Multirater Agreement of Arthroscopic Meniscal Lesions

Warren R. Dunn,*[†] MD, MPH, Brian R. Wolf,[‡] MD, Annunziato Amendola,[‡] MD, Jack T. Andrish,[§] MD, Christopher Kaeding,[||] MD, Robert G. Marx,[†] MD, MSc, FRCSC, Eric C. McCarty,[¶] MD, Richard D. Parker,[§] MD, Rick W. Wright,[#] MD, and Kurt P. Spindler,** MD

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

Background: Establishing the validity of classification schemes is a crucial preparatory step that should precede multicenter studies. There are no studies investigating the reproducibility of arthroscopic classification of meniscal pathology among multiple surgeons at different institutions.

Hypothesis: Arthroscopic classification of meniscal pathology is reliable and reproducible and suitable for multicenter studies that involve multiple surgeons.

Study Design: Multirater agreement study.

Methods: Seven surgeons reviewed a video of 18 meniscal tears and completed a meniscal classification questionnaire. Multirater agreement was calculated based on the proportion of agreement, the kappa coefficient, and the intraclass correlation coefficient.

Results: There was a 46% agreement on the central/peripheral location of tears ($\kappa = 0.30$), an 80% agreement on the depth of tears ($\kappa = 0.46$), a 72% agreement on the presence of a degenerative component ($\kappa = 0.44$), a 71% agreement on whether lateral tears were central to the popliteal hiatus ($\kappa = 0.42$), a 73% agreement on the type of tear ($\kappa = 0.63$), an 87% agreement on the location of the tear ($\kappa = 0.61$), and an 84% agreement on the treatment of tears ($\kappa = 0.66$). There was considerable agreement among surgeons on length, with an intraclass correlation coefficient of 0.78, 95% confidence interval of 0.57 to 0.92, and $P < .001$.

Conclusions: Arthroscopic grading of meniscal pathology is reliable and reproducible.

Clinical Relevance: Surgeons can reliably classify meniscal pathology and agree on treatment, which is important for multicenter trials.

Keywords: multicenter; meniscus; multirater agreement; Multicenter Orthpaedic Outcomes Network (MOON)

Meniscal injuries are treated with partial meniscectomy, several repair techniques, transplantation, or neglect depending on physician evaluation during arthroscopy. To determine the correlation of different types of meniscal tears with long-term patient outcome, it is imperative to have reliable and reproducible arthroscopic evaluations and documentation of meniscal lesions.

The ability for surgeons to agree on meniscal tear grading is crucially important for treatment considerations when they are involved in multicenter studies. If agreement between surgeons is poor or if grading for a particular surgeon is not consistent over time, then a more reliable grading system should be sought.

The type of meniscal injury is believed to be related to patient prognosis after knee injury.[7] Therefore, consistency in evaluation and documentation of meniscal tears is vital to meaningful analysis of long-term outcomes of treatment. The purpose of this study was to document the interobserver reliability of evaluation, grading, and deci-

*Address correspondence to Warren R. Dunn, MD, MPH, Vanderbilt Sports Medicine, 2601 Jess Neely Drive, Nashville, TN 37212 (e-mail: dunnw@hss.edu).

No potential conflict of interest declared.

The American Journal of Sports Medicine, Vol. 32, No. 8
DOI: 10.1177/0363546504264586

sion making for a variety of meniscal tears by 7 senior arthroscopists (all members of the Multicenter Orthopaedic Outcomes Network [MOON]) using digital surgical video.

METHODS

Digital recordings of knee arthroscopies during which a meniscal tear was identified were reviewed after submission by the participating authors. Three authors (WRD, BRW, and RGM) who did not participate in the grading selected 18 cases for the study from a total of 24 that were reviewed. The chosen video clips demonstrated a variety of medial or lateral meniscal tears. The videos were selected based on variety of tear appearance and the adequacy of the digitally recorded evaluation of the tear. There were 12 medial and 6 lateral tears determined to be of adequate quality to be included for review. The video clip of each tear was edited to show a concise arthroscopic evaluation of each tear using an arthroscopic probe. On the basis of prior experience with an interrater study,[11] each video clip was edited to approximately 30 seconds in length to avoid respondent fatigue. A compilation of these 18 selected cases was then produced and saved onto a CD in mpeg format and distributed to the 7 experienced orthopaedic surgeons.

