ARTICLE IN PRESS

J Shoulder Elbow Surg (2017) ■■, ■■–■■

ELSEVIER

Journal of Shoulder and Elbow Surgery

www.elsevier.com/locate/ymse

ORIGINAL ARTICLE

Radiographs and computed tomography scans show similar observer agreement when classifying glenoid morphology in glenohumeral arthritis

Jessica G. Aronowitz, MD, W. Scott Harmsen, MS, Cathy D. Schleck, BS, John W. Sperling, MD, MBA, Robert H. Cofield, MD, Joaquin Sánchez-Sotelo, MD, PhD*

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

Background: Glenohumeral subluxation and glenoid morphology are commonly evaluated in primary osteoarthritis by use of the Walch classification. The reliability of this classification system has been analyzed only by computed tomography (CT). The purpose of this study was to determine the reliability of plain axillary radiographs compared with CT scans.

Methods: Three shoulder surgeons blindly and independently evaluated the radiographs and CT scans of 75 consecutive shoulders with primary glenohumeral osteoarthritis. Each observer classified all shoulders according to Walch in 4 separate sessions, each 6 weeks apart. There were 2 sessions using only radiographs and 2 using only CT scans. The order of shoulders evaluated was randomized.

Results: The first reading by the most senior observer based on CT was arbitrarily used as the "gold standard" (A1, 21; A2, 13; B1, 12; B2, 28; C, 1). The average intraobserver agreement for radiographs was 0.66 (substantial; 0.66, 0.59, and 0.74 for each observer). The average intraobserver agreement for CT scans was 0.60 (moderate; 0.53, 0.61, and 0.65). Pairwise comparisons between observers showed higher agreement for radiographs than for CT scans (0.48 vs. 0.39). The average agreement for observations on radiographs and CT scans was 0.42 (moderate; 0.40, 0.37, and 0.50).

Conclusion: In this study, intraobserver agreement using the Walch classification based on axillary radiographs was substantial and compared favorably with agreement based on CT scans. The Walch classification provides a useful frame of reference when assessing subluxation and glenoid morphology in primary glenohumeral osteoarthritis, but not unlike other classification systems, it does not allow perfect agreement among observers.

Level of evidence: Level III; Diagnostic Study

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

Approval for this study was provided by the Mayo Clinic Institutional Review Board: 12-008023.

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

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

Classification systems are widely used in orthopedic surgery to characterize the nature of a problem, to guide clinical decision-making, and to allow comparison between patients and the outcome of various treatment options. Ideally, a useful classification system should be easy to understand and to apply,

1058-2746/$ - see front matter © 2017 Journal of Shoulder and Elbow Surgery Board of Trustees. All rights reserved.
http://dx.doi.org/10.1016/j.jse.2017.02.015

be of practical clinical usefulness, and be both valid and reliable.[1]

The original Walch classification was introduced in 1999 to describe arthritic glenoid morphology.[10] This system is widely used by shoulder surgeons and was developed on the basis of axial computed tomography (CT) scans to classify glenoid version, glenoid bone loss, and humeral head subluxation. In their article, Walch et al[10] reported an overall substantial intraobserver and interobserver agreement with κ values ranging from 0.65 to 0.70. These data were based on the readings of 2 observers analyzing 113 shoulders using the 3 main categories (A, B, and C) and not the subdivided categories (A1, A2, B1, B2, and C) that are more commonly used.

Two additional studies have analyzed the reliability of the Walch classification based on CT scans. In 2008, Scalise et al[6] evaluated the Walch classification using the CT scans of 23 patients (24 shoulders) and 4 experienced orthopedic shoulder surgeons as the observers. They found that the overall interobserver agreement and intraobserver agreement in this study were both fair, with κ values of 0.37 and 0.34, respectively. In 2010, Nowak et al[3] described their evaluation of the Walch classification using the CT scans of 23 patients (26 shoulders). The observers in this study included 3 attending shoulder surgeons and 5 shoulder/sports medicine fellows. They found that the overall interobserver agreement was moderate ($\kappa = 0.508$) and that the intraobserver reproducibility was substantial ($\kappa = 0.611$).

