We investigated whether training doctors to classify proximal fractures of the humerus according to the Neer system could improve interobserver agreement. Fourteen doctors were randomised to two training sessions, or to no training, and asked to categorise 42 unselected pairs of plain radiographs of fractures of the proximal humerus according to the Neer system. The mean kappa difference between the training and control groups was 0.30 (95% CI 0.10 to 0.50, p = 0.006). In the training group the mean kappa value for interobserver variation improved from 0.27 (95% CI 0.24 to 0.31) to 0.62 (95% CI 0.57 to 0.67). The improvement was particularly notable for specialists in whom kappa increased from 0.30 (95% CI 0.23 to 0.37) to 0.79 (95% CI 0.70 to 0.88). These results suggest that formal training in the Neer system is a prerequisite for its use in clinical practice and research.

Received 14 November 2001; Accepted after revision 3 April 2002

Fractures of the proximal humerus account for approximately 4% of all fractures.1,2 The incidence is approximately 70 per 100 000 and will probably increase;3,4 the fractures are age- and osteoporosis-related.3-6 Several systems for classifying them have been proposed in an attempt to improve the description of the fractures and to provide guidelines for treatment and outcome.7-15 The Neer system was introduced in 197016 (Fig. 1) and has been the most widely used in clinical practice and research. It is based on the displacement and angulation of four anatomical parts of the proximal humerus, namely, the greater tuberosity, the lesser tuberosity, the articular segment and the shaft of the humerus (Fig. 2).

Within the last 15 years several studies have reported poor observer variation for the Neer system.17-25 Lack of consistency in classification is one possible reason for the discrepancies in outcome after treatment of complex fractures of the proximal humerus, especially of three- and four-part fractures.28

Previous studies have reported no improvement in observer variation using the Neer system despite the exclusive availability of radiographs of high quality,19,22 CT and three-dimensional reconstructions,22-25 a reduction in the number of classification units,19,22 and more formal training.19,22 We have been unable to find any randomised trial of the effect of training on the interobserver variation of the Neer system and have therefore undertaken such a trial.

Patients and Methods

Between August 1 and October 15, 1999 we identified 55 patients with a fracture of the proximal humerus who had been discharged from the Bispebjerg University Hospital (Copenhagen). We excluded 14 patients: four children, two patients with pathological fractures, one with pseudarthrosis, one with a healed fracture, three who had been miscoded, and three whose radiographs had been lost. The final study group consisted of 41 patients with a total of 42 pairs of radiographs. There was no selection according to the quality of the radiographs. Plain anteroposterior (AP) and lateral radiographs were available for all the fractures.

More than six months after the last patient had had radiography we invited the orthopaedic medical staff to a teaching session. They had a brief introduction before being given a diagram of the Neer classification system, a written definition of displacement and angulation, a ruler and a goniometer. They were asked to study each pair of radiographs and to choose one of the 16 possibilities on the diagram. They were informed that their observations would be reported anonymously. No communication between observers was allowed during the classification. There was no time limit for making decisions.
They were then randomly allocated to a training group and a control group. The training group was taught the Neer system for 45 minutes. This was based on the original reports of the system. Two weeks later, the training group received a second teaching session of 45 minutes, before all participating doctors were presented with the same pairs of radiographs, but in a different and unstructured order. None of the authors was an observer.

The classification data were then transferred to a statistical program.

**Statistical methods.** Within each group of observers the kappa value for interobserver variation were calculated for each possible pair of observers, before calculating the mean kappa value.  

Kappa statistics adjust simple observed agreement for chance agreement. The kappa coefficients range from 1 (perfect agreement) to 0 (chance agreement) and <0 (systematic disagreement). According to Landis and Koch values <0 indicate poor reliability, 0.00 to 0.20 slight reliability, 0.21 to 0.40 fair reliability, 0.41 to 0.60 moderate reliability, 0.61 to 0.80 substantial agreement and 0.81 to 1.00 excellent agreement.

Before calculating the kappa values the 16 groups of the Neer system were reduced to six as follows: one-part (undisplaced), two-part, three-part and four-part fractures, fracture-dislocations, and articular surface fractures.

We assessed the effect of training on interobserver variation by comparing the change in the mean kappa values between the two groups of observers. The level of statistical significance was assessed by an unpaired t-test applied to the differences in kappa values for each pair of observers. A predefined subgroup analysis was performed to investigate if any change in interobserver variation was related to whether the observers were specialists.

