JSES International 8 (2024) 1051–1054

Contents lists available at ScienceDirect

JSES International

journal homepage: www.jsesinternational.org

Inter-rater and intrarater reliability of superior labrum anterior to posterior lesion classification using magnetic resonance arthrography

Austin W. Bowering, BScN*, Brittany N. Bolt, MD, Conall G. Donaghy, MB, BCh, BAO, Nicholas C. Smith, MD, MSc, FRCSC

Department of Orthopedic Surgery, Memorial University of Newfoundland, St. John's, Newfoundland, Canada

ARTICLE INFO

Keywords:
Orthopedic surgery
Radiology
SLAP tear
Magnetic resonance arthrography
Snyder classification
Maffet subclassification

Level of evidence: Level IV; Diagnostic Study

Background: The glenoid labrum is a fibrocartilaginous ring that affixes the joint capsule and ligaments of the glenohumeral joint. Superior labrum anterior to posterior (SLAP) lesions are a subset of injuries that affect the superior glenoid labrum, most common in laborers and overhead-throwing athletes. In 1990, Snyder et al classified SLAP lesions into one of four types. Later, Maffet et al expanded this scale to include three additional subclassifications. At present, arthroscopy is considered the gold standard for SLAP tear diagnosis. Classification under arthroscopy has demonstrated low to moderate inter-rater reliability. Magnetic resonance arthrography (MRa) is an alternate, less invasive test for diagnosing SLAP lesions. The reliability of MRa for diagnosing slap tears is uncertain.

Methods: Magnetic resonance arthrograms were identified using the Picture Archiving and Communication System (PACS). In total, 273 shoulder arthrograms were reviewed, and 20 were selected with the desired pathology. Three orthopedic surgeons and three musculoskeletal radiologists were asked to classify the SLAP lesions into one of seven categories (Snyder & Maffet classification systems). Data was collected on two separate occasions at an interval of at least two months. Inter-rater and intrarater reliability were calculated using Fleiss Kappa and Cohen's Kappa, respectively.

Results: Between all raters, there was poor inter-rater reliability for each round of data collection ($\kappa = .177$, $\kappa = .124$ for rounds 1 and 2, respectively). Between orthopedic surgeons, there were poor levels of agreement ($\kappa = -.056$, $\kappa = .114$), whereas, between radiologists, there was fair to moderate agreement ($\kappa = 0.479$, $\kappa = 0.340$). Within orthopedic raters, κ values ranged from -0.059 to 0.125, indicating, at best, poor intrarater reliability. Within radiologists, κ values ranged from 0.545 to 0.553, indicating moderate agreement within raters. The analysis determined that none of the orthopedic values for inter or intrarater reliability could be deemed statistically different from zero.

Conclusion: Overall, classification using MRa resulted in significant disagreement between and within raters. Trained radiologists demonstrated higher overall levels of agreement than orthopedic surgeons. In summary, when using MRa to assess SLAP lesions, Snyder and Maffet classification demonstrates poor reliability by orthopedic surgeons and moderate reliability when used by musculoskeletal radiologists.

The stability of the glenohumeral joint depends partly on the glenoid labrum, a fibrocartilaginous ring that deepens the socket and affixes the joint capsule and ligaments of the shoulder.[2] Superior labrum anterior to posterior (SLAP) lesions are a subset of injuries that affect the superior glenoid labrum.[11] This pattern of injury was first described in 1985 by Andrews et al and in 1990, Snyder et al established the first classification system. Snyder suggested that SLAP lesions could be placed in one of four distinct categories: In Type I lesions, there is fraying with a degenerative appearance of the superior labrum; In Type II injuries, the superior labrum and attached biceps tendon tear from the inferior glenoid; In Type III injuries, the central aspect of the tear displaces into the joint; and in Type IV injuries, the tear extends into the bicep tendon.[10] Later, Maffet et al expanded this system to include three additional subclassifications (Types V-II); this system sought to characterize combined-type lesions not adequately categorized by the original Snyder scale.[8]

Overall, SLAP lesions are uncommon, accounting for less than 5% of all shoulder injuries seen in clinics. Nevertheless, they are

Newfoundland and Labrador Health Research Board approved this study (HREB #2021.109).

*Corresponding author: Austin W. Bowering, BScN, Department of Orthopedic Surgery, Memorial University of Newfoundland, 18 Simcoe Dr., Mount Pearl, NL A1N 4W2, Canada.