These previous studies analyzed the reliability of the classification system as originally described using only CT scans. The purposes of this study were to determine the reliability of reading the 5 categories of the Walch classification system using CT scans and axillary radiographs, to compare the readings of each observer between CT scans and radiographs, and to assess the CT scan vs. radiograph readings of each of the observers for the subcategories of bone loss and eccentric wear/subluxation.

Materials and methods

Seventy consecutive patients (75 shoulders) with the diagnosis of primary glenohumeral osteoarthritis who underwent total shoulder arthroplasty between June 2011 and August 2012 were selected for the study. All patients had CT scans and axillary radiographs performed as part of their routine preoperative evaluation. There were 40 men (57%) and 30 women (43%). The average age of the patients was 71 years (range, 54-86 years). Thirty-seven (49%) shoulders were right and 38 (51%) shoulders were left.

A total of 5 axial images from the CT scans for each patient were selected by the lead author of the study, including the midaxial cut and 4 images equally spaced above and below the center of the glenoid. The axillary radiograph obtained during the patient's preoperative evaluation was also used.

Three highly experienced shoulder surgeons independently evaluated each shoulder. All evaluators were blinded to readings of the others as well as to any patient information. All observers were familiar with the Walch classification and were each provided a description (Table I) and a pictorial representation of the classification system as a reference with each reading (Fig. 1). There were no time limitations imposed on any of the readings. There were 4 separate readings: 2 using only radiographs and 2 using only CT scans. The observers first evaluated radiographs, followed by CT scans, then radiographs again, and last CT scans a second time. Each reading was separated by at least 6 weeks, and the order of images was randomized for each reading to minimize any recall bias.

Table I Walch classification

Walch category	Description
A	Humeral head centered on the glenoid fossa
Subgroup A1	No or minor central erosion
Subgroup A2	Major central erosion
B	Posterior subluxation of the humeral head
Subgroup B1	No posterior bone loss
Subgroup B2	Posterior bone loss resulting in a biconcave glenoid
C	Glenoid retroversion >25°

The intraobserver reliability was determined by comparison of the classification of each subject by the observers for both the axillary radiographs and the CT scans. Pairwise comparisons between each observer were also performed to determine interobserver reliability. The κ values were calculated for both interobserver and intraobserver reliability. The κ value adjusts for the proportion of agreement among observers that could have occurred by chance. Landis and Koch[2] 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, and 1.00 indicates perfect agreement. Additional analysis was performed to compare the initial reading of radiographs and CT scans of each observer.

Results

The first CT scan reading performed by the most senior observer was arbitrarily used as the "gold standard" to define the sample analyzed. Of the 75 observations in this reading, the most senior observer classified 21 shoulders as type A1 (28%), 13 as type A2 (17.3%), 12 as type B1 (16%), 28 as type B2 (37.3%), and 1 as type C (1.3%).

Agreement for each observer can be found in Table II. The average intraobserver agreement for the axillary radiographs was substantial ($\kappa = 0.66$). The average intraobserver agreement for the CT scans was moderate ($\kappa = 0.60$).

The interobserver agreement between each physician for both radiographs and CT scans is shown in Table III. The average interobserver agreement for the axillary radiographs was moderate ($\kappa = 0.48$). The average interobserver agreement for the CT scans was fair ($\kappa = 0.39$).

Table IV shows the first CT scan and radiograph readings of each of the 3 observers. The diagonal from top left to bottom right indicates the agreement between CT scans and radiographs. Observers 1, 2, and 3 demonstrated agreement in 35 cases (46.7%), 35 cases (46.7%), and 43 cases (57.3%), respectively. In total, there was agreement among