Comparison of the inter- and intra-observer reproducibility of the Crowe, Hartofilakidis and modified Cochin classification systems for the diagnosis of developmental dysplasia of the hip

A. Clavé a, L. Kerboull b, T. Musset c, X. Flecher d, D. Huten e, C. Lefèvre a, F. Gaucher f, E. Stindel a,∗, The French Hip & Knee Society (SFHG) g

a Service de chirurgie orthopédique et traumatologique, faculté de médecine, université de Bretagne-Occidentale, hôpital la Cavale-Blanche, CHRU de Brest, boulevard Tanguy-Prigent, 29200 Brest, France
b Centre médico-chirurgical Paris V, 36, boulevard Saint-Marcel, 75005 Paris, France
c Clinique mutualiste de la Porte-de-l'Orient, 3, rue Robert-de-la-Croix, 56324 L'Orient, France
d Institut du mouvement et de l'appareil locomoteur, faculté de médecine, université d'Aix-Marseille, hôpital Sainte Marguerite, Marseille, France
e Service de chirurgie orthopédique, réparatrice et traumatologique, faculté de médecine Rennes, université de Rennes, hôpital Sud, CHRU de Rennes, 16, boulevard de Bulgarie, 35203 Rennes, France
f Service de chirurgie orthopédique, Hôpital-Dieu-de-Pont-l'Abbé, rue Signor, 29120 Pont-l'Abbé, France
g 56, rue Boissonade, 75014 Paris, France

A R T I C L E  I N F O

Article history:
Accepted 22 July 2014

Keywords:
Hip
Dysplasia
Classification
Congenital hip disease

A B S T R A C T

Introduction: Developmental dysplasia of the hip (DDH) leads to multiple treatment challenges during adulthood. Surgical treatment is mainly based on radiographic evaluation of the anatomical alterations. Several classification systems have been described in the published English scientific literature, but the French Cochin classification has not been used very much. Its primary advantage lies in its ability to intricately describe the DDH alterations with a large number of grades. We hypothesized that the inter- and intra-observer reproducibility of the SOFCOT-modified Cochin classification system was equal to that of the Crowe and Hartofilakidis classifications.

Material and methods: Five French orthopaedic surgeons who were DDH experts classified 94 A/P pelvis radiographs (179 hips) using the Crowe (Cr), Hartofilakidis (Ha) and modified Cochin (Co) systems. This evaluation was repeated a second time one month later. The intra-observer reproducibility was determined with weighted Kappa and concordance coefficients. The inter-observer reproducibility was performed by calculating the multirater Kappa coefficient on each of the two data series.

Results: For the intra-observer reliability, the average weighted concordance coefficients (95% CI) were 88.62–94.52 for Cr, 89.43–93.80 for Ha and 92.14–95.71 for Co. The average weighted Kappa coefficients (95% CI) were 0.70–0.85 for Cr, 0.67–0.82 for Ha and 0.75–0.83 for Co. For the inter-observer reliability, the Kappa for each assessment round was 0.57 and 0.48 for Cr, 0.43 and 0.44 for Ha, and 0.43 and 0.37 for Co.

Discussion: The intra- and inter-observer reliability for the modified Cochin classification system is the same as the one for the Crowe and Hartofilakidis classifications. The theoretical advantage of this classification system should be confirmed by comparing the findings with intra-operative anatomical observations.
Level of proof, type of study: IV.

© 2014 Published by Elsevier Masson SAS.

1. Introduction

Development dysplasia of the hip (DDH) in adults, which is becoming less prevalent because of better prevention [1], is a condition where treatment requires complex surgery. The surgical treatment must take into account the combination of muscular and bone (femur, acetabulum) abnormalities [2–5]. Pre-operative
evaluation of these alterations is a key component when defining the treatment strategy. This will help the surgeon anticipate the intra-operative challenges and potential complications, determine which implants to use and to a lesser degree, predict the expected outcome [2,3,6–8]. This evaluation uses radiographic classification systems that define groups of typical alterations. These are either objective (based on measuring a radiographic index) or subjective (based on evaluating descriptive anatomical elements). The Crowe (objective) and Hartofilakidis (subjective) classification systems are used most commonly in published English-language scientific publications. Although the reproducibility of these classifications has been validated [9–11], their ability to predict surgical problems has been questioned by many authors [7,12,13]. Alternative classification systems include the one proposed and evaluated by Gaston et al. in 2009 [14] and the one published by Kerboull back in 1987 [2,15,16]. The latter has also been called the Cochin classification. Its five-level version was modified by the SOFCOT (French Society of Orthopaedic and Trauma Surgery) in 2012, but its reproducibility has never been evaluated nor compared to the commonly-used Crowe and Hartofilakidis classifications.

We hypothesized that the inter- and intra-observer reproducibility of the SOFCOT-modified Cochin classification system was equal to that of the Crowe and Hartofilakidis classifications.

