CONSISTENCY OF AO FRACTURE CLASSIFICATION FOR THE DISTAL RADIUS

HANS J. KREDER, DOUGLAS P. HANEL, MICHAEL McKEE, JESSE JUPITER, GARY McGILLIVARY, MARC F. SWIONTKOWSKI
From Sunnybrook Health Science Center and the University of Toronto, Canada

We sought to quantify agreement by different assessors of the AO classification for distal fractures of the radius. Thirty radiographs of acute distal radial fractures were evaluated by 36 assessors of varying clinical experience.

Our findings suggest that AO ‘type’ and the presence or absence of articular displacement are measured with high consistency when classification of distal radial fractures is undertaken by experienced observers. Assessors at all experience levels had difficulty agreeing on AO ‘group’ and especially AO ‘subgroup’. To categorise distal radial fractures according to joint displacement and AO type is simple and reproducible.

Our study examined only whether distal radial fractures could be consistently classified according to the AO system. Validation of the classification as a predictor of outcome will require a prospective clinical study.

Received 2 February 1996; Accepted 1 April 1996

Variables which affect patient outcome, independent of treatment, are known as potential confounding factors. When they are distributed unevenly across study populations, conclusions may be biased unless these factors are considered. Dorey, Grigoris and Amstutz (1994) urged that results be presented in relation to important risk factors such as age and the aetiology of the disease concerned. Ideally, confounding factors are controlled by randomly assigning treatment, or by carefully adjusting for them by statistical techniques. Even in prospective randomised trials, major confounding factors must be considered, to provide a description of the patient population, and allow a deeper understanding of the interplay between the particular factor, the treatment and the outcome.

Severity of injury represents one potential confounding factor, although other variables such as age, intercurrent injury, medical illness, and socioeconomic status may be more important (The Coronary Drug Project Research Group 1980; Caine et al. 1991; Glower et al. 1992). The AO/ASIF classification system of fractures of long bones is intended to provide a measure of severity of injury: higher grades indicate more severe fractures (Müller 1991). The Orthopaedic Trauma Association has recently adopted this scheme as the backbone of its complete fracture classification system and thus it will probably be widely used for clinical research in the future. Interobserver and intraobserver consistency is a prerequisite for the effective use of any classification system.

Our aim was to quantify agreement in applying the AO classification to distal fractures of the radius and in recognising articular displacement.

PATIENTS AND METHODS

We chose 30 patients from a trauma database to provide a wide range of severe fractures. All 27 possible classification ‘subgroups’ were ultimately chosen by at least four of the assessors which suggested that we had indeed covered the complete spectrum of severity of injury. Before the first grading session, a 30-minute review of the AO classification system (Müller 1987, 1991) of the distal radius was given, emphasising in particular the key points with regard to the grading of a specific ‘type’, ‘group’ and ‘subgroup’. There are three AO categories:

1) ‘Extra-articular’ fractures which do not involve the radiocarpal joint surface at all.
2) ‘Partial articular’ fractures which involve this joint, but a portion of the articular surface remains in continuity with the diaphysis.
3) ‘Complete articular’ fractures which are distinguished by complete separation of the involved articular surface and the diaphysis.

AO groups and subgroups involve a progressively more detailed description of the fracture patterns within these three categories.

Assessors were subdivided into the following levels: attending (surgeons with a permanent faculty appointment), fellow (training completed but no permanent faculty position), resident (postgraduate year one to six in an orthopaedic residency programme), and non-clinical (physiotherapy, nursing, non-orthopaedic physician, medical student). After the review, a group of 8 attending surgeons, 9 fellows, 15 residents, and 4 non-clinicians from four different centres were asked to classify each fracture using a handout describing in detail and illustrating all 27 AO fracture subgroups (Muller 1987) (Fig. 1). Articular involvement is inherent in the assignment of AO type. The AO classification, however, does not differentiate as to whether or not the intra-articular fracture component is displaced. Thus we asked each assessor to decide whether the articular surface was displaced or not. The same radiographs were then reclassified two to four weeks later by the same assessor.

**Statistics.** It is important to use the appropriate statistics to reflect observer agreement accurately. Percentage agreement (the percentage of assessors who agree with one another) is a flawed measure which should not be used because it does not consider that some agreement will occur purely by chance (Posner et al 1990; Sackett et al. 1991). In 1960, Cohen described a so-called ‘kappa’ statistic which would adjust for chance agreement. The original kappa statistic was applicable only to the situation in which two assessors classified the same subjects into mutually exclusive categories. Since this original description, additional statistics have been described with more flexibility (Fleiss 1971; Fleiss, Nee and Landis 1979; O’Connell and Dobson 1984; Posner et al 1990).

