Inter-observer reliability and intra-observer reproducibility of the Weber classification of ankle fractures

I. A. Malek, B. Machani, A. M. Mevcha, N. H. Hyder

From Leighton Hospital, Crewe, England

Our aim was to assess the reproducibility and the reliability of the Weber classification system for fractures of the ankle based on anteroposterior and lateral radiographs. Five observers with varying clinical experience reviewed 50 sets of blinded radiographs. The same observers reviewed the same radiographs again after an interval of four weeks. Inter- and intra-observer agreement was assessed based on the proportion of agreement and the values of the kappa coefficient.

For inter-observer agreement, the mean kappa value was 0.61 (0.59 to 0.63) and the proportion of agreement was 78% (76% to 79%) and for intra-observer agreement the mean kappa value was 0.74 (0.39 to 0.86) with an 85% (60% to 93%) observed agreement. These results show that the Weber classification of fractures of the ankle based on two radiological views has substantial inter-observer reliability and intra-observer reproducibility.

Classification systems for fractures are often based on the mechanism of injury and are used to devise a plan of management or to predict prognosis. They give clarity to communication and facilitate comparison of published results. Several authors have attempted to evaluate the reliability and reproducibility of these classification systems, some of which have better reproducibility and reliability than others. Since Cohen established the kappa coefficient in 1960, it has been widely used to assess the level of agreement between observers.

Weber classified fractures of the ankle into categories A (fracture of the fibula is distal to the syndesmosis), B (fracture is at the level of the syndesmosis) and C (fracture is proximal to the syndesmosis). Thomsen et al showed that there was an acceptable inter- and intra-observer agreement of both the Weber and the Lauge-Hansen classification systems using the kappa coefficient. However, several studies have shown that the use of the kappa coefficient value alone is not sufficient to assess agreement between multiple observers.

We have assessed the inter- and intra-observer agreement for the Weber classification using standard anteroposterior and lateral radiographs of the ankle rather than the frontal, lateral and oblique views as mentioned in the study of Thomsen et al.

Patients and Methods

Using the hospital coding system, 50 patients with ankle fractures were randomly selected. The first author (IAM) reviewed the radiographs, and those which did not show a definite fibular fracture were excluded. Patients who had special radiographs such as stress views or further imaging such as CT scans were also excluded.

The selected radiographs were blinded and then reviewed by five different observers with different levels of clinical experience. Two were consultants (including NHH) and three were orthopaedic registrars (including BM) in their second, fourth and sixth years of training. The observers were asked to classify the fractures according to the Weber classification. The first author (IAM) and a co-author (AMM) recorded the responses from the observers. To minimise bias they did not take part in the process of classification. After an interval of four weeks, the observers were again requested to classify the same set of radiographs, which had been mixed in order to minimise any chance of recollection.

Statistical analysis. The proportion of agreement (observed agreement), the kappa statistics for agreement between multiple observers and those for intra-observer reproducibility were calculated using Statsdirect software (Statsdirect Ltd, Sale, United Kingdom). The observed agreement is the portion of cases for
which the observers agree and is defined as the number of occasions of complete agreement divided by the total number of occasions. The expected agreement is the probability that two persons will provide the same response to a question for any given patient (chance agreement).

The kappa coefficient is the observed agreement which is above and beyond that due to chance, as follows:

\[
\text{Kappa} = \frac{(\text{observed agreement} - \text{expected agreement})}{(1 - \text{expected agreement})}
\]

A kappa value of 1.00 means perfect agreement and that of 0.00, agreement equal to that of chance alone. A negative kappa value implies agreement worse than that of chance alone. Landis and Koch\(^1\) have characterised different ranges of values for kappa with respect to suggested degrees of agreement. A kappa coefficient value of less than 0.00 suggests poor agreement, 0.00 to 0.20 slight agreement, 0.21 to 0.40 fair, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial and 0.81 to 1.00 almost perfect agreement.

### Results

**Inter-observer reliability.** The proportion of agreement on the two occasions was 76% (64% to 84%) and 79% (64% to 90%), respectively, with an overall proportion of agreement value of 78% (76% to 79%).

The generalised kappa coefficient values for multi-observer agreement on the two occasions were as follows:

1. Kappa\(_1\) = 0.59 (SEM 0.045), 95% confidence interval (CI) 0.50 to 0.68, \(p < 0.00001\).
2. Kappa\(_2\) = 0.63 (SEM 0.049), 95% CI 0.53 to 0.73, \(p < 0.00001\).

The mean of the generalised kappa value for multi-observer agreement was 0.61 (0.59 to 0.63) which implies substantial agreement.

Kappa coefficients for category-specific agreement for subgroups were: type A, 0.44; type B, 0.60; and type C, 0.67 which means moderate, moderate and substantial agreement, respectively.

**Intra-observer reproducibility.** Intra-observer agreement was substantial with a kappa coefficient value of 0.74 (0.39 to 0.86) and a mean proportion of agreement of 85% (60% to 93%). Four observers had almost perfect agreement, except for the junior orthopaedic registrar (observer 2; Table I).

