Intra and Interobserver Agreement Regarding the Walch Classification System for Shoulder Joint Arthritis*

Concordância intra e interobservador com relação ao sistema de classificação de Walch para artrose da articulação do ombro

Lauro José Rocchetti Pajolli[1] Marcelo Casciato Carlini[1] Isabella Ferrari[1] Fábio Teruo Matsunaga[1] Nicola Archetti Netto[1] Marcel Jun Sugawara Tamaoki[1]

[1] Orthopedics and Traumatology Department, Escola Paulista de Medicina, Universidade Federal de São Paulo (Unifesp), São Paulo, SP, Brazil

Address for correspondence Lauro José Rocchetti Pajolli, MD, Departamento de Ortopedia e Traumatologia, Escola Paulista de Medicina, Universidade Federal de São Paulo (Unifesp), São Paulo, SP, Brasil (e-mail: laurojrpajolli@gmail.com).

Rev Bras Ortop 2019;54:644–648.

Abstract

Objective To evaluate the inter- and intraobserver agreement regarding the Walch classification system for shoulder arthritis.

Methods Computed tomography scans of the shoulder joint of adult patients were selected between 2012 and 2016, and they were classified by physicians with different levels of expertise in orthopedics. The images were examined at three different times, and the analyses were evaluated by the Fleiss Kappa index to verify the intra- and interobserver agreement.

Results The Kappa index for the intraobserver agreement ranged from 0.305 to 0.545. The inter-observer agreement was very low at the end of the three evaluations ($\kappa = 0.132$).

Conclusion The intraobserver agreement regarding the modified Walch classification varied from moderate to poor. The interobserver agreement was low.

Keywords
- shoulder joint
- osteoarthritis/ classification
- reproducibility of results
- x-ray computed tomography

Resumo

Objetivo Avaliar a concordância inter e intraobservadores com relação ao sistema de classificação de Walch para artrose do ombro.

Materiais e Métodos Foram selecionadas tomografias computadorizadas da articulação do ombro de pacientes adultos entre 2012 e 2016, que foram classificadas por médicos com diferentes níveis de experiência em ortopedia. As imagens foram examinadas em três momentos distintos, e a análise foi avaliada pelo índice Kappa de Fleiss para verificar a concordância intra e interobservador.

* *Work developed at Hospital da Pontifícia Universidade Católica de Campinas, Campinas, SP, Brazil. Originally Published by Elsevier.*

received
October 13, 2017
accepted
April 3, 2018

DOI https://doi.org/10.1016/j.rbo.2018.04.002.
ISSN 0102-3616.

Palavras-chave
- articulação do ombro
- osteoartrite/ classificação
- reprodutibilidade dos testes

Resultados O índice Kappa na concordância intraobservador variou entre 0,305 e 0,545. A concordância interobservador se mostrou muito baixa no fim das três avaliações ($\kappa = 0{,}132$).

Conclusão A concordância intraobservador com relação à classificação de Walch modificada mostrou-se variável, entre moderada e baixa. A concordância interobservador foi baixa.

Introduction

Osteoarthritis (OA) is defined as joint degeneration of primary and secondary origin. Such a limitation causes difficulty to perform daily activities, and can become disabling.

Shoulder arthrosis can affect up to 20% of the elderly population.[1] The primary form is insidious, with no previous shoulder disorders, and it usually affects other joints. In the secondary form, however, there is a previous history.[1]

The initial treatment for OA is based on clinical and drug management. The surgical treatment is frequently indicated to patients with impairments to perform their daily activities who did not respond to the medical treatment.

The number of shoulder arthroplasties and hemiarthroplasties has been growing over the past few decades. Previous studies show a 10.6% and 6.7% increase in the number of shoulder total arthroplasties and hemiarthroplasties respectively, between 1993 and 2007.[2]

Imaging scans aid in the diagnosis and staging of the disease, as well as in the indication of the treatment. Radiographs are routinely used in three views – the anteroposterior, scapular and axillary views.[1] The main objective of computed tomography (CT) scans is to show glenoid anteversion and to provide a detailed view of joint involvement.[3]

The main purpose of the classifications is to enable the communication among professionals studying a certain disease, in order to standardize diagnoses and treatments in clinical research. Thus, a good classification must be reproducible and have the ability to predict the prognosis of a particular condition.[4]

One method to evaluate the reproducibility of a classification system is the analysis of the intra- and interobserver agreement. Intraobserver agreement refers to the concordance in the observations made by the same observer in different observation intervals, whereas interobserver agreement refers to the concordance between different observers.

