2017

Interrater and intrarater reliability of arthroscopic measurements of articular cartilage defects in the knee

David C. Flanigan
Ohio State University

James L. Carey
Ohio State University

Robert H. Brophy
Washington University School of Medicine in St. Louis

William C. Graham
Ohio State University

Alex C. DiBartola
Ohio State University

See next page for additional authors

A commentary by Mark R. Hutchinson, MD, is linked to the online version of this article at jbjs.org.

Interrater and Intrarater Reliability of Arthroscopic Measurements of Articular Cartilage Defects in the Knee

David C. Flanigan, MD, James L. Carey, MD, MPH, Robert H. Brophy, MD, William C. Graham, MD, Alex C. DiBartola, BS, David Hamilton, MD, Haikady N. Nagaraja, PhD, and Christian Lattermann, MD

Investigation performed at the Department of Orthopaedics, The Ohio State University, Columbus, Ohio

Background: Cartilage lesions of the knee are difficult to treat. Lesion size is a critical factor in treatment algorithms, and the accurate, reproducible sizing of lesions is important. In this study, we evaluated the interrater and intrarater reliability of, and correlations in relation to, various arthroscopic sizing techniques.

Methods: Five lesions were created in each of 10 cadaveric knees (International Cartilage Repair Society grade 3C). Three orthopaedic surgeons used 4 techniques (visualization and use of a 3-mm probe, a simple metal ruler, and a sliding metallic ruler tool) to estimate lesion size. Repeated-measures data were analyzed using a mixed-effect linear model. The differences between observed and gold-standard (plastic mold) values were used as the response. Intraclass and interclass correlation coefficient (ICC) values for intrarater and interrater reliability were computed, as were overall correlation coefficients between measurements and gold standards.

Results: The mean lesion size was 2.37 cm^2 (range, 0.36 to 6.02 cm^2). Rater, lesion location and size, and measurement method all affected the cartilage defect measurements. Surgeons underestimated lesion size, and measurements of larger lesions had a higher percentage of error compared with those of smaller lesions. When compared with plastic molds of lesions, 60.5% of surgeon measurements underestimated lesion size. Overall, the correlation between measurements and gold standards was strongest for the simple metal ruler method and weakest for the visualization method.

Conclusions: Several factors may influence arthroscopic estimation of cartilage lesion size: the lesion location, measurement tool, surgeon, and defect size itself. The intrarater and interrater reliability was moderate to good using a 3-mm probe, sliding metallic ruler tool, or simple metal ruler and was fair to moderate using visualization only.

Clinical Relevance: There is a need for more accurate methods of determining the size of articular cartilage lesions.

Peer Review: This article was reviewed by the Editor-in-Chief and one Deputy Editor, and it underwent blinded review by two or more outside experts. It was also reviewed by an expert in methodology and statistics. The Deputy Editor reviewed each revision of the article, and it underwent a final review by the Editor-in-Chief prior to publication. Final corrections and clarifications occurred during one or more exchanges between the author(s) and copyeditors.

Cartilage lesions are common in the general and athletic populations[1-3]. Current algorithms rely on the size and location of the defect to direct treatment recommendations[4-8]. Specifically, cartilage lesion size may be an important factor in determining clinical outcomes[9,10]. Preoperative magnetic resonance imaging (MRI) yields inaccurate estimates of lesion size and may overlook other defects in the knee[11-14]. Cartilage defects often are larger than appreciated on MRI because of hidden delaminations or unstable cartilage flaps surrounding the defect. Therefore, the gold standard for sizing cartilage defects of the knee

Disclosure: No external funding was received for this study. On the **Disclosure of Potential Conflicts of Interest** forms, *which are provided with the online version of the article,* one or more of the authors checked "yes" to indicate that the author had a relevant financial relationship in the biomedical arena outside the submitted work (http://links.lww.com/JBJS/D410).

TABLE I Randomization Process*

Visualization		
Surgeon 1	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Surgeon 2	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Surgeon 3	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Simple metal ruler		
Surgeon 1	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Surgeon 2	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Surgeon 3	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
3-mm probe		
Surgeon 1	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Surgeon 2	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Surgeon 3	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Sliding metallic ruler		
Surgeon 1	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Surgeon 2	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10
Surgeon 3	Randomization	Arthroscopy of knees 1-10
	Re-randomization	Arthroscopy of knees 1-10

*A total of 240 arthroscopies were performed (80 per surgeon).

has been the direct measurement of defect size during arthroscopic knee surgery, commonly based on post-debridement arthroscopic sizing. Novel sizing methods, such as ultrasound examination, have been introduced as clinically feasible mechanisms for the diagnosis of cartilage lesions at the time of arthroscopy, but direct visual measurement still remains the gold standard[15].

TABLE II Summary Statistics for Gold-Standard (GS) Defect Sizes and Related Estimates by Measurement Method*

	GS, N = 10 (cm^2)			Least-Squares Estimate of the GS Mean Value (95% CI) (cm^2)			
Location	Mean	Std. Dev.	Range	3-mm Probe	Sliding Metallic Ruler	Simple Metal Ruler	Visualization
LFC	2.351	1.2693	0.74-4.61	1.578 (1.184-1.972)	1.915 (1.305-2.526)	1.822 (1.334-2.309)	2.417 (1.722-3.112)
MFC	2.897	1.4083	1.13-6.02	1.803 (1.339- 2.267)	2.044 (1.428-2.661)	2.318 (1.651-2.985)	2.462 (1.751-3.228)†
Patella	2.552	1.0262	0.93-4.29	2.089 (1.526-2.652)	2.196 (1.748-2.643)	2.331 (1.734-2.928)	2.230 (1.619-2.839)
Tibia	0.725	0.2337	0.36-1.08	0.594 (0.477-0.710)	0.792 (0.654-0.929)	0.680 (0.565-0.794)	0.931 (0.791-1.070)
Trochlea	3.320	1.2265	1.40-5.19	1.951 (1.614-2.287)	2.257 (1.867-2.647)	2.481 (1.979-2.982)	2.177 (1.771-2.584)

*Samples from the trochlea had the largest defect areas, followed by the medial femoral condyle (MFC), the lateral femoral condyle (LFC), and the patella. The tibial defects were the smallest. †Excluding 1 outlier.

