

Multirater Agreement of Arthroscopic Grading of Knee Articular Cartilage

Robert G. Marx,^{*†} MD, MSc, FRCSC, Jason Connor,[‡] MS, Stephen Lyman,[†] PhD, Annunziato Amendola,[§] MD, Jack T. Andrish,[‡] MD, Christopher Kaeding,^{||} MD, Eric C. McCarty,^{||} MD, Richard D. Parker,[‡] MD, Rick W. Wright,[#] MD, and Kurt P. Spindler,^{||} MD, for the Multicenter Orthopaedic Outcomes Network

From the [†]Hospital for Special Surgery, New York, New York, the [‡]Cleveland Clinic Foundation, Cleveland, Ohio, the [§]University of Iowa Hospitals and Clinics, Iowa City, Iowa, the ^{||}Ohio State Sports Medicine Center, Columbus, Ohio, the ^{||}Vanderbilt Sports Medicine Center, Nashville, Tennessee, and the [#]Washington University Orthopedic & Sports Medicine Center, St. Louis, Missouri

Background: Acute and chronic cartilage injury of the knee has an important impact on prognosis. The validity of the classification of such injuries is critical for prospective multicenter studies. The agreement among multiple surgeons at different institutions for articular cartilage lesions has not been established.

Hypothesis: Arthroscopic classification of articular cartilage lesions is reliable and reproducible and can be used for multicenter studies involving multiple surgeons.

Study Design: Cohort study (diagnosis); Level of evidence, 1.

Methods: A total of 6 surgeons from 5 centers reviewed 31 videos of articular cartilage lesions. With grade 2 and grade 3 combined for the analysis, observed agreement ranged from 81% to 94%, and kappa ranged from 0.34 to 0.87. An additional 22 videos comprising grade 2 and grade 3 lesions were analyzed, and the observed agreement was 80%, with an overall kappa of 0.47.

Conclusion: Arthroscopic grading of articular cartilage lesions is reproducible among surgeons at different centers.

Clinical Relevance: Articular cartilage lesions can be reliably classified among surgeons at different sites. Such reliability is important for multicenter clinical research studies involving arthroscopic knee surgery.

Keywords: multicenter; cartilage; arthroscopy; agreement

Acute and chronic cartilage injury of the knee has an important impact on prognosis.¹⁰ The gold standard for the diagnosis of articular cartilage and meniscus injury of the knee is arthroscopic evaluation.⁴ The presence and severity of such injuries may be the most important factors in the long-term prognosis of acute knee injuries in general and ACL injuries in particular.⁶ Therefore, to determine the

impact of partial and full thickness cartilage injuries on the long-term outcome for these patients, accurate arthroscopic evaluations and classification are critical.⁸

Currently, there is little information regarding the reliability of arthroscopic knee grading of articular cartilage injury using current classification systems.^{1,2,5} If there is poor agreement among surgeons, arthroscopic evaluations should not be compared between surgeons, or a more reliable classification system should be sought. Likewise, if grading is not consistent with one surgeon over time, a more repeatable classification system should be identified.

Because chondromalacia is a key prognostic variable after ACL reconstruction,¹⁰ accurate documentation for multicenter studies is required. Our hypothesis is that fellowship-trained sports medicine orthopaedic surgeons will give similar grading of chondromalacia as seen on arthroscopic video. This study measures the interobserver

*Address correspondence to Robert G. Marx, MD, MSc, FRCSC, Sports Medicine and Shoulder Service and the Foster Center for Clinical Outcome Research, Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021 (e-mail: MarxR@hss.edu).

No potential conflict of interest declared.

variability in the grading of articular cartilage injury using surgical video.

METHODS

A single orthopaedic surgeon (K.P.S.) videotaped the knee arthroscopy of 20 knees that underwent ACL reconstruction. Two surgeons (R.G.M. and K.P.S.) reviewed these arthroscopic videotapes and elected to add 11 additional knee arthroscopy videos of patients with more significant articular cartilage damage because the ACL reconstructions did not include many severe chondral defects. After the data were collected and analyzed, an additional 22 videotapes were made, including only lesions that were believed to be grade 2 or grade 3, to determine whether surgeons could differentiate between the grades. All views of articular cartilage were from the anterolateral portal. The area of interest was probed repeatedly, and different views were obtained. Copies of the videotapes were distributed for classification to 6 other experienced sports medicine surgeons within the Multicenter Orthopaedic Outcomes Network.

