J Shoulder Elbow Surg (2021) 30, 736–746

ELSEVIER

Journal of Shoulder and Elbow Surgery

www.elsevier.com/locate/ymse

Reliability of the modified Walch classification for advanced glenohumeral osteoarthritis using 3-dimensional computed tomography analysis: a study of the ASES B2 Glenoid Multicenter Research Group

ASES B2 Glenoid Multicenter Research Group, Eric T. Ricchetti, MD*, Michael S. Khazzam, MD, Patrick J. Denard, MD, David M. Dines, MD, T. Bradley Edwards, MD, Vahid Entezari, MD, MMSc, Richard J. Friedman, MD, FRCSC, Grant E. Garrigues, MD, Robert J. Gillespie, MD, Brian M. Grawe, MD, Andrew Green, MD, Armodios M. Hatzidakis, MD, J. Gabriel Horneff, MD, Jason E. Hsu, MD, Andrew Jawa, MD, Yuxuan Jin, MS, Peter S. Johnston, MD, Bong-Jae Jun, PhD, Jay D. Keener, MD, James D. Kelly II, MD, Young W. Kwon, MD, PhD, Anthony Miniaci, MD, Brent J. Morris, MD, Surena Namdari, MD, MSc, Edwin E. Spencer, MD, Gregory Strnad, MS, Gerald R. Williams Jr, MD, Joseph P. Iannotti, MD, PhD

ASES B2 Glenoid Multicenter Research Group

Background: Variations in glenoid morphology affect surgical treatment and outcome of advanced glenohumeral osteoarthritis (OA). The purpose of this study was to assess the inter- and intraobserver reliability of the modified Walch classification using 3-dimensional (3D) computed tomography (CT) imaging in a multicenter research group.
Methods: Deidentified preoperative CTs of patients with primary glenohumeral OA undergoing anatomic or reverse total shoulder arthroplasty (TSA) were reviewed with 3D imaging software by 23 experienced shoulder surgeons across 19 institutions. CTs were separated into 2 groups for review: group 1 (96 cases involving all modified Walch classification categories evaluated by 12 readers) and group 2 (98 cases involving posterior glenoid deformity categories [B2, B3, C1, C2] evaluated by 11 readers other than the first 12). Each case group was reviewed by the same set of readers 4 different times (with and without the glenoid vault model present), blindly and in random order. Inter- and intraobserver reliabilities were calculated to assess agreement (slight, fair, moderate, substantial, almost perfect) within groups and by modified Walch classification categories.
Results: Interobserver reliability showed fair to moderate agreement for both groups. Group 1 had a kappa of 0.43 (95% confidence interval [CI]: 0.38, 0.48) with the glenoid vault model absent and 0.41 (95% CI: 0.37, 0.46) with it present. Group 2 had a kappa of

Institutional Review Board approval was obtained for this study from the Cleveland Clinic (IRB study no. 17-1206).

*Reprint requests: Eric T. Ricchetti, MD, Department of Orthopaedic Surgery, Orthopedic and Rheumatologic Institute, Cleveland Clinic, 9500 Euclid Avenue, Mail Code A40, Cleveland, OH 44195, USA.
E-mail address: ricchee@ccf.org (E.T. Ricchetti).

https://doi.org/10.1016/j.jse.2020.07.013

0.38 (95% CI: 0.33, 0.43) with the glenoid vault model absent and 0.37 (95% CI: 0.32, 0.43) with it present. Intraobserver reliability showed substantial agreement for group 1 with (0.63, range 0.47-0.71) and without (0.61, range 0.52-0.69) the glenoid vault model present. For group 2, intraobserver reliability showed moderate agreement with the glenoid vault model absent (0.51, range 0.30-0.72), which improved to substantial agreement with the glenoid vault model present (0.61, range 0.34-0.87).

Discussion: Inter- and intraobserver reliability of the modified Walch classification were fair to moderate and moderate to substantial, respectively, using standardized 3D CT imaging analysis in a large multicenter study. The findings potentially suggest that cases with a spectrum of posterior glenoid bone loss and/or dysplasia can be harder to distinguish by modified Walch type because of a lack of defined thresholds, and the glenoid vault model may be beneficial in determining Walch type in certain scenarios. The ability to reproducibly separate patients into groups based on preoperative pathology, including Walch type, is important for future studies to accurately evaluate postoperative outcomes in TSA patient cohorts.

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

Keywords: Walch classification; glenohumeral osteoarthritis; glenoid morphology; computed tomography; 3-dimensional; reliability; shoulder arthroplasty

The Walch classification is the most commonly used method for defining pathology in glenohumeral osteoarthritis (OA). The original classification consisted of 5 different morphologic subtypes (A1, A2, B1, B2, C) based on the pattern of glenoid morphology and humeral head alignment on preoperative computed tomography (CT) scans in patients undergoing shoulder arthroplasty.[26] The classification has been the basis to stratify outcomes of surgical interventions and make recommendations regarding the type of shoulder arthroplasty to perform for varying types of pathology, as clinical outcomes and implant survivorship have been shown to be impacted by certain pathologic patterns.[5,7,11,17,27] However, controversy exists regarding the optimal surgical treatment of patients with more severe glenoid pathology, specifically patients with B and C glenoid types. In addition, several studies assessing the original Walch classification have shown only fair to moderate interobserver reliability and fair to substantial intraobserver reliability on CT even when used by experienced shoulder surgeons,[18,22] and conclude that improvements in the classification could provide further utility.

The Walch classification has subsequently undergone modifications that include refinement of the definition of certain existing Walch types, as well as the addition of new Walch types (B3, C2, D) describing morphologic patterns not in the original classification.[1,10] These modifications have shown improved inter- and intraobserver reliability, although the findings have been mixed across recent studies, which have been limited to a small number of raters at single institutions.[1,10,25] Assessment of the reliability of the modified Walch classification by a larger number of raters across multiple institutions may provide further information on the reproducibility of the classification when applied in more widespread clinical use. This has important implications, as the conclusions of a study on the clinical outcomes of shoulder arthroplasty may be dependent on the type of pathology and specific arthroplasty treatment investigated within that clinical cohort. A classification with poor reliability has the potential to result in misguided or ambiguous treatment selection and misinterpretation of clinical outcomes.

