Intra- and interobserver reliability of classification scores in calcific tendinitis using plain radiographs and CT scans

Markus Maier, Johanna Schmidt-Ramsin, Christian Glaser, Anne Kunz, Helmut Küchenhoff, Thomas Tischer

From the Ludwig-Maximillians University, Munich, Germany and the Technical University, Munich, Germany

Calcific tendinitis is a common cause of shoulder pain ; its treatment is based on the stage of the disease. Unfortunately the existing classification based on radiographs has low reliability. The aim of this study was to evaluate the contribution of CT scans to this matter. Fifty-six consecutive patients with calcific tendinitis were included in this study. Radiographs and CT scans were taken and were evaluated independently by six observers and classified according to the following systems : Gärtner, De Palma, Patte and Mole. After four months the same observers repeated their evaluation. The inter- and intraobserver reliability was calculated using Cohen's kappa analysis. The intraobserver reliability was sufficient using both CT and radiographs. Minimally better (not statistically significant) results were found for CT scans, especially for the Gärtner classification. Interobserver reliability was also better with CT scans but most of the time still hardly satisfactory. The classification scores for calcific tendinitis show insufficient reliability and reproducibility. This can be improved somewhat when using CT scans, but still remains unsatisfactory.

Keywords : calcific tendonitis ; classification ; Gärtner ; De Palma ; Patte ; Mole.

INTRODUCTION

Calcific tendinitis is a common cause of shoulder pain in the middle aged population, usually self-limiting and presenting at different stages *(9,10,21,*

22). Its pathogenesis is still discussed and various treatment options exist, depending on the stage of the disease : conservative treatment, needling *(5)*, shock wave application *(1)* or surgical removal (open versus arthroscopic) *(19)* of the calcific deposit, mostly from within the supraspinatus tendon. The progression of the disease can be divided into three main phases : the first stage is the pre-calcific stage (stage I), where chondroid metaplasia takes place. This is usually asymptomatic. Next, during the calcific stage (stage II), calcium is being deposited in the extracellular matrix and calcific

■ Markus Maier, MD, Professor, Consultant.

■ Johanna Schmidt-Ramsin, MD, Resident.

Department of Orthopaedic Surgery, Ludwig-Maximillians Universität, München, Germany.

■ Anne Kunz, MD, Resident.

■ Helmut Küchenhoff, MD, Professor, Consultant.

Statistical Consulting Unit, Department of Statistics, Ludwig-Maximillians Universität, München, Germany.

■ Christian Glaser, MD, Consultant.

Department of Radiology, Ludwig-Maximillians Universität, München, Germany.

■ Thomas Tischer, MD, Fellow.

Department of Orthopaedic Surgery, Technical University, München, Germany.

Correspondence : Thomas Tischer, Department of Orthopaedic Surgery, Technical University of Munich, Ismaninger Str. 22, D-81675 Munich, Germany.

E-mail : thomas.tischer@gmx.net

deposits are being formed (formation phase), replacing the fibrocartilage tissue. After a certain time period, the process turns into a resting phase and patients are free of pain. On radiographs, the deposits are now usually well demarcated. After a while, the resorption of the deposits is triggered by some unknown mechanism (resorption phase), which can be very painful to the patients. The calcific deposits appear rather blurred on radiographs. In the postcalcific stage (stage III), the defects are filled with granulation tissue and remodelling of the tendon takes place (22).

Different classification systems can be used to assess the radiological appearance of the calcific deposits. Reliable scores are important when deciding on patient therapy because calcific deposits are treated differently according to their actual stage of disease (22). Furthermore, the scores are important to precisely evaluate the outcome of treatment and compare different studies. The two most widely used classification scores for the appearance of the calcific deposits on radiographs are the De Palma and Gärtner classifications (3,6,14). It was shown recently that neither intra- nor interobserver reliability is acceptable using the De Palma, Gärtner, Patte or Mole classification on X-ray images (12,13). Therefore, other imaging modalities like ultrasound and MRI were evaluated for their use to classify calcific deposits, but they could not reach widespread clinical use so far (2,23). CT scans were reported to be more reliable than plain radiographs to predict the consistency of the deposit, but no further information was given about the inter- or intraobserver agreement (4). Because of this lack of reliable classification systems, recent studies sometimes do not apply radiological scores anymore (1,18), making it difficult to compare the outcomes between studies and to evaluate new treatment options.

Therefore the aim of this study was to investigate whether a more precise classification of the calcific deposits can be achieved using CT scans. The intra- and interobserver reliability of the De Palma, Gärtner, Patte and Mole classifications were evaluated by six independent observers grading radiographs and CT images of calcific deposits according to these classifications and repeating this process again a few months later.