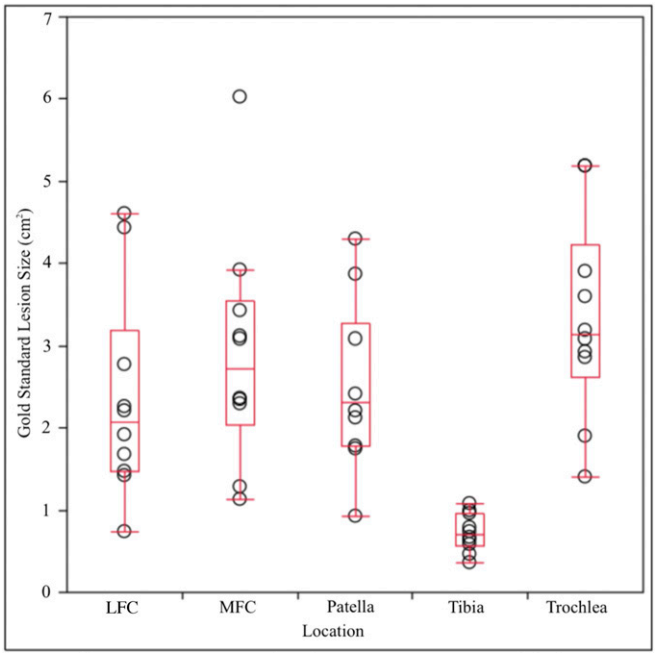

Fig. 1

Box plot of the gold-standard values for defects created in the lateral femoral condyle (LFC) (range, 0.74 to 4.61 [as shown in red: 10th percentile, 0.808; 25th percentile, 1.465; median, 2.065; 75th percentile, 3.185; and 90th percentile, 4.592]), the medial femoral condyle (MFC) (range, 1.13 to 6.02 [10th percentile, 1.145; 25th percentile, 2.04; median, 2.72; 75th percentile, 3.545; and 90th percentile, 5.81]), the patella (range, 0.93 to 4.29 [10th percentile, 1.01; 25th percentile, 1.77; median, 2.31; 75th percentile, 3.275; and 90th percentile, 4.25]), the tibia (range, 0.36 to 1.08 [10th percentile, 0.37; 25th percentile, 0.5575; median, 0.695; 75th percentile, 0.9625; and 90th percentile, 1.072]), and the trochlea (range, 1.40 to 5.19 [10th percentile, 1.45; 25th percentile, 2.6125; median, 3.13; 75th percentile, 4.23; and 90th percentile, 5.189]).

The reported accuracy of arthroscopic sizing of defects varies in the current literature[16]. Arthroscopic assessment was found to overestimate lesion size when compared with arthrotomy for the sizing of defects in 2 European studies[17,18]. These studies comprised a total of 440 lesion measurements at the time of initial arthroscopic assessment using an arthroscopic probe and subsequent open measurement with a ruler at the time of arthrotomy. In contrast, a recent study found that arthroscopy generally underestimated the true size of defects based on a plastic mold of the lesion[19]. Furthermore, in that study, the size of the defect and the measurement device used affected the accuracy of measurement.

Limitations of the previous studies prevented conclusive application to clinical practice. All involved single measurements by surgeons with different ranges of experience, and various unmeasured factors may have influenced the accuracy of sizing and potentially biased the conclusions[17-19]. Additionally, these studies did not evaluate interrater or intrarater reliability. Current clinical trials often utilize a single sizing mechanism for measuring cartilage defects, and a better understanding of which sizing mechanism is most reproducible and accurate is needed[20,21].

We evaluated the interrater and intrarater reliability of the sizing of cartilage defects using common arthroscopic measuring devices. Our hypothesis was that the interrater and intrarater reliability would be high.

Materials and Methods

The experiments were performed on 10 fresh human cadaveric knee specimens from mid-femur to mid-tibia. No previous surgery had been performed on any of the knees. Knees were mounted to an arthroscopic leg holder. Through a 10-cm midline incision, a medial parapatellar incision was made. Through this incision, 5 discrete lesions were created in each knee, 1 in each of the following locations: the medial femoral condyle, the lateral femoral condyle, the medial tibial plateau, the trochlea, and the patella. An independent team made all of the study cartilage lesions. In each location, the lesions were randomly assigned a size and shape (circular, oval, or amorphous), such that lesion shapes and sizes were equally distributed among locations (tibial lesion size was limited due to anatomical limitations). Core punches of known size were used to make the initial defects, followed by the use of curets to complete full-thickness cartilage defects (International Cartilage Repair Society grade 3C). As previously established, a plastic mold (Friendly Plastic; AMACO) was created for each defect[19]. This mold represented our gold standard for accurate sizing of each defect, and no lesion debridement was performed by surgeons prior to lesion measurements. Molds then were painted black using permanent marker and subsequently scanned into a computer. Using commercially available software (Adobe Acrobat 7.0 Professional; Adobe Systems) the area of each gold-standard plastic mold was determined. All study knees were then closed in layers.

