J Shoulder Elbow Surg (2019) 28, 625–630

ELSEVIER

Journal of Shoulder and Elbow Surgery

www.elsevier.com/locate/ymse

Intraobserver and interobserver reliability of the modified Walch classification using radiographs and computed tomography

Dave R. Shukla, MB, BCh, BAO, Richard J. McLaughlin, MD, Julia Lee, MD, Robert H. Cofield, MD, John W. Sperling, MD, MBA, Joaquin Sánchez-Sotelo, MD, PhD*

Department of Orthopedic Surgery, Mayo Clinic, Rochester, MN, USA

Background: The Walch classification was introduced to classify glenoid morphology in primary glenohumeral osteoarthritis. A modified Walch classification was recently proposed, with 2 additional categories, B3 (monoconcave glenoid with posterior bone loss leading to retroversion > 15° or subluxation > 70%) and D (excessive anterior subluxation), as well as a more precise definition of subtypes A2 and C. The purpose of this study was to evaluate the intraobserver and interobserver agreement of the modified Walch classification system using both plain radiographs and computed tomography (CT).

Methods: Three fellowship-trained shoulder surgeons blindly and independently evaluated radiographs and CT scans of 100 consecutive shoulders (98 patients) with primary glenohumeral osteoarthritis and classified all shoulders according to the modified Walch classification in 4 separate sessions, each 4 weeks apart. Statistical analysis with the κ coefficient was used to evaluate reliability.

Results: The first reading by the most senior observer on the basis of CT scans was used as the gold standard (distribution: A1, 18; A2, 12; B1, 20; B2, 25; B3, 22; C, 1; and D, 2). The average intraobserver agreement for radiographs and CT scans was 0.73 (substantial; 0.72, 0.74, and 0.72) and 0.73 (substantial; 0.77, 0.69, and 0.72), respectively. The average interobserver agreement was 0.55 (moderate; 0.61, 0.51, and 0.53) for radiographs and 0.52 (moderate; 0.63, 0.50, and 0.43) for CT scans.

Conclusion: Intraobserver agreement of the modified Walch classification was substantial both for axillary radiographs and for CT scans. Interobserver agreement was fair. Although the modified Walch classification represents an improvement over the original classification, automated computer-based analysis of CT scans may be needed to further improve the value of this classification.

Level of evidence: Basic Science Study; Validation of Classification System

Keywords: Walch classification; glenoid morphology; glenohumeral osteoarthritis; computed tomography (CT); radiography; shoulder arthroplasty

The Mayo Clinic Institutional Review Board approved this study (No. 12-008023).

*Reprint requests: Joaquin Sánchez-Sotelo, MD, PhD, Department of Orthopedic Surgery, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA.

E-mail address: sanchezsotelo.joaquin@mayo.edu (J. Sánchez-Sotelo).

Classification systems in orthopedic surgery can be useful tools to characterize pathology and facilitate communication. They can serve to compare disease severity and guide management. However, the effectiveness of any given

1058-2746/$ - see front matter © 2018 Journal of Shoulder and Elbow Surgery Board of Trustees. All rights reserved.
https://doi.org/10.1016/j.jse.2018.09.021

classification system depends to some extent on its ease of understanding, and its use should yield similar results among persons using it. Intraobserver and interobserver reliabilities are commonly used to measure how reproducible a classification system is.

The original Walch classification of glenoid morphology in primary osteoarthritis was introduced in 1999.[12] Two observers analyzed computed tomography (CT) images of 113 shoulders and found "substantial" intraobserver and interobserver reliabilities (Cohen κ coefficient = 0.65-0.70). Several additional studies on the reliability of the original Walch classification followed. Scalise et al[10] reported fair intraobserver ($\kappa = 0.34$) and interobserver ($\kappa = 0.37$) agreement, whereas Nowak et al[6] found that their intraobserver agreement was substantial ($\kappa = 0.61$) and interobserver agreement was moderate ($\kappa = 0.51$). Aronowitz et al[1] included an assessment of radiographs as well as CT scans and noted that intraobserver and interobserver agreements using radiographs were substantial ($\kappa = 0.66$) and moderate ($\kappa = 0.48$), respectively, whereas intraobserver and interobserver agreements using CT scans were moderate ($\kappa = 0.60$) and fair ($\kappa = 0.39$), respectively.

