Interobserver and intraobserver variation in classification systems for fractures of the distal humerus

A. M. Wainwright, J. R. Williams, A. J. Carr

From the John Radcliffe Hospital and Nuffield Orthopaedic Centre, Oxford, England

We assessed the inter- and intraobserver variation in classification systems for fractures of the distal humerus. Three orthopaedic trauma consultants, three trauma registrars and three consultant musculoskeletal radiologists independently classified 33 sets of radiographs of such fractures on two occasions, each using three separate systems.

For interobserver variation, the Riseborough and Radin system produced ‘moderate’ agreement (kappa = 0.513), but half of the fractures were not classifiable by this system. For the complete AO system, agreement was ‘fair’ (kappa = 0.343), but if only AO type and group or AO type alone was used, agreement improved to ‘moderate’ and ‘substantial’, respectively (kappa = 0.52 and 0.66). Agreement for the system of Jupiter and Mehne was ‘fair’ (kappa = 0.295). Similar levels of intraobserver variation were found.

Systems of classification are useful in decision-making and evaluation of outcome only if there is agreement and consistency among observers. Our study casts doubt on these aspects of the systems currently available for fractures of the distal humerus.

Received 25 November 1998; Accepted after revision 6 October 1999
Type IV. There is severe comminution of the articular surface and wide separation of the humeral condyles.

The AO classification\(^5\) arose from efforts to produce a comprehensive system for fractures of the long bones. The distal segment ‘13’ of the humerus is divided into three types as follows:

Type A. Extra-articular fractures.

Type B. Partial articular fractures; these maintain some continuity between the shaft and the articular surface.

Type C. Complete articular fractures.

These are subdivided into groups 1, 2 and 3 on the basis of the location of the line of the fracture and the amount of fragmentation. They are then subdivided into subgroups •1, •2 and •3 by specific further qualifications.

Jupiter and Mehne\(^6\) developed a different comprehensive system based on the surgical anatomy and the techniques used for reconstruction. There are three basic categories as follows:

Grade I. Intra-articular fractures.

Grade II. Extra-articular, intracapsular fractures.

Grade III. Extracapsular fractures.

Intracapsular fractures are subdivided into: A, single column; B, bicolumnar; C, capitellar; and D, trochlear.

The distance of the fracture from the joint and the direction of the transverse limb of the fracture determine further categories.

Statistical methods. It would be ideal to have an absolute standard with which to compare the observations but any study of this type is limited by our ability to define accurately the pattern of the fracture. Even if findings at the time of operation were taken as the standard, there may have been disagreement on classification at this stage, and not all fractures require an operation. Lacking an absolute standard the categories of classification were treated as nominal data and we used the kappa statistic for evaluation of inter- and intraobserver agreement. The kappa statistic is
Summary of the AO system for classification of distal humeral fractures (adapted with permission from Müller ME, et al. *The comprehensive classification of fractures in long bones.* Berlin: Springer-Verlag, 1990.)
a chance corrected measure of agreement for nominal data, first described by Cohen.\textsuperscript{7} It compares an observed measure of agreement $P(A)$, with the level of agreement expected by chance alone $P(E)$.\textsuperscript{8,9} The maximum value of 1.0 means that every assessor agrees with every other on every case. A value of 0 indicates no more agreement than expected by chance alone, and the value may be less than 0. Interpretation of the kappa value was based on the guidelines proposed by Landis and Koch,\textsuperscript{10} as shown in Table I.

**Results**

All 33 sets of radiographs were classified by all nine assessors. In the Riseborough and Radin classification all the groups were used for at least one observation. In the AO classification all except two of the subgroups were used; the exceptions were A1.1 and A1.2. In the Jupiter and Mehne classification all except three categories, ID, IIIA and IIIB, were used.

**Interobserver variation.** The interobserver agreement for the classification systems is summarised in Table II. In 48% of the assessments, the fractures could not be classified using the Riseborough and Radin system and there was some disagreement as to which of the fractures could be classified using this method. At least one of the observers thought that a fracture could not be classified by the system in 24 out of the 33 cases. In only seven was there complete agreement among observers that the fracture could not be classified. Overall, the interobserver agreement was ‘moderate’ (kappa = 0.513).

The AO system has 27 subgroups and had ‘fair’ interobserver agreement (kappa = 0.343). The agreement improved when assessment was by the nine AO groups (kappa = 0.521, ‘moderate’), or by the three AO types alone (kappa = 0.661, ‘substantial’).