MATERIAL AND METHODS

Patient selection

Fifty-six consecutive patients (37 women, 19 men ; mean age : 50.8 years ; range : 31 to 82) who presented with calcific tendinitis between 06/2000 and 09/2001 were included. Inclusion criteria were the presence of only one radiologically discernible calcification greater than 5 mm in diameter in the supraspinatus tendon. Patients were only included if they had not received a corticosteroid injection in the last 12 weeks before presentation. Radiographs in three planes (true AP, axial, scapula-Y) and axial CT scans with 1mm slice thickness were obtained of each shoulder. The images were randomised and numbered.

Classification

Two weeks before classification, the observers (four experienced orthopaedic surgeons and two radiologists) received the original articles explaining the various classification systems. The following classifications were used :

1. Gärtner : type I (well demarcated, dense deposit), type II (soft contour and dense deposit or sharp contours and transparent deposit) and type III (soft contours and transparent) (6).
2. De Palma : type I (fluffy and amorphous) and type II (defined and homogeneous) (3).
3. Patte : type I (localised and homogeneous) and type II (diffuse, disseminated and heterogeneous) (17).
4. Mole : type A (dense, homogeneous, sharp contours), type B (dense, cloudy, sharp contours), type C (inhomogeneous, soft contours) and type D (dystrophic calcifications of the insertion zone of the rotator cuff tendons) (15).

The time for making the diagnosis was not limited. However, once the next film was presented a decision could not be changed. Radiographs and CT scans were analysed independently in a randomised order by six independent observers. This was replicated 16 to 18 weeks later. The observers were not informed of any results until the study was completed.

Statistical analysis

Statistical analysis was carried out based on Cohen's kappa statistics, in order to determine observer variability in the interpretation of morphological findings (16,20).

The weighted kappa statistics using a quadratic weighting scheme were calculated for each situation. Intra rater kappa was calculated for the two measurements for each rater. Inter rater kappa was determined for all pairs of raters. Means and standard deviation are reported for the different possible pairs of observers. For comparing different methods paired t-tests were used. To achieve normality, the kappa values had to be transformed by the Fisher transformation [20]. The data analysis for this publication was generated using SAS/STAT software, version 9.1 of the SAS System (SAS Institute Inc.). Interpretation of kappa values was performed according to Landis and Koch [11] : insufficient (0.0 – 0.2), satisfactory (0.21 – 0.4), sufficient (0.41 – 0.6), good (0.61 – 0.8) and excellent (0.81 – 1.0) agreement.

RESULTS

Intraobserver reliability

CT

The intraobserver agreement of the different observers using the four classifications on CT scans was very low (table I). Sufficient agreement was only achieved using the Gärtner (mean kappa value 0.42 ± 0.17) classification. De Palma (mean kappa value 0.34 ± 0.11) and Mole (mean kappa value 0.23 ± 0.21) classifications were satisfactory and Patte (mean kappa value 0.18 ± 0.24) classification was insufficient. When comparing orthopaedic surgeons (n = 4) with radiologists (n = 2), orthopaedic surgeons achieved higher kappa values for Gärtner classification (0.50 ± 0.07 (ortho) and 0.26 ± 0.22 (radio)) and radiologists higher reliability for De Palma classification (0.33 ± 0.14 (ortho) and 0.36 ± 0.04 (radio)).

Radiographs

No statistically significant difference (p = 0.789) was noted between plain radiographs and CT with respect to reliability, although CT reliability was generally better with the Gärtner and De Palma classifications and worse with the Mole and Patte scores (table I). With evaluation of plain radiographs, orthopaedic surgeons achieved higher kappa values using De Palma's classification than radiologists (0.27 ± 0.12 vs. 0.17 ± 0.06), whereas

radiologists achieved nearly the same results using Gärtner's classification (0.36 ± 0.28 vs. 0.36 ± 0.22).

Interobserver reliability

CT

The interobserver agreement was low for CT scans and the best agreement was noted for Gärtner, followed by De Palma (table II). For De Palma, which only has two classes (Type I and II), the agreement between the different observers was assessed in detail : at the first point of time, the gradings of 25/56 (44.6%) deposits were in agreement between all six observers, versus 13/56 (23.2%) between five observers, 12/56 (21.2%) between four observers and 6/56 (10.7%) between three observers. At the second point of time, agreement was achieved in 14/56 (25%) cases between all six observers, between five in 18/56 (32.1%), between four in 13/56 (23.2%) and between three in 11/56 (19.6%).

Radiographs

The interobserver reliability in radiographs was worse when using the Gärtner (p = 0.014) classification (table II). In detail, for the De Palma classification, the agreement between the different observers was also calculated : at the first time point, 25/56 (44.6%) deposit gradings were in agreement between all six observers, 12/56 (21.2%) between five observers, 12/56 (21.2%) between four observers and 7/56 (12.5%) between three observers. At the second time point, agreement between all six observers was achieved in 6/56 (10.7%) cases, between five in 21/56 (37.5%), between four in 23/56 (41.1%) and between three in 6/56 (10.7%) cases.