As a result of the varying degrees of reliability reported when using the original Walch classification, a modified version has been proposed by Walch and colleagues (ie, Bercik et al[2]) with additional subtypes and further clarification of the definitions of specific categories. Although this new modified classification scheme is promising, its reliability needs to be assessed by additional independent researchers. As such, the purposes of this study were to assess the reliability of the modified Walch classification[2] using both axillary radiographs and CT scans and to analyze the subcategorization of bone loss and eccentric wear or subluxation.

Materials and methods

For this study, we selected 98 consecutive patients (100 shoulders) with a diagnosis of primary glenohumeral osteoarthritis who underwent shoulder arthroplasty at a single institution between November 2013 and November 2016. We included patients who had a diagnosis of primary osteoarthritis and who then underwent shoulder arthroplasty of some type. Patients were excluded if there was any other diagnosis (ie, avascular necrosis, rheumatoid arthritis, or cuff tear arthropathy); any prior surgical procedure, including any arthroscopic instability or rotator cuff procedure; or any prior open procedure. All patients had preoperative CT scans and axillary radiographs obtained routinely prior to surgery. There were 50 men (51%) and 48 women (49%). The mean patient age was 69 years (range, 44-91 years). There were 46 left shoulders (46%) and 54 right shoulders (54%).

For the purpose of this study, all axillary radiographs were blinded and provided to the observers in sequential random order. Similarly, five 1-mm axial cuts from each CT scan were selected, including the midaxial cut and 4 images equally spaced cranial to and caudal to the center of the glenoid. These selected images for each shoulder were also blinded and provided to the observers in sequential random order.

Three highly experienced fellowship-trained shoulder surgeons with special interest in shoulder arthroplasty (J.W.S., R.H.C., and J.S.-S.) independently evaluated all of the images. As mentioned before, the evaluators were blinded to the readings of the other evaluators, as well as to any patient information. All observers were provided with a description and pictorial representation of the modified Walch classification from the original article[2] as a reference to be used when performing each reading (Fig. 1, Table I). No time limitations were imposed on any of the readings. There were 4 separate readings: 2 using only radiographs and 2 using only CT scans. The observers evaluated the radiographs first, followed by the CT scans and then the radiographs again, and finally, the CT scans were assessed a second time. Each reading was separated by at least 4 weeks.

The intraobserver reliability was determined by comparison of the classification of each subject by the observers for both the axillary radiographs and CT scans. Pair-wise comparisons between observers were also performed to determine interobserver and intraobserver reliability. We calculated κ values for both interobserver and intraobserver reliability; κ values adjust for the proportion of agreement among observers that could have occurred by chance. Landis and Koch[5] previously categorized κ values of 0.00 to 0.20 as slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 or greater, almost perfect agreement. A value of 0.00 indicates agreement no better than chance whereas 1.00 indicates perfect agreement. Additional analysis was performed to compare the initial reading of radiographs and CT scans of each observer, as well as agreement when the classification was collapsed into bone loss and subluxation categories.

Results

The first CT scan reading was performed by the most senior observer (R.H.C.) and served as the gold standard. The most senior surgeon classified 18 shoulders as A1 (18%), 12 as A2 (12%), 20 as B1 (20%), 25 as B2 (25%), 22 as B3 (22%), 1 as C (1%), and 2 as D (2%). Intraobserver agreement for all 3 observers is summarized in Table II. The average intraobserver agreement for both CT scans and radiographs was substantial ($\kappa = 0.72$ and $\kappa = 0.73$, respectively).