The Jupiter and Mehne system\textsuperscript{6} has 19 categories and gave an overall ‘fair’ agreement (kappa = 0.295).

**Intraobserver agreement.** This showed a similar pattern to interobserver agreement (Table III). Only seven observers repeated the classification and one did not reclassify the AO subgroup.

The system of Riseborough and Radin\textsuperscript{2} produced wide ranges of intraobserver agreement (kappa scores between 0.364 and 0.696).

Agreement in the AO classification by subgroup was ‘fair to moderate’ (kappa = 0.254 to 0.536). When classified by AO group agreement was ‘moderate to substantial’ (kappa = 0.466 to 0.75), and by AO type was higher (kappa = 0.489 to 0.835). Two observers were consistent in their intraobserver agreement for type (kappa > 0.81).

Using the Jupiter and Mehne system, the intraobserver agreement varied widely (kappa = 0.001 to 0.550). The lowest level occurred with one observer who favoured one class on the initial assessment and another on the subsequent occasion, putting almost half (15/33) of the readings into the IB3 class and only one in the IB4b class on the first assessment, and then reversed this on the second.

**Discussion**

Fractures of the distal humerus account for 2% of all fractures and about one-third of those at the elbow. A review of the history of the classification of fractures of the elbow is given elsewhere.\textsuperscript{11}

**Table I.** Interpretation of the kappa statistic according to Landis and Koch\textsuperscript{10}

<table>
<thead>
<tr>
<th>Kappa statistic</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.00</td>
<td>Poor</td>
</tr>
<tr>
<td>0.00 to 0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21 to 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 to 0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61 to 0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81 to 1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>

**Table II.** Kappa scores for interobserver agreement

<table>
<thead>
<tr>
<th>Observers</th>
<th>Riseborough and Radin</th>
<th>AO subgroup</th>
<th>AO group</th>
<th>AO type</th>
<th>Jupiter and Mehne</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consultant trauma surgeons</td>
<td>0.521</td>
<td>0.335</td>
<td>0.476</td>
<td>0.693</td>
<td>0.348</td>
</tr>
<tr>
<td>Consultant radiologists</td>
<td>0.518</td>
<td>0.320</td>
<td>0.618</td>
<td>0.802</td>
<td>0.262</td>
</tr>
<tr>
<td>Registrars</td>
<td>0.568</td>
<td>0.272</td>
<td>0.458</td>
<td>0.613</td>
<td>0.199</td>
</tr>
<tr>
<td>Overall</td>
<td>0.513</td>
<td>0.343</td>
<td>0.521</td>
<td>0.661</td>
<td>0.295</td>
</tr>
</tbody>
</table>

**Table III.** Kappa scores for intraobserver agreement

<table>
<thead>
<tr>
<th>Observers</th>
<th>Riseborough and Radin</th>
<th>AO subgroup</th>
<th>AO group</th>
<th>AO type</th>
<th>Jupiter and Mehne</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.696</td>
<td>0.536</td>
<td>0.750</td>
<td>0.788</td>
<td>0.475</td>
</tr>
<tr>
<td>B</td>
<td>0.364</td>
<td>0.254</td>
<td>0.562</td>
<td>0.690</td>
<td>0.001</td>
</tr>
<tr>
<td>C</td>
<td>0.664</td>
<td></td>
<td>0.566</td>
<td>0.811</td>
<td>0.550</td>
</tr>
<tr>
<td>D</td>
<td>0.362</td>
<td>0.378</td>
<td>0.470</td>
<td>0.703</td>
<td>0.239</td>
</tr>
<tr>
<td>E</td>
<td>0.680</td>
<td>0.428</td>
<td>0.622</td>
<td>0.835</td>
<td>0.400</td>
</tr>
<tr>
<td>F</td>
<td>0.422</td>
<td>0.336</td>
<td>0.466</td>
<td>0.489</td>
<td>0.140</td>
</tr>
<tr>
<td>G</td>
<td>0.601</td>
<td>0.440</td>
<td>0.646</td>
<td>0.556</td>
<td>0.275</td>
</tr>
<tr>
<td>Mean</td>
<td>0.541</td>
<td>0.395</td>
<td>0.583</td>
<td>0.695</td>
<td>0.297</td>
</tr>
</tbody>
</table>
It has been proposed that there should be two types of system of classification,12 one (taxonomy) for research purposes to make valid comparisons between groups, and the other (treatment) for clinical use to determine management and prognosis. For either of these an ideal system should be easy to use, widely accepted and produce good inter- and intraobserver agreement.1 Our aim was to assess the last and no attempt was made to comment on the ease of use or acceptability.