Three sports medicine fellowship-trained orthopaedic surgeons participated in this study. Each surgeon used a standard arthroscopic camera and irrigation system (Smith & Nephew). Four unique techniques were used to estimate cartilage defect area. First, each surgeon visually estimated the size of the defects through the arthroscopic camera, without the aid of mechanical instrumentation (referred to as

TABLE III Overall Intrarater and Interrater Reliability*

Measurement Method	Intrarater ICC (95% CI)	Interrater ICC (95% CI)
3-mm probe	0.6371 (0.4062-0.7915)	0.5612 (0.3895-0.6952)
Sliding metallic ruler	0.7001 (0.4970-0.8305)	0.6170 (0.4596-0.7368)
Simple metal ruler	0.6352 (0.4035-0.7903)	0.5882 (0.4232-0.7155)
Visualization	0.5067 (0.2318-0.7067)	0.3096 (0.0964-0.4956)

*The ICC values represent the intrarater and interrater reliability of all measurements of femoral defects across all tools, excluding the gold-standard measurements.

TABLE IV Spearman Rank Correlation Analysis*

Measurement Method	Spearman ρ Coefficient	
	Median (95% CI)	Range†
3-mm probe	0.7412 (0.5587-0.8552)	0.6591-0.7847
Sliding metallic ruler	0.7647 (0.5949-0.8692)	0.6080-0.9207
Simple metal ruler	0.7728 (0.6075-0.8739)	0.6708-0.8719
Visualization	0.6150 (0.3754-0.7776)	0.4034-0.8032

*Overall Spearman rho correlation coefficients comparing all ratings of femoral measurements with gold-standard measurements, excluding tibial data. Pairs generated by each replicate of each rater for each method (n = 40) were compared with gold-standard measurements to generate correlation coefficients. †Range of 6 values for each measurement method; see Table V.

visualization). This step was repeated for each knee in a random fashion (Latin square randomization), before measurements with instruments, such that surgeons were randomly assigned knees for measurement.

Surgeons then measured all defects in all 10 knees (50 defects per surgeon [5 defects per knee × 10 knees per surgeon]) with each measurement instrument (3 techniques) on separate occasions. This sequence was repeated a second time. In addition to visualization, the 3 measurement techniques used by the surgeons were as follows: use of a mechanical probe (Smith & Nephew) with a 3-mm hook perpendicular to the shaft of the tool and with lines along the shaft of the tool marked in 5-mm increments (*3-mm probe*), use of a simple metal ruler (Smith & Nephew) with millimeter demarcations (*simple metal ruler*), and use of a tool with a retractable and flexible end and markings every 2 mm (Arthrex) (*sliding metallic ruler tool*). A random pattern of knees and measurement tools was determined for each surgeon before investigation. Thus, no surgeon knew which technique they were to use or which knee they were to measure until measurement took place. Table I depicts the randomization process.

Each surgeon sized defects in single knees using each of the 3 measurement devices in random order, in an effort to prevent measurement bias between instruments. Surgeons reported defect size to a research team member in 1 of 3 ways: (1) reporting the area directly during visualization (e.g., "3 cm^2"), (2) describing the shape and size of the defect, allowing for subsequent calculation of the area (e.g., "a circle with a 3-cm radius"), or (3) providing 2 measurement values that then were multiplied to determine the area (e.g., "2 cm by 1.5 cm"). Each surgeon, using each of the measuring techniques, measured all 10 knees once. The process was repeated, but with a different randomized knee and instrument order (Latin square randomization). Thus, each surgeon arthroscopically measured the defects in the knees on 4 separate occasions, including visualization measurement. The entire

TABLE V Individual Spearman Rho Values and 95% CIs*

Replicate	Rater	Measurement Method	Spearman ρ	95% CI
1	1	3-mm probe	0.7051	0.5044-0.8336
2	1	3-mm probe	0.6591	0.4374-0.8053
1	2	3-mm probe	0.7786	0.6167-0.8773
2	2	3-mm probe	0.6973	0.4928-0.8288
1	3	3-mm probe	0.7847	0.6263-0.8809
2	3	3-mm probe	0.7773	0.6147-0.8766
1	1	Sliding metallic ruler	0.6080	0.3657-0.7731
2	1	Sliding metallic ruler	0.6676	0.4495-0.8105
1	2	Sliding metallic ruler	0.8281	0.6962-0.9059
2	2	Sliding metallic ruler	0.7013	0.4987-0.8312
1	3	Sliding metallic ruler	0.8517	0.7352-0.9193
2	3	Sliding metallic ruler	0.9207	0.8542-0.9576
1	1	Simple metal ruler	0.6708	0.4542-0.8125
2	1	Simple metal ruler	0.7202	0.5269-0.8426
1	2	Simple metal ruler	0.7897	0.6342-0.8838
2	2	Simple metal ruler	0.8719	0.7693-0.9306
1	3	Simple metal ruler	0.8652	0.7580-0.9269
2	3	Simple metal ruler	0.7558	0.5812-0.8639
1	1	Visualization	0.5949	0.3478-0.7647
2	1	Visualization	0.4789	0.1968-0.6878
1	2	Visualization	0.6352	0.4036-0.7903
2	2	Visualization	0.8032	0.6557-0.8916
1	3	Visualization	0.4034	0.1051-0.6351
2	3	Visualization	0.6507	0.4254-0.8000

*Values presented are a breakdown of the values summarized in Table IV.