Which of these should be used depends on the purpose of the analysis as well as on the number of assessors and the nature of the data being analysed (Posner et al 1990). A value of 1.0 means that every assessor agrees with every other on every single case. A value of zero suggests that all observed agreement can be explained purely by chance, while values below zero imply a systematic disagreement in that agreement was worse than would be expected by chance. Landis and Koch (1977) have developed some arbitrary but widely-quoted criteria for the interpretation of agreement statistic scores. Values of 0.00 to 0.20 represent slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, and 0.61 to 0.80 substantial agreement. Above 0.80 is considered almost perfect agreement. Most clinical measures fall into the ‘moderate’ agreement range (Koran 1975a,b).

Our results were analysed using the SAV statistic of O’Connell and Dobson (1984). Like the original kappa statistic described by Cohen (1960), the SAV statistic allows assessor agreement to be quantified beyond that expected by chance alone, but it may be extended to situations involving multiple assessors. The SAV value represents the level of agreement averaged over all subjects. Agreement for each individual subject may also be quantified. In our study, agreement was graded by the clinician’s experience of each variable. When computing chance-corrected agreement statistics, a weighting scheme may be used whereby adjacent responses are given partial credit. This approach is applied if the classification is ordinal, but has the unwanted effect of false elevation of the statistical value as the number of rating categories increases. This is analogous to the inflation of the correlation coefficient as the range of values increases. By utilizing the unweighted SAV values, we are able to compare directly classification subsets with differing numbers of categories, that is, AO type with three categories v AO subgroup with 27 possibilities. We were also interested in learning whether assessors experienced particular difficulties with any specific category of the AO type classification. For this reason we computed the AO category specific kappa values after the method of Fleiss (1971). This computation does not require a consensus or ‘gold standard’, but is related to the conditional probability that an assessor would classify an injury into a certain category given that another has used that category. Thus information can be obtained regarding the relative agreement for each of the specific three categories, types A, B and C.

**RESULTS**

Three of the 36 assessors (two attending surgeons and one fellow) indicated that they regularly use the AO classification system (Muller 1987) for the distal radius, and 23 others were familiar with it. Fifteen regularly used the Frykman (1967) classification, and six the universal system or a modification (Missakian et al 1992), while 12 observers did not regularly have any classification system. Thus, after excluding the four non-clinical assessors, most (26 of 32, 81%) were familiar with the AO classification before the review session.

**AO categories.** Agreement was best for AO type and decreased progressively for AO group and subgroup (Fig. 2). For AO type, interobserver agreement improved with increasing experience. For AO group and subgroup, it was less related to experience. In other words, the attending surgeons were just as likely to disagree on the specific subgroup as were the non-clinicians. Interobserver agreement among those clinicians who indicated prior familiarity or regular use of the AO classification was no better than that in clinicians without such experience. Attending surgeons achieved the highest interobserver agreement scores for the AO classification categories with an SAV value of 0.68 (substantial agreement) for the AO type; 0.48 (moder-
An information sheet was given to each assessor to aid categorisation of fractures according to the AO classification system. The illustrations and descriptions are taken directly from the AO classification system (Müller 1987). The assessors had an initial review session at which the AO classification system was discussed in detail. They were encouraged to pursue a three-step decision process as outlined by Müller (1987, 1991) regarding which type, group and subgroup best described the fracture (reproduced with permission).
ate agreement) for the nine possible AO groups; and 0.33 (fair agreement) for the 27 potential AO subgroups. In general, agreement regarding AO type A was equal to that for AO type B and AO type C. There was no evidence to suggest that assessors had particular difficulty with a specific AO type designation. We did not perform category specific testing for the expanded classification of groups and subgroups.

A review of average agreement scores by individual radiograph revealed that there was some case dependency to the classification, with certain radiographs subject to more disagreement than others. It is perhaps to be expected that there are certain cases which are simply more difficult to categorise. Obviously, it would be desirable to identify the specific features which account for this difficulty. For acute distal radial fractures, we were unable to identify prospectively any radiological features which might account for such a variation by case. The spread of individual radiological agreement values was lower among the attending surgeons which suggests less case-specific variation with increasing experience.