2. Material and methods

The radiography databases from five French hospitals were used to select a group of standing A/P pelvis views. Radiographs were eligible to be selected if they were from adult patients with hip dysplasia that had never been surgically treated. The following inclusion and exclusion criteria were used.

Inclusion criteria:

- radiographs with signs of DDH as defined by Crowe on at least one hip.

Exclusion criteria:

- radiographs with no signs of DDH;
- radiographs that do not show the entire pelvis (anterosuperior iliac spines to ischium);
- radiographs not taken in full frontal view (defined as asymmetry of the iliac crests and obturator foramen and/or coccyx projection that is not centred relative to the pubic symphysis).

To avoid recall bias, the sequence in which the radiographs were analysed was randomly set by making up two reading lists (list A and list B).

All of the following documents were saved to DVD and sent to five French orthopaedic surgeons throughout France who are DDH experts (R1, R2, R3, R4, R5):

- reading list A;
- digitized version of the radiographs;
- description of the three classification systems (Cr, Ha, Co)–(Figs. 1–3 and Tables 1 and 2);
- sheet used for standardized recording of the grades.

Each analysable hip was classified in the three classification systems (Cr, Ha, Co) by each of the five surgeons in the order shown in list A. The recording sheets were frozen and then the same method reapplied 30 days later using list B. The classification results were combined into a single file for statistical analysis (STATA 12.1, StataCorp, 4905 Lakeway Drive, Texas, USA).

The intra-observer reproducibility was evaluated by calculating two weighed coefficients: concordance and Kappa (as defined by

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>R &lt; 0.1</td>
</tr>
<tr>
<td>II</td>
<td>0.1 &lt; R &lt; 0.15</td>
</tr>
<tr>
<td>III</td>
<td>0.16 &lt; R &lt; 0.2</td>
</tr>
<tr>
<td>IV</td>
<td>R &gt; 0.2</td>
</tr>
</tbody>
</table>

Table 1
Crowe classification.
Table 2
SOFCOT-modified Cochin classification.

<table>
<thead>
<tr>
<th>Cochin grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dysplasia: centred in the true acetabulum with or without break in Shenton’s line</td>
</tr>
<tr>
<td>2</td>
<td>Subluxation: break in Shenton’s line and femoral head partially outside true acetabulum</td>
</tr>
<tr>
<td>3</td>
<td>Anterior or anterolateral dislocation or low fixed dislocation: in front and above true acetabulum, but some overlap and sometimes close to subluxation limit</td>
</tr>
<tr>
<td>4</td>
<td>Intermediate dislocation or high fixed dislocation: balanced on true acetabulum, between this structure and gluteal tuberosity</td>
</tr>
<tr>
<td>5</td>
<td>Posterior dislocation or high non-fixed dislocation: backwards, high in the buttock, often no contact with ilium</td>
</tr>
</tbody>
</table>

Table 3
Weighing matrix for calculation of weighted Kappa coefficient.

<table>
<thead>
<tr>
<th></th>
<th>1.0000</th>
<th>0.6667</th>
<th>0.3333</th>
<th>0.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6667</td>
<td>0.3333</td>
<td>0.0000</td>
</tr>
<tr>
<td>0.6667</td>
<td>0.6667</td>
<td>1.0000</td>
<td>0.3333</td>
<td>0.0000</td>
</tr>
<tr>
<td>0.3333</td>
<td>0.3333</td>
<td>0.6667</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Cohen [17]). An example of the weighing matrix used to take into account the magnitude of the disagreements is shown in Table 3.

The inter-observer reproducibility was evaluated by calculating the multirater Kappa coefficient on each of the two data series [18]. Table 4 shows the coefficients for each of the two data series.

The Kappa coefficients were interpreted according to the recommendations compiled in Table 5 (based on Landis and Koch [19]).

3. Results

The data set consisted of 94 preoperative standing A/P pelvis radiographs (179 hips) that were classified twice by five surgeons (raters).

For each of the three classification systems:

Table 4
Measurement of intra-rater reproducibility.

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowe</td>
<td>89.11</td>
<td>95.25</td>
<td>90.69</td>
<td>90.36</td>
<td>92.46</td>
</tr>
<tr>
<td>Weighted Kappa</td>
<td>0.70</td>
<td>0.87</td>
<td>0.78</td>
<td>0.75</td>
<td>0.80</td>
</tr>
<tr>
<td>Hartofilakidis</td>
<td>89.57</td>
<td>91.85</td>
<td>90.78</td>
<td>93.02</td>
<td>90.88</td>
</tr>
<tr>
<td>Weighted Kappa</td>
<td>0.65</td>
<td>0.80</td>
<td>0.78</td>
<td>0.76</td>
<td>0.75</td>
</tr>
<tr>
<td>Modified Cochin</td>
<td>92.85</td>
<td>94.53</td>
<td>95.81</td>
<td>92.18</td>
<td>94.30</td>
</tr>
<tr>
<td>Weighted Kappa</td>
<td>0.76</td>
<td>0.83</td>
<td>0.78</td>
<td>0.77</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 5
Interpretation of Kappa coefficient (based on Landis and Koch [19]).

