] R.H. Brook, The RAND/UCLA Appropriateness Method, 1994.
[11] D.L. Riddle, W.A. Jiranek, C.W. Hayes, Use of a validated algorithm to judge the appropriateness of total knee arthroplasty in the United States: a multicenter longitudinal cohort study, Arthritis Rheumatol. 66 (8) (2014) 2134–2143, https://doi.org/10.1002/art.38685.
[12] A. Escobar, J.M. Quintana, I. Arostegui, et al., Development of explicit criteria for total knee replacement, Int. J. Technol. Assess. Health Care 19 (1) (2003) 57–70, https://doi.org/10.1017/s0266462303000060.
[13] A. Escobar-Martinez, R.A. Perera, D.L. Riddle, Development and underlying structure of a second-generation appropriateness classification system for total knee arthroplasty, Arthritis Care Res. (2020), https://doi.org/10.1002/acr.24169 in press:doi: 10.1002/acr.24169.
[14] A. Escobar, A. Bilbao, M. Bertrand, et al., Validation of a second-generation appropriateness classification system for total knee arthroplasty: a prospective cohort study, J. Orthop. Surg. Res. 16 (1) (2021) 227, https://doi.org/10.1186/S13018-021-02371-Z.
[15] D.L. Riddle, R.A. Perera, Appropriateness and total hip arthroplasty: determining the structure of the American academy of orthopaedic surgeons system of classification, J. Rheumatol. 46 (9) (2019), https://doi.org/10.3899/jrheum.180911.
[16] D.L. Riddle, R.A. Perera, Appropriateness and total knee arthroplasty: an examination of the American Academy of Orthopaedic Surgeons appropriateness rating system, Osteoarthritis Cartilage 25 (12) (2017) 1994–1998, https://doi.org/10.1016/j.joca.2017.08.018.
[17] D.L. Riddle, H. Ghomrawi, W.A. Jiranek, L. Dumenci, R.A. Perera, A. Escobar, Appropriateness criteria for total knee arthroplasty: additional comments and considerations, J Bone Joint Surg AM 100 (4) (2018) e22, https://doi.org/10.2106/JBJS.17.00405.
[18] L. Dumenci, R. Perera, F. Keefe, et al., Model-based pain and function outcome trajectory types for patients undergoing knee arthroplasty: a secondary analysis from a randomized clinical trial, Osteoarthritis Cartilage 27 (6) (2019) 878–884, https://doi.org/10.1016/j.joca.2019.01.004.
[19] D.L. Riddle, G.J. Macfarlane, D.F. Hamilton, M. Beasley, L. Dumenci, Cross-validation of good versus poor self-reported outcome trajectory types following knee arthroplasty, Osteoarthritis Cartilage 30 (1) (2022) 61–68, https://doi.org/10.1016/j.joca.2021.09.004.
[20] G. Lester, The osteoarthritis initiative: a NIH public-private partnership, HSS J. 8 (1) (2012) 62–63, https://doi.org/10.1007/s11420-011-9235-y.
[21] N.A. Segal, M.C. Nevitt, K.D. Gross, et al., The Multicenter Osteoarthritis Study: opportunities for rehabilitation research, Pharm. Manag. PM R 5 (8) (2013) 647–654, https://doi.org/10.1016/j.pmrj.2013.04.014.
[22] R.K. Patten, A. Tacey, M. Bourke, R. Lane, M.N. Woessner, I. Levinger, The impact of waiting time for orthopaedic consultation on pain levels in individuals with osteoarthritis: a systematic review and meta-analysis, Osteoarthritis Cartilage 30 (12) (2022) 1561–1574, https://doi.org/10.1016/j.joca.2022.07.007.
[23] J.H. Kellgren, J.S. Lawrence, Radiological assessment of osteo-arthrosis, Ann. Rheum. Dis. 16 (4) (1957) 494–502, https://doi.org/10.1136/ard.16.4.494.
[24] T. Neogi, D. Felson, J. Niu, et al., Association between radiographic features of knee osteoarthritis and pain: results from two cohort studies, BMJ 339 (2009) b2844, https://doi.org/10.1136/bmj.b2844.
[25] I.F. Petersson, T. Boegard, T. Saxne, A.J. Silman, B. Svensson, 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. 56 (1997) 493–496, 0003-4967 (Print).
[26] S. Karunaratne, I. Harris, L. Trevena, M. Horsley, M. Solomon, Observing the use of knee arthroplasty appropriateness tools in clinical practice: do appropriateness criteria tools predict surgeon decision-making? Osteoarthritis Cartilage 29 (9) (2021) 1275–1281, https://doi.org/10.1016/J.JOCA.2021.06.009.
[27] M.M. Vissers, J.B. Bussmann, J.A.N. Verhaar, J.J.V. Busschbach, S.M.A. Bierma-Zeinstra, M. Reijman, Psychological factors affecting the outcome of total hip and knee arthroplasty: a systematic review, Semin. Arthritis Rheum. 41 (4) (2012) 576–588, https://doi.org/10.1016/j.semarthrit.2011.07.003.
[28] L.S. Radloff, The CES-D scale: a self report depression scale for research in the general population, Appl. Psychol. Meas. 1 (1977) 385–401.
[29] J.E. Ware, M. Kosinski, D.M. Turner-Bowker, B. Gandek, How to Score Version 2 of the SF-12 Health Survey, QualityMetric Incororporated, 2002.