The purpose of this study was to assess the inter- and intraobserver reliability of the modified Walch classification using 3-dimensional (3D) CT imaging in a large multicenter research group (ASES B2 Glenoid Multicenter Research Group), including assessment of a subset of cases with posterior glenoid deformity where controversy on the optimal surgical treatment may exist. Secondarily, the study evaluated which pathologic features or modified Walch types may lower reliability, and if use of the glenoid vault model improved the inter- or intraobserver reliability of the modified Walch classification. The glenoid vault is a highly consistent and conserved shape across normal individuals, and prior studies have shown the glenoid vault model, a standardized 3D model of the normal glenoid vault, to be predictive of premorbid glenoid version, inclination, and joint line position when placed into a pathologic glenoid, as well as helpful in defining certain modified Walch classification types.[4,8,10,19,21,24].

Materials and methods

A library of deidentified preoperative CTs of patients with primary glenohumeral OA undergoing anatomic or reverse total shoulder arthroplasty (TSA) was created with a 3D imaging analysis software (OrthoVis Shoulder Research software, Cleveland Clinic; Cleveland, OH, USA). Within the software, the plane of the scapula was defined in 3D by the glenoid center point, inferior angle of the scapula, and scapula trigonum (Fig. 1), with 2-dimensional (2D) orthogonal planes (axial, coronal, and sagittal) then referenced relative to the scapular plane, as previously described.[6,10,12-14,20,28] The glenoid plane was defined by 3 representative points placed on the surface of the glenoid fossa, thereby defining average glenoid version and inclination in 3D relative to the scapular plane (Fig. 1), as previously described.[6,10,12-14,20,28] The glenoid vault model, a standardized 3D

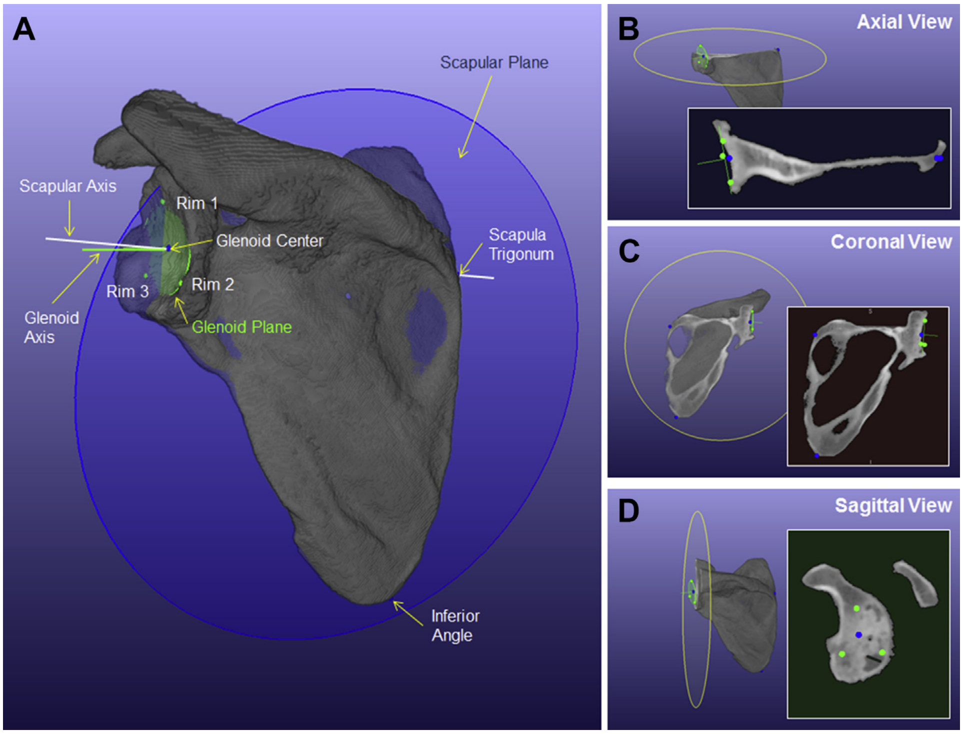

Figure 1 (**A**) The scapular plane is defined using the glenoid center point, the inferior angle of the scapula, and the scapula trigonum. The glenoid plane is defined using 3 representative points placed on the surface of the glenoid fossa. Pathologic glenoid version and inclination are then measured in 3D from the scapular and glenoid fossa planes. Representative 2-D orthogonal axial (**B**), coronal (**C**), or sagittal (**D**) views of the plane of the scapula can then be selected to view pathology. Reproduced from Iannotti et al.[10]

model of the normal glenoid vault, was then placed into the glenoid of each CT scan as a tool to define the patients' premorbid glenoid anatomy, as previously described (Fig. 2).[4,8,19,21,24] Cases were uploaded into a secure web portal (CCF Shoulder Research Portal, Cleveland Clinic; Cleveland, OH, USA) and reviewed by a group of 23 experienced shoulder surgeons across 19 institutions (ASES B2 Glenoid Multicenter Research Group). The imaging software in the secure web portal allowed for the assessment of individual CT scans in 3D and in the 2D axial and coronal orthogonal planes, with 3D measurements of glenoid version and inclination relative to the pathologic bone (''native version'' and ''native inclination'') and relative to the premorbid glenoid vault model (''vault version'' and ''vault inclination'') provided in each case (Fig. 3).