There are several classifications for shoulder OA. The most used OA classification system was proposed by Walch et al[3] in 1999, which was modified in 2016.[4] This system stages and assesses the progression of shoulder OA based on CT scans of the patients' joints. It considers glenoid morphology, its retroversion angle, and its relationship with the humeral head. These data enable the determination of the best type of arthroplasty to be performed to treat the condition.

However, there is little information on reproducibility and agreement, especially regarding the 2016 modification.

The present study aims to evaluate the intra- and interobserver agreement regarding the modified Walch classification for shoulder OA.

Materials and Methods

The present is a retrospective, cross-sectional, analytical study of the agreement regarding classifications. The research project was approved by the Ethics in Research Committee of Plataforma Brasil (under C.A.A.E nº 66863817.3.0000.5505).

Sample size determination

Initially, 62 was determined as the required number of scans to obtain Kappa values greater than 0.70, with a significance level of 5% and 80% of power.

Sample selection

The selected images were obtained between 2012 and 2016 at the Shoulder and Elbow Surgery sector, and they were from adults older than 18 years of age with shoulder OA. In order to assure the good quality of the images, they were selected by two orthopedists who did not participate in the disease classification process.

Scans from patients with proximal humeral fractures, glenoid fractures, scapular body fractures, and shoulder joint dislocations were excluded, as were all images showing any synthesis material.

Initially, 62 shoulder scans were analyzed. After applying the exclusion criteria, ten exams were excluded from the study. Thus, 52 scans were evaluated for shoulder OA classification.

Image classification process

The scans were classified by five examiners with different levels of experience.

Two expert-level examiners (ELE1 and ELE2, with more than six years of experience as shoulder and elbow orthopedists), one advanced-level examiner (ALE, with one year of experience as a shoulder and elbow orthopedist), one basic level examiner (BLE, an orthopedics resident) and one undergraduate medical student (UMS).

To minimize the bias due to interpretation difficulties and inexperience with the classification system, the observers underwent a previous training regarding the Walch classification. In addition, during the classification process, a brochure with the full Walch classification system was available to the examiner.

The images were organized in a closed digital file. The classifications were made by the observers in three moments,

with a three-week interval between them. In the first evaluation (T1), the images were visualized in numerical order. In the second (T2) and third (T3) evaluations, three and six weeks later respectively, the image sequence was randomized. For each evaluation, the image sequence was randomized by a person unrelated to the analysis and not directly linked to the study; this sequence was revealed only during the final statistical analysis.

Each examiner classified the images independently. There was no time limit for the evaluation.

The examiners were instructed not to discuss the systems until the end of the classification stage. In addition, they had no access to the patients' history or any clinical data about them.

Statistical analysis

The statistical analysis was performed by a specialist in medical statistics. The Fleiss Kappa test was used to assess the intra- and interobserver agreement for each scale. The Fleiss Kappa coefficient is considered the most appropriate to analyze situations in which there are multiple examiners involved or in which many different evaluations are performed, and when the evaluated scale has many categories.[5]

The test was interpreted according to Altman[6] as "proportional agreement with chance correction." Kappa is the agreement coefficient ranging from +1 (perfect agreement) to 0 (agreement equal to chance) to -1 (complete disagreement). There are no definitions as to the accepted levels of agreement, but some studies suggest that results from 0 to 0.2 show very little agreement; from 0.21 to 0.40, small agreement; from 0.41 to 0.60, moderate agreement; and from 0.61 to 0.80, substantial agreement. Values higher than 0.80 are considered virtually perfect agreements.[6,7]

Shoulder arthrosis classification system

According to the Walch classification, shoulder OA is divided into four types and their subdivisions:

(A) arthrosis with a centralized humeral head (with no displacement); (A1) small erosion; (A2) large erosion; (B) arthrosis with posterior subdislocation of the humeral head; (B1) decreased joint space, presence of osteophytes and subchondral sclerosis; (B2) glenoid retroversion and posterior lip involvement (biconcave glenoid); (B3) retroversion $> 14\circ$, with or without subdislocation; (C) glenoid retroversion $> 25\circ$, regardless of erosion; (D) glenoid anteversion and/or anterior humeral head subdislocation.[4]

Results

There was no correct answer, just the observation of intra- and interobserver agreement (the greatest agreement and greatest disagreement).

►**Figure 1** shows the Kappa index for the intraobserver agreement at three distinct assessments using seven levels (A1, A2, B1, B2, B3, C and D). The best result revealed a moderate agreement ($\kappa = 0.545$).

►**Figure 2** shows the Kappa index for the interobserver agreement for separate assessments, as well as the overall agreement at the completion of the three assessments using the same seven levels. The best agreement was obtained at the first evaluation, but it was deemed small ($\kappa = 0.214$). After the three assessments, there was very little interobserver agreement ($\kappa = 0.132$).