[30] O. Sangha, G. Stucki, M.H. Liang, A.H. Fossel, J.N. Katz, The Self-Administered Comorbidity Questionnaire: a new method to assess comorbidity for clinical and health services research, Arthritis Rheum. 49 (2) (2003) 156–163, https://doi.org/10.1002/art.10993.
[31] N.F. Woolacott, M.S. Corbett, S.J.C. Rice, The use and reporting of WOMAC in the assessment of the benefit of physical therapies for the pain of osteoarthritis of the knee: findings from a systematic review of clinical trials, Rheumatology 51 (2012) 1440–1446, https://doi.org/10.1093/rheumatology/kes043.
[32] E.M. Roos, S. Toksvig-Larsen, Knee injury and Osteoarthritis Outcome Score (KOOS) - validation and comparison to the WOMAC in total knee replacement, Health Qual. Life Outcome 1 (2003) 17, https://doi.org/10.1186/1477-7525-1-17.
[33] J.E. Ware, M. Kosinski, D.M. Turner-Bowker, B. Gandek, SF-12v2 how to score version 2 of the SF-12 health Survey, QualityMetric Incorporated (2002).
[34] E.R. Vina, D. Ran, E.L. Ashbeck, C.K. Kwoh, Widespread pain is associated with increased risk of No clinical improvement after TKA in women, Clin. Orthop. Relat. Res. 478 (7) (2020) 1453, https://doi.org/10.1097/CORR.0000000000001001.
[35] S.M. Goodman, M.L. Parks, K. McHugh, et al., Disparities in outcomes for african Americans and whites undergoing total knee arthroplasty: a systematic literature review, J. Rheumatol. 43 (4) (2016) 765–770, https://doi.org/10.3899/jrheum.150950.
[36] L.C. Nguyen, D.C. Sing, K.J. Bozic, Preoperative reduction of opioid use before total joint arthroplasty, J. Arthroplasty 31 (9) (2016) 282–287, https://doi.org/10.1016/j.arth.2016.01.068.

[37] M.M. Dowsey, T. Spelman, P.F.M. Choong, Development of a prognostic nomogram for predicting the probability of nonresponse to total knee arthroplasty 1 Year after surgery, J. Arthroplasty 31 (8) (2016) 1654–1660, https://doi.org/10.1016/j.arth.2016.02.003.
[38] M.M. Dowsey, A.J. Smith, P.F.M. Choong, Latent Class Growth Analysis predicts long term pain and function trajectories in total knee arthroplasty: a study of 689 patients, Osteoarthritis Cartilage 23 (12) (2015) 2141–2149, https://doi.org/10.1016/j.joca.2015.07.005.
[39] N.D. Clement, D.J. Weir, J. Holland, D.J. Deehan, Contralateral knee pain reduces the rate of patient satisfaction but does not clinically impair the change in WOMAC score after total knee arthroplasty, Bone and Joint Journal 102 (1) (2020), https://doi.org/10.1302/0301-620X.102B1.BJJ-2019-0328.R1.
[40] B. Gandek, J.E. Ware, Aaronson, et al., NK Cross-validation of item selection and scoring for the SF-12 health Survey in nine countries:\results from the IQOLA project. International quality of life assessment, J. Clin. Epidemiol. 51 (11) (1998) 1171–1178.
[41] M.E. Charlson, P. Pompei, K.L. Ales, C.R. MacKenzie, A new method of classifying prognostic comorbidity in longitudinal studies: development and validation, JChronicDis 40 (5) (1987) 373–383, https://doi.org/10.1016/0021-9681(87)90171-8.
[42] M.M. Weden, L.S. Zabin, Gender and ethnic differences in the co-occurrence of adolescent risk behaviors, Ethn. Health 10 (3) (2005) 213–224.
[43] B.O. Muthén, L.K. Muthén, Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes, Alcohol Clin. Exp. Res. 24 (6) (2000), https://doi.org/10.1111/j.1530-0277.2000.tb02070.x.
[44] R.C. MacCallum, M. Roznowski, L.B. Necowitz, Model modifications in covariance structure analysis: the problem of capitalization on chance, Psychol. Bull. 111 (3) (1992), https://doi.org/10.1037/0033-2909.111.3.490.
[45] P.W. Stratford, D.M. Kennedy, Does parallel item content on WOMAC's pain and function subscales limit its ability to detect change in functional status? BMC Muscoskel. Disord. 9 (5) (2004) 17.
[46] S.K. Sterba, Fitting nonlinear latent growth curve models with individually varying time points, Struct. Equ. Model. 21 (4) (2014), https://doi.org/10.1080/10705511.2014.919828.
[47] Y. Liu, H. Liu, H. Li, Q. Zhao, The effects of individually varying times of observations on growth parameter estimations in piecewise growth model, J. Appl. Stat. 42 (9) (2015), https://doi.org/10.1080/02664763.2015.1014884.
[48] Muthen LK, Muthen BO. Mplus user's guide. https://www.statmodel.com/html_ug.shtml.
[49] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1) (1977) 159–174.
[50] S.M. Sayah, S. Karunaratne, P.R. Beckenkamp, et al., Clinical course of pain and function following total knee arthroplasty: a systematic review and meta-regression, J. Arthroplasty 36 (12) (2021) 3993–4002, https://doi.org/10.1016/j.arth.2021.06.019.