All orthopaedic surgeons involved in this study had completed fellowship training in sports medicine and had a minimum of 3 years of experience in practice. These surgeons reviewed the video clips for the 18 cases and filled out a questionnaire for each meniscal tear. The content of this questionnaire was familiar to the study participants as it is already in use as part of a standardized scheme of prospective data collection used by MOON. The surgeons were asked in this questionnaire (Table 1) if the tear was partial or complete, the location of the tear, the type of tear, the length of the tear, whether the tear was degenerative in nature, whether lateral tears were central to the popliteal hiatus, and the appropriate treatment for each tear. There were 3 possible answers to the location question: *anterior*, *posterior*, or *anterior and posterior*. A question regarding the central versus peripheral location of the tear had 6 potential responses: *central one third*, *middle one third*, *peripheral one third*, *central and middle one third*, *middle and peripheral one third*, or *central and middle and peripheral one third*. There were 6 possible answers to tear type: *radial*, *oblique*, *longitudinal (vertical)*, *bucket handle (displaced)*, *horizontal*, or *complex*. Tear length was measured in millimeters and was limited to 13 of 18 cases that were felt to have adequate enough probing of the tear to allow estimation of length. Depth was recorded as *partial* or *complete*. Degenerative characteristics (including cavitations, multiple cleavage planes, or other degenerative features) were dichotomous responses (yes/no). Whether lateral tears were central to the popliteal hiatus was also dichotomous. Last, there were 5 potential responses to treatment options: *no treatment*, *excision*, *repair*, *abrasion and trephination*, or *meniscal transplantation*.

TABLE 1
Meniscal Pathology Evaluation Form

Case______

Limit your answer to one response per item:

Item 1: Depth
___Partial
___Complete

Item 2: Location
___Anterior
___Posterior
___Anterior and posterior

Item 3: Central vs peripheral
___Central one third
___Middle one third
___Peripheral one third
___Central + middle one third
___Central + middle + peripheral one third
___Middle + peripheral one third

Item 4: Is tear central to the popliteal hiatus? (applies to lateral tears only)
___No
___Yes

Item 5: Type
___Radial
___Oblique (flap tear)
___Longitudinal (vertical)
___Displaced bucket handle
___Horizontal
___Complex (more than one of the above)

Item 6: Length in mm_____

Item 7: Degenerative component (cavitations, multiple cleavage planes, etc)
___No
___Yes

Item 8: Treatment
___No treatment for tear
___Excision
___Repair
___Abrade + trephine
___Meniscus transplant

Statistical Analysis

Interobserver agreement was analyzed by calculating the observed agreement, as well as multirater kappa statistics for categorical data, whereas an intraclass correlation coefficient was calculated for continuous data. A 95% confidence interval was calculated for each kappa. Cohen introduced kappa as a chance-adjusted statistic used to evaluate the agreement present between raters.[6] *Observed agreement* is the probability that 2 surgeons provided the same response to a question for a specific patient.

TABLE 2
Proportion of Agreement (Observed) and Expected Agreement (Chance) for Categorical Items

Item[a]	Observed Agreement	Expected Agreement	Kappa	95% Confidence Interval	P Value
Type	0.73	0.27	0.63	0.55-0.71	<.001
Location	0.87	0.67	0.61	0.29-0.91	<.001
Central/peripheral	0.46	0.23	0.30	0.23-0.37	<.001
Depth	0.80	0.63	0.46	0.21-0.70	<.001
Degenerative	0.72	0.50	0.44	0.33-0.54	<.001
Popliteal hiatus	0.71	0.51	0.42	0.22-0.61	<.001
Treatment	0.84	0.52	0.66	0.47-0.85	<.001

[a] See Table 1 for categorical items. All questions apply to all 18 cases except popliteal hiatus, which applies only to the 6 lateral meniscal cases.