TABLE VI Bias Relative to Gold-Standard (GS) Median Values and 95% Prediction Limits*

Location	Measurement Method	GS Median Value (cm^2)	Bias (cm^2)	95% Prediction Limit (cm^2)	
				Lower	Upper
LFC					
	3-mm probe	2.07	−0.64	−1.05	−0.22
	Sliding metallic ruler	2.07	−0.33	−0.75	0.08
	Simple metal ruler	2.07	−0.44	−0.85	−0.02
	Visualization	2.07	0.16	−0.26	0.58
MFC					
	3-mm probe	2.72	−1.01	−1.43	−0.59
	Sliding metallic ruler	2.72	−0.79	−1.21	−0.37
	Simple metal ruler	2.72	−0.52	−0.94	−0.11
	Visualization	2.72	−0.34	−0.75	0.08
Patella					
	3-mm probe	2.31	−0.35	−0.76	0.07
	Sliding metallic ruler	2.31	−0.27	−0.69	0.15
	Simple metal ruler	2.31	−0.14	−0.56	0.27
	Visualization	2.31	−0.24	−0.66	0.17
Tibia					
	3-mm probe	0.70	−0.12	−0.53	0.30
	Sliding metallic ruler	0.70	0.08	−0.34	0.49
	Simple metal ruler	0.70	−0.04	−0.45	0.38
	Visualization	0.70	0.22	−0.20	0.63
Trochlea					
	3-mm probe	3.13	−1.28	−1.70	−0.86
	Sliding metallic ruler	3.13	−1.00	−1.41	−0.58
	Simple metal ruler	3.13	−0.78	−1.19	−0.36
	Visualization	3.13	−1.08	−1.50	−0.66

*Median values represent the median defect size by location. Bias indicates the predicted mean difference between the estimate for the given measurement tool at the given anatomical location and the corresponding GS value. LFC = lateral femoral condyle, and MFC = medial femoral condyle.

process was then repeated a second time, for an assessment of intrarater reliability, resulting in a total of 1,200 measurements.

Statistical Analysis

Our statistical objectives were to (1) assess the reliability and reproducibility of cartilage defect measurements and (2) assess systematic bias inherent in the 4 measurement methods. Repeated-measures data were analyzed using a mixed-effect linear model in which the 3 surgeons and 10 knees were treated as random effects and the gold-standard values for defect sizes were used as covariates. The differences between the observed and gold-standard values were used as the response (see Appendix). Statistical models were fitted for each anatomical location and measurement method separately, because variability for individual raters and between raters (variance component) differed with both of these factors. Prediction bias and 95% prediction limits were computed for the median gold-standard values using these models. Intraclass and interclass correlation coefficient (ICC) values for intrarater and interrater reliability were computed using the estimates of relevant variance components. Intraclass and interclass correlation is an assessment of the reproducibility of quantitative measurements of the same measure, made more than once by the same observer or made by different observers[22]. In addition, overall ICC values were calculated for all ratings using a separate statistical model (see Appendix). Spearman rank correlation analysis was used to evaluate the strength of the association between the gold-standard femoral measurements and the measurements of raters. Using pairs generated by each replicate of each rater for each method, correlations were computed for each measurement method. The 95% confidence interval (CI) for ICC and Spearman rank correlation values were computed using Fisher z-transforms of relevant correlation coefficients.

Statistical analyses were carried out using JMP software (version 10; SAS Institute). A power analysis was not possible because of the nature of the study objectives and study design (interrater and intrarater reliability analysis involving multiple anatomical locations).

Results

Five lesions were successfully created in all 10 knees. Lesion size ranged from 0.36 to 6.02 cm^2 (mean, 2.369 cm^2) (Table II and Fig. 1). Each surgeon performed 80 arthroscopies. A total of 1,200 data points were collected: 300 from visualization and 900 from sizing with the measurement instruments. Rater, location of lesion, and measurement

TABLE VII Standard Deviation Between Replicate Measurements by the Same Rater*

	Std. Dev. (cm^2)			
Location	3-mm Probe	Sliding Metallic Ruler	Simple Metal Ruler	Visualization
LFC	0.4452	0.5578	0.4218	1.1802
MFC	0.4909	0.4076	0.5996	1.1362†
Patella	0.6290	0.6158	0.8410	1.0124
Tibia	0.1971	0.2122	0.2369	0.4494
Trochlea	0.5808	0.5620	0.6936	0.8538

*The replication error variability was substantially higher for the visualization method. LFC = lateral femoral condyle, and MFC= medial femoral condyle. †Excluding 1 outlier.

method all affected the cartilage defect measurements/estimates.

The overall intrarater and interrater reliability data are shown in Table III. Among all femoral lesions, the intrarater reliability (ICC, 0.7001) and the interrater reliability (ICC, 0.6170) were highest for the sliding metallic ruler tool. The intrarater reliability (ICC, 0.5067) and the interrater reliability (ICC, 0.3096) were lowest for the visualization method. According to the Altman guidelines, reliability of 0.0 to 0.20 = poor, >0.20 to 0.40 = fair, >0.40 to 0.60 = moderate, >0.60 to 0.80 = good, and >0.80 to 1.00 = very good[23]. On the basis of these guidelines, good reliability was demonstrated by the sliding metallic ruler tool for both values (intrarater and interrater). The visualization measurement method demonstrated moderate intrarater reliability and fair interrater reliability.

Tables IV and V present Spearman rho correlation coefficients for each measurement method. We assessed the association between the gold-standard femoral measurements and the measurements made by the raters using the different