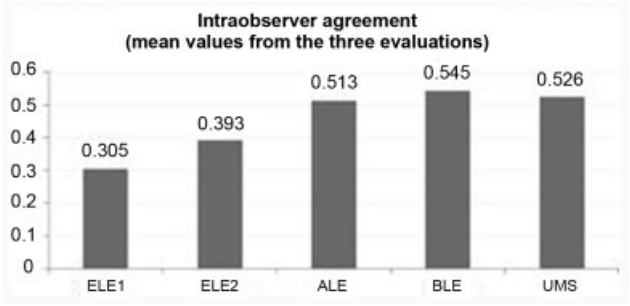

Fig. 1 Mean intraobserver agreement at the end of the three evaluations. Abbreviations: ELE1 and ELE2, expert level examiners; ALE, advanced-level examiner; BLE, basic level examiner; UMS, undergraduate medical student.

The agreement calculations were made using only the four basic levels of the Walch classification (A, B, C, and D). Images rated as A1 and A2 were grouped as A; images classified as B1, B2 and B3 were grouped as B.

►**Figure 3** shows the Kappa index for the intraobserver agreement using only the four basic levels. In this scenario, the best result was substantial, a virtually perfect agreement ($\kappa = 0.798$).

F**igure 4** presents the comparison of the interobserver Kappa indices when the seven levels (A1, A2, B1, B2, B3, C and D) were used, after the grouping regarding the four basic levels. Although the classification system was simplified, the best interobserver agreement remained very small ($\kappa = 0.172$).

Discussion

The Walch classification was chosen because it is widely used by orthopedists to determine shoulder joint involvement in patients with primary arthrosis. Intra- and interobserver agreement is very important to the evaluation of any orthopedic classification system.

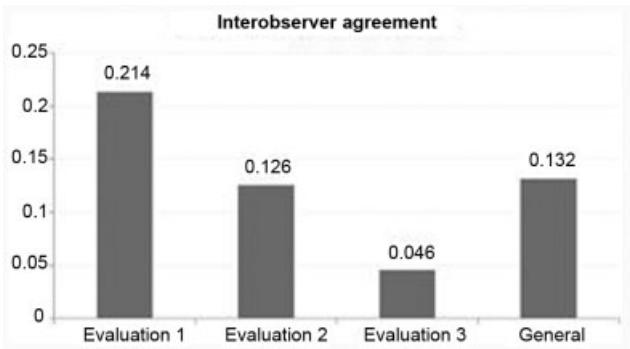

Fig. 2 Interobserver agreement regarding the three evaluations and at general agreement evaluation.

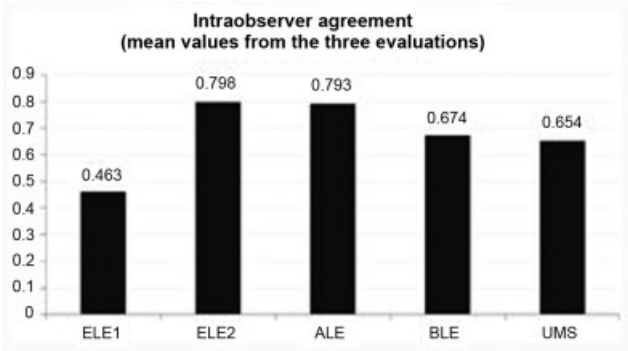

Fig. 3 Mean intraobserver agreement using four levels. Abbreviations: ELE1 and ELE2, expert level examiners; ALE, advanced level examiner; BLE, basic level examiner; UMS, undergraduate medical student.

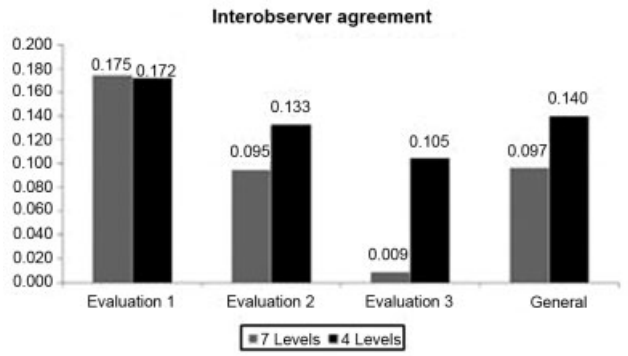

Fig. 4 Comparison of the interobserver agreement at each of the three evaluations and general agreement evaluation using seven and four levels.

