Inter- and intra-observer variability of the Crowe and Hartofilakidis classification systems for congenital hip disease in adults

C. K. Yiannakopoulos, A. Chougle, A. Eskelinen, J. P. Hodgkinson, G. Hartofilakidis

From the Orthopaedic Department, University of Athens, Greece

Our study evaluated the reliability of the Crowe and Hartofilakidis classification systems for developmental dysplasia of the hip in adults. The anteroposterior radiographs of the pelvis of 145 patients with 209 osteoarthritic hips were examined twice by three experienced hip surgeons from three European countries and the abnormal hips were rated using both classifications. The inter- and intra-observer agreement was calculated.

Interobserver reliability was evaluated using weighted and unweighted kappa coefficients and for the Crowe classification, among the three pairs there was a minimum kappa coefficient with linear weighting of 0.90 for observers A and C and a maximum kappa coefficient of 0.92 for observers B and C. For the Hartofilakidis classification, the minimum kappa value was 0.85 for observers A and B, and the maximum value was 0.93 for observers B and C. With regard to intra-observer reliability, the kappa coefficients with linear weighting between the two evaluations of the same observer ranged between 0.86 and 0.95 for the Crowe classification and between 0.80 and 0.93 for the Hartofilakidis classification.

The reliability of both systems was substantial to almost perfect both for serial measurements by individual readers and between different readers, although the information offered was dissimilar.

Several classification systems for developmental dysplasia of the hip have been proposed. The most popular are those of Crowe et al and Hartofilakidis et al (Table I). The former relies on the assumption that the normal ratio of the diameter of the femoral head to the height of the pelvis is 1:5 and that any proximal migration of the femoral head can be expressed either as a percentage of the height of the pelvis or of the height of the femoral head-neck junction. It relies on the determination of four landmarks on an anteroposterior (AP) radiograph of the whole pelvis (Fig. 1). The vertical distance of the femoral head-neck junction from the interteardrop line which connects the lower part of the radiological teardrops is defined as the femoral head height. Four types of dislocation of the femoral head (I to IV) are described and while it is assumed that there is a direct relationship between the severity of the hip disease and the degree of proximal migration of the femoral head, acetabular involvement is not considered.

The classification system of Hartofilakidis et al recognises three types of congenital hip disease: dysplasia and low and high dislocation (Fig. 2). It also recognises three types of acetabulum, as verified during total hip replacement. Our study was designed to assess the inter- and intra-observer agreement of both classification systems.

Patients and Methods

We reviewed 145 AP radiographs of the pelvis from adult patients with 209 osteoarthritic hips secondary to developmental dysplasia, selected from the senior author’s (GH) digital database. We excluded those in which the whole pelvis was not shown. The radiographs were randomly assigned a number between 1 and 145 and evaluated twice by three experienced hip surgeons (CKY, AC, AE), performing more than 200 cases every year, from different countries (England, Finland and Greece). They were familiar with the two classifications and had used them in clinical practice or in research. Before embarking on the study each reviewer was provided with the same descriptions and diagrams of the two systems and with a DVD containing all the radiographs. They were asked to rate all affected hips independently using both systems. All the identification labels were covered. A second DVD was sent to each observer one month later with the radiographs presented in a different order and they were asked to perform a
new evaluation. The observers were blinded to the previous order of the radiographs and assessment was performed independently. Interobserver variability was measured by comparing the ratings of all observers on each occasion, while intra-observer reliability was determined by comparing the two assessments of each observer.

**Sample-size estimation.** The intraclass correlation coefficient is often used as an index of reliability in a measurement study. It may be thought of as the correlation between any two observations made on the same subject. When this correlation is high, the observations on a subject tend to match, and the measurement reliability is high. Calculation of the sample size was based on the primary outcome with the aim of showing a reliability which was at least substantial (kappa value > 0.7) and the power was set to 80%. Additionally, the minimum value of kappa deemed to be clinically important was 0.4 with a proportion of positive ratings of 0.9. The sample size required for two raters was 180 according to Sim and Wright.

**Statistical analysis.** Assessment of inter- and intra-observer consistency was accomplished by the use of the kappa coefficient (κ) as proposed by Fleiss. Interobserver agreement was assessed by calculating kappa coefficients for each possible pair of observers. Intra-observer agreement was assessed by calculating kappa coefficients for each pair of evaluations.