All CTs were first classified using the modified Walch classification by 2 of the authors (E.T.R, J.P.I.) by consensus using criteria described by Bercik et al[1] and Iannotti et al.[10] CTs were then separated into 2 case groups for review by 2 sets of readers; group 1, consisting of 96 cases involving all modified Walch classification categories (A1, A2, B1, B2, B3, C1, C2, D) evaluated by 12 readers from the ASES B2 Glenoid Multicenter Research Group; and group 2, consisting of 98 cases involving only posterior glenoid deformity categories (B2, B3, C1, C2) evaluated by 11 readers other than the first 12 from the ASES B2 Glenoid Multicenter Research Group. Each case group was reviewed by the same set of readers 4 different times, blindly and in random order each time. The first 2 reviews for a given case group were performed without the presence of the glenoid vault model, whereas the third and fourth reviews for a given case group were performed with the glenoid vault model present. Readers were required to review a training manual that defined the criteria for each modified Walch type (Table I) and included 3D training case examples that were available for review in the imaging software on the secure web portal during classification of the study cases. The 3D training case examples were also blindly and randomly included in the cases reviewed by each reader in groups 1 and 2 to assess the reliability of these training cases. A 3-week period was given to complete each of the reviews (total of 4 reviews), with each reviewing period separated by 5 weeks (washout period). Readers selected a Walch type for every reviewed case.

Interobserver reliability was determined for groups 1 and 2 using Fleiss kappa statistics with 95% confidence intervals (CIs). Kappa values were calculated for the cases in each group both with and without the glenoid vault model present, as well as with and without the training cases present. Interobserver reliability was also calculated by modified Walch classification using Gwet agreement coefficient 1 (AC1) method with 95% CI, both for comparing across the major groups (A, B, C, D) and for comparing across all of the categories (A1, A2, B1, B2, B3, C1, C2, D), again both with and without the glenoid vault model present. Fleiss kappa statistics are frequently criticized for their tendency to be low in situations where raw agreement is high. As a result, when calculating interobserver reliability by modified Walch classification, an alternative approach proposed by Gwet, the AC1 method, had been used.[9] Gwet AC1 adopts a more lenient chance correction compared with Fleiss kappa statistics.

Intraobserver reliability was calculated for each reader in groups 1 and 2 using Cohen kappa statistics, as well as overall for each group. This included determining overall intraobserver reliability in cases with and without the glenoid vault model present. The overall intraobserver reliability for each group was obtained

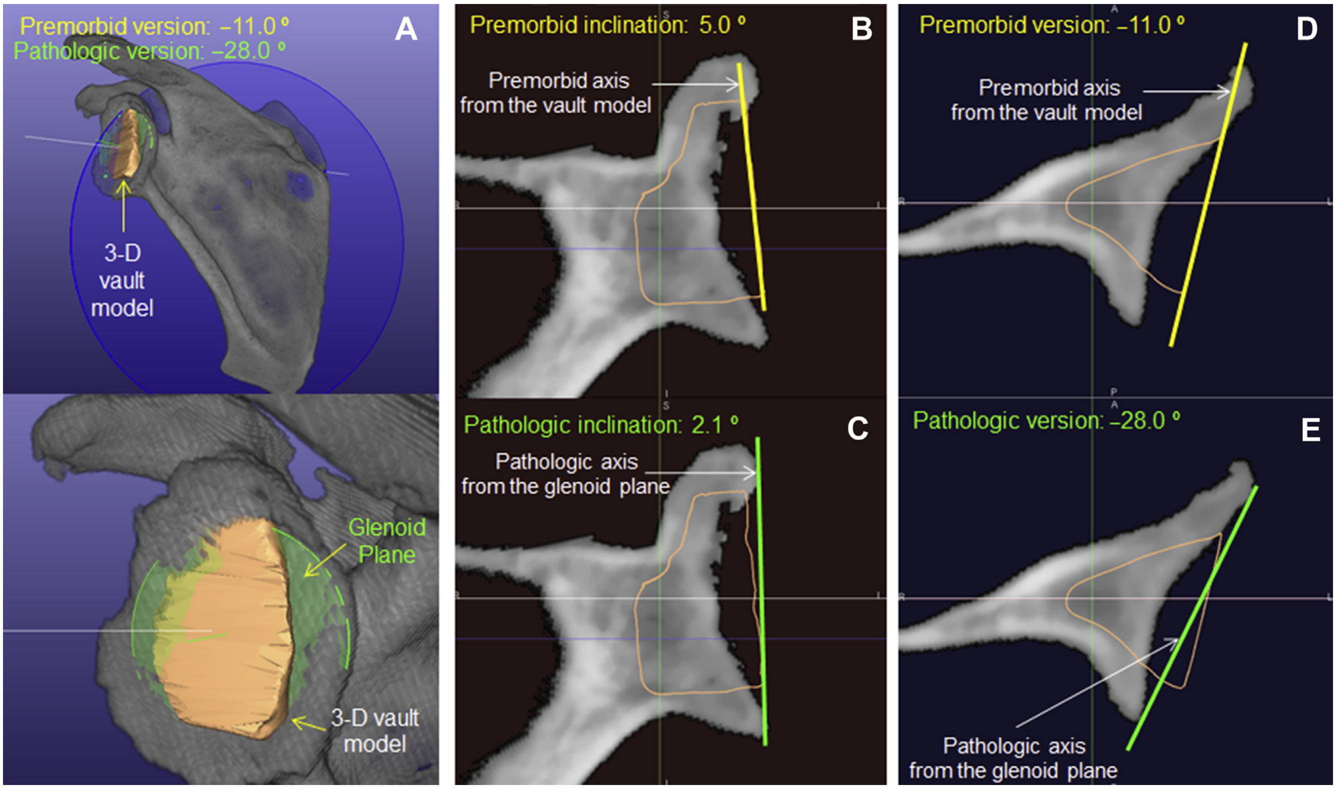

Figure 2 The glenoid vault model (**A**) is virtually placed into the glenoid of each preoperative CT scan as a tool to define premorbid glenoid anatomy. Premorbid (vault) glenoid version and inclination are measured in 3D from the scapula plane and the plane of the glenoid vault surface. (**B-E**) Representative 2D orthogonal views relative to the scapular plane are used to confirm vault model position and demonstrate pathologic and premorbid (vault) inclination (**B, C**) and version (**D, E**). Joint line medialization of the pathologic glenoid surface relative to the vault model surface can also be assessed on the 2D orthogonal views. Reproduced from Iannotti et al.[10]

by dividing the summed numerators of the individual agreement measures for each reader by the summed denominators of the same measures. A Cohen kappa value of 0.6 or higher for intraobserver reliability on a given set of case group reviews was also used to define a "good" reader in order to evaluate if "good" readers had better interobserver reliability than readers overall.