The average interobserver agreement for both CT scans and radiographs was fair ($\kappa = 0.52$ and $\kappa = 0.55$, respectively). Values for each observer are summarized in Table III. Of note, interobserver agreement was moderate between the gold-standard observer and observer 2 for both CT scans and radiographs ($\kappa = 0.63$ and $\kappa = 0.61$, respectively).

Table IV shows the first CT scan and radiograph readings of each of the 3 observers. From top left to bottom right, the cells marked with asterisks arranged diagonally indicate the agreements between CT scans and radiographs for each individual surgeon. Observers 1, 2, and 3 showed intraobserver agreement in 39 cases (39%), 51 cases (51%), and 58 cases (58%), respectively. In total, complete agreement occurred in 148 of 300 paired readings (49%).

The classification system was then collapsed according to bone loss (Table V): no bone loss (A1, B1, or D) versus bone loss (A2, B2, B3, or C). The cells marked with asterisks

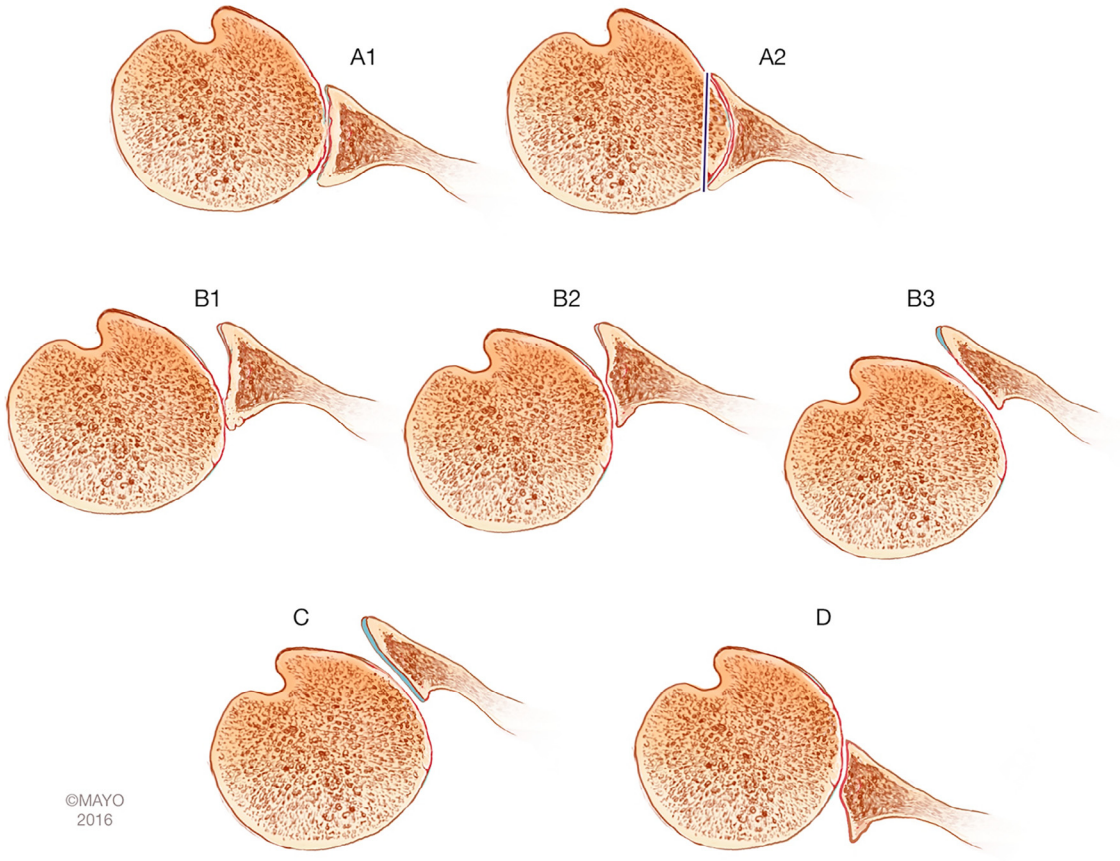

Figure 1 Walch classification. Used with permission of Mayo Foundation for Medical Education and Research. All rights reserved.