All orthopaedic surgeons involved in this study had completed fellowship training in sports medicine and had a minimum of 2 years (mean, 13 years; range, 2-30 years) of experience in practice. The surgeons who reviewed the arthroscopy videos completed a questionnaire incorporating questions from a modified Outerbridge classification system.⁹ This system is the most commonly used classification system, with more than 31 000 cases documented in the literature, according to Curl et al.³ Written criteria regarding the staging of chondral and meniscal injury were sent to each study surgeon, but not discussed with him or her, to ensure that such criteria would be applicable in the real-world setting, that is, generalizable. The surgeons were asked to watch the videotape and score each articular surface and meniscus. Articular surfaces were graded as follows: normal, softening (grade 1); fissures and superficial changes (grade 2); fragmentation and deep changes (grade 3); and exposed bone (grade 4). For the purposes of analysis, grade 2 and grade 3 were combined for the first 31 videos to obtain adequate sample sizes for each grade. For the additional 22 videos in phase 2 of the study, grade 2 and grade 3 lesions were not combined.

Statistical Analysis

Multirater kappa statistics were used to measure agreement among the surgeons with the Outerbridge classification system (grade 2 and grade 3 lesions were combined for the first 31 videos but not the additional 22 videos). Kappa statistics measure agreement beyond the agreement because of chance alone. Two measures contribute to the kappa statistic: expected agreement and observed agreement. Expected agreement is the probability that 2 surgeons provided the same response to a question for any given patient (chance agreement). Observed agreement is the probability that 2 surgeons provided the same response to a question for a specific patient. Kappa, κ , is

TABLE 1
Landis and Koch⁷ Kappa Interpretation Criteria

κ	Interpretation
Below 0.0	Poor agreement
0.00-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Almost perfect agreement

the amount of observed agreement that is beyond the agreement expected because of chance alone.

$$\kappa = \frac{\text{Expected Agreement} - \text{Observed Agreement}}{1 - \text{Expected Agreement}}$$

A κ of 0.00 represents agreement equivalent with random chance alone, whereas a κ of 1.00 represents perfect agreement. A negative κ represents agreement worse than what would be expected because of chance alone. A κ of -1.0 would represent complete discordance between observers. The interpretation criteria for the κ statistic by Landis and Koch are presented in Table 1.⁷

RESULTS

Based on the distribution of the data, grade 2 and grade 3 lesions were combined for the first 31 videos to obtain statistically stable estimates. As well, it was felt to be clinically acceptable to combine these grades because partial thickness lesions are pathologically similar. Kappa measures were in the substantial agreement to almost perfect agreement range for all medial and lateral articular lesions, with the exception of the lateral ($\kappa = 0.51$) and medial ($\kappa = 0.34$) tibial plateau (Table 2). The medial tibial plateau articular lesions had low interobserver reliability because of high expected agreement (Table 2). In summary, based on Landis and Koch criteria⁷ for the patella and femoral condyles, there was almost perfect agreement. For the trochlea, there was substantial agreement; there was moderate agreement for the lateral tibial plateau. However, for the medial tibial plateau, there was only fair agreement.

For the additional 22 videos of lesions that were believed to be either grade 2 or grade 3, the expected agreement was 38.9% and the observed agreement was 67.2%, for an overall κ of 0.47 (Table 3). A total of 132 grades were given by the 6 surgeons (22 videos \times 6 surgeons = 132 grades). Of these, all were grade 2 or grade 3 except for 9 that were listed as grade 1 and 8 that were listed as grade 4. Of the lesions that were listed as grade 1, 5 were on a single video; that is, 5 of the 6 surgeons felt the lesion was grade 1 for one of the videos). Similarly, of the 8 occasions in which a surgeon listed a video as being grade 4, 5 of the lesions were on a single video.

TABLE 2
Interrater Agreement for 31 Lesions With
Grade 2 and Grade 3 Lesions Combined

Location	Expected Agreement	Observed Agreement	κ
Lateral articular lesions			
Femoral condyle	0.55	0.94	0.86
Tibial plateau	0.61	0.81	0.51
Patellar	0.58	0.93	0.80
Trochlear	0.67	0.90	0.71
Medial articular lesions			
Femoral condyle	0.56	0.93	0.84
Tibial plateau	0.79	0.87	0.34
Patellar	0.56	0.94	0.87
Trochlear	0.67	0.92	0.76

TABLE 3
Interrater Agreement for 22 Lesions Without
Grade 2 and Grade 3 Lesions Combined

Cartilage Grade	Expected Agreement, %	Observed Agreement, %	κ
1	0.5	3.4	0.45
2	13.2	22.3	0.41
3	24.9	38.1	0.52
4	0.4	3.4	0.52
Overall	38.9	67.2	0.47

DISCUSSION

This study demonstrated that many aspects of knee cartilage grading are repeatable and reliable between surgeons, even when using video only without the tactile sensation of probing. All of the surgeons were experienced in knee arthroscopy. They had not previously seen the patients who were videotaped, and none of the surgeons were present at the time of the operation. No discussion occurred between the surgeons about the cases or their individual method for grading cartilage.