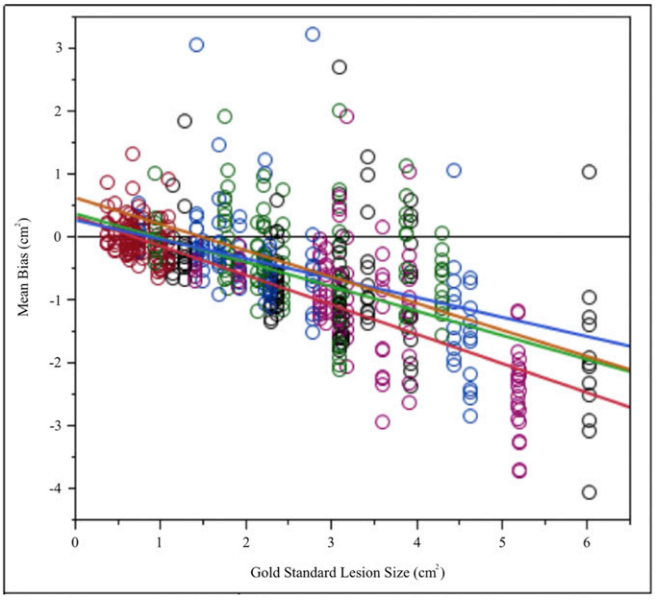

Fig. 2

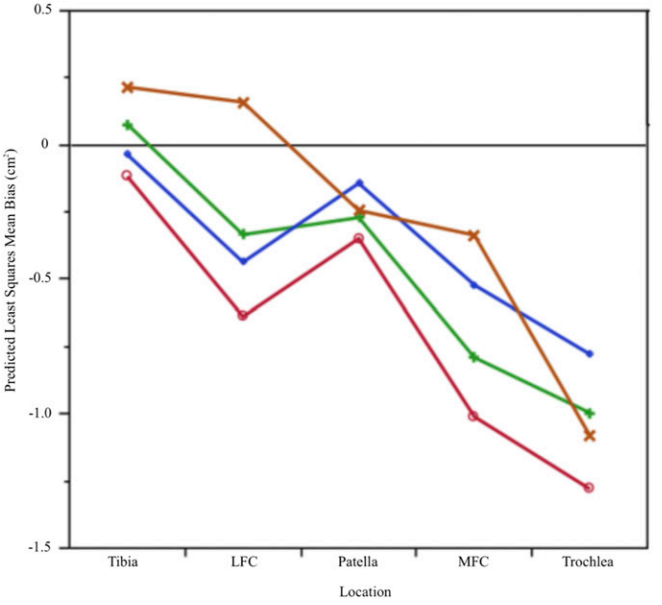

Fig. 3

Fig. 2 Bivariate fit of the difference between gold-standard lesion size and surgeon estimates. The average estimates of defect size are plotted against the 50 gold-standard values. The regression lines drawn make use of these estimates and use slope and intercept for each of the 4 measurement methods to represent summary information about the slopes and intercepts averaged over the 5 anatomical locations. Blue line = simple metal ruler, red line = 3-mm probe, green line = sliding metallic ruler, and orange line = visualization. Red circle = tibia, blue circle = lateral femoral condyle, green circle = patella, black circle = medial femoral condyle, and purple circle = trochlea. **Fig. 3** Predicted bias (in cm^2) of the median values at the 5 anatomical locations for each of the measurement methods in relation to the gold-standard value (the 0.0 cm^2 line). Blue closed dot = simple metal ruler, red open circle = 3-mm probe, green "+" sign = sliding metallic ruler, and orange "x" = visualization. LFC = lateral femoral condyle, and MFC = medial femoral condyle.

TABLE VIII Intraclass/Interclass Correlation of Estimates*

		Same Knee	
Location	Measurement Method	Intraclass Correlation Coefficient (ICC-1)	Interclass Correlation Coefficient (ICC-2)
LFC			
	3-mm probe	0.6691	0.6631
	Sliding metallic ruler	0.7607	0.7157
	Simple metal ruler	0.7736	0.7541
	Visualization	0.6037	0.3096
MFC			
	3-mm probe	0.6911	0.6527
	Sliding metallic ruler	0.8607	0.8098
	Simple metal ruler	0.7914	0.6420
	Visualization	0.5197	0.3563
Patella			
	3-mm probe	0.7113	0.5687
	Sliding metallic ruler	0.5985	0.4986
	Simple metal ruler	0.5497	0.5365
	Visualization	0.5138	0.3938
Tibia			
	3-mm probe	0.6359	0.2973
	Sliding metallic ruler	0.5775	0.4018
	Simple metal ruler	0.5804	0.1981
	Visualization	0.2709	0.0757
Trochlea			
	3-mm probe	0.4778	0.3866
	Sliding metallic ruler	0.6362	0.4113
	Simple metal ruler	0.6089	0.4883
	Visualization	0.4848	0.2361

*For the lateral femoral condyle (LFC) and the medial femoral condyle (MFC), the ICCs were similar for the 3-mm probe, the sliding metallic ruler, and the simple metal ruler. ICC values being close to 1 implies that raters are interchangeable. The best agreement was achieved for the MFC, and the lowest agreement was for the tibia.

measuring methods. The simple metal ruler method demonstrated the strongest correlation with the gold-standard measurements (median Spearman rho, 0.77), whereas the visualization method demonstrated the weakest correlation (median Spearman rho, 0.62). On the basis of the Cohen standards for evaluating correlation coefficients, high correlation (>0.50) existed between all measurement techniques and the gold standard[24].

Surgeons consistently underestimated lesion size, and the difference between gold-standard size and measured size was larger (more inaccurate) for larger gold-standard lesions compared with smaller gold-standard lesions (Tables VI and VII). Figure 2 shows the difference between gold-standard lesion size and all surgeon estimates and demonstrates the general trend of the undersizing of lesions as the gold-standard lesion size increased. In addition, bias (in cm^2) in relation to the gold standard among all surgeons for each anatomical location and measurement method is depicted in Figure 3. When comparing estimates of lesion size with the gold standard, 60.5% of the surgeon measurements underestimated lesion size. The average difference between the gold standard and the surgeon estimate among all locations and all tools was 0.5253 cm^2 (range, 0.0239 to 1.1036 cm^2). When defects were underestimated, defect size influenced the amount of underestimation. Bias from true size tended to increase as lesion size increased.

Table VI presents the median lesion size among all 10 lesions at a particular location among the 10 knees (gold-standard median) and the average bias (among all surgeon estimates) from the gold standard at each location and for each tool. By location, measurements of tibial cartilage lesions

were closest to true size. The tibia also had the smallest created lesions, averaging 0.725 cm^2 (range, 0.36 to 1.08 cm^2). In contrast, measurements of lesions at the trochlear site were, on average, the farthest from true size. The trochlea had the largest created lesions, averaging 3.32 cm^2 (range, 1.40 to 5.19 cm^2). Visualization had the largest variation (standard deviation) between replicate measurements by the same rater (Table VII).