Table I Modified Walch classification

Walch category	Description
A	Humeral head centered on glenoid fossa
Subgroup A1	No or minor central erosion
Subgroup A2	Major central erosion
B	Posterior subluxation of humeral head
Subgroup B1	No posterior bone loss
Subgroup B2	Posterior bone loss resulting in biconcave glenoid
Subgroup B3	Monoconcave bone loss with 15° of retroversion and/or 70% posterior subluxation
C	Glenoid retroversion > 25°
D	Any glenoid anteversion or <40% subluxation

Table II Intraobserver agreement for CT scans and radiographs separately

	κ Value	
	CT Scans	Radiographs
Observer 1	0.77	0.72
Observer 2	0.69	0.75
Observer 3	0.71	0.72
Average	0.72	0.73

CT, computed tomography.
The results are based on 2 readings separated by at least 4 weeks.

arranged diagonally from top left to bottom right represent agreement between CT scans and radiographs. Agreement was shown in 81 shoulders (81%) for observer 1, 79 (79%) for observer 2, and 93 (93%) for observer 3. Overall, the presence or absence of bone loss was agreed on in 254 of 300 shoulders (85%).

Similarly, the classification system was collapsed according to the presence of subluxation (B1, B2, B3, C, or D) or absence of subluxation (A1 or A2) (Table VI). Again, the cells marked with asterisks and arranged diagonally indicate agreement between CT scans and radiographs. Agreement was shown in 60 shoulders (60%) for observer 1, 73 (73%) for observer 2, and 78 (78%) for observer 3. Overall, the presence or absence of eccentric wear or subluxation was agreed on in 211 of 300 shoulders (70%).

Discussion

The results of this study indicate that the "modified" Walch classification[2] provides fair and substantial agreement within and between observers when applied using axillary radiographs or the appropriate axial CT images. In fact, on the basis

Table III Interobserver agreement for pairs of 3 observers for CT scans and axillary radiographs

Observer	κ Value	
	CT Scans	Radiographs
Observer 1 vs observer 2	0.63	0.61
Observer 1 vs observer 3	0.50	0.51
Observer 2 vs observer 3	0.43	0.53
Average	0.52	0.55

CT, computed tomography.

Table IV First CT and radiographic readings for all 3 observers

Radiographs	CT							Total
	A1	A2	B1	B2	B3	C	D	
Reader 1								
A1	10*	0	9	5	0	0	2	26
A2	4	9*	5	5	7	0	0	30
B1	1	0	3*	1	2	0	0	7
B2	0	1	0	8*	3	1	0	13
B3	1	2	1	5	9*	0	0	18
C	0	0	0	1	1	0*	0	2
D	2	0	2	0	0	0	0*	4
Total	18	12	20	25	22	1	2	100
Reader 2								
A1	11*	6	1	1	0	0	1	20
A2	3	17*	3	2	6	0	0	31
B1	2	2	1*	0	1	0	0	6
B2	0	1	3	9*	4	1	0	18
B3	1	5	0	2	13*	0	0	21
C	0	0	0	1	0	0*	0	1
D	1	1	1	0	0	0	0*	3
Total	18	32	9	15	24	1	1	100
Reader 3								
A1	7*	0	2	1	0	0	2	12
A2	0	8*	1	6	7	0	0	22
B1	0	0	5*	0	0	0	0	5
B2	0	0	3	18*	7	0	1	29
B3	0	2	0	7	18*	1	0	28
C	0	0	0	0	0	2*	0	2
D	1	0	0	1	0	0	0*	2
Total	8	10	11	33	32	3	3	100

CT, computed tomography.
* Agreement between CT scans and radiographs for each individual surgeon.