Three recent studies have evaluated the interobserver and/or intraobserver reliability in arthroscopic knee classifications. A study by Javed et al⁵ compared findings from 2 trainees with findings from a senior orthopaedic surgeon with regard to the arthroscopic evaluation of injured knees. One trainee evaluated 45 knees, and the other evaluated 33 different knees. The same senior surgeon evaluated all 78 knees. The details of the rating system were not discussed, although it appears that an open-ended grading method was used. No formal statistical tests were mentioned, although 95% confidence intervals surrounding the percentage of disagreement between the observers demonstrated that disagreement on a variety of measures ranged from less than 10% to more than 30%. This study concluded that variation in experience was the primary determinant of disagreement.

Another study by Brismar et al¹ studied 19 videotaped knee arthroscopies in 19 patients with mild to moderate osteoarthritis. These knees were classified using the Outerbridge, Collins, and French Society of Arthroscopy measures. Four orthopaedic surgeons evaluated each videotape on 2 occasions at least 2 months apart. Comparisons were made for intraobserver and interobserver variation using κ statistics. The reliability was similar for all 3 rating measures with κ statistics, ranging between 0.42 and 0.66 for intrarater reliability and 0.46 and 0.58 for interrater reliability.

The third study by Cameron et al² studied 6 cadaveric knees that had undergone diagnostic arthroscopy and subsequent confirmatory arthrotomy. The arthroscopy was videotaped and then evaluated by 9 different orthopaedic surgeons using the Outerbridge classification system. Only chondral lesions were graded. The mean interobserver κ was 0.52. Surgeons with more than 5 years in practice had a κ of 0.72 compared to a κ of 0.50 among fellows and surgeons with fewer than 5 years in practice.

Because videotapes of knee arthroscopy were used in this study, the surgeons cannot make their diagnoses based on the tactile sensation that is used at actual arthroscopy. This situation limits their ability to make a diagnosis to what they can see rather than what they can feel. The ability to feel is especially important in diagnosing the cartilage lesions because the smoothness or roughness of the surface will provide tactile feedback to the surgeon for cartilage grading. Given this limitation, there appears to be very good reliability in most of these measures. In an attempt to minimize this limitation, the surgeon who performed the arthroscopic procedures probed each lesion repeatedly and obtained multiple views of the cartilage.

The results of this study do not differ greatly from the findings of the previous interrater reliability studies of knee lesions, which found moderate to good agreement between orthopaedic surgeons for a variety of injury categories. Arthroscopic grading of articular cartilage in the knee has acceptable reproducibility among surgeons at different clinical sites. This variable has sufficient reproducibility for multicenter studies that involve multiple surgeons.

This is the largest number of arthroscopies in a study evaluating the interobserver agreement of chondromalacia grading using κ statistics. The ability to perform multicenter studies involving multiple surgeons is dependent on close agreement in the classification of intra-articular lesions, such as chondromalacia, because these are believed to be associated with long-term outcome. This study demonstrated sufficient reproducibility overall to pool data for such studies.

REFERENCES

1. Brismar BH, Wredmark T, Movin T, Leandersson J, Svensson O. Observer reliability in the arthroscopic classification of osteoarthritis of the knee. *J Bone Joint Surg Br.* 2002;84:42-47.
2. Cameron ML, Briggs KK, Steadman JR. Reproducibility and reliability of the Outerbridge classification for grading chondral lesions of the knee arthroscopically. *Am J Sports Med.* 2003;31:83-86.

3. Curl WW, Krome J, Gordon ES, Rushing J, Smith BP, Poehling GG. Cartilage injuries: a review of 31,516 knee arthroscopies. *Arthroscopy*. 1997;13:456-460.
4. Fife RS, Brandt KD, Braunstein EM, et al. Relationship between arthroscopic evidence of cartilage damage and radiographic evidence of joint space narrowing in early osteoarthritis of the knee. *Arthritis Rheum*. 1991;34:377-382.
5. Javed A, Siddique M, Vaghela M, Hui AC. Interobserver variations in intra-articular evaluation during arthroscopy of the knee. *J Bone Joint Surg Br*. 2002;84:48-49.
6. Keene GC, Bickerstaff D, Rae PJ, Paterson RS. The natural history of meniscal tears in anterior cruciate ligament insufficiency. *Am J Sports Med*. 1993;21:672-679.
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
8. Noyes FR, Stabler CL. A system for grading articular cartilage lesions at arthroscopy. *Am J Sports Med*. 1989;17:505-513.
9. Outerbridge RE. The etiology of chondromalacia patellae. 1961. *Clin Orthop Relat Res*. 2001;389:5-8.
10. Shelbourne KD, Gray T. Results of anterior cruciate ligament reconstruction based on meniscus and articular cartilage status at the time of surgery: five- to fifteen-year evaluations. *Am J Sports Med*. 2000;28:446-452.