According to Landis and Koch, agreement was graded as slight (κ = 0 to 0.2), fair (κ = 0.21 to 0.40), moderate (κ = 0.41 to 0.60), substantial (κ = 0.61 to 0.80) or almost perfect (κ = 0.81 to 1.0).

### Table I. Description of the Crowe1 and Hartofilakidis3 classification systems of congenital hip disease in adults

<table>
<thead>
<tr>
<th>Classification</th>
<th>Type</th>
<th>Description</th>
<th>Anatomy of the acetabulum as verified during surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowe</td>
<td>I</td>
<td>Proximal displacement &lt; 0.1% of pelvic height or less than 50% subluxation</td>
<td>Segmental deficiency of the superior wall</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>Displacement of 0.10% to 0.15% or subluxation 50% to 75%</td>
<td>Secondary shallowness due to fossa-covering osteophyte</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>Displacement of 0.15% to 0.20% or subluxation 75% to 100%</td>
<td>Anterior and posterior segmental deficiency</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>Displacement &gt; 0.20% or subluxation &gt; 100%</td>
<td>Narrow opening and inadequate depth of the true acetabulum</td>
</tr>
<tr>
<td>Hartofilakidis</td>
<td>Dysplasia</td>
<td>The femoral head is contained within the original acetabulum despite the degree of subluxation</td>
<td>Segmental deficiency of the entire acetabulum with narrow opening</td>
</tr>
<tr>
<td></td>
<td>Low dislocation</td>
<td>The femoral head articulates with a false acetabulum which partially covers the true acetabulum to a varying degree</td>
<td>Inadequate depth</td>
</tr>
<tr>
<td></td>
<td>High dislocation</td>
<td>The femoral head is completely out of the true acetabulum and migrated superiorly and posteriorly to a varying degree</td>
<td>Excessive anteversion</td>
</tr>
</tbody>
</table>

Radiograph showing measurements for the Crowe classification system (A, vertical distance between the reference interteardrop line (line 1) and the head-neck junction (line 2); B, vertical distance between the line connecting the ischial tuberosities (line 3) and the line connecting the iliac crests (line 4).
All tests were two-sided and statistical significance was set at $p \leq 0.05$. All analyses were carried out using the statistical package SPSS version 13.0 (SPSS Inc., Chicago, Illinois) and the STATA statistical software (StataCorp LP, College Station, Texas).

**Results**

All hips were classified and the agreement among the reviewers is shown in Table II. At the first examination, all agreed on the rating in 83.73% of cases with regard to the Crowe classification and in 84.68% with regard to the Hartofilakidis classification. At the second examination agreement was 78.47% and 79.9%, respectively. In no case was there complete disagreement among the reviewers.

In the first evaluation, paired comparisons showed an interobserver unweighted kappa coefficient ranging from 0.84 to 0.86 for the Crowe classification and 0.79 to 0.90 for the Hartofilakidis classification (Table III). The kappa coefficient with linear weighting ranged from 0.90 to 0.92 for the Crowe classification and 0.85 to 0.93 for the Hartofilakidis classification.

All radiographs showing the Hartofilakidis classification system describing three types of increasing severity of the deformity, i.e. dysplasia, low and high dislocation. a) In type-A deformity (dysplasia) the femoral head is contained within the original acetabulum, b) in type-B deformity (low dislocation) it articulates with a false acetabulum which partially covers the true acetabulum and c) in type-C deformity (high dislocation) the femoral head has migrated superiorly and posteriorly to the hypoplastic true acetabulum.
In the second evaluation, paired comparisons showed an interobserver unweighted kappa coefficient ranging from 0.71 to 0.82 for the Crowe classification and 0.66 to 0.77 for the Hartofilakidis classification (Table III). The kappa coefficient with linear weighting ranged from 0.83 to 0.90 for the Crowe classification and 0.75 to 0.84 for the Hartofilakidis classification.

The unweighted kappa coefficients between the two evaluations of the same observer ranged from 0.77 to 0.91 for the Crowe classification and 0.76 to 0.91 for the Hartofilakidis classification. The kappa values with linear weighting were 0.86 to 0.95 for the Crowe classification and 0.80 to 0.93 for the Hartofilakidis classification (Table IV).