The scale of kappa or Gwet AC1 statistics was defined as follows: 0-0.2 slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1 almost perfect agreement.[16] Data management and analysis was performed in R software (version 3.5; R Foundation for Statistical Computing; Vienna, Austria). All tests were 2-sided, assuming an alpha level of 0.05.

Results

Table II shows the interobserver reliability results for groups 1 and 2 across the different scenarios evaluated. Interobserver reliability for both groups of cases reached the level of fair to moderate agreement (0.21-0.60) by the kappa statistic regardless of the scenario, including with or without the glenoid vault model present and in training cases compared with nontraining cases. "Good" readers, as defined by a kappa value of 0.6 or higher for intraobserver reliability, also performed no better by agreement level on interobserver reliability as a group when compared to readers overall. Group 1 (12 readers), representing cases from all categories of the modified Walch classification (A1, A2, B1, B2, B3, C1, C2, D), had a kappa value of 0.43 (95% CI: 0.38, 0.48) with the glenoid vault model absent, compared to 0.41 (95% CI: 0.37, 0.46) with the glenoid vault model present when evaluating all of the cases in the group. Group 2 (11 readers), representing cases from categories of the modified Walch classification associated with posterior glenoid deformity (B2, B3, C1, C2), had a kappa value of 0.38 (95% CI: 0.33, 0.43) with the glenoid vault model absent, compared to 0.37 (95% CI: 0.32, 0.43) with the glenoid vault model present when evaluating all of the cases in the group.

Table III demonstrates the overall intraobserver reliability results for groups 1 and 2, when evaluated with and without the glenoid vault model present. Intraobserver reliability reached the level of substantial agreement (0.61-0.80) by kappa statistic for group 1 both with (0.63, range 0.47-0.71) and without (0.61, range 0.52-0.69) the presence of the glenoid vault model. For group 2, intraobserver reliability only reached the level of moderate agreement (0.41-0.60) with the glenoid vault model absent (0.51,

Figure 3 Images of a preoperative 3D CT scan uploaded into the secure web portal (CCF Shoulder Research Portal) used by observers for classification of cases by modified Walch type. The imaging software in the secure web portal allows for the assessment of individual CT scans in 3D and in the 2D axial and coronal orthogonal planes, with 3D measurements of glenoid version and inclination relative to the pathologic bone (“Native Version” and “Native Inclination,” noted in the upper right-hand corner of the webpage) and relative to the premorbid glenoid vault model (“Vault Version” and “Vault Inclination,” noted in the “Info” tab of the webpage) provided in each case. Panel **A** shows the 3D CT scan without the glenoid vault model present, and without the measurements of “Vault Version” and “Vault Inclination” shown. The first 2 reviews for a given case group were performed without the presence of the glenoid vault model and its associated measurements. Panel **B** shows the 3D CT scan with the glenoid vault model present and includes the measurements of “Vault Version” and “Vault Inclination.” The third and fourth reviews for a given case group were performed with the glenoid vault model and its associated measurements present.

range 0.30-0.72), which improved to substantial agreement with the glenoid vault model present (0.61, range 0.34-0.87).

Table IV shows the interobserver reliability by modified Walch classification type when all of the cases from groups 1 and 2 were combined. Interobserver reliability ranged from fair to moderate agreement (0.21-0.60) by Gwet AC1 across most of the modified Walch classification types, without consistent improvement in reliability when the glenoid vault model was present. B1 glenoids had the lowest interobserver reliability of the modified Walch classification types, both with the glenoid vault model absent (0.35, 95% CI: 0.24, 0.45) and present (0.25, 95% CI: 0.16, 0.35). In contrast, B2 glenoids had the highest

Table I Definitions for each Walch type, including descriptions of the case example(s) for each Walch type

Modified Walch classification type	Definition
A1	Centered humeral head, minor glenoid erosion. Case example A1_Case1 demonstrates these traits, with minimal glenoid erosion relative to the vault model.
A2	Centered humeral head, major central glenoid erosion defined by a line drawn from the anterior to posterior rims of the glenoid transecting the humeral head (Bercik et al) and/or the vault model (Iannotti et al). Case examples A2_Case1 and A2_Case2 both show evidence of central glenoid wear relative to the vault model, with A2_Case1 (top) showing milder wear and A2_Case2 (bottom) showing more advanced wear.
B1	Posteriorly subluxated humeral head, with no or minor posterior glenoid erosion. Case example B1_Case1 demonstrates these traits, with signs of mild posterior glenoid erosion associated with a posteriorly subluxated humeral head.
B2	Posteriorly subluxated humeral head, posterior glenoid erosion with biconcavity and no dysplasia. Case examples B2_Case1 and B2_Case2 both show these findings, with B2_Case1 (top) showing milder posterior glenoid wear with a biconcavity and B2_Case2 (bottom) showing more advanced posterior glenoid wear with a biconcavity.
B3	Posteriorly worn glenoid that is monoconcave with little or no biconcavity due to posterior and central glenoid erosion, without dysplasia. Thresholds of at least 15° glenoid retroversion and at least 70% posterior humeral head subluxation relative to the scapular line are proposed in Bercik et al as part of the definition, but these should not be used as hard cut-offs. Case examples B3_Case1 (top), B3_Case2 (middle), and B3_Case3 (bottom) show these findings, with evidence of posterior and central glenoid wear relative to the vault model. B3_Case1 shows milder wear relative to B3_Case2 and B3_Case3. B3_Case3 shows only a small remaining paleoglenoid (<25% glenoid diameter) and is, therefore, considered a B3 rather than a B2 glenoid.
C1	Dysplastic glenoid with high degrees of retroversion due to dysplasia rather than glenoid erosion. A threshold of at least 25° glenoid retroversion is proposed in Bercik et al as part of the definition, but this should not be used as a hard cut-off. Case example C1_Case1 demonstrates these traits, with hypoplasia of the posterior glenoid creating high degrees of glenoid retroversion. The vault model may be difficult to fit in a C1 glenoid because of the dysplasia and hypoplasia that is present, as shown in C1_Case1, or may show higher than normal vault (premorbid) retroversion.
C2	Dysplastic glenoid with acquired posterior glenoid erosion creating glenoid biconcavity and posterior subluxation of the humeral head. The vault model is used to help define a C2 relative to a B2 glenoid, as a C2 glenoid will have both high pathologic retroversion and high vault (premorbid) retroversion, whereas a B2 glenoid will have high pathologic retroversion and normal vault (premorbid) retroversion. Case examples C2_Case1 and C2_Case2 both show these findings, with both cases showing increased vault (premorbid) retroversion. C2_Case1 (top) shows less advanced posterior glenoid wear and biconcavity compared to C2_Case2 (bottom).
D	Glenoid anteversion or anterior humeral head subluxation. Case examples D_Case1 and D_Case2 both show these findings, with D_Case1 (top) showing milder anterior glenoid wear than D_Case2 (bottom) relative to the vault model.