On analysis of measurement methods, use of the 3-mm probe resulted in lesion sizes that were farthest from true size (average of median biases for the 5 locations, -0.68 cm^2) and thus was most inaccurate (Table VI). Visualization had the least average median bias (-0.256 cm^2). Average median bias for the sliding metallic ruler tool was -0.462 cm^2 and for the simple metal ruler was -0.384 cm^2.

The intrarater and interrater reliability was fair to good in most areas of the knee. The ICC values calculated for intraobserver and interobserver reliability showed distinct trends by location and measurement method (Table VIII). The ICC values for both intrarater reliability (ICC-1, the level of agreement of repeated measurements by a single observer) and interrater reliability (ICC-2, the level of agreement between observers) were generally lowest for the measurement of lesions at the tibia. The highest ICC-1 and ICC-2 values (best correlation) occurred for the measurement of lesions at the medial and lateral femoral condyles. ICC-1 and ICC-2 values were higher for the sliding metallic ruler tool and the simple metal ruler for measurements made at the 2 condyles compared with the other locations. Overall, ICC-1 and ICC-2 values were best with the sliding metallic ruler tool. ICC-1 and ICC-2 values were slightly lower for the simple metal ruler and 3-mm probe compared with those for the sliding metallic ruler tool. Overall, visualization had the lowest reliability of all methods.

Discussion

Several factors may influence arthroscopically obtained estimates of defect size: the lesion location, measurement tool, surgeon, and defect size itself. Most importantly, we found that intrarater and interrater reliability was fair to moderate for the evaluation of cartilage lesions using visualization only, but moderate to good when using a simple metal ruler, sliding metallic ruler tool, or 3-mm probe. In addition, the determination of lesion size was dependent on location, was less accurate as lesion size increased, and was dependent on the method of measurement. Finally, visualization of cartilage lesions demonstrated the weakest correlation with the gold-standard measurements, and the simple metal ruler demonstrated the strongest correlation. Thus, because cartilage treatment algorithms universally use lesion size as a branch criterion, the surgeon should measure with a probe or a graduated measurement device (simple metal ruler or sliding metallic ruler) to optimize reliability and correlation to gold-standard measurements.

Our findings have important implications. Algorithms for treatment choice often base decision points on the size and location of cartilage defects[4,5]. Surgeons frequently rely on post-debridement arthroscopic measurement of defects, and a variety of methods are commonly used to estimate size. Despite new imaging modalities, many surgeons believe that use of measuring devices during arthroscopic surgery is necessary for achieving accurate lesion measurement[15,25-27].

Our study is, to our knowledge, the first to evaluate the intrarater and interrater reliability of cartilage defect measurements. The intrarater and the interrater reliability were best for the measurement of femoral condylar lesions. Intrarater reliability was better than interrater reliability in the patellofemoral joint. The use of visualization demonstrated less reliability compared with use of a measuring tool. Use of a graduated measuring tool (simple metal ruler or sliding metallic ruler) had the best interclass coefficients in all areas of the knee except the patella, where a probe performed better. This information may be useful in the design of prospective and multicenter studies. Many studies use a 2-cm^2 cutoff to determine treatment choice[4] or use size as a method of evaluating study results[28-32]. Clinical studies should indicate a consistent method for measuring lesions to ensure minimal variation among measurement devices. Measurement tools have moderate to good reliability, and we found best reliability in the femoral condyles. A more appropriate and reliable instrument to measure patellofemoral defects is needed to improve the reproducibility of sizing.

Surgeons consistently underestimated defect size in our study, and the underestimation and variability increased as the size of the defect increased. The inaccuracy of arthroscopic measurements has been highlighted previously[17-19]. Two clinical studies showed an overestimation of sizing by arthroscopic probe with 4 or 5-mm increments at the time of arthroscopic surgery[17,18]. These studies compared measurements of size with a gold standard using a ruler to measure defect size during the open procedure. In contrast, in a previous study using several devices, surgeons typically overestimated small lesions but underestimated larger lesions (>2 cm^2)[19]. In that study, surgeons performed single measurements with different measuring tools on 1 occasion for the same knee. This measurement then was compared with a plastic mold of the defect that accurately reflected the true size of the defect. Our study supports the finding that arthroscopic measurement underestimates the size of the defect, and if not considered, may bias treatment. Differences among previous clinical and laboratory studies and our laboratory study could arise from regional biases in sizing of defects or differences in study design, including the ability to mold the true size of the defect in laboratory studies.

Interestingly, lesion location influenced the surgeon's ability to accurately size a lesion. Lesions in the trochlea had the highest bias in measurement, as previously reported[19]. Variability in the contour of the trochlea[33,34] makes manipulating instruments arthroscopically to obtain accurate measurement especially difficult and may explain this variation. In contrast, the sizing of patellar defects had the smallest variability. This

difference could be due to the limited amount of curvature in comparison with the trochlea and femoral condyles, allowing for more precise anterior-posterior and medial-lateral measurement. Most importantly, this finding highlights the difficulty of cartilage defect measurements along curved surfaces with large defects.

All measurement tools underestimated the size of the defects, but the 3-mm probe had the highest measurement bias at all lesion locations. In contrast, the simple metal ruler and sliding metallic ruler tool more accurately measured defect size. These 2 devices also had the best intrarater and interrater reliability. Although visualization had lower bias, the interrater and intrarater reliability was fair to moderate overall, suggesting visualization may not be a reliable sizing method. Determining the best way to measure a defect is crucial because clinical studies use lesion size as an outcome determinant[28-32]. However, how accurate lesion measurement needs to be to impact treatment outcomes in patient care remains to be clarified. Nevertheless, on the basis of our findings, prospective studies may benefit from the use of a clearly line-marked measuring tool for measuring defects. Reporting of the measurement technique should be obligatory. In addition, when measurements are in "gray zones" of treatment algorithms, it may be prudent to remeasure cartilage lesions prior to making treatment decisions, given the clear possibility of inaccurate measurement.