The Kappa index regarding the intraobserver agreement ranged from 0.305 (ELE1) to 0.545 (BLE), showing that there was small to moderate agreement for the same evaluator. The wide variation between the results probably results from the complexity of this classification system. Professional experience did not have the expected effect on intra-observer agreement, since the highest index was obtained by the BLE, and the lowest index was obtained by the ELE1.

Interobserver agreement was very low at the completion of the three evaluations ($\kappa = 0.132$). The index decreased between the three evaluation moments. This reduction showed that time and familiarization with the classification system had no relevant effect at the end of the evaluations; in addition, the training performed prior to the first evaluation may have influenced the results.

Our work showed lower Kappa indices compared to studies assessing the agreement regarding different classification systems, as well as lower intra- and interobserver agreement concerning the Walch classification when compared to other studies. Matsunaga et al,[8] analyzing the Mason classification for proximal radial fractures, demonstrated satisfactory intra- ($\kappa = 0.582$) and interobserver ($\kappa = 0.429$–0.560) agreement.

The use of the four basic levels of assessment resulted in a better intraobserver agreement, with substantial values obtained for most evaluators. This finding highlights the difficulty in evaluating the Walch classification subdivisions, and it shows that a simplification of the classification leads to a better agreement.

Belotti et al[9] demonstrated that intra- and interobserver agreement for distal radial classifications was higher if there were fewer variables. This fact is in line with the present study, in which there was an increase in agreement when fewer variables were used.

Our results reveal an important difference compared to those reported by Bercik et al,[4] who demonstrated very good interobserver and virtually perfect intraobserver agreement. This difference may be explained by the use of specialized software to determine the version angle of the glenoid and three-dimensional (3D) reconstructions of CT scans in the abovementioned studies, which were not employed by us.

The use of 3D reconstruction images seems to improve the understanding of glenoid morphology. Osteoarthritis can cause bone degeneration in the sagittal, coronal and axial planes, thus presenting itself as a 3D defect that is difficult to see in two-dimensional images.

Scalise et al[10] and Budge et al[11] used CT with 3D reconstruction. Both showed that there was a better morphological understanding of the glenoid and, thus, a better agreement between the evaluators when 3D images were analyzed.

It is worth noting that the present study was limited to evaluating the opinions of the examiners; it did not have the goal of establishing a correct answer for each scan evaluated. Therefore, the accuracy of each observer was not assessed. This would require analyzing each observer's responses and comparing them with a golden standard method (with high specificity and sensitivity) for diagnosis.

Conclusion

The intraobserver agreement of the modified Walch classification varied from moderate to poor. The interobserver agreement, however, was low.

Conflicts of Interest
The authors have none to declare.

References

1 Cofield RH, Briggs BT. Glenohumeral arthrodesis. Operative and long-term functional results. J Bone Joint Surg Am 1979;61(05): 668–677
2 Day JS, Lau E, Ong KL, Williams GR, Ramsey ML, Kurtz SM. Prevalence and projections of total shoulder and elbow arthroplasty in the United States to 2015. J Shoulder Elbow Surg 2010;19 (08):1115–1120
3 Walch G, Badet R, Boulahia A, Khoury A. Morphologic study of the glenoid in primary glenohumeral osteoarthritis. J Arthroplasty 1999;14(06):756–760
4 Bercik MJ, Kruse K II, Yalizis M, Gauci MO, Chaoui J, Walch G. A modification to the Walch classification of the glenoid in primary glenohumeral osteoarthritis using three-dimensional imaging. J Shoulder Elbow Surg 2016;25(10):1601–1606
5 Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(05):360–363
6 Altman DG. Practical statistics for medical research. 3rd ed. London: Chapman and Hall; 1995

7 Svanholm H, Starklint H, Gundersen HJ, Fabricius J, Barlebo H, Olsen S. Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. APMIS 1989;97(08):689–698
8 Matsunaga FT, Tamaoki MJ, Cordeiro EF, et al. Are classifications of proximal radius fractures reproducible? BMC Musculoskelet Disord 2009;10:120
9 Belloti JC, Tamaoki MJ, Franciozi CE, et al. Are distal radius fracture classifications reproducible? Intra and interobserver agreement. Sao Paulo Med J 2008;126(03):180–185
10 Scalise JJ, Codsi MJ, Bryan J, Brems JJ, Iannotti JP. The influence of three-dimensional computed tomography images of the shoulder in preoperative planning for total shoulder arthroplasty. J Bone Joint Surg Am 2008;90(11):2438–2445
11 Budge MD, Lewis GS, Schaefer E, Coquia S, Flemming DJ, Armstrong AD. Comparison of standard two-dimensional and three-dimensional corrected glenoid version measurements. J Shoulder Elbow Surg 2011;20(04):577–583