These definitions incorporated the key features of the modified Walch classification, as defined by Bercik et al[1] and Iannotti et al.[10]

interobserver reliability of the modified Walch classification types with the glenoid vault model absent (0.61, 95% CI: 0.51, 0.70) and present (0.54, 95% CI: 0.45, 0.64). D glenoids had the most improvement in interobserver reliability with the vault model, improving from 0.25 (95% CI: 0.08, 0.49) with the glenoid vault model absent to 0.49 (95% CI: 0.15, 0.84) with the glenoid vault model present. "Good" readers, as defined by a kappa value of 0.6 or higher for intraobserver reliability, did not consistently perform better by agreement level on interobserver reliability across the modified Walch classification types when

Table II Interobserver reliability by Fleiss kappa statistic with 95% confidence intervals (95% CI) for groups 1 and 2 when compared across case types

Case group and type	Glenoid vault model absent		Glenoid vault model present	
	All readers (95% CI)	Readers ≥0.6 (95% CI)	All readers (95% CI)	Readers ≥0.6 (95% CI)
Group 1: All cases (n = 96)	0.43 (0.38, 0.48)	0.47 (0.40, 0.54)	0.41 (0.37, 0.46)	0.45 (0.37, 0.52)
Group 1: Training cases only (n = 6)	0.38 (0.12, 0.65)	0.46 (0.014, 0.90)	0.41 (0.16, 0.66)	0.43 (−0.033, 0.89)
Group 1: Training cases excluded (n = 90)	0.42 (0.37, 0.48)	0.46 (0.39, 0.54)	0.41 (0.36, 0.46)	0.44 (0.36, 0.52)
Group 2: All cases (n = 98)	0.38 (0.33, 0.43)	0.42 (0.36, 0.48)	0.37 (0.32, 0.43)	0.37 (0.30, 0.43)
Group 2: Training cases only (n = 8)	0.32 (0.11, 0.54)	0.36 (0.10, 0.52)	0.43 (0.26, 0.59)	0.49 (0.22, 0.75)
Group 2: Training cases excluded (n = 90)	0.39 (0.33, 0.44)	0.43 (0.36, 0.49)	0.37 (0.31, 0.43)	0.36 (0.29, 0.42)

CI, confidence interval.
Results are shown with and without the presence of the glenoid vault model, as well as for all readers and those with intrarater reliability by Cohen kappa statistic of 0.6 or higher.

Table III Overall intraobserver reliability by Cohen kappa statistic with the range of intraobserver reliabilities per reader for groups 1 and 2

Case group and type	Glenoid vault model absent		Glenoid vault model present	
	Cohen kappa	Range of readers	Cohen kappa	Range of readers
Group 1: All cases (n = 96), 12 readers	0.61	0.52-0.69	0.63	0.47-0.71
Group 2: All cases (n = 98), 11 readers	0.51	0.30-0.72	0.61	0.34-0.87

Results are shown with and without the presence of the glenoid vault model.

compared to readers overall, but they did consistently perform better with A2 cases. For A2 cases, "good" readers showed substantial agreement both with (0.66, 95% CI: 0.45, 0.88) and without (0.65, 95% CI: 0.45, 0.85) the glenoid vault model present, compared with only moderate agreement for readers overall (Table IV). When reducing comparisons of interobserver reliability to the major modified Walch classification groups (A, B, C, D), the Gwet AC1 values did not substantially improve from the comparisons across all of the modified Walch categories (A1, A2, B1, B2, B3, C1, C2, D), with interobserver reliability ranging from fair to moderate agreement (0.21-0.60) in similar patterns by major group (Table V).

Cases not assigned a consistent modified Walch type across readers were qualitatively reviewed for factors related to low reliability. The lack of definitive thresholds for central and/or anterior glenoid bone loss created difficulty in distinguishing transitions between A1, A2, and D glenoids in certain cases. Similarly, the lack of definitive thresholds for central and/or posterior glenoid bone loss created difficulty in distinguishing transitions between A1, A2, B1, B2, and B3 glenoids in certain cases. When looking specifically at B2 and B3 glenoids, the absence of a defined cutoff for the minimum remaining paleoglenoid length that constitutes a B2 glenoid created difficulty in distinguishing transitions between B2 and B3 glenoids in certain cases. Finally, the lack of definitive criteria for the definition of glenoid dysplasia created difficulty in distinguishing B2, B3, C1, and C2 glenoids in certain cases.