### Discussion

Thomsen et al\(^9\) assessed multiobserver agreement of the Weber\(^1\) and Lauge-Hansen\(^1\) classifications based on frontal, lateral and oblique views. Four observers reviewed 94 sets of radiographs at an interval of three months. As far as the Weber classification was concerned, the observers were asked to classify the radiographs into four categories: A, B, C and ‘non-classifiable’. Inter-observer agreement was found to be acceptable with agreement of 0.737 (73.7%) and 0.740 (74%) on two occasions, and kappa values of 0.58 and 0.56, respectively.\(^9\) The observed agreement for intra-observer variation was 0.732 to 0.856 with a kappa value of 0.60 to 0.76 for four observers.\(^9\) We feel that the addition of the fourth category of ‘not-classifiable’ affected the results inappropriately, since it was not part of the Weber classification.

There were two reasons for repeating this study, first to measure the impact of using standard anteroposterior and lateral radiographs alone on the level of agreement and, secondly, to eliminate the category of ‘not-classifiable’ by selecting only ankles which had a definite fibular fracture.

In our study inter-observer agreement was substantial with an observed agreement of 76% and 79% and kappa coefficient values of 0.59 and 0.63 on two occasions. For intra-observer agreement overall agreement was 60% to 93% with kappa values of 0.39 to 0.86. Four observers had almost perfect agreement. The most junior orthopaedic registrar was the exception. Overall, these results were comparable to those of Thomsen et al.\(^9\)

A review of the orthopaedic literature showed that these values were higher than those achieved by many other classification systems which have been assessed (Table II).

The validity of the kappa coefficient in assessing multiobserver agreement requires discussion. Several studies have described problems and paradoxes with the kappa coefficient.\(^13\)–\(^16\) Sometimes, a low kappa value is possible despite higher observed agreement and this is due to an ‘unfair’ correction of chance agreement. In simple terms, a low kappa coefficient value does not necessarily mean that there is poor agreement. To overcome this phenomenon, Feinstein and Cicchetti\(^13\) suggested the use of observed agreement alone without imposing the unfair corrections

### Table I. Intra-observer agreement for each observer

<table>
<thead>
<tr>
<th>Observer</th>
<th>Kappa</th>
<th>SEM</th>
<th>95% CI*</th>
<th>Percentage of agreement</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.80</td>
<td>0.12</td>
<td>0.56 to 1.00</td>
<td>90</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>2</td>
<td>0.39</td>
<td>0.09</td>
<td>0.21 to 0.58</td>
<td>60</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>3</td>
<td>0.86</td>
<td>0.12</td>
<td>0.61 to 1.00</td>
<td>93</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>4</td>
<td>0.86</td>
<td>0.11</td>
<td>0.63 to 1.00</td>
<td>92</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>5</td>
<td>0.80</td>
<td>0.11</td>
<td>0.56 to 1.00</td>
<td>89</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

* 95% CI, 95% confidence interval

### Table II. Review of studies on the agreement of classification systems for fractures

<table>
<thead>
<tr>
<th>Author/s</th>
<th>Classification system</th>
<th>Kappa coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomsen et al(^1)</td>
<td>Garden’s femoral neck</td>
<td>0.39</td>
</tr>
<tr>
<td>Kreder et al(^2)</td>
<td>AO: distal radius</td>
<td>0.48</td>
</tr>
<tr>
<td>Siebenrock and Gerber(^3)</td>
<td>AO: proximal humerus</td>
<td>0.53</td>
</tr>
<tr>
<td>Siebenrock and Gerber(^3)</td>
<td>Neer: proximal humerus</td>
<td>0.40</td>
</tr>
<tr>
<td>Swiontkowski et al(^4)</td>
<td>AO/OTA: pilon fracture</td>
<td>0.57</td>
</tr>
<tr>
<td>Brumbach and Jones(^5)</td>
<td>Gustillo: compound fracture</td>
<td>0.60</td>
</tr>
<tr>
<td>Sidor et al(^6)</td>
<td>Neer: proximal humerus</td>
<td>0.52</td>
</tr>
<tr>
<td>Smith et al(^7)</td>
<td>Ficat: osteonecrosis of the head of the femur</td>
<td>0.46</td>
</tr>
</tbody>
</table>

* AO, Arbeitsgemeinschaft für Osteosynthesefragen; OTA, Orthopaedic Trauma Association
for chance associated with the kappa coefficient when assessing multiobserver agreement. The reason for our inclusion of the kappa value was to compare our results with those of other similar studies applied to different classification systems.

The interpretation of the kappa value for poor, moderate or strong agreement is also controversial. Two common systems are used in the medical literature. The criteria of Landis and Koch\textsuperscript{17} have been described earlier. Svanholm et al\textsuperscript{18} had different criteria and stated that a kappa coefficient value of greater than 0.75 suggested excellent agreement, while that of less than 0.50 indicated poor agreement.

We suggest that future studies to test the reliability and reproducibility of any classification system should mention the observed agreement value along with the kappa coefficient value, and should state which system is used to reach the conclusion so that unnecessary confusion may be avoided.

We conclude that the Weber classification system of fractures of the ankle based on standard anteroposterior and lateral radiographs has substantial inter-observer and intra-observer reliability and reproducibility.

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