<table>
<thead>
<tr>
<th>Weighed Kappa value</th>
<th>Interpretation of concordance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Poor</td>
</tr>
<tr>
<td>0.00–0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21–0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41–0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61–0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81–1</td>
<td>Nearly perfect</td>
</tr>
</tbody>
</table>

Table 6
Measurement of inter-rater reproducibility.

<table>
<thead>
<tr>
<th></th>
<th>List A</th>
<th>List B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowe</td>
<td>0.57</td>
<td>0.48</td>
</tr>
<tr>
<td>Hartofilakidis</td>
<td>0.43</td>
<td>0.44</td>
</tr>
<tr>
<td>Modified Cochin</td>
<td>0.43</td>
<td>0.37</td>
</tr>
</tbody>
</table>

- Table 4 shows the weighted concordance and Kappa coefficients for each observer;
- Table 6 shows the weighted multirater Kappa coefficient for the image series performed in the sequence defined in list A (column 2) and list B (column 3).

The average weighed concordance coefficients (95% confidence intervals) were the following:

- [88.62–94.52] for Cr;
- [89.43–93.80] for Ha;
- [92.14–95.71] for Co.

The average weighed Kappa coefficients (95% confidence intervals) were the following:

- [0.70–0.85] for Cr;
- [0.67–0.82] for Ha;
- [0.75–0.83] for Co.
4. Discussion

The results show that the modified Cochin classification has the same intra-observer reproducibility as the Crowe and Hartofilakidis classification systems. According to Landis and Koch [19], this reproducibility can be qualified as “substantial” for raters R1, R3 and R4 and “nearly perfect” for raters R2 and R5. The high number of grades in the Cochin classification does not decrease the intra-observer reproducibility.

In the current study, the Crowe and Hartofilakidis classifications had lower intra-observer reproducibility than in a previous study by Yiannakopoulos et al. [9] but better than in the Kose et al. [11] and Gaston et al. [14] studies. No matter which classification system was used, an expert DDH surgeon is able to grade the radiographs with at least “substantial” reproducibility. However, a surgeon’s ability to grade radiographs depends on experience. R1 always had lower scores than the other raters and R2 and R5 always had higher scores than the other four raters, no matter which classification system was used. Other than for Rater 2, all the raters obtained their best scores with the Cochin classification system, which makes it the most robust classification.

The modified Cochin classification, which has been described in its original format in several French journals and books [2,15,16,20], is a subjective classification taking into account femoral and acetabular abnormalities in the frontal and sagittal planes. The modified version used here has five grade levels; dysplasia and subluxation grades were added to the three stages described initially (anterior or low fixed dislocation, intermediate or high fixed dislocation, posterior or high non-fixed dislocation). This system allows for a discriminating classification of anatomical alterations. However, having a larger number of grades in a classification system increases the risk of poor reproducibility.

The analysis of inter-observer reproducibility showed that the concordance in the three classification systems varies between “moderate” and “fair.” The results for the intra-rater variability with five surgeons shown in Table 6 are logically lower than the ones reported by Hartofilakidis, where the variability between only two raters was analyzed. During this test, the Crowe classification had slightly better scores [9]. This can be explained by the use of an objective endpoint that corresponds to the sampling result. The phenomenon was only observed during the analysis of intra-rater concordance.

However, reproducibility is not the only criterion used to determine which classification system is better than the others—accuracy and predictive ability must also be taken into account. To our knowledge there are no published studies comparing the accuracy of the classification systems evaluated here to actual intra-operative anatomical findings. Predictive ability is a function of the number of organs considered (femur and/or acetabulum) and how well various three-dimensional displacements are taken into consideration. The Crowe classification would seem to be limited in this aspect, as it only takes into consideration femur displacement in the frontal plane, without taking into consideration either femoral or acetabulum morphological abnormalities. It only provides a limited assessment of the potential intra-operative challenges. The Hartofilakidis classification is based on a three-stage anatomical description of femoral ascension in the frontal plane relative to the false or true acetabulum and the bone morphology. As a consequence, it is better able to accurately predict intra-operative findings than the Crowe classification. However, it only captures femur displacements in one plane and only has three grades. The contents of each grade seem heterogeneous and do not allow a surgeon to differentiate between potentially simple or complex cases.

The Cochin classification seems to more accurately describe the changes in anatomical relationships present in DDH and to anticipate unique surgical features. Its limitations revolve around the narrow boundaries between two grades, with the requirement to have lateral radiographs to better evaluate the anteroposterior deformities and displacements.

5. Conclusion

The intra- and inter-observer reliability of the modified Cochin classification system is the same as the one for the Crowe and Hartofilakidis classifications. Its primary advantage lies in its ability to intricately describe the alterations with a large number of grades. This theoretical advantage in terms of preoperative planning must be confirmed by comparing the classification results with intra-operative anatomical findings.

Disclosure of interest

The authors declare that they have no conflicts of interest concerning this article.

References