Discussion

The purpose of the current study was to assess the inter- and intraobserver reliability of the modified Walch classification using 3D CT imaging in a large multicenter research group (ASES B2 Glenoid Multicenter Research Group), including assessment of a subset of cases with posterior glenoid deformity where controversy on the optimal surgical treatment may exist. In addition to calculating overall inter- and intraobserver reliability, the study evaluated which pathologic features or modified Walch types may lower reliability, and if use of the glenoid vault model improved the inter- or intraobserver reliability of the modified Walch classification. We found that inter- and intraobserver reliability of the modified Walch classification were fair to moderate (kappa = 0.37-0.43) and moderate to substantial (kappa = 0.51-0.63), respectively, when using standardized 3D CT imaging analysis in this large multicenter study (23 readers across 19 institutions), both when evaluating the classification as a whole and when looking at a subset of cases associated with posterior glenoid deformity. Slightly lower interobserver reliability and lower overall intraobserver reliability were seen in group 2, representing only modified Walch categories associated with posterior glenoid deformity (B2, B3, C1, C2), when compared to group 1, which included all categories of the modified Walch classification. The intraobserver reliability of the individual readers also varied over a wider range in the case group representing only cases with posterior glenoid deformity (group 2). Although the presence of the

Table IV Interobserver reliability by Gwet AC1 method with 95% confidence intervals (95% CI) when all group 1 and 2 cases were combined and assessed by modified Walch classification type

Modified Walch classification type	Glenoid vault model absent		Glenoid vault model present	
	All readers (95% CI)	Readers ≥0.6 (95% CI)	All readers (95% CI)	Readers ≥0.6 (95% CI)
A1	0.44 (0.30, 0.57)	0.52 (0.29, 0.75)	0.43 (0.35, 0.52)	0.46 (0.23, 0.69)
A2	0.43 (0.30, 0.55)	0.65 (045, 0.85)	0.49 (0.37, 0.61)	0.66 (0.45, 0.88)
B1	0.35 (0.24, 0.45)	0.42 (0.20, 0.64)	0.25 (0.16, 0.35)	0.28 (0.07, 0.49)
B2	0.61 (0.51, 0.70)	0.60 (0.51, 0.70)	0.54 (0.45, 0.64)	0.54 (0.45, 0.64)
B3	0.49 (0.42, 0.56)	0.59 (0.50, 0.68)	0.56 (0.47, 0.64)	0.64 (0.54, 0.74)
C1	0.53 (0.44, 0.63)	0.52 (0.36, 0.67)	0.37 (0.20, 0.53)	0.33 (0.15, 0.51)
C2	0.39 (0.30, 0.47)	0.34 (0.22, 0.46)	0.35 (0.25, 0.45)	0.32 (0.22, 0.42)
D	0.25 (0.08, 0.42)	0.22 (−0.008, 0.45)	0.49 (0.15, 0.84)	0.53 (0.042, 1)

CI, confidence interval.
Results are shown with and without the presence of the glenoid vault model, as well as for all readers and those with intrarater reliability by Cohen kappa statistic of 0.6 or higher.

Table V Interobserver reliability by Gwet AC1 method with 95% confidence intervals (95% CI) when all group 1 and 2 cases were combined and assessed by modified Walch classification major groups

Modified Walch classification major groups	Glenoid vault model absent		Glenoid vault model present	
	All readers (95% CI)	Readers ≥0.6 (95% CI)	All readers (95% CI)	Readers ≥0.6 (95% CI)
A	0.43 (0.35, 0.52)	0.60 (0.48, 0.73)	0.46 (0.39, 0.53)	0.59 (0.46, 0.72)
B	0.52 (0.46, 0.57)	0.57 (0.50, 0.63)	0.50 (0.44, 0.56)	0.54 (0.48, 0.61)
C	0.44 (0.37, 0.50)	0.42 (0.33, 0.51)	0.36 (0.29, 0.44)	0.33 (0.26, 0.41)
D	0.25 (0.08, 0.42)	0.22 (−0.01, 0.45)	0.49 (0.15, 0.84)	0.53 (0.04, 1)

CI, confidence interval.
Results are shown with and without the presence of the glenoid vault model, as well as for all readers and those with intrarater reliability by Cohen kappa statistic of 0.6 or higher.

glenoid vault model did not improve interobserver reliability in groups 1 and 2, it did improve intraobserver reliability in group 2 from a level of moderate agreement to a level of substantial agreement.

These findings potentially suggest that cases with a spectrum of posterior glenoid bone loss and/or dysplasia can be harder to distinguish from one another by modified Walch type because of a lack of defined thresholds, including cutoffs for glenoid bone loss (joint line medialization) glenoid version, humeral head subluxation, and the amount of remaining paleoglenoid that still constitutes a B2 glenoid. The glenoid vault model may help with interpreting such thresholds. Iannotti et al[10] developed a decision tree algorithm in their study of the modified Walch classification based on 4 preoperative 3D CT measures: (1) presence of glenoid biconcavity, (2) joint line medialization as defined by the glenoid vault model, (3) pathologic glenoid version, and (4) premorbid glenoid version as defined by the glenoid vault model. The decision tree algorithm was found to be 87.1% accurate in classifying cases by modified Walch type by defining cutoff thresholds for each of these measures to maximize algorithm accuracy. The current study did not find that the presence of the glenoid vault model universally improved reliability of the modified Walch classification, including when looking at the case groups separately and when combining all of the cases to evaluate by each modified Walch type. However, the glenoid vault model was beneficial in certain scenarios. As noted above, intraobserver reliability in group 2 improved from moderate to substantial agreement with the glenoid vault model present. In addition, D glenoids had the most improvement in interobserver reliability when the glenoid vault model was present, likely because of the ability to better appreciate anterior glenoid bone loss and glenoid anteversion. The vault model may also help with distinguishing transitions between A1, A2, B1, B2, and B3 glenoids in certain cases by allowing a better appreciation of central and/or posterior glenoid bone loss, as well as in distinguishing B2, B3, C1, and C2 glenoids in certain cases by demonstrating normal or abnormal premorbid glenoid version. However, this was not formally evaluated in this study, and future work is needed to quantitatively assess this.

The individual cases with the lowest reliability in classifying by modified Walch type in this study were not always those with severe pathology, however, but were often cases with milder pathology, where the transition or distinction between one modified Walch type and another may be subtle. This finding highlights that transition points between modified Walch types can be difficult to distinguish with both milder and more severe pathology. The lack of improvement in interobserver reliability when comparisons were reduced to the major modified Walch classification groups (A, B, C, D) also highlights that transition points between modified Walch types can be difficult to distinguish both within and across these major groups. Interestingly, B2 glenoids had the highest interobserver reliability of all of the modified Walch types in the current study. This finding may relate to the characteristic glenoid biconcavity that defines the B2 glenoid and is consistently present even if the other pathologic features of a B2 glenoid can vary from mild to severe.