Although systems of classification are widely used, several studies have shown that many in current use do not have consistency among or within observers. For the Gustilo and Anderson classification of open fractures, Brumback and Jones13 found moderate to poor interobserver agreement in a survey of 245 orthopaedic surgeons. Previously, Horn and Rettig14 had shown moderate interobserver reliability.

Johnstone, Radford and Parnell15 assessed interobserver variation using the AO/ASIF classification of fractures of the long bones. They compared individual classification with a consensus view and found agreement of only 32%. There was moderate agreement for the Evans classification of trochanteric fractures.16 Observers agreed on the stage in Garden’s classification of fractures of the femoral neck in only 22%.17 Analysis of the interobserver reliability and intraobserver reproducibility in the classification of fractures of the distal radius found moderate agreement between observers for the Mayo and fair agreement for the Frykman, Melone and AO classifications.18 Another study showed only fair agreement of the AO system for the distal radius when assessed for the potential 27 subgroups.19

In a study of fractures of the proximal humerus there was fair or poor inter- and intraobserver reliability of classification using Neer’s system and the AO system,20 and in another study there was moderate interobserver reliability, but higher intraobserver reproducibility of the Neer classification.21 These studies prompted an Editorial by Burstein22 which recommended that the Neer classification system was unwarranted as a means of differentiating treatment and correlating outcomes in clinical series. He concluded that all studies using a classification scheme to assess clinical outcome should show that there had been acceptable intraobserver repeatability and interobserver reliability.

To our knowledge, there have been no other studies published of the interobserver and intraobserver reliability of the classification of fractures of the distal humerus. Johnstone et al15 found that classification of a sole fracture of the distal humerus gave the greatest amount of variation from the consensus classification of all the fractures which they analysed.

It may be that the process of classification can only be accurately performed after the fracture has been assessed by direct vision at operation. However, part of the function of a useful system of classification is to determine the best method of treatment, so that preoperative planning can be performed before embarking on an operation.

In our study, observers commented that further radiographs with traction applied to the limbs are often made in the operating theatre. This may aid in classifying the fractures, but we did not assess this. It has also been suggested that CT may clarify the patterns of fracture. This theory has been investigated for the proximal humerus but two studies of the Neer classification system23,24 showed that CT added little to improving reproducibility for those fractures.

One reason why the agreement may be higher in our assessment compared with a general population is that the observers were part of a study and took special care. They were also from the same unit and had common experience of the patients in many cases. The order of presentation was not rerandomised, although there was a considerable period between reviews. Reasons for the lower agreements may be that the systems of classification were not clearly understood, or there was fatigue after reviewing many similar radiographs. Bias may have also been introduced by the fact that the assessors in our study were more familiar with the AO system. Only two stated that they would routinely use the AO system to describe a fracture, most preferring to describe the fracture in a sentence.

There are also intrinsic differences among the three systems. The AO system employs a stepwise classification, which becomes more detailed to use with each stage. The Risborough and Radin method is a single assessment into four categories of increasing severity. The Jupiter and Mehne system has elements of both of these approaches. One observer commented that there appear to be two phases in classification, an initial interpretation of radiographs to produce a mental image of a fracture in three dimensions, and subsequently, a matching of the three-dimensional image to the system of classification. When developing future systems it may be useful to analyse at which stage the agreement is lost.

Systems of classification are useful in decision-making and comparison of outcomes only if there is a high level of agreement among observers and consistency by them. Our study has cast doubt on these features of current systems for fractures of the distal humerus.

We found that the Radin and Risborough classification was not applicable to half of these fractures and produced moderate agreement. There was only fair agreement in the AO and the Jupiter and Mehne systems, but when only the basic outline classification was used there was substantial agreement.

We would like to thank Mr Clasper, Mr Handley, Mr Kamborouglou, Mr Keys, Dr McNally, Mr Oliver, and Dr Ostlere for taking the time to classify these fractures. We are grateful to the authors and publishers of the three classification systems for allowing us to reproduce diagrams taken from their original texts. We would also like to thank the staff of the radiology and photographic departments and the Cairns and Girdlestone Memorial Libraries for their help.

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