THE CATTERALL CLASSIFICATION OF PERTHES’ DISEASE:
AN ASSESSMENT OF RELIABILITY

FINN CHRISTENSEN, KJELD SØBALLE, ROLF EJSTED, TORBEN LUXHØJ

From Kolding Hospital, Denmark

The reliability of the Catterall grouping of Perthes’ disease was examined by determining the agreement between pairs of observers using weighted kappa statistics. Anteroposterior and lateral radiographs of 100 hip joints were grouped independently by four experienced observers. There was a low, and in our opinion, unacceptable degree of inter-observer agreement even when Groups 2 and 3 were combined.

In 1971, Catterall classified Perthes’ disease into four groups in an attempt to improve selection for treatment and to provide a prognosis earlier in the course of the disease. A number of authors have used the Catterall classification in the interpretation of their work (Kamhi and MacEwen 1975; Lloyd-Roberts, Catterall and Salamon 1976; Harrison, Turner and Smith 1982; Reinker and Larsen 1983) but only one previous study of the degree of inter-observer error has been reported (Hardcastle et al. 1980).

Our study was undertaken to determine the reliability of the Catterall grouping by examining the agreement between pairs of observers.

MATERIAL AND METHODS
Radiographs of 100 hips with Perthes’ disease were studied. The films had been taken at the time of diagnosis, and were selected because good quality anteroposterior and lateral views were available. Of the four observers taking part, three (A, B and C), had each had more than five years’ experience in orthopaedic surgery, while the fourth (D) had had two years’ experience. All four had studied the relevant literature and had been trained in using the Catterall classification. The radiographs were studied independently by each observer and were grouped. “Head at risk” signs were not recorded.

Statistical methods. The reliability of the Catterall grouping was expressed in terms of inter-observer agreement in pairs and calculated using weighted kappa statistics (Cohen 1968). This statistical method, as compared with the more commonly used chi-square method, has the advantages that the degree of disagreement is taken into account, and that allowance is made for chance agreement. Weighted kappa values can vary from −1.0 (complete disagreement) through 0.0 (chance agreement) to +1.0 (complete agreement). The degree of disagreement between pairs of observers in assigning the group of each case was given a disagreement weight, \( v \), which was 0, 1, 2 or 3. Weighted kappa was then calculated from the formula:

\[
\kappa_w = 1 - \frac{\sum v_i \times p_{oi} - \sum v_i \times p_{oi}}{\sum v_i \times p_{oi} - \sum v_i \times p_{oi}}
\]

where \( p_{oi} \) is the observed and \( p_{oi} \) the chance agreement in each group.

RESULTS
All observers classified the largest number of hips as being in Group 2 and the fewest in Group 4 (Table I).

<table>
<thead>
<tr>
<th>Groups</th>
<th>Observers</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>

The assessments of observers A and B are set out in full in Table II as an example of the disagreements. The percentage of inter-observer agreement varied between 64% and 79%, whereas kappa values varied from 0.50 to 0.67. (Table III). Considering each Catterall group separately we found that the observed agreement (in relation to the agreement expected by chance) was higher in Groups I.
DISCUSSION

Catterall’s classification of hip joints with Perthes’ disease has been used widely to help provide a prognosis and indications for treatment. Despite this only one paper has dealt with the degree of agreement between different observers (Hardcastle et al. 1980). In our opinion, Hardcastle’s investigation fails to evaluate fully the reliability of Catterall’s grouping. In the first place the authors do not compare the agreement of observers in pairs, and in the second place they use percentage values and take no account of chance agreement.

We found that inter-observer agreement between pairs of observers ranged from 0.50 to 0.67; this level of agreement was not improved when Groups 2 and 3 were considered as one. Agreement of this order is, in our opinion, quite unacceptable, and it is questionable whether the Catterall classification should form the basis for decisions on treatment unless the radiological hallmarks of the various groups can be more clearly defined.

REFERENCES


Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70:213-20.