Group consensus methods to define thresholds or more automated analysis techniques may be needed to better differentiate transition points between Walch types and improve the reliability of the modified Walch classification when applied to individual cases. Even when "good" readers, as defined by those with a kappa value of 0.6 or higher for intraobserver reliability, were analyzed separately in this study, there was no better agreement level on interobserver reliability when compared to readers overall in nearly all comparisons. This finding suggests that even internally consistent readers are applying criteria to select a particular modified Walch type in different ways when compared to each other and supports the need to clarify the definitions of the modified Walch classification to improve reliability. The ability to reproducibly separate patients into groups based on preoperative pathology, including modified Walch type, is important for future studies to accurately evaluate postoperative outcomes in TSA patient cohorts. It remains unknown, however, whether lack of agreement with regard to Walch classification is associated with a lack of agreement with regard to treatment decision making. Surgical decision making in cases with milder pathology, for example, may be less controversial even if the modified Walch classification does not have universal agreement. In contrast, a moderate-to-severe B2 glenoid has a characteristic appearance that may be easily classified but may have a wide variation on the recommended surgical treatment based on patient age, activity level, goals of pathologic correction, etc.

The inter- and intraobserver reliability both with the original and modified Walch classification has been variable across prior studies. When using the original classification, Scalise et al[22] showed fair inter- and intraobserver agreement in CT analysis of 23 patients by 4 experienced shoulder surgeons with kappa statistics of 0.37 and 0.37, respectively. Nowak et al[18] showed moderate interobserver (kappa = 0.508) and substantial intraobserver (kappa = 0.611) agreement using CT scans of 26 patients classified by 3 attending shoulder surgeons and 5 shoulder/sports medicine–trained fellows. When Bercik et al[1] first modified the Walch classification, they found that interobserver reliability improved from fair agreement (kappa = 0.391) with the original classification to substantial agreement (kappa = 0.703) using the modified classification, whereas intraobserver reliability improved from moderate agreement (kappa = 0.605) with the original classification to nearly perfect agreement (kappa = 0.882) using the modified classification in 3D CT analysis of 129 patients by 3 fellowship-trained shoulder surgeons. Shukla et al[25] also recently evaluated the modified Walch classification, comparing axillary plain radiographs to CT for reliability. In 100 shoulders with glenohumeral OA, the mean intra- and interobserver agreements were kappa = 0.73 and kappa = 0.55 for plain radiographs and kappa = 0.72 and kappa = 0.52 for CT scans, respectively. However, the agreement for the 3 fellowship-trained shoulder surgeons between CT and radiographs after the first read was only 35 of 60 (58%).[25] Although the current study shows lower inter- and intraobserver reliability than other recent studies, a strength of this study is the large number of readers (all experienced shoulder surgeons) that participated from a large number of institutions. These results may, therefore, be more generalizable for clinical practice. Another strength of the study is the use of 3D CT imaging analysis, in which the 2D axial and coronal images were corrected to the plane of the scapula. This method has been shown to provide more accurate assessment of glenoid morphology and pathology and humeral head alignment compared with plain radiographs and uncorrected 2D CT imaging,[1-3,15,23] and 3D CT imaging analysis software is now widely commercially available for preoperative planning in TSA.[1,13,14]

There were several limitations to the current study. First, there is no gold standard of determining the correct modified Walch classification for a given case. Two of the study authors initially classified all of the cases by consensus using the criteria described by Bercik et al[1] and Iannotti et al,[10] and it is possible that they misclassified cases compared to the readers in each case group. However, kappa statistics to determine inter- and intraobserver reliabilities for each case group are based only on agreement across or within readers, not on a correct choice. Second, the reasons for low reliability or agreement in modified Walch type for a given case were not extensively evaluated in the current study, including assessing the factors readers may have prioritized to select a particular Walch type for a given case or the association of certain pathologic CT measurements (glenoid version and bone loss, humeral head subluxation, etc.) with high or low reliability across cases. Readers may also have varying abilities to visually interpret CT images, including differing ability to interpret 2D vs. 3D images, that were not evaluated. For example, "good" readers in this study showed higher interobserver reliability for selecting A2 cases than readers overall, a difference that may in part be related to an ability to better visually distinguish bone loss in these cases. Further understanding the reasons for low reliability in certain cases is a topic of future study and may help to refine the definitions of the modified Walch classification to improve reliability. Third, the study did not evaluate surgical decision making across readers. As discussed above, it remains unknown whether lack of agreement with regard to modified Walch classification is associated with a lack of agreement with regard to treatment choice. This is also a topic of future study.

Conclusions

Inter- and intraobserver reliability of the modified Walch classification were fair to moderate and moderate to substantial, respectively, using standardized 3D CT imaging analysis in a large multicenter study, both when evaluating the classification as a whole and when looking at a subset of cases associated with posterior glenoid deformity. Future studies should seek to further refine thresholds or criteria of the modified Walch types to improve the reliability of the modified Walch classification when applied to individual cases. In addition, the impact of such classification as it relates to treatment selection should be further evaluated to determine the critical thresholds between Walch types. The ability to reproducibly separate patients into groups based on preoperative pathology, including modified Walch type, is important for future studies to accurately evaluate postoperative outcomes in TSA patient cohorts. This may also help to determine the importance of quantitative measures of preoperative pathology compared to the Walch classification on postoperative outcomes.

Acknowledgment

The authors thank Catherine Shemo, BS, for her assistance with IRB approval and research compliance.