For all kappa values 95% confidence intervals (CI) were calculated. The prevalence of each Neer group was calculated in order to evaluate the practice of classification within the different levels of experience. The level of statistical significance of post-training differences in prevalences was assessed by a sign-rank test. Kappa values.
were calculated for agreement within each of the six Neer groups.

Results

Nineteen doctors attended the first session, were randomised, and carried out the baseline classification. There were ten in the training group and nine in the control group. In the second session 14 returned, eight specialists in orthopaedic surgery and six non-specialists. The five doctors who dropped out were two non-specialists and one specialist from the training group, and two non-specialists from the control group. The final distribution of observers was four specialists and three non-specialists in each group.

The mean kappa value at baseline was 0.27 (95% CI 0.23 to 0.31) in the training group and 0.28 (95% CI 0.24 to 0.31) in the control group (Table I). Thus, the two groups were initially comparable. The difference in the post-training mean kappa value between the two groups was 0.30 (95% CI 0.10 to 0.50, p = 0.006, df = 12). In the training group the overall mean kappa coefficient for pairwise agreement increased from 0.27 (95% CI 0.24 to 0.31) at baseline to 0.62 (95% CI 0.57 to 0.67) after training (Table I). The increase in agreement after teaching in the training group was approximately equally distributed among all pairs of observers, while agreement between all pairs of observers in the control group remained distributed around baseline kappa (Fig. 3). In the control group the overall mean kappa coefficient for pairwise agreement increased from 0.28 (95% CI 0.24 to 0.31) at baseline to 0.33 (95% CI 0.29 to 0.36) in the second classification session.

After teaching, the mean kappa value in the training group increased from 0.27 (95% CI 0.23 to 0.31) to 0.62 (95% CI 0.57 to 0.67) with a significant difference of 0.30 (95% CI 0.10 to 0.50, p = 0.006, df = 12). The increase in agreement was approximately equally distributed among all pairs of observers, while in the control group the overall mean kappa coefficient for pairwise agreement increased from 0.28 (95% CI 0.24 to 0.31) at baseline to 0.33 (95% CI 0.29 to 0.36) in the second classification session.

Table I. Mean (95% CI) kappa values for the classification of fractures of the humerus according to the Neer system

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall (n = 7)</td>
<td>Overall (n = 7)</td>
</tr>
<tr>
<td></td>
<td>0.28 (0.24 to 0.31)</td>
<td>0.27 (0.23 to 0.31)</td>
</tr>
<tr>
<td></td>
<td>Specialists (n = 4)</td>
<td>Specialists (n = 4)</td>
</tr>
<tr>
<td></td>
<td>0.30 (0.23 to 0.37)</td>
<td>0.30 (0.23 to 0.37)</td>
</tr>
<tr>
<td></td>
<td>Non-specialists (n = 3)</td>
<td>Non-specialists (n = 3)</td>
</tr>
<tr>
<td></td>
<td>0.19 (0.10 to 0.29)</td>
<td>0.26 (0.16 to 0.36)</td>
</tr>
</tbody>
</table>

Table II. The proportional distribution of Neer groups

<table>
<thead>
<tr>
<th></th>
<th>Control 1st class*</th>
<th>Control 2nd class</th>
<th>Training 1st class</th>
<th>Training 2nd class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undisplaced</td>
<td>0.25</td>
<td>0.28</td>
<td>0.19</td>
<td>0.45</td>
</tr>
<tr>
<td>Two-part</td>
<td>0.35</td>
<td>0.40</td>
<td>0.44</td>
<td>0.40</td>
</tr>
<tr>
<td>Three-part</td>
<td>0.21</td>
<td>0.18</td>
<td>0.20</td>
<td>0.06</td>
</tr>
<tr>
<td>Four-part</td>
<td>0.09</td>
<td>0.06</td>
<td>0.10</td>
<td>0.02</td>
</tr>
<tr>
<td>Fracture-dislocation</td>
<td>0.09</td>
<td>0.08</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>Articular surface</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Total</td>
<td>1.00</td>
<td>1.01</td>
<td>1.02</td>
<td>0.99</td>
</tr>
</tbody>
</table>

*classification

Fig. 2
Diagram showing the four anatomical parts of the proximal humerus (a, greater tuberosity; b, lesser tuberosity; c, articular segment; d, shaft of the humerus) (reprinted with permission from Codman EA. The shoulder: rupture of the supraspinatus tendon and other lesions in or about the subacromial bursa. Melbourne: Kreger, 1984 (first published 1934)).