Expected agreement is the probability that 2 surgeons will provide the same response to a question for any given patient (chance agreement). The kappa statistic, κ, is the observed agreement that is above and beyond that due to chance:

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

A kappa value of 1.00 represents perfect agreement, whereas a kappa of 0.00 constitutes agreement equal to that of chance alone. A negative kappa value implies agreement worse than that of chance alone. Landis and Koch have provided criteria by which to evaluate kappa agreement statistics.[10] A kappa value below 0.0 suggests poor agreement, a kappa value of 0.00 to 0.20 constitutes slight agreement, 0.20 to 0.40 is fair agreement, 0.41 to 0.60 is moderate agreement, 0.61 to 0.80 is substantial agreement, and 0.81 to 1.00 is almost perfect agreement. Statistical analyses were performed using SAS for Windows version 9.0 (SAS Institute, Cary, NC).

RESULTS

All 7 surgeons viewed the video on a personal computer and returned the questionnaire. There was only 1 missing datum point. The observed agreements were consistently above those expected by chance alone for all categorical variables. These values, as well as the point estimates and 95% confidence intervals, are listed in Table 2. Using the Landis and Koch criteria,[10] there was fair agreement on the central/peripheral location of tears; moderate agreement on the depth of tears, whether there was a degenerative component, and whether lateral tears were central to the popliteal hiatus; and substantial agreement on the type, location, and treatment of tears. There was considerable agreement among surgeons on length, with an intraclass correlation coefficient of 0.78, 95% confidence interval of 0.57 to 0.92, and $P < .001$.

DISCUSSION

This study demonstrates that meniscal pathology can be reliably and reproducibly graded between surgeons at different institutions even when using only video. Although all of the surgeons were experienced in knee arthroscopy, no discussion occurred between the surgeons about the cases or their individual methods for grading meniscal tears.

The proportion of agreement was above 70%, and the associated kappa coefficients using the Landis and Koch[10] interpretation were moderate to substantial for all but 1 item (central/peripheral location). The low observed agreement (46%) for the central/peripheral location question is likely influenced by the presence of 6 possible responses. The low kappa coefficient is influenced by the prevalence effect, which is the fact that kappa is affected by prevalence just as positive and negative predictive values are influenced by prevalence.[12] In fact, numerous authors have described trouble and paradoxes with the kappa coefficient.[2,3,5,8,13] The prevalence of the observations can alter the kappa in spite of constant values of accuracy for each rater. This paradox of a low kappa despite high observed agreement arises because of the "unfair" correction of chance agreement, which is based on the marginal totals.[8] If the marginal totals are symmetrically unbalanced, the observed agreement can be associated with a low kappa. Several questions in this study were affected by this phenomenon (ie, in which a high observed agreement was associated with a relatively low kappa). Feinstein and Cicchetti suggested that one way of addressing this problem is simply to use the proportion of agreement and not impose the unfair corrections for chance associated with the kappa coefficient.[8] Furthermore, the underlying assumption that is made to correct for chance (expected) agreement is that the raters are independent. Given that raters are clearly not independent, the relevance of this term, as well as its appropriateness as a correction to actual agreement, is questionable. Despite these shortcomings, the kappa coefficient is firmly rooted in the medical literature, and it is for this reason that it is included in the current study; however, our conclusion that grading of meniscal pathology is reproducible and reliable is based on the proportion of agreement (observed agreement) and not the kappa coefficients.