DISCUSSION

In order to determine the best possible treatment for the patient, it is important to be able to reliably predict the consistency of the deposit and hereby the stage of the disease by characterizing the radiological image in one of the classification systems in

Table I. — Intraobserver reliability CT / radiograph : Kappa values for intraobserver agreement
for the different classification scores using CT and plain radiographs. The single values represent
the mean kappa values between the six different observers

	Classification	Observers [n]	Kappa Value [Mean ± SD]	Minimum	Maximum
CT	DePalma	6	0.34 ± 0.11	0.12	0.44
X-ray		6	0.24 ± 0.11	0.12	0.38
CT	Gärtner	6	0.42 ± 0.17	0.10	0.61
X-ray		6	0.36 ± 0.21	0.16	0.63
CT	Mole	6	0.23 ± 0.21	0.03	0.54
X-ray		6	0.34 ± 0.24	0.13	0.75
CT	Patte	6	0.18 ± 0.24	-0.17	0.51
X-ray		6	0.28 ± 0.13	0.07	0.48

Table II. — Interobserver reliability CT / radiograph : Kappa values of interobserver agreement
for the different classification scores using CT and radiographs. The single values represent the mean
of the 30 kappa values of the 15 pairs of different observers measured at the two points of time

	Classification	Observer [n]	Kappa Value [Mean ± SD]	Minimum	Maximum
CT	DePalma	30	0.32 ± 0.22	-0.08	1.00
X-ray		30	0.34 ± 0.27	-0.01	1.00
CT	Gärtner	30	0.40 ± 0.15	0.15	0.78
X-ray		30	0.33 ± 0.16	-0.05	0.68
CT	Mole	30	0.20 ± 0.15	-0.10	0.47
X-ray		30	0.18 ± 0.15	-0.09	0.54
CT	Patte	30	0.22 ± 0.15	-0.08	0.54
X-ray		30	0.24 ± 0.14	-0.07	0.51

clinical use at present. A good classification system further has two prerequisites : each observer should grade the calcific deposit identically at successive evaluations (intraobserver agreement). Second, identical classifications should be obtained by different observers (interobserver agreement).

For the investigation of the common systems' reliability, recent studies by Maier *et al* determined the intra- and interobserver agreement of the classification systems by Patte, Mole (*13*), and De Palma (*12*) using X-ray images of calcific deposits. The results from these studies show that the currently used classification systems based on X-ray images are far from satisfactory in terms of inter- and intraobserver reliability according to the Landis and Koch analysis as described above. Compared to our study they reported only slightly higher kappa values despite higher patient numbers. The De Palma classification showed the best results for intraobserver (kappa value 0.487) and interobserver reliability (kappa value 0.234) on radiographs. To improve the reliability, Farin *et al* (*4*) investigated the value of CT scans in terms of predicting the consistency of the calcifications as later identified by needling. In this study, CT scans were in better agreement with the needling results than ultrasound or conventional radiographs. However the patient number was very small (n = 20) and none of the commonly used classification systems was used. In order to investigate whether the reliability of a classification could be improved by more accurate images of the calcific deposits, our study was planned with both radiographs and CT scans taken of every patient with calcific tendinitis as described

in the material and methods section. As our results show, the reliability (both inter- and intraobserver) of classification systems was marginally higher using CT scans when compared to plain radiographs, especially regarding the Gärtner classification. The difference however is not statistically significant. With respect to our results, the Mole and Patte classifications cannot be recommended in terms of intra- and interobserver reliability, neither using CT nor plain radiographs. This is in some contrast to our previous work, where at least for the Patte classification a kappa value of 0.458 for intraobserver agreement and 0.4 for interobserver agreement was achieved (*13*).

Moreover CT scans expose patients to more radiation, are more time consuming and more expensive. Therefore we cannot recommend using CT scans for the classification of calcific deposits, based on our analysis of the classification systems according to the Landis and Koch analysis (*11*). Most likely, the applied method cannot be the reason for a lack of correlation between scores. There may be different reasons for the obstacles encountered. Most difficulties for an exact classification arise from the continuous progression of the disease. There is no clear cut step from one stage to another, neither clinically nor radiographically. Even the mineral composition does not change between acute and chronic calcific deposits (*7,8*). This suggests that the progression of the disease must be influenced by other mechanisms.

Nonetheless, for the best treatment option, the determination of the exact stage of the disease should be warranted. Unfortunately, until now, there is no classification system that can serve for this purpose. Even when using imaging modalities other than plain radiographs, no significant improvement in reliability of the different classification scores has been made until today. Studies based on currently used classification systems thus have to be interpreted with caution.

