WRIST AND HAND

Is it time to revisit the AO classification of fractures of the distal radius?

INTER- AND INTRA-OBSERVER RELIABILITY OF THE AO CLASSIFICATION

We conducted an observational radiographic study to determine the inter- and intra-observer reliability of the AO classification of fractures of the distal radius. Plain posteroanterior and lateral radiographs of 456 patients with an acute fracture of the distal radius were classified by a consultant orthopaedic hand specialist and two specialist trainees, and the k coefficient for the inter- and intra-observer reliability of the type, group and subgroup classification was calculated.

Only the type of fracture (A, B or C) was found to provide substantial intra-observer reliability (k_{type} 0.65). The inclusion of ‘group’ and ‘subgroup’ into the classification reduced the inter-observer reliability to fair (k_{group} 0.29, k_{subgroup} = 0.28) and the intra-observer reliability to moderate (k_{group} 0.53, k_{subgroup} 0.49). Disagreement was found to arise between specific subgroups, which may be amenable to clarification.

Cite this article: Bone Joint J 2015; 97-B:818–23.

The classification of fractures facilitates diagnosis and communication in both clinical and academic settings.\(^1,2\) There are many systems for classifying fractures of the distal radius, such as those of Frykman\(^3\) and Fernández,\(^4\) which are based broadly on either the anatomical characteristics of the fracture or the mechanism of injury. Historically, the Frykman system, which classifies fractures according to the presence of an ulnar styloid fracture and involvement of the radiocarpal and radioulnar joints, used to be favoured.\(^5,6\) This system, however, like many others for the distal radius, is flawed as it fails to include either displacement, or fragmentation, of the fracture.\(^6\)

In comparison, the Müller Arbeitsgemeinschaft für Osteosynthesefragen (AO) classification of fractures of the long bone provides the most comprehensive system for the distal radius. It was developed in 1986 as a universal system applicable to all skeletal fractures, not only as a tool for communicating the configuration of a fracture, but also providing a format that can be used to teach the intricacies of diagnosis.\(^7\) An alphanumeric system is used to assign a code to the fracture based on its location and morphology.\(^8\) The complexity of the fracture is reflected by an increase in the alphanumeric order.\(^7\) The code has since been adopted by the Orthopaedic Trauma Association and now constitutes the diagnostic component of the AO surgical reference tool, providing the surgeon with a framework in which they can diagnose the fracture and be presented with operative and non-operative treatments specific to that classification.\(^3,8\)

This additional functionality is unique to this classification system.

Although the AO system provides a significant amount of information, its complexity allows only a moderate degree of intra- and inter-observer reliability, at best.\(^5,6\) Simplification of the system to exclude the group and subgroup components has been shown to provide a more reliable alternative.\(^1,9\) However, only a limited number of patients have been included in investigations of agreement, preventing meaningful analysis of where disagreement arises and whether particular subgroups are problematic.\(^1,9\) In this study we, therefore, assessed the inter- and intra-observer reliability of the type, group and subgroup components of the system in patients suitable for operative fixation, and analysed the areas where disagreement arose.

Materials and Methods
We assessed standard digitised posteroanterior and lateral radiographs taken at presentation for 456 patients with an acute fracture of the distal radius. Patients were recruited as part of a United Kingdom National Institute for Health Research clinical trial, between August 2010 and June 2012.\(^10\) Adult patients were eligible to enter the study if they had sustained a dorsally displaced fracture of the distal radius,
defined as within 3 cm of the radiocarpal joint, which the consulting surgeon considered suitable for fixation. Patients were excluded if they had an open fracture (Gustilo grade > 1), if the fracture could not be reduced closed, if there were contraindications to surgery or if the patient was unable to comply with the main procedures in the trial.11

A senior specialist trainee (CP) assessed all radiographs on two occasions, one week apart, facilitating an analysis of intra-observer agreement and the identification of specific areas of disagreement in the subgroup classifications. An analysis of the agreement between sequential blocks of 50 radiographs was also performed, to detect the presence of fatigue between the classification of subsequent radiographs. A consultant orthopaedic hand specialist (HH) and a second orthopaedic specialist trainee (CH) of limited experience assessed the radiographs of 50 patients selected from the main group using a computer-generated random number sequence, to allow further assessment of
both intra- and inter-observer agreement. Observations were carried out on two separate occasions a minimum of one week apart.