E-mail address: awb411@mun.ca (A.W. Bowering).

https://doi.org/10.1016/j.jseint.2024.06.009

prevalent in those engaged in repetitive overhead activities. Arthroscopy is considered the gold standard of SLAP tear diagnosis.[5] In 2001, Jee et al were the first to examine inter-rater reliability of magnetic resonance arthrography (MRa) for SLAP lesion diagnosis and classification using arthroscopy. Their research found that inter-rater agreement between radiologists using the Snyder classification system was moderate to substantial (κ = 0.44 - 0.77).[6] In 2008, Gobezie et al examined interobserver and intraobserver variability in diagnosing and treating SLAP tears using the Snyder Classification during arthroscopy. Their study found only moderate agreement between and within expert surgeons.[3] In 2010, Holzafel et al were the first to examine intrarater reliability of MRa for SLAP lesion classification under arthroscopy. Using the Snyder system, their study found that all three radiologists demonstrated excellent intrarater agreement (κ = 0.94, 0.84, 0.93).[4]

Overall, the classification of SLAP lesions remains a diagnostic challenge. Despite the use of the Snyder and Maffet classification systems in clinical practice, only one study has examined the reliability of the Maffet classification system under arthroscopic conditions. In 2018 Wang et al found that Maffet classification under arthroscopic conditions by trained orthopedic surgeons yielded poor inter-rater reliability (k = 0.26). To our knowledge, no research has yet been conducted to examine the reliability of MRa in the classification of SLAP lesions using the Maffet subclassification system.[12] It is, therefore, our hope that this study can add substance to the current body of literature related to the reliability of MRa in SLAP lesion classification.

Materials and methods

MR arthrograms were identified using the Picture Archiving and Communication System (PACS). Arthrograms performed between 2016 and 2020 were identified by selecting the MR filter and searching "EH SHOULDER Arthrograms". A total of 273 shoulder arthrograms were reviewed and 20 were selected for the desired pathology. As MR arthrography has a specificity of greater than 90% for SLAP lesions, arthrograms were selected based on radiology reports. A convenience sample of three orthopedic surgeons and four radiologists were asked to review the 20 shoulder arthrograms. The inclusion criterion for orthopedic surgeons was a current shoulder subspecialty practice, and for radiologists, the completion of a musculoskeletal fellowship program. These practitioners were asked to independently classify the SLAP lesions into one of seven categories using the Snyder and Maffet classification systems. A second round of ratings was requested from the same practitioners at least two months after the initial survey. To ensure anonymity, each patient MRa and physician-rater was assigned a unique numerical code and documentation was stripped of identifying information – this information was then stored under two-way encryption and kept by the research team. Participants were required to sign a consent form and were provided the opportunity to withdraw from the study at any point in time. One participant withdrew from the study prior to data collection, bringing the final number of physician-participants to six, with three of each specialty. Approval for this study was granted by the Newfoundland and Labrador Health Research Board.

Statistical analysis was performed using Statistical Package for Social Sciences (Version: 28.0.1.0; IBM Corp., Armonk, NY, USA). Inter-rater reliability was calculated using Fleiss Kappa, whereas intrarater reliability was calculated using Cohen's Kappa. Statistical interpretation followed the benchmark set by Landis & Koch and modified by Altman, where kappa values less than 0.2 indicate poor agreement, 0.21 to 0.40 indicate fair agreement, 0.41 to 0.60 indicate moderate agreement, 0.61 to 0.80 indicate good agreement, and 0.81 to 1.0 indicate very good agreement.[1,7] In general, Low negative values (0 to −0.10) can be interpreted as "no agreement". However, any large negative kappa values represent great disagreement among raters.[9] Reliability was examined at the level of all-raters, as well as by medical specialty. At the conclusion of data collection, it was found that rater 5 omitted a single point of data on the initial survey. As such, the sample data used to calculate reliability at the levels of all-raters and orthopedic surgeons for round 1 contains 19 effective subjects rather than the expected 20.

Results

Fleiss κ was run to determine whether there was agreement between all practitioners (Table I). Analysis determined there was poor agreement between the six practitioners for both rounds of data collection, κ_1 = .177 (95% confidence interval [CI], .113 to .240), $P < .001$; κ_2 = 0.124 (95% CI, .0.064 to .184), $P < .001$. Subgroup analysis was then done for orthopedic surgeons and radiologists as independent groups (Table II). Between orthopedic surgeons, there was found to be no agreement during round 1 (κ = −.056 (95% CI, −.188 to .075), P = .400) and poor agreement during round 2 (κ = 0.114 (95% CI, −.009 to .238), P = .069). As both P values exceed the significance criteria ($P < .05$) these values cannot be considered statistically different from zero. Between radiologists there was moderate agreement during round 1 (κ = .479 (95% CI, .313 to .615), $P < .001$) and fair agreement during round 2 (κ = .340 (95% CI, .190 to .490), $P < .001$).