**Joint status.** Agreement scores for joint displacement were above the SAV value of 0.61 (substantial agreement) for all assessors with the exception of the four non-clinicians (SAV = 0.49) (Fig. 2). Agreement was universally higher for joint displacement as compared with classification of the AO type, group and subgroup. For some radiographs with an SAV value of 0.41 to 0.60 (moderate agreement) regarding joint displacement, assessors showed major disagreement regarding the AO classification (Fig. 3).

**Intraobserver consistency.** In general non-clinicians were the least consistent, while attending surgeons agreed with themselves most often. Intraobserver agreement was highest for articular displacement with SAV values ranging from 0.43 (moderate agreement) for non-clinicians to 0.88 (nearly perfect agreement) for fellows and attending sur-
The SAV values for AO type ranged from 0.67 (substantial agreement) for residents to 0.86 (nearly perfect agreement) for attending surgeons with a value of 0.40 (fair agreement) for non-clinicians. Intraobserver agreement decreased progressively with the more detailed AO categorisation of fractures into groups and subgroups beyond the three possible AO types. Agreement values for the full 27-category classification ranged between 0.25 and 0.42 (fair to moderate agreement).

DISCUSSION

To the best of our knowledge, of the many classification systems applied to the distal radius only Older’s classification has been tested for consistency (Older, Stabler and Cassebaum 1965; Andersen et al 1991) with kappa values of 0.75 and 0.69 for intraobserver and interobserver agreement, respectively. This applies mainly to extra-articular injuries, and thus may not be useful for discriminating severity in intra-articular fractures. In another study, a single distal fracture of the radius was included in an overall evaluation of interobserver variation using the AO classification (Johnstone, Radford and Parnell 1993) and the study involved a consensus opinion by the assessors after individual ratings. Seven of the 18 assessors agreed with the consensus opinion regarding the AO type, while only three graded the subgroup correctly. Unfortunately, no statements can be made regarding chance-corrected agreement based on a single radiograph graded at a single observation.

The AO system allows collapse down to three types for grading by severity, with the possibility of expansion for what was hoped would provide more detailed information on fracture pattern useful in planning treatment. In theory this is an attractive concept. For fractures of the distal radius, however, there is no prospective research to suggest that each of the expanded 27 subgroups correlates with important treatment considerations. In general, we feel that treatment should be based on sound basic principles of fracture care and anatomy as well as consideration of the patient and his circumstances, rather than relying on a specific fracture classification system. Moreover, simple classification schemes such as the universal classification (Missakian et al 1992) point out important features which must be considered when planning fracture treatment. Our main interest in using the AO classification scheme is to grade fractures by severity in a consistent manner.

After identification of the fracture as one involving the distal radius the use of the full AO classification requires three sequential decisions. At each step in the decision-making process, classification error due to disagreement is compounded. Substantial agreement was shown by the experienced assessors in categorising the fracture as type A, B or C. For distal radial and other periarticular fractures, evidence suggests that displaced intra-articular fractures would generally be expected to have a worse prognosis than extra-articular or undisplaced articular injuries (Knirk and Jupiter 1986; Porter and Stockley 1987; Jakim, Pieterse and Sweet 1991; Trumble, Schmitt and Vedder 1994). Thus the presence or absence of articular displacement should be included in any classification which is intended to reflect severity of injury. Unfortunately, the AO classification does not consider articular displacement at all for defining fracture types, groups or subgroups.

Our study shows that fractures of the distal radius may be categorised by articular displacement and AO type with high interobserver agreement, whereas subclassification into AO groups and subgroups is unreliable. In order to grade the severity, we suggest supplementing the AO grouping by classifying the articular surface into displaced or undisplaced groups as follows:

I. Joint undisplaced
   A Extra-articular
   B Partial articular
   C Complete articular

II. Joint displaced
   A Partial articular
   B Complete articular.

A fracture is considered to be displaced if there is any detectable displacement of the radiocarpal articular surface on plain radiographs. The agreement for this combination of AO type and joint displacement status is shown in Figure 2. We have shown that these categories can be applied with moderate consistency, even by junior clinicians. For research purposes, classification should be undertaken by a consensus of experienced clinicians, as both interobserver and intraobserver consistency varies with experience. While we have shown that joint status and AO type represent consistently measurable parameters, prospective research is needed to demonstrate the validity of these measures in predicting patient-oriented functional outcome.

This work was supported in part by a grant from the Orthopaedic Trauma Association and Orthopaedic Research and Education Foundation.

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


Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.


Older TM, Stabler EV, Cassebaum WH. Colles fracture: evaluation and selection of therapy. *J Trauma* 1965;5:469-76.