Discussion

The results of our study confirm that both the Crowe and the Hartofilakidis classification systems show excellent reliability, as judged by the inter- and intra-observer variability. They are simple to use and are reliable and reproducible. Several factors may affect inter- and intra-observer reliability such as the experience of the raters, ambiguity of the system, interval between readings and additional imaging modalities. The use of standard evaluation forms and radiographs may reduce the extent of variability. Similarly, agreement may be improved by using advanced imaging techniques, such as CT or MRI.

The limitations of the Crowe classification are the requirement of a radiograph of the whole pelvis and the variability of locating the femoral head-neck junction with rotation of the limb. The main limitation of the Hartofilakidis classification is the difficulty in classifying borderline cases of dysplasia and low dislocation, and low and high dislocation.

The Crowe classification describes the proximal migration of the femoral head regardless of the acetabular deformity and assumes that there is a direct relationship between the amount of migration and the severity of disease. Although this hypothesis is usually valid this is not always so. The Hartofilakidis classification relies on the anatomy of the acetabulum as encountered during surgery. If there is difficulty in classifying borderline cases, CT can clarify the anatomy of the acetabulum and the amount of migration is not a definite criterion by which to judge the type of dysplasia.

Both classifications were compared recently by Decking et al. Three observers of different experience (one consultant, one resident and one medical student) evaluated the radiographs of 51 patients (62 hips) on two occasions. A high inter- and intra-observer reliability of both systems was demonstrated. In our study a significantly higher number of hips was included (209 hips) and the three raters were experienced hip surgeons who understood the two systems and were more competent to estimate the bone deficiencies encountered intra-operatively. The reliability of a classification system depends on the consistency of measurements or observations and relates to the quality of the measurement or observation. It also depends on the extent of agreement between repeated measurements. A reliable system classifies a disease or fracture consistently, but does not necessarily show what is happening. A valid classification system gives the underlying pathology of the disease or configuration of the fracture and is therefore closer to the ‘truth’ of what is being measured or described. Reliability and validity are not independent, but are related. A method may be reliable if it measures something consistently, but is valid only if the results of measurement approximates to the true value. A classification system therefore may be reliable but not valid and cannot be valid without being reliable. Systems relying on poorly-described features may inherently perform poorly on reliability analyses because the observers cannot identify them accurately. The Hartofilakidis system can predict in a reliable and reproducible manner the bone deficiencies encountered during total hip replacement from pre-operative radiographs.

The interpretation of kappa values is based on convention. Landis and Koch proposed that a value > 0.80 represented excellent agreement between observers, but Fleiss and Svanholm et al suggested that a value > 0.75 indicated excellent agreement. In our study both systems showed high inter- and intra-observer agreement, with weighted kappa values > 0.75. An inherent characteristic of the evaluation of classification systems using the kappa coefficient is that it depends on the prevalence of each type of classification, that is, more common types are expected to show greater agreement. Also, kappa depends on the

<table>
<thead>
<tr>
<th>Observer pair</th>
<th>Crowe classification</th>
<th>Hartofilakidis classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unweighted kappa coefficient</td>
<td>Kappa with linear weighting</td>
</tr>
<tr>
<td>A/A</td>
<td>0.77 (0.03)</td>
<td>0.86 (0.02)</td>
</tr>
<tr>
<td>B/B</td>
<td>0.90 (0.02)</td>
<td>0.95 (0.01)</td>
</tr>
<tr>
<td>C/C</td>
<td>0.91 (0.02)</td>
<td>0.95 (0.01)</td>
</tr>
</tbody>
</table>
number of categories in each system, with a tendency for a lower value as the number of categories increases, as in the Crowe system. The interobserver variation must be considered when designing protocols for multicentre clinical trials.

In conclusion, both systems are reproducible and repeatable for experienced observers and can be used for the evaluation of the dysplastic hip. The quantitative Crowe classification does not require knowledge of the anatomy of a dysplastic hip whereas the qualitative Hartofilakis classification provides insight into the structural changes which will be encountered at operation. The two systems evaluate the hip from a different perspective and can be combined for research or operating purposes.

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