Three recent studies have evaluated the interobserver and/or intraobserver reliability in arthroscopic knee classifications. A study by Javed et al compared findings from 2 trainees with findings from a senior orthopaedic surgeon with regard to arthroscopic evaluation of pathologic knees.[9] This study concluded that variation in experience was the primary determinant of disagreement. Another study by Brismar et al examined 19 videotaped knee arthroscopies in 19 patients with mild to moderate osteoarthritis.[1] These knees were classified using the Outerbridge, Collins, and French Society of Arthroscopy measures.[1] The reliability was similar for all 3 rating measures, with kappa statistics ranging between 0.42 and 0.66 for intrarater reliability and between 0.46 and 0.58 for interrater reliability. The third study by Cameron et al examined 6 cadaveric knees that had undergone diagnostic arthroscopy and subsequent confirmatory arthrotomy.[4] Only chondral lesions were graded using the Outerbridge classification system. The average interobserver kappa was 0.52. Surgeons with more than 5 years in practice had a kappa of 0.72 compared to a kappa of 0.50 among fellows and surgeons with fewer than 5 years in practice.

Only one previous study has evaluated the interobserver reliability of meniscal tear evaluation using arthroscopy.[9] This study evaluated the agreement between 2 senior specialist registrars and a consultant orthopaedic surgeon, all with a special interest in knee surgery. Disagreement between the trainees and the consultant was 17% and 21%, respectively, for meniscal tears evaluated arthroscopically. No kappa coefficients were calculated.

Because videos of knee arthroscopy were used in this study, the surgeons could not classify the pathology in the typical dynamic fashion that includes the tactile feedback of probing during arthroscopy. This limits their ability to classify the pathology to what they can see rather than what they can feel. Despite this limitation, there appears to be very good reliability among raters. In an attempt to minimize this limitation, only videos with adequate visualization of the lesion with extensive probing were included for review. However, the exclusion of cases with poor visualization is also a limitation of the study because it is a best-case scenario that might not be the situation in a clinical trial in which all meniscal tears are included and subsequently classified.

The results of this study do not differ greatly from the findings of the previous interrater reliability studies of knee pathology, which found moderate to good agreement between orthopaedic surgeons for a variety of pathologic categories. However, this study differs in that it is the first multicenter study investigating interrater agreement of meniscal pathology in the literature. Arthroscopic grading of meniscal pathology has acceptable reproducibility among surgeons. Although there was fair agreement on one item, all other items had moderate to substantial agreement. Because measurement error can bias and/or confound the results of a study, particularly a multicenter study, this is a crucial preparatory step to a multicenter trial involving meniscal pathology. In fact, the final common pathway for any trial that involves meniscal pathology involves treatment decisions, and this study demonstrated that there is substantial agreement on how particular tears should be treated. To that end, the classification of meniscal pathology used in the current study has sufficient reproducibility for multicenter studies that involve multiple surgeons.

ACKNOWLEDGMENT

The late Dr Sandy Kirkley was an integral member of the Multicenter Orthopaedic Outcomes Network, and her efforts on this work are acknowledged. Her enthusiasm, dedication, and energy pushed us all to work to answer research questions in orthopaedic sports medicine. She is missed by us, her colleagues, her students, and her patients. Dr Dunn is supported by the American Academy of Orthopaedic Surgeons/Orthopedic Research and Education Foundation Health Services Research Fellowship.

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. Byrt T. Problems with kappa. *J Clin Epidemiol*. 1992;45:1452.
3. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46:423-429.
4. 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.
5. Cicchetti DV, Feinstein AR. High agreement but low kappa, II: resolving the paradoxes. *J Clin Epidemiol*. 1990;43:551-558.
6. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
7. Fairbank TJ. Knee joint changes after mensiscectomy. *J Bone Joint Surg Br*. 1948;30:664-670.
8. Feinstein AR, Cicchetti DV. High agreement but low kappa, I: the problems of two paradoxes. *J Clin Epidemiol*. 1990;43:543-549.
9. 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.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
11. Marx R, Connor J, Amendola A, et al. Inter-observer agreement for the assessment of intra-articular pathology in knee arthroscopy. Paper presented at: AOSSM Annual Meeting, San Diego, Calif; 2003.
12. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol*. 1988;41:949-958.
13. Zwick R. Another look at interrater agreement. *Psychol Bull*. 1988;103:374-378.