We employed a randomization protocol for the evaluation of cartilage defects with various devices. Despite efforts to eliminate potential biases and allow broad application of results, our study had some limitations. First, the range of lesion size was limited in certain areas of the knee, specifically the patella and the tibia. All tibial lesions were created arthroscopically with a curet that limited possible sizes and shapes in the uncovered portion of the tibia (weight-bearing portion of the tibia not covered by the meniscus). Second, lesions of the patella had a limited range of sizes because of the limited surface area. However, this limited range of sizes likely mimics in vivo cartilage lesions on the patellar surface.

Despite these limitations, we demonstrated that arthroscopic evaluation of articular cartilage lesions in the knee consistently undersizes lesions, with variable reliability based on lesion size and location. More accurate methods are needed to determine the size of articular cartilage lesions, whether developing new arthroscopic instruments or augmenting existing ones. Until better options are available, decisions regarding the clinical treatment of lesions, as well as studies of treatments for articular cartilage lesions, should reflect this reality.

Appendix

eA Additional details of the statistical models used are available with the online version of this article as a data supplement at jbjs.org (http://links.lww.com/JBJS/D411). ■

NOTE: The authors acknowledge the contributions of Joshua Mitchell, MD, MPH, in the planning and execution of this study.

David C. Flanigan, MD[1]
James L. Carey, MD, MPH[2]
Robert H. Brophy, MD[3]
William C. Graham, MD[1,4]
Alex C. DiBartola, BS[1]
David Hamilton, MD[5]
Haikady N. Nagaraja, PhD[1]
Christian Lattermann, MD[5]

[1]Division of Sports Medicine Cartilage Repair Center, Department of Orthopaedics (D.C.F., W.C.G., and A.C.D.), Division of Biostatistics, College of Public Health (H.N.N.), and Wexner Medical Center (D.C.F. and A.C.D.), The Ohio State University, Columbus, Ohio

[2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

[3]Department of Orthopaedic Surgery, Washington University, St. Louis, Missouri

[4]OrthoCarolina, Pineville, North Carolina

[5]University of Kentucky, Lexington, Kentucky

E-mail address for D.C. Flanigan: david.flanigan@osumc.edu

References

1. Arøen A, Løken S, Heir S, Alvik E, Ekeland A, Granlund OG, Engebretsen L. Articular cartilage lesions in 993 consecutive knee arthroscopies. Am J Sports Med. 2004 Jan-Feb;32(1):211-5.

2. Flanigan DC, Harris JD, Trinh TQ, Siston RA, Brophy RH. Prevalence of chondral defects in athletes' knees: a systematic review. Med Sci Sports Exerc. 2010 Oct;42(10):1795-801.

3. Widuchowski W, Lukasik P, Kwiatkowski G, Faltus R, Szyluk K, Widuchowski J, Koczy B. Isolated full thickness chondral injuries. Prevalance and outcome of treatment. A retrospective study of 5233 knee arthroscopies. Acta Chir Orthop Traumatol Cech. 2008 Oct;75(5):382-6.

4. Cole BJ, Pascual-Garrido C, Grumet RC. Surgical management of articular cartilage defects in the knee. J Bone Joint Surg Am. 2009 Jul;91(7):1778-90.

5. Behery O, Siston RA, Harris JD, Flanigan DC. Treatment of cartilage defects of the knee: expanding on the existing algorithm. Clin J Sport Med. 2014 Jan;24(1):21-30.

6. Carey JL, Grimm NL. Treatment algorithm for osteochondritis dissecans of the knee. Orthop Clin North Am. 2015 Jan;46(1):141-6.

7. Ozmeriç A, Alemdaroğlu KB, Aydoğan NH. Treatment for cartilage injuries of the knee with a new treatment algorithm. World J Orthop. 2014 Nov 18;5(5):677-84.

8. Behery OA, Harris JD, Karnes JM, Siston RA, Flanigan DC. Factors influencing the outcome of autologous chondrocyte implantation: a systematic review. J Knee Surg. 2013 Jun;26(3):203-11. Epub 2012 Nov 12.

9. Flynn S, Ross KA, Hannon CP, Yasui Y, Newman H, Murawski CD, Deyer TW, Do HT, Kennedy JG. Autologous osteochondral transplantation for osteochondral lesions of the talus. Foot Ankle Int. 2016 Apr;37(4):363-72. Epub 2015 Dec 14.

10. Koh YG, Choi YJ, Kwon OR, Kim YS. Second-look arthroscopic evaluation of cartilage lesions after mesenchymal stem cell implantation in osteoarthritic knees. Am J Sports Med. 2014 Jul;42(7):1628-37. Epub 2014 Apr 17.

11. Campbell AB, Knopp MV, Kolovich GP, Wei W, Jia G, Siston RA, Flanigan DC. Preoperative MRI underestimates articular cartilage defect size compared with findings at arthroscopic knee surgery. Am J Sports Med. 2013 Mar;41(3):590-5. Epub 2013 Jan 16.

12. Campbell AB, Quatman CE, Schmitt LC, Knopp MV, Flanigan DC. Is magnetic resonance imaging assessment of the size of articular cartilage defects accurate? J Knee Surg. 2014 Feb;27(1):67-75. Epub 2013 Jul 24.

13. Gomoll AH, Yoshioka H, Watanabe A, Dunn JC, Minas T. Preoperative measurement of cartilage defects by MRI underestimates lesion size. Cartilage. 2011 Oct;2(4):389-93.