Table V No bone loss versus bone loss

Radiographs	CT	
	A1, B1, or D	A2, B2, B3, or C
Reader 1		
A1, B1, or D	29*	8
A2, B2, B3, or C	11	52*
Reader 2		
A1, B1, or D	18*	11
A2, B2, B3, or C	10	61*
Reader 3		
A1, B1, or D	17*	2
A2, B2, B3, or C	5	76*

CT, computed tomography.
* Agreement between CT scans and radiographs.

Table VI No subluxation versus subluxation

Radiographs	CT	
	A1 or A2	B1, B2, B3, C, or D
Reader 1		
A1 or A2	23*	33
B1, B2, B3, C, or D	7	37*
Reader 2		
A1 or A2	37*	14
B1, B2, B3, C, or D	13	36*
Reader 3		
A1 or A2	15*	19
B1, B2, B3, C, or D	3	63*

CT, computed tomography.
* Agreement between CT scans and radiographs.

of a comparison of our group's similar assessment of the original Walch classification,[12] there appears to have been an improvement in all categories, which seems to indicate that the stated goal of the authors, namely to improve the reliability of the classification, has been achieved. Compared with the work of Aronowitz et al[1] in which an almost identical methodology was used, the mean intraobserver agreement for radiographs improved from 0.66 to 0.73 (both substantial), and the mean intraobserver agreement for CT scans improved from 0.60 (moderate) to 0.72 (substantial). The pair-wise comparisons between observers improved both for radiographs (0.48 to 0.55) and for CT scans (0.39 to 0.52). In addition, the mean agreement among observers on the presence or absence of bone loss increased from 70%[1] to 85% in our study, although the mean agreement on the presence or absence of subluxation was similar between the study by Aronowitz et al (72%) and our study (70%).

The original Walch classification[12] was introduced in 1999 and provided a widely adopted method by which surgeons could assess glenoid morphology in primary osteoarthritis using 2-dimensional CT scans. An accurate assessment of glenoid morphology prior to shoulder arthroplasty is of critical importance to optimize implant positioning, select the correct implant type (ie, anatomic vs reverse shoulder arthroplasty), and anticipate management of subluxation or bone loss.

Several aspects of these data are of clinical importance. One purpose of this classification system is to provide an accurate assessment of glenoid retroversion. Once the type B glenoid was described and characterized, surgeons gained a greater understanding of glenoid morphology in osteoarthritis. Walch et al[13] reported revision rates of up to 16% in

patients who underwent anatomic arthroplasty with biconcave glenoids. In addition, glenoid loosening has been correlated with excessive retroversion of the glenoid component,[4,8] as well as with Walch type B glenoids.[3,11,12,14] These determinations of concavity and version are based on surgeon assessment and measurements, and the Walch classification is one of the most widely accepted systems for such determination. The body of literature on this topic continues to expand, and the surgical community benefits from authors' reported outcomes on how implant design (ie, anatomic vs reverse) correlates with success in certain glenoid morphologies. For example, the use of a reverse arthroplasty in the setting of a type B2 glenoid with humeral head posterior subluxation of over 80% has become an accepted treatment option. However, the accuracy of these determinations and clinical outcome reports are predicated on our understanding that this system has a certain degree of intraobserver and interobserver reliability. The original Walch classification underwent this process of evaluation to assess reliability through several publications, and now, a thorough evaluation of the agreement of the new modified system is critical for future outcome reporting.

Another clinically relevant contribution of these data relates to the evaluation of different imaging types. This study showed that the modified classification can be applied to both CT images and axillary radiographs. This is useful for surgeons who do not, or cannot, routinely obtain CT images prior to shoulder replacement.