Disclaimer

The 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. Bercik MJ, Kruse K 2nd, 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. https://doi.org/10.1016/j.jse.2016.03.010
2. Bryce CD, Davison AC, Lewis GS, Wang L, Flemming DJ, Armstrong AD. Two-dimensional glenoid version measurements vary with coronal and sagittal scapular rotation. J Bone Joint Surg Am 2010;92:692-9. https://doi.org/10.2106/JBJS.I.00177
3. 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;20:577-83. https://doi.org/10.1016/j.jse.2010.11.003
4. Codsi MJ, Bennetts C, Gordiev K, Boeck DM, Kwon Y, Brems J, et al. Normal glenoid vault anatomy and validation of a novel glenoid implant shape. J Shoulder Elbow Surg 2008;17:471-8. https://doi.org/10.1016/j.jse.2007.08.010
5. Denard PJ, Walch G. Current concepts in the surgical management of primary glenohumeral arthritis with a biconcave glenoid. J Shoulder Elbow Surg 2013;22:1589-98. https://doi.org/10.1016/j.jse.2013.06.017
6. Donohue KW, Ricchetti ET, Ho JC, Iannotti JP. The association between rotator cuff muscle fatty infiltration and glenoid morphology in glenohumeral osteoarthritis. J Bone Joint Surg Am 2018;100:381-7. https://doi.org/10.2106/JBJS.17.00232
7. Donohue KW, Ricchetti ET, Iannotti JP. Surgical management of the biconcave (B2) glenoid. Curr Rev Musculoskelet Med 2016;9:30-9. https://doi.org/10.1007/s12178-016-9315-1
8. Ganapathi A, McCarron JA, Chen X, Iannotti JP. Predicting normal glenoid version from the pathologic scapula: a comparison of 4 methods in 2- and 3-dimensional models. J Shoulder Elbow Surg 2011;20:234-44. https://doi.org/10.1016/j.jse.2010.05.024
9. Gwet KL. Handbook of Inter-Rater Reliability. Gaithersburg, MD: Advanced Analytics, LLC; 2014.
10. Iannotti JP, Jun BJ, Patterson TE, Ricchetti ET. Quantitative measurement of osseous pathology in advanced glenohumeral osteoarthritis. J Bone Joint Surg Am 2017;99:1460-8. https://doi.org/10.2106/JBJS.16.00869
11. Iannotti JP, Norris TR. Influence of preoperative factors on outcome of shoulder arthroplasty for glenohumeral osteoarthritis. J Bone Joint Surg Am 2003;85:251-8.
12. Iannotti JP, Ricchetti ET, Rodriguez EJ, Bryan JA. Development and validation of a new method of 3-dimensional assessment of glenoid and humeral component position after total shoulder arthroplasty. J Shoulder Elbow Surg 2013;22:1413-22. https://doi.org/10.1016/j.jse.2013.01.005
13. Iannotti JP, Walker K, Rodriguez E, Patterson TE, Jun BJ, Ricchetti ET. Accuracy of 3-dimensional planning, implant templating, and patient-specific instrumentation in anatomic total shoulder arthroplasty. J Bone Joint Surg Am 2019;101:446-57. https://doi.org/10.2106/JBJS.17.01614
14. Iannotti JP, Weiner S, Rodriguez E, Subhas N, Patterson TE, Jun BJ, et al. Three-dimensional imaging and templating improve glenoid implant positioning. J Bone Joint Surg Am 2015;97:651-8. https://doi.org/10.2106/JBJS.N.00493
15. Kwon YW, Powell KA, Yum JK, Brems JJ, Iannotti JP. Use of three-dimensional computed tomography for the analysis of the glenoid anatomy. J Shoulder Elbow Surg 2005;14:85-90. https://doi.org/10.1016/j.jse.2004.04.011
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
17. Mizuno N, Denard PJ, Raiss P, Walch G. Reverse total shoulder arthroplasty for primary glenohumeral osteoarthritis in patients with a biconcave glenoid. J Bone Joint Surg Am 2013;95:1297-304. https://doi.org/10.2106/JBJS.L.00820
18. 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. https://doi.org/10.1016/j.jse.2009.08.003
19. Ricchetti ET, Hendel MD, Collins DN, Iannotti JP. Is premorbid glenoid anatomy altered in patients with glenohumeral osteoarthritis? Clin Orthop Relat Res 2013;471:2932-9. https://doi.org/10.1007/s11999-013-3069-5

20. Ricchetti ET, Jun BJ, Cain RA, Youderian A, Rodriguez EJ, Kusin D, et al. Sequential 3-dimensional computed tomography analysis of implant position following total shoulder arthroplasty. J Shoulder Elbow Surg 2018;27:983-92. https://doi.org/10.1016/j.jse.2017.12.012
21. Scalise JJ, Bryan J, Polster J, Brems JJ, Iannotti JP. Quantitative analysis of glenoid bone loss in osteoarthritis using three-dimensional computed tomography scans. J Shoulder Elbow Surg 2008;17:328-35. https://doi.org/10.1016/j.jse.2007.07.013
22. 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. https://doi.org/10.1016/j.jse.2007.12.006
23. 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;90:2438-45. https://doi.org/10.2106/JBJS.G.01341
24. Scalise JJ, Codsi MJ, Bryan J, Iannotti JP. The three-dimensional glenoid vault model can estimate normal glenoid version in osteoarthritis. J Shoulder Elbow Surg 2008;17:487-91. https://doi.org/10.1016/j.jse.2007.09.006
25. Shukla DR, McLaughlin RJ, Lee J, Cofield RH, Sperling JW, Sanchez-Sotelo J. Intraobserver and interobserver reliability of the modified Walch classification using radiographs and computed tomography. J Shoulder Elbow Surg 2019;28:625-30. https://doi.org/10.1016/j.jse.2018.09.021
26. Walch G, Badet R, Boulahia A, Khoury A. Morphologic study of the glenoid in primary glenohumeral osteoarthritis. J Arthroplasty 1999;14:756-60.
27. 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. https://doi.org/10.1016/j.jse.2011.11.030
28. Walker KE, Simcock XC, Jun BJ, Iannotti JP, Ricchetti ET. Progression of glenoid morphology in glenohumeral osteoarthritis. J Bone Joint Surg Am 2018;100:49-56. https://doi.org/10.2106/JBJS.17.00064