The AO classification for a fracture of the distal radius categorises fractures into three types: extra-articular, partial-articular and complete articular, based on the involvement of the radiocarpal joint. Groups and subgroups are then assigned, based on the morphology of the fracture, to provide an increasingly intricate description of its pattern. There are nine groups for fractures of the distal radius describing the presence of metaphyseal and articular fragmentation. Each group then has three subgroups describing the angulation of the fracture and the plane of the line of fracture, with the exception of 23-C3 multifragmentary intra-articular fractures, which have no subgroups (Fig. 1).

Prior to the classification sessions, the observers familiarised themselves with the system using the AO surgical reference tool (www2.aofoundation.org), which provides schematic and written descriptions of the system with radiographic examples. In addition, the system was discussed among the observers and they were given the opportunity to practise classifying the fractures. During the testing sessions observers were able to access the online tool and were provided with additional paper copies of the tool for reference.

**Statistical analysis.** We assessed the inter- and intra-observer reliability of the type, group and subgroup components with Cohen’s k test using SPSS v21.0 (IBM, Armonk, New York) for each of the observers. Landis and Koch’s guidance for the interpretation of the k coefficient was then used to assign a ‘strength’ of agreement (Table I).

**Table I. Landis and Koch’s interpretation of k values**

<table>
<thead>
<tr>
<th>Value of k coefficient</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>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. Inter-observer k scores at pre-operative times T1 and T2**

<table>
<thead>
<tr>
<th>Observer</th>
<th>Time point</th>
<th>AO classification (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
<td>Group</td>
</tr>
<tr>
<td>OHS versus OST1</td>
<td>T1 0.60 (0.39:0.82)</td>
<td>0.22 (0.07:0.37)</td>
</tr>
<tr>
<td></td>
<td>T2 0.65 (0.46:0.85)</td>
<td>0.46 (0.29:0.62)</td>
</tr>
<tr>
<td>OHS versus OST2</td>
<td>T1 0.47 (0.25:0.68)</td>
<td>0.19 (0.03:0.34)</td>
</tr>
<tr>
<td></td>
<td>T2 0.59 (0.40:0.79)</td>
<td>0.26 (0.10:0.41)</td>
</tr>
<tr>
<td>OST1 versus OST2</td>
<td>T1 0.66 (0.47:0.85)</td>
<td>0.45 (0.29:0.62)</td>
</tr>
<tr>
<td></td>
<td>T2 0.39 (0.19:0.59)</td>
<td>0.17 (0.03:0.31)</td>
</tr>
</tbody>
</table>

OHS, consultant orthopaedic hand specialist; OST1, senior orthopaedic specialist trainee; OST2, junior orthopaedic specialist trainee

**Results**

The study group consisted of predominantly complete articular (n = 140) and extra-articular fractures (n = 308), with a similar distribution of displacement and metaphyseal fragmentation.

**Inter-observer agreement.** Moderate inter-observer agreement was observed for the type of fracture, with a k score of 0.56 (0.39 to 0.66) (Table II).

The inclusion of the ‘group’ and ‘subgroup’ components of the system reduced the agreement to fair, with k scores ranging from 0.17 to 0.45 for both, and mean k scores of 0.29 (95% CI 0.16 to 0.43) and 0.28 (95% CI 0.15 to 0.40), respectively (Table II). The distribution of fractures among the categories for the observers is shown in Figure 2.

**Intra-observer reliability.** Substantial intra-observer reliability was observed for the type of fracture, with a mean k score of 0.65 (95% CI 0.37 to 0.94) (Table III).
reduction in reliability was again observed with the inclusion of the group and subgroup components, with mean k scores of 0.53 (95% CI 0.25 to 0.81) and 0.49 (95% CI 0.20 to 0.79), respectively (Table III). The consultant demonstrated the highest level of agreement, with substantial agreement for all components of the classification system.

No observer fatigue was detected for the classification of all the radiographs. Analysis of groups of 50 radiographs revealed no reduction in intra-observer reliability with subsequent groups.

**Subgroup analysis.** There were disagreements in the classification of 204 fractures (44.7%) in the intra-observer data. Most of the disagreement arose between the 23-A2.2 and 23-A3.2 subgroups (n = 50), followed by the 23-C1.2 versus 23-C2.1 (n = 23), and 23-A2.1 versus 23-A2.2 (n = 21) subgroups. The distribution of the disagreement between the groups is shown in Figure 3.

**Discussion**
In this study, the AO classification system for fractures of the distal radius was found to have only fair inter-observer and moderate intra-observer reliability. Exclusion of the subgroup components led to a slight improvement in the k value, but no change in the strength of the reliability. Substantial intra-observer and moderate inter-observer reliability was, however, achieved by simplifying the system to just the type of fracture.