For the Walch classification, 3 broad categories were initially created, and Walch et al[12] reported that intraobserver reproducibility and interobserver reliability were in the "good" range based on κ indices (range, 0.65-0.70). Subsequently, several reports re-evaluated these values with varying results. Nowak et al reported moderate interobserver agreement ($\kappa = 0.51$) and substantial intraobserver agreement ($\kappa = 0.61$) based on the measurements of 8 observers. Scalise et al,[10] however, instituted the classification with 4 observers and reported only fair interobserver ($\kappa = 0.37$) and intraobserver ($\kappa = 0.34$) reliabilities.

Given the varying levels of interobserver and intraobserver reliability reported with the original Walch classification, the original authors modified the system in an effort to increase the consistency among observers.[2] The new classification was also created to include previously unrecognized morphologies that the authors identified with time and experience over 17 additional years of performing shoulder arthroplasty. Specifically, the type B3 and D categories were added, and a more detailed definition of the type A2 glenoid was included. The type B3 glenoid was defined as monoconcave with posterior wear and 15° of retroversion or more and/or 70% posterior humeral head subluxation. The type D glenoid was defined as any amount of anteversion or subluxation of the humeral head of less than 40%. The definition of the type A2 glenoid, previously termed the "cupula," was adjusted to include morphology in which an anteroposterior line that connects the glenoid rims transects the humeral head.

The improvements noted in our study support the findings of Bercik et al,[2] as these authors reported substantial improvements in interobserver reliability from a mean of 0.39 (fair agreement) to 0.70 (substantial agreement) whereas intraobserver reliability improved from a mean of 0.61 (moderate agreement) to 0.88 (nearly perfect agreement). However, the mean interobserver reliability found in this study was still 1 category below that of Bercik et al (ie, moderate agreement in our study vs substantial agreement in their study). There may be several possible explanations for this. First, the observers in the publication of Bercik et al had the benefit of conceptualizing and developing this modification, so those observers would be more familiar with the system than any other observers. Second, as stated in the article, the new modification was developed by those authors over many years and likely continually refined prior to testing and publication. However, the observers in our study were introduced to this modification for the first time once that publication was available. Therefore, it is possible that with increased practice and application of the modified system, the reliability values may improve to levels that more closely approximate those reported by Bercik et al. As occurred after the introduction of the original Walch classification, other authors instituted the classification and studied the reliabilities, the results of which ultimately contributed to this refinement in the system.

One might argue that the experience level of each observer might have affected these results, as the interobserver reliabilities were fair between the CT scans and radiographs. However, although the gold-standard observer (seniormost surgeon) has practiced for the longest period, we are confident that the other observers would meet the criteria to be considered experts in their field; therefore, we do not believe that any differential in surgeon experience confounded these results.

This study found that both axillary radiographs and CT scans can be used reliably with the modified Walch classification to deliver a reproducible assessment of glenoid morphology, as well as to broadly subcategorize the presence or absence of bone loss and eccentric wear or subluxation. However, multiple studies have confirmed that CT scans are more accurate for specific measurements beyond classification. Nyffeler et al[7] found that glenoid assessment using CT scans showed "excellent" interobserver reliability but that the reliability using axillary radiographs was "poor," based exclusively on images from patients who had anterior glenohumeral instability or a history of arthroplasty, not images from patients with primary arthritis. Another study found that glenoid version measured on radiographs was an average of 7.4° greater than that measured on magnetic resonance imaging.[9]

Conclusions

The modified Walch classification represents an improvement to the original Walch classification, as it has

incorporated knowledge gained with the use of 3-dimensional glenoid reconstructions, as well as close to 2 decades of additional experience in performing shoulder arthroplasty. These improvements have resulted in greater intraobserver and interobserver reliabilities and have reduced the differential values of reliabilities obtained from radiographs versus CT scans.

Although CT offers a clearly superior imaging tool prior to shoulder arthroplasty, in particular regarding accurate measurements of version, inclination, and subluxation, as well as computer-based preoperative planning, it is reassuring to learn that shoulder surgeons can reach reasonable agreement using both radiographs and CT scans. This information would be of particular use when counseling patients in the absence of CT scans, as well as when retrospective research needs to include shoulders for which preoperative CT scans were not available.