Fig. 3a
Agreement between all pairs of observers within a) the training group and b) the control group at both sessions of classification.
group increased from 0.30 (95% CI 0.23 to 0.37) to 0.79 (95% CI 0.70 to 0.88) in the specialist group as compared with 0.26 (95% CI 0.16 to 0.36) to 0.51 (95% CI 0.39 to 0.62) in the non-specialist group.

In the training group the proportion of fractures classified as undisplaced increased from 0.19 to 0.45 (p = 0.016) (Table II). At both sessions the non-specialists tended to classify more fractures as displaced than the specialists. After teaching the mean kappa value for classifying fractures as undisplaced was 0.86 (95% CI 0.73 to 0.98) for the specialists and 0.56 (95% CI 0.39 to 0.74) for the non-specialists.

Discussion

We found that two teaching sessions improved the overall mean kappa value from 0.27 (95% CI 0.23 to 0.31) to 0.62 (95% CI 0.57 to 0.67) for the Neer system. The improvement was particularly noticeable for the specialists in whom kappa increased from 0.30 (95% CI 0.23 to 0.37) to 0.79 (95% CI 0.70 to 0.88)

The sample. One potential problem for interobserver studies is an insufficient number of observers and patients. We were unable to find sample-size formulae for observer-variation studies and therefore we followed the advice of authors of previous articles and included more than 40 patients. The number of patients and observers in our study was sufficient to detect a statistically significant rise in the mean kappa value after training.

Our sample of 14 observers consisted of doctors working in the Department of Orthopaedic Surgery. The distribution of specialists and non-specialists reflects a typical Danish university hospital.

Bias. The observer variation at the first classification session was similar in both groups which were therefore comparable at baseline. The drop-out was 5/19 (26%), but was the same in both groups, and did not alter the distribution between specialists and non-specialists.

The radiographs included reflected the clinical population. The trial was performed more than six months after they were taken and therefore the risk of staff remembering them was small.

For pragmatic reasons we reduced the number of groups in the Neer system from 16 to six before carrying out statistical calculations. Two recent studies of the reduction of the number of classes showed no significant change in the kappa values.

Most interobserver studies on the classification of fractures use settings which differ from clinical situations and this may affect the results. Observers may concentrate particularly when they participate in a study. However, the large number of cases and the lack of clinical consequence in an experimental situation could imply less rigorous attention to detail. We imitated several aspects of the real clinical situation by including a consecutive series of patients, an unselected sample of radiographs, and unselected and unprepared observers exclusively.

Kappa and statistics. Interobserver variation is usually reported as kappa statistics instead of crude observer agreement since the influence of chance agreement is accounted for. However, the level of chance agreement is estimated on the basis of the prevalence of the various categories of observations. This has two important implications for the interpretation of kappa values. First, even with a fixed level of agreement between observers, kappa will be reduced when the prevalence of a category is very high or very low. Secondly, prevalences of categories can differ from one clinical setting to another and crude comparisons of kappa values between different clinical studies may therefore be meaningless.

Imaging of fractures. We did not include CT scans or three-dimensional reconstructions in our study because their use is not part of standard daily routine and recent studies have shown no improvement in observer variation by adding such techniques.

Difficulties in translating information about images into classification groups has been suggested as a major reason for the poor interobserver variation which has been reported in previous studies. The significant effect of training in our study indicates that basic difficulties in translating information cannot fully explain the poor previous results.

Specialists and non-specialists. Uncertain observers are probably more anxious not to overlook a fracture which requires surgery, than to misclassify a fracture which only needs conservative treatment. This could be a reason for the dramatic increase, after teaching, in the proportion of undisplaced fractures and the decrease in displaced fractures.

In the typical clinical situation, the non-specialist will inspect the radiographs and if in doubt, consult a more experienced colleague. This element of doubt was not incorporated in our study. We found only a moderate interobserver agreement among non-specialists. However, we cannot exclude the possibility that non-specialists when in doubt and having consulted with a specialist, may achieve similar interobserver variations as specialists alone.

The low baseline kappa values reproduced earlier dismal reports for the Neer system. We found that when the Neer system is used by untrained doctors it is too unreliable for clinical practice and research. However, when specialists were trained in the system the kappa value increased to good or excellent levels. These results suggest that formal training in the Neer system should be a prerequisite for its use both in clinical practice and research.

We thank Thomas Gjørup, Peter Gøtzsche and Lene Skovgaard for valuable comments on earlier versions of this paper. S. Brorson was supported by a grant from the Danish Medical Research Council.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

References