Cohen's κ was run for each independent rater to determine the intrarater reliability. Reliability was found to vary significantly between raters (Table III). Radiology-specific values were higher on average (mean κ = 0.549) than their orthopedic counterparts (mean κ = 0.014). None of the orthopedic values could be deemed statistically different from zero.

Discussion

Analysis found that between raters, MRa yielded poor inter-rater reliability for both rounds of data collection (k = 0.177, k = 0.124, for rounds 1 and 2, respectively). As the diagnosis of SLAP lesion using MRa may consider the input of both orthopedic surgeon and radiologist, these values more likely represent clinical reality. However, the analysis of specialty-specific inter-rater reliability produced noticeable differences. Radiologist-specific inter-rater reliability was on average, higher than orthopedic surgeons. We speculate that these differences result from specialty-specific training differences and a standardized process by which radiologists view MR images compared to their surgical colleagues. As none of the orthopedic values could be considered statistically different from zero, limited inference can be made to the practical implications of these values.

Limitations for the study include the retrospective study design. Although the specificity of arthrography is high, the inability to corroborate arthrography with arthroscopy precludes any certainty that the arthrograms were representative of true SLAP pathology. Although a minimum of 2 months was afforded between surveys, practitioners were free to respond at their convenience, meaning the exact time between completion varied between practitioners. One of the raters identified that the image quality of MRa 5 was substandard; however, as this study does not seek to examine the validity of MRa as a tool for SLAP lesion classification, the confounding effects of image quality are likely to be negligible as all practitioners were provided the same collection of MRa images. Regardless, future studies should employ strategies to ensure image quality prior to survey

Table I
Inter-rater reliability (all-raters).

Round 1*						
	Kappa	Asymptotic			Asymptotic 95% confidence interval	
		Standard error	z	Sig.	Lower bound	Upper bound
Overall agreement	.177	.033	5.422	<.001	.113	.240
Agreement on individual categories						
Rating category	Conditional probability	Kappa	Asymptotic		Asymptotic 95% confidence interval	
			Standard error	z	Sig.	Lower bound	Upper bound
1	.000	−.056	.059	−.938	.348	−.172	.061
2	.543	.101	.059	1.713	.087	−.015	.218
3	.145	.054	.059	.915	.360	−.062	.170
4	.178	.107	.059	1.811	.070	−.009	.223
5	.624	.518	.059	8.751	.000	.402	.634
6	.000	−.027	.059	−.456	.648	−.143	.089
7	.000	−.036	.059	−.614	.539	−.152	.080
Round 2†						
	Kappa	Asymptotic			Asymptotic 95% confidence interval	
		Standard error	z	Sig.	Lower bound	Upper bound
Overall agreement	.124	.031	4.053	<.001	.064	.184
Agreement on individual categories						
Rating category	Conditional probability	Kappa	Asymptotic		Asymptotic 95% confidence interval	
			Standard error	z	Sig.	Lower bound	Upper bound
1	.360	.302	.058	5.228	<.001	.189	.415
2	.461	.126	.058	2.178	.029	.013	.239
3	.057	−.001	.058	−.022	.983	−.114	.112
4	.229	.127	.058	2.194	.028	.014	.240
5	.353	.097	.058	1.682	.093	−.016	.210
6	.133	.111	.058	1.925	.054	−.002	.224
7	.133	.088	.058	1.519	.129	−.025	.201

*Sample data contains 19 effective subjects and 6 raters.
†Sample data contains 20 effective subjects and 6 raters.

Table II
Inter-rater (specialty).

Orthopedic surgery – round 1*						
	Kappa	Asymptotic			Asymptotic 95% confidence interval	
		Standard error	z	Sig.	Lower bound	Upper bound
Overall agreement	−.056	.067	−.843	.400	−.188	.075
Orthopedic surgery – round 2†						
	Kappa	Asymptotic			Asymptotic 95% confidence interval	
		Standard error	z	Sig.	Lower bound	Upper bound
Overall agreement	.114	.063	1.821	.069	−.009	.238
Radiology – round 1†						
	Kappa	Asymptotic			Asymptotic 95% confidence interval	
		Standard error	z	Sig.	Lower bound	Upper bound
Overall agreement	.479	.085	5.656	<.001	.313	.645
Radiology – round 2†						
	Kappa	Asymptotic			Asymptotic 95% confidence interval	
		Standard error	z	Sig.	Lower bound	Upper bound
Overall agreement	.340	.076	4.448	<.001	.190	.490

*Sample data contains 19 effective subjects and 3 raters.
†Sample data contains 20 effective subjects and 3 raters.

administration. Importantly, this study does not speak on the validity of MRa as a tool for SLAP lesion diagnosis. That is to say, the intention of this study was not aimed at arriving at an accurate classification, but rather to assess the agreement and reproducibility of classification (accurate or not) between and within individual practitioners. As this study did not collect demographic data, the effects of years in practice and other physician-specific traits cannot be determined. Furthermore, as

Table III
Intrarater reliability.