Previous studies have also questioned the reliability of the AO classification system.\(^5,6\) In this study, the mean k score for the inter-observer reliability of the full classification system was 0.28 (95% CI 0.15 to 0.40), which is within the range reported previously of k 0.19 to 0.34.\(^1,5,6\) A range of k values of 0.23 to 0.64 has been reported for intra-observer reliability, which again corresponds with the k value of 0.49 reported here.\(^1,5,6\)

Belloti et al\(^5\) suggest that the intra-observer reliability of the AO classification system is subject to an expertise effect, with greater reliability achieved by those with greater orthopaedic and radiological expertise.\(^5\) A similar finding was shown in this study, with an increase in the k value with increasing seniority of the observer. Most noticeably, the consultant hand specialist demonstrated substantial intra-observer reliability with all components of the classification, compared with only moderate reliability for the junior trainee.

Simplification of the AO system to just ‘type of fracture’ was also considered in two previous studies, with a comparable improvement in the reliability.\(^1,9\) Kreder et al\(^9\) demonstrated a significant improvement when both the subgroup and group components were excluded separately. Andersen et al,\(^5\) however, found only a marginal improvement in reliability with the exclusion of the subgroup component, but a significant improvement with the exclusion of the group component. However, excluding
the groups and subgroups clearly reduced the ability of the classification to discriminate between different patterns of injury. A detailed analysis of disagreements between observers has not been considered previously.

In this study, an analysis of the intraobserver classification for the complete group of patients found that disagreement tended to arise between specific groups, namely those where the observer had to make a decision regarding the degree of metaphyseal comminution or the degree of angulation. In light of these findings, instead of excluding the groups and subgroups to improve reliability, the groups could be amalgamated, or their descriptions refined, to provide greater clarity. This would clearly require further investigation.

Reliability is only one of several components a classification system should display when assessing a fracture.6,14 Although the AO system offers a comprehensive description and a seemingly straightforward classification, both its usefulness and its clinical applicability have been criticised.9 For example, the 27 subgroups are difficult to recall and have not been shown to correspond with specific forms of treatment, hence several subgroups may be redundant.9 In addition, the AO system fails to consider other clinically relevant features, such as the presence of articular displacement, the mechanism of the injury and the extent of soft-tissue damage.9

These criticisms can be applied to the other prominent classification systems. For example, the frequently used Frykman classification does not include radiological features such as displacement, shortening and comminution.6,15,16 In addition, Keerthi et al16 found no correlation between the grade of the fracture and the patient’s functional outcome one year after fixation, indicating that the classification may have limited prognostic value for surgically managed patients. The Universal system is similarly limited, despite being simplified and easier to use.6 Belloti et al15 found the assessment of instability, which is one of only four considerations for the system, difficult without a clear consensus in the literature, resulting in poor interobserver reliability. The Fernandez2 and Melone17 classifications have also demonstrated poor inter- and intraobserver reliability.1,18 Furthermore, Andersen et al1 found that most intra-articular fractures could be placed in only two classes with the Melone classification, due to the greater emphasis on the medial articular facet, regardless of other variations in the pattern of the fracture.

Patients who underwent fixation of the fracture were chosen specifically for this study to determine whether the AO classification system could be reliably applied to those with more complicated injuries that require specific decisions on their management. Certain categories were, therefore, not considered, including undisplaced fractures and isolated fractures of the distal ulna. Previous studies typically included 30 to 50 patients only, with all types of pattern of fracture, thereby preventing a detailed analysis. The subgroup analysis offers another potential limitation of this study. The analysis was based solely on the classification performed by one observer, and observations were separated by only one week. Therefore, bias may have been introduced, resulting in artificially high intraobserver reliability.

In conclusion, the AO classification system provided only fair to moderate reliability when applied to radiographs of patients requiring fixation of a fracture of the distal radius. This suggests that, in its current form, this system of classification is inadequate for clinical and academic use. An analysis of observer disagreements suggests that simplification of the system or further clarification of the descriptions may improve its reliability.

Author contributions:
C.E. Plant: Data collection; Statistical analysis; Writing the paper.
C. Hickson: Data collection; Review of the paper.
H. Hedley: Data collection; Review of the paper.
N. R. Parsons: Statistical analysis; Writing the paper.
M. L. Costa: Statistical analysis; Writing the paper.

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.

This article was primary edited by G. Scott and first proof edited by J. Scott.

References