Disclaimer

Joaquin Sánchez-Sotelo reports that he contributed to an implant design for Stryker and receives royalties.

Robert H. Cofield reports that he has a patent with Smith & Nephew, with royalties paid to the Mayo Foundation.

John W. Sperling reports that he contributed to an implant design for Biomet and receives royalties.

The other authors, their immediate families, and any research foundations 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. Aronowitz JG, Harmsen WS, Schleck CD, Sperling JW, Cofield RH, Sánchez-Sotelo J. Radiographs and computed tomography scans show similar observer agreement when classifying glenoid morphology in glenohumeral arthritis. J Shoulder Elbow Surg 2017;26:1533-8. http://dx.doi.org/10.1016/j.jse.2017.02.015
2. 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;25:1601-6. http://dx.doi.org/10.1016/j.jse.2016.03.010
3. Farron A, Terrier A, Büchler P. Risks of loosening of a prosthetic glenoid implanted in retroversion. J Shoulder Elbow Surg 2006;15:521-6. http://dx.doi.org/10.1016/j.jse.2005.10.003
4. Ho JC, Sabesan VJ, Iannotti JP. Glenoid component retroversion is associated with osteolysis. J Bone Joint Surg Am 2013;95:e82. http://dx.doi.org/10.2106/JBJS.L.00336
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
6. Nowak DD, Gardner TR, Bigliani LU, Levine WN, Ahmad CS. Interobserver and intraobserver reliability of the Walch classification in primary glenohumeral arthritis. J Shoulder Elbow Surg 2010;19:180-3.
7. Nyffeler RW, Jost B, Pfirrmann CW, Gerber C. Measurement of glenoid version: conventional radiographs versus computed tomography scans. J Shoulder Elbow Surg 2003;12:493-6. https://doi.org/10.1016/S1058-2746(03)00181-2
8. Raiss P, Schmitt M, Bruckner T, Kasten P, Pape G, Loew M, et al. Results of cemented total shoulder replacement with a minimum follow-up of ten years. J Bone Joint Surg Am 2012;94:e1711-10. http://dx.doi.org/10.2106/JBJS.K.00580
9. Raymond AC, McCann PA, Sarangi PP. Magnetic resonance scanning vs axillary radiography in the assessment of glenoid version for osteoarthritis. J Shoulder Elbow Surg 2013;22:1078-83. http://dx.doi.org/10.1016/j.jse.2012.10.036
10. Scalise JJ, Codsi MJ, Brems JJ, Iannotti JP. Inter-rater reliability of an arthritic glenoid morphology classification system. J Shoulder Elbow Surg 2008;17:575-7. http://dx.doi.org/10.1016/j.jse.2007.12.006
11. Shapiro TA, McGarry MH, Gupta R, Lee YS, Lee TQ. Biomechanical effects of glenoid retroversion in total shoulder arthroplasty. J Shoulder Elbow Surg 2007;16:S90-5. http://dx.doi.org/10.1016/j.jse.2006.07.010
12. Walch G, Badet R, Boulahia A, Khoury A. Morphologic study of the glenoid in primary glenohumeral osteoarthritis. J Arthroplasty 1999;14:756-60.
13. Walch G, Moraga C, Young A, Castellanos-Rosas J. Results of anatomic nonconstrained prosthesis in primary osteoarthritis with biconcave glenoid. J Shoulder Elbow Surg 2012;21:1526-33. http://dx.doi.org/10.1016/j.jse.2011.11.030
14. Walch G, Young AA, Boileau P, Loew M, Gazielly D, Molé D. Patterns of loosening of polyethylene keeled glenoid components after shoulder arthroplasty for primary osteoarthritis: results of a multicenter study with more than five years of follow-up. J Bone Joint Surg Am 2012;94:145-50. http://dx.doi.org/10.2106/JBJS.J.00699