14. Reed ME, Villacis DC, Hatch GF 3rd, Burke WS, Colletti PM, Narvy SJ, Mirzayan R, Vangsness CT Jr. 3.0-Tesla MRI and arthroscopy for assessment of knee articular cartilage lesions. Orthopedics. 2013 Aug;36(8):e1060-4.

15. Penttilä P, Liukkonen J, Joukainen A, Virén T, Jurvelin JS, Töyräs J, Kröger H. Diagnosis of knee osteochondral lesions with ultrasound imaging. Arthrosc Tech. 2015 Sep 14;4(5):e429-33.

16. Spahn G, Klinger HM, Baums M, Pinkepank U, Hofmann GO. Reliability in arthroscopic grading of cartilage lesions: results of a prospective blinded study for evaluation of inter-observer reliability. Arch Orthop Trauma Surg. 2011 Mar;131 (3):377-81. Epub 2011 Jan 20.

17. Årøen A, Røtterud JH, Sivertsen EA. Agreement in arthroscopic and arthrotomy assessment of full-thickness articular cartilage lesions of the knee in a clinical setting in 33 consecutive patients. Cartilage. 2013 Jul;4 (3):214-8.

18. Niemeyer P, Pestka JM, Erggelet C, Steinwachs M, Salzmann GM, Südkamp NP. Comparison of arthroscopic and open assessment of size and grade of cartilage defects of the knee. Arthroscopy. 2011 Jan;27(1):46-51. Epub 2010 Oct 13.

19. Siston RA, Geier D, Bishop JY, Jones GL, Kaeding CC, Granger JF, Skaife T, May M, Flanigan DC. The high variability in sizing knee cartilage defects. J Bone Joint Surg Am. 2013 Jan 02;95(1):70-5.

20. Randsborg PH, Brinchmann J, Løken S, Hanvold HA, Aae TF, Årøen A. Focal cartilage defects in the knee - a randomized controlled trial comparing autologous chondrocyte implantation with arthroscopic debridement. BMC Musculoskelet Disord. 2016 Mar 08;17(1):117.

21. Wong KL, Lee KB, Tai BC, Law P, Lee EH, Hui JH. Injectable cultured bone marrow-derived mesenchymal stem cells in varus knees with cartilage defects undergoing high tibial osteotomy: a prospective, randomized controlled clinical trial with 2 years' follow-up. Arthroscopy. 2013 Dec;29 (12):2020-8.

22. Koch GG. Encyclopedia of statistical sciences. 4th ed. New York: John Wiley & Sons; 1982. p 213-217.

23. Altman D. Practical statistics for medical research. London: Chapman and Hall; 1991.

24. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Erlbaum; 1988.

25. Spahn G, Klinger HM, Hofmann GO. How valid is the arthroscopic diagnosis of cartilage lesions? Results of an opinion survey among highly experienced arthroscopic surgeons. Arch Orthop Trauma Surg. 2009 Aug;129(8):1117-21. Epub 2009 Apr 15.

26. Meftah M, Katchis SD, Scharf SC, Mintz DN, Klein DA, Weiner LS. SPECT/CT in the management of osteochondral lesions of the talus. Foot Ankle Int. 2011 Mar;32 (3):233-8.

27. Kok AC, Terra MP, Muller S, Askeland C, van Dijk CN, Kerkhoffs GM, Tuijthof GJ. Feasibility of ultrasound imaging of osteochondral defects in the ankle: a clinical pilot study. Ultrasound Med Biol. 2014 Oct;40(10):2530-6. Epub 2014 Jul 9.

28. Knutsen G, Engebretsen L, Ludvigsen TC, Drogset JO, Grøntvedt T, Solheim E, Strand T, Roberts S, Isaksen V, Johansen O. Autologous chondrocyte implantation compared with microfracture in the knee. A randomized trial. J Bone Joint Surg Am. 2004 Mar;86-A(3):455-64.

29. Nawaz SZ, Bentley G, Briggs TW, Carrington RW, Skinner JA, Gallagher KR, Dhinsa BS. Autologous chondrocyte implantation in the knee: mid-term to long-term results. J Bone Joint Surg Am. 2014 May 21;96(10): 824-30.

30. Saris D, Price A, Widuchowski W, Bertrand-Marchand M, Caron J, Drogset JO, Emans P, Podskubka A, Tsuchida A, Kili S, Levine D, Brittberg M; SUMMIT study group. Matrix-applied characterized autologous cultured chondrocytes versus microfracture: two-year follow-up of a prospective randomized trial. Am J Sports Med. 2014 Jun;42(6):1384-94. Epub 2014 Apr 8.

31. Zak L, Albrecht C, Wondrasch B, Widhalm H, Vekszler G, Trattnig S, Marlovits S, Aldrian S. Results 2 years after matrix-associated autologous chondrocyte transplantation using the Novocart 3D scaffold: an analysis of clinical and radiological data. Am J Sports Med. 2014 Jul;42(7):1618-27. Epub 2014 May 9.

32. Crawford DC, DeBerardino TM, Williams RJ 3rd. NeoCart, an autologous cartilage tissue implant, compared with microfracture for treatment of distal femoral cartilage lesions: an FDA phase-II prospective, randomized clinical trial after two years. J Bone Joint Surg Am. 2012 Jun 06;94 (11):979-89.

33. Iranpour F, Merican AM, Dandachli W, Amis AA, Cobb JP. The geometry of the trochlear groove. Clin Orthop Relat Res. 2010 Mar;468 (3):782-8.

34. Pinskerova V, Nemec K, Landor I. Gender differences in the morphology of the trochlea and the distal femur. Knee Surg Sports Traumatol Arthrosc. 2014 Oct;22(10):2342-9. Epub 2014 Aug 6.