REFERENCES

1. **Albert JD, Meadeb J, Guggenbuhl P *et al.*** High-energy extracorporeal shock-wave therapy for calcifying tendinitis of the rotator cuff : A randomised trial. *J Bone Joint Surg* 2007 ; 89-B : 335-341.

2. **Chiou HJ, Chou YH, Wu JJ *et al.*** Evaluation of calcific tendonitis of the rotator cuff : role of color Doppler ultrasonography. *J Ultrasound Med* 2002 ; 21 : 289-295 ; quiz 96-97.

3. **De Palma AF, Kruper JS.** Long-term study of shoulder joints afflicted with and treated for calcific tendinitis. *Clin Orthop* 1961 ; 20 : 61-72.

4. **Farin PU.** Consistency of rotator-cuff calcifications. Observations on plain radiography, sonography, computed tomography, and at needle treatment. *Invest Radiol* 1996 ; 31 : 300-304.

5. **Farin PU, Rasanen H, Jaroma H, Harju A.** Rotator cuff calcifications : treatment with ultrasound-guided percutaneous needle aspiration and lavage. *Skeletal Radiol* 1996 ; 25 : 551-554.

6. **Gartner J, Simons B.** Analysis of calcific deposits in calcifying tendinitis. *Clin Orthop* 1990 ; 254 : 111-120.

7. **Hamada J, Ono W, Tamai K *et al.*** Analysis of calcium deposits in calcific periarthritis. *J Rheumatol* 2001 ; 28 : 809-813.

8. **Hamada J, Tamai K, Ono W, Saotome K.** Does the nature of deposited basic calcium phosphate crystals determine clinical course in calcific periarthritis of the shoulder ? *J Rheumatol* 2006 ; 33 : 326-332.

9. **Harvie P, Pollard TC, Carr AJ.** Calcific tendinitis : natural history and association with endocrine disorders. *J Shoulder Elbow Surg* 2007 ; 16 : 169-173.

10. **Hurt G, Baker CL Jr.** Calcific tendinitis of the shoulder. *Orthop Clin North Am* 2003 ; 34 : 567-575.

11. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. *Biometrics* 1977 ; 33 : 159-174.

12. **Maier M, Maier-Bosse T, Schulz CU *et al.*** Inter and intraobserver variability in DePalma's classification of shoulder calcific tendinitis. *J Rheumatol* 2003 ; 30 : 1029-1031.

13. **Maier M, Maier-Bosse T, Veihelmann A *et al.*** Observer variabilities of radiological classifications of calcified deposits in calcifying tendinitis of the shoulder. *Acta Orthop Belg* 2003 ; 69 : 222-225.

14. **McKendry RJ, Uhthoff HK, Sarkar K, Hyslop PS.** Calcifying tendinitis of the shoulder : prognostic value of clinical, histologic, and radiologic features in 57 surgically treated cases. *J Rheumatol* 1982 ; 9 : 75-80.

15. **Mole D, Kempf JF, Gleyze P *et al.*** Results of endoscopic treatment of non-broken tendinopathies of the rotator cuff. 2. Calcifications of the rotator cuff. *Rev Chir Orthop* 1993 ; 79 : 532-541.

16. **Ostergaard M, Klarlund M, Lassere M *et al.*** Interreader agreement in the assessment of magnetic resonance images of rheumatoid arthritis wrist and finger joints – an international multicenter study. *J Rheumatol* 2001 ; 28 : 1143-1150.

17. **Patte D, Goutallier D.** Periarthritis of the shoulder. Calcifications. *Rev Chir Orthop* 1988 ; 74 : 277-278.

18. **Sabeti-Aschraf M, Dorotka R, Goll A, Trieb K.** Extracorporeal shock wave therapy in the treatment of calcific tendinitis of the rotator cuff. *Am J Sports Med* 2005 ; 33 : 1365-1368.

19. **Seil R, Litzenburger H, Kohn D, Rupp S.** Arthroscopic treatment of chronically painful calcifying tendinitis of the supraspinatus tendon. *Arthroscopy* 2006 ; 22 : 521-527.

20. **Shouki M.** *Measures of Interobserver Agreement.* Chapman & Hall 2004.

21. **Speed CA, Hazleman BL.** Calcific tendinitis of the shoulder. *New Engl J Med* 1999 ; 340 : 1582-1584.

22. **Uhthoff HK, Loehr JF.** Calcifying Tendinitis, In : Rockwood CA *et al.* (eds). *The Shoulder.* Saunders, Philadelphia 1998 : pp 989-1008.

23. **Zubler C, Mengiardi B, Schmid MR *et al.*** MR arthrography in calcific tendinitis of the shoulder : diagnostic performance and pitfalls. *Eur Radiol* 2007 ; 17 : 1603-1610.