		Value	Asymptotic standard error*	Approximate T†	Approximate significance
Rater 1					
Measure of agreement	Kappa	.125	.115	1.245	.213
N of valid cases		20			
Rater 2					
Measure of agreement	Kappa	−.023	.221	−.102	.919
N of valid cases		20			
Rater 3					
Measure of agreement	Kappa	.545	.160	2.887	.004
N of valid cases		20			
Rater 4					
Measure of agreement	Kappa	.549	.136	3.856	<.001
N of valid cases		20			
Rater 5					
Measure of agreement	Kappa	−.059	.074	−.626	.531
N of valid cases		19			
Rater 6					
Measure of agreement	Kappa	.553	.129	5.136	<.001
N of valid cases		20			

*Not assuming the null hypothesis.
†Using the asymptotic standard error assuming the null hypothesis.

this study uses the Maffet subclassification scale, the disagreement between raters could be artificially inflated as Snyder Type II and Maffet subclassification injuries are considered wholly different for the purposes of statistical analysis.

Conclusion

This study demonstrates that there is inconsistency between and within practitioners when using MRa as a tool for SLAP lesion classification. At best, trained radiologists have the potential to demonstrate “moderate” levels of inter and intrarater reliability when classifying this type of injury. There has been some suggestion in the literature that accurate classification may assist in the selection of appropriate therapy; however, as the current classification system has demonstrated limited reproducibility using both MRa and arthroscopy, the role of classification in clinical practice is, at best, uncertain. Rather, it could be suggested that an entirely new approach to SLAP lesion classification is indicated to ensure more consistent SLAP lesion diagnosis.

Disclaimers:

Funding: No funding was disclosed by the authors.
Conflicts of interest: The authors, their immediate families, and any research foundation with which they are affiliated have not received any financial payments or other benefits from any commercial entity related to the subject of this article.

References

1. Altman DG. Practical statistics for medical research. 1 ed. Chapman and Hall/CRC; 1990.
2. Clavert P. Glenoid labrum pathology. Orthop Traumatol Surg Res 2015;101:S19-24. https://doi.org/10.1016/j.otsr.2014.06.028.
3. Gobezie R, Zurakowski D, Lavery K, Millett P, Cole B, Wamer J. Analysis of interobserver and intraobserver variability in the diagnosis and treatment of SLAP tears using the snyder classification. Am J Sports Med 2008;36:1373-9. https://doi.org/10.1177/0363546508314795.
4. Holzapfel K, Waldt S, Bruegel M, Paul J, Heinrich P, Imhoff A, et al. Inter- and intraobserver variability of MR arthrography in the detection and classification of superior labral anterior posterior (SLAP) lesions: evaluation in 78 cases with arthroscopic correlation. Eur Radiol 2010;20:666-73. https://doi.org/10.1007/s00330-009-1593-1.
5. Ireland M, Hatzenbuehler J. Superior labrum anterior posterior (SLAP) tears. UpToDate. https://www.uptodate.com/contents/superior-labrum-anterior-posterior-slap-tears/print#:~:text=Arthroscopy%20is%20the%20gold%20standard,cuff%20or%20biceps%20tendon. Accessed August 6, 2023.
6. Jee W, McCauley T, Katz L, Matheny J, Ruwe P, Daigneault J. Superior labral anterior posterior (SLAP) lesions of the glenoid labrum: reliability and accuracy of MR arthrography for diagnosis. Radiology 2001;218:127-32.
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1997;33:159-74.
8. Maffet MW, Gartsman GM, Moseley B. Superior labrum-biceps tendon complex lesions of the shoulder. Am J Sports Med 1995;23:93-8.
9. McHugh M. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276-82. https://doi.org/10.11613/BM.2012.031.
10. Snyder SJ, Karzel RP, Pizzo WD, Ferkel RD, Friedman MJ. SLAP lesions of the shoulder. Arthrosc J Arthrosc Relat Surg 1990;6:274-9.
11. Varacallo M, Tapscott DC, Mair SD. Superior labrum anterior posterior lesions. In: StatPearls. Treasure Island, FL: StatPearls Publishing; 2023.
12. Wang K, Yalizis M, Hoy G, Ek E. Current trends in the evaluation and treatment of SLAP lesions: analysis of a survey of specialist shoulder surgeons. JSES Open Access 2018;2:48-53. https://doi.org/10.1016/j.jses.2017.12.002.