Inter and Intraobserver Variability in DePalma’s Classification of Shoulder Calcific Tendinitis

MARKUS MAIER, TAMARA MAIER-BOSSE, CHRISTOPH U. SCHULZ, MARC STEINBORN, CLEMMENS A. SCHMITT, MARTIN KLEEN, and AXEL STÄBLER

ABSTRACT. Objective. To determine the intraobserver reproducibility and interobserver reliability of DePalma and Kruer’s classification of calcific tendinitis of the rotator cuff. The result of this classification influences the choice of therapeutic procedures in patients with symptomatic calcific tendinitis.

Methods. Plain anteroposterior radiographs of shoulders from 100 patients with symptomatic calcified deposits of the rotator cuff were classified according to the criteria of DePalma and Kruer by 6 independent observers at 2 different time points within 4 months. The kappa values of intraobserver reproducibility and interobserver reliability were calculated.

Results. Kappa values of intraobserver reproducibility had a mean of 0.487 (SD 0.094); kappa values of interobserver reliability were 0.234 for the first test and 0.273 for the second test.

Conclusion. Determination of intraobserver reproducibility gave satisfactory to sufficient results and interobserver reliability was satisfactory for both tests indicating that studies based on the classification of DePalma and Kruer should be interpreted cautiously. (J Rheumatol 2003;30:1029–31)

Key Indexing Terms:
CALCIFIC TENDINITIS MORPHOLOGY CLASSIFICATION

MATERIALS AND METHODS

Patient selection. Out of a total of 124 anteroposterior digital radiographic images showing a single calcified deposit of the supraspinatus tendon of the rotator-cuff, 100 were randomly selected to be included in the study. The location of the calcified deposit in the supraspinatus tendon was confirmed by ultrasound. No calcified deposit displayed overlay with bone structures and all deposits were larger than 5 mm in diameter. The patients included 61 women and 39 men (mean-age 54 years, range 40 to 75 years, SD 7.5 years). Forty-nine calcified deposits were located in left shoulders and 51 in right shoulders. Deposits were classified according to DePalma and Kruer to distinguish between 2 morphologic types: type I, fluffy and amorphous; type II, defined and homogenous.

Classification. Two weeks before classification, all 6 observers (orthopedic surgeons and radiologists with different levels of clinical expertise and familiar with this classification) received a copy of the original article and were also allowed to use this article during the classification procedure. After removing all personal data, films were randomly numbered and presented to individual observers alone. Examination time per film was not restricted. Once a decision was made it could not be changed and the next film was presented. This rule was followed until the evaluation of the 100 deposits was completed by all observers. All deposits were classified for a second time after a latency of 16 to 17 weeks. In the meantime films were not available to the observers, and no results of the first classification were presented to individual observers alone. Examination time per film was not restricted. Once a decision was made it could not be changed and the next film was presented. This rule was followed until the evaluation of the 100 deposits was completed by all observers. All deposits were classified for a second time after a latency of 16 to 17 weeks. In the meantime films were not available to the observers, and no results of the first classification were released. The second classification was performed in an identical manner after films were randomized and renumbered.

Statistical analysis. Observer variability was determined by kappa statistics, established to determine observer variability in the interpretation of morphologic findings. For calculation, a proportion of possible agreements between observers that occurred by chance (p_c) is calculated from the proportion of possible agreements (p_p) and p_c is divided by 1 minus p_c (Formula 1).

For classification schemes using more than 2 levels, we used a quadratic weighted variation of kappa. For this, observations were weighted according to the extent of disagreement before calculation of kappa: the larger the disagreement, the more weight was assigned to a given
Calcific tendinitis of the shoulder is a self-limiting disease usually found in acute cases and indicate the stage of resorption only pain reduction and needling of the deposit become therapeutic alternatives. In contrast, during the stage of resorption only pain reduction and needling of the deposit are recommended11.

Due to the artificial test setting in our present study, the decision making process of the observers might differ from an actual clinical situation. Therefore, the present study has a number of limitations, even if the study design we used was established to test observer variability12-14: (1) the observers were aware that the interpretation of the radiography images had no direct clinical consequence for a patient’s treatment; (2) in a real clinical situation observers know about the clinical symptoms of each individual patient; (3) observers might have interpreted radiography images more freely and easily than in a clinical situation, and as a result, the study may have been biased towards a higher observer variability; or (4) conversely the observers might have assessed the radiographs more accurately than in a clinical situation. This could denote that observer variability would be better under more regular clinical conditions.

More classifications concerning the morphological aspect of calcifications of the rotator cuff have been described15-17. However, we decided to test the classification of DePalma and Kruper1 as all observers knew the original description of this classification and were familiar with its use in a routine clinical assessment. Those readers using the classification of DePalma and Kruper1 for a long time (i.e., consultants) achieved a smaller observer variability than observers in training. However, there is no obvious explanation why radiologists showed poorer observer variability than surgeons.

A morphologic classification of calcified deposits commonly used in clinical rheumatology and orthopedic surgery must comply with certain prerequisites: first, any given observer should classify identically the morphologic aspect of a deposit at various time-points (intraobserver reproducibility); second, identical classifications should be

### Table 1. Kappa values of intraobserver reproducibility.

<table>
<thead>
<tr>
<th>Observer</th>
<th>Kappa Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.440</td>
</tr>
<tr>
<td>B</td>
<td>0.570</td>
</tr>
<tr>
<td>C</td>
<td>0.566</td>
</tr>
<tr>
<td>D</td>
<td>0.322</td>
</tr>
<tr>
<td>E</td>
<td>0.512</td>
</tr>
<tr>
<td>F</td>
<td>0.511</td>
</tr>
</tbody>
</table>
obtained from different observers (interobserver reliability)\textsuperscript{19}. These prerequisites are not completely met by the classification of DePalma and Kruper\textsuperscript{1}, whereas former studies using morphologic classifications already pointed to that problem. Interobserver reliability was only sufficient when testing the classification of ankle fractures according to Lauge-Hansen (kappa index 0.50) and Weber (kappa index 0.57)\textsuperscript{19}. Neer’s classification of proximal humeral fractures achieved had only insufficient to satisfactory interobserver reliability (kappa indices ranging from 0.26 to 0.50)\textsuperscript{20-22}.

Our results could lead to the conclusion that the classification of DePalma and Kruper\textsuperscript{1} needs more precise criteria for reproducible and reliable classification of calcified deposits of the rotator-cuff tendons. This would be of great clinical use since their classification is in common clinical use and contributes to the decision to treat or not to treat patients with symptomatic calcific tendinitis of the shoulder, and also suggests the type of therapeutic procedure that should be applied\textsuperscript{3,4}. However, the value of a classification system must be questioned if its reproducibility and reliability are limited. Therefore, studies based on this classification should be interpreted cautiously.

REFERENCES