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.

■ C. K. Yiannakopoulos, MD, Consultant Orthopaedic Surgeon
Nikis 2, 14561, Athens, Greece.

■ A. Chougle, FRCS(Trauma & Orth), Consultant Orthopaedic Surgeon
North Manchester General Hospital, Crumpsall, Manchester M8 5RB, UK.

■ A. Eskelinen, MD, PhD, Consultant Orthopaedic Surgeon
Coxa Hospital for Joint Replacement, Biokatu 6b, 33250 Tampere, Finland.

■ J. P. Hodgkinson, FRCS, Consultant Orthopaedic Surgen
The Centre for Hip Surgery Wrightington Hospital, Hall Lane, Appley Bridge, Wigan WN6 9EP, UK.

■ G. Hartofilakidis, MD, FACS, Emeritus Professor of Orthopaedics
University of Athens Medical School, Fotiou Patriarchou 21, Athens 11471, Greece.

Correspondence should be sent to Dr C. K. Yiannakopoulos; e-mail: cky@ath.forthnet.gr

doi:10.1302/0301-620X.90B5.19724 $2.00

J Bone Joint Surg [Br] 2008;90-B:579-83.
Received 21 May 2007; Accepted after revision 14 January 2008

Several classification systems for developmental dysplasia of the hip have been proposed.[1-5] The most popular are those of Crowe et al[1] and Hartofilakidis et al[3] (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[3] 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.[4]

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.[4,6-8] 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

Table I. Description of the Crowe[1] and Hartofilakidis[3] classification systems of congenital hip disease in adults

Classification	Type	Description	Anatomy of the acetabulum as verified during surgery
Crowe	I	Proximal displacement < 0.1% of pelvic height or less than 50% subluxation	
	II	Displacement of 0.10% to 0.15% or subluxation 50% to 75%	
	III	Displacement of 0.15% to 0.20% or subluxation 75% to 100%	
	IV	Displacement > 0.20% or subluxation > 100%	
Hartofilakidis	Dysplasia	The femoral head is contained within the original acetabulum despite the degree of subluxation	Segmental deficiency of the superior wall Secondary shallowness due to fossa-covering osteophyte
	Low dislocation	The femoral head articulates with a false acetabulum which partially covers the true acetabulum to a varying degree	Complete absence of the superior wall Anterior and posterior segmental deficiency Narrow opening and inadequate depth of the true acetabulum
	High dislocation	The femoral head is completely out of the true acetabulum and migrated superiorly and posteriorly to a varying degree	Segmental deficiency of the entire acetabulum with narrow opening Inadequate depth Excessive anteversion Abnormal distribution of bone stock, mainly located superioposteriorly in relation to the true acetabulum

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.[9] 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)[10] 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.[9]

Statistical analysis. Assessment of inter- and intra-observer consistency was accomplished by the use of the kappa coefficient (κ) as proposed by Fleiss.[10] 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,[11] 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).

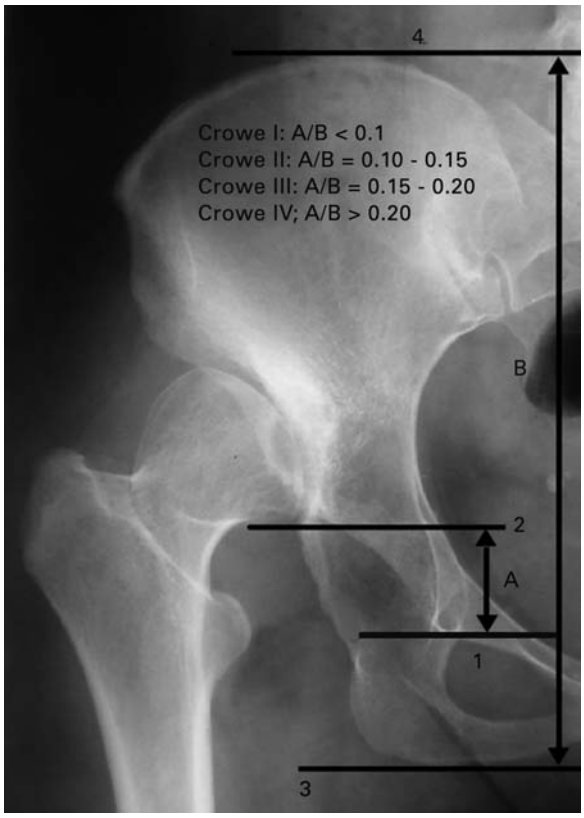

Fig. 1

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).

Fig. 2a

Fig. 2b

Fig. 2c

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.

Table II. Agreement among the reviewers with regard to the Crowe[1] and Hartofilakidis[3] classifications by number and *percentage*

	First examination		Second examination	
Agreement	**Crowe**	**Hartofilakidis**	**Crowe**	**Hartofilakidis**
All three reviewers	175 (*83.73*)	177 (*84.68*)	164 (*78.47*)	167 (*79.9*)
Two reviewers	34 (*16.26*)	32 (*15.31*)	45 (*21.53*)	42 (*20.1*)
No agreement	0	0	0	0

Table III. Paired comparisons of interobserver agreement for the Crowe and Hartofilakidis classifications giving the unweighted and linearly-weighted kappa coefficient values and the SEM in parentheses

	First evaluation			Second evaluation		
Observer pair	**Unweighted kappa coefficient**	**Kappa with linear weighting**	**Observed agreement (%)**	**Unweighted kappa coefficient**	**Kappa with linear weighting**	**Observed agreement (%)**
Crowe						
A/B	0.84 (0.03)	0.91 (0.01)	*89.5*	0.75 (0.04)	0.85 (0.02)	*82.0*
A/C	0.83 (0.03)	0.90 (0.02)	*89.8*	0.71 (0.04)	0.83 (0.02)	*79.7*
B/C	0.86 (0.03)	0.92 (0.02)	*91.0*	0.82 (0.03)	0.90 (0.02)	*88.0*
Hartofilakidis						
A/B	0.79 (0.03)	0.85 (0.02)	*86.6*	0.77 (0.03)	0.84 (0.02)	*87.0*
A/C	0.80 (0.03)	0.90 (0.01)	*87.0*	0.66 (0.04)	0.75 (0.03)	*77.5*
B/C	0.90 (0.02)	0.93 (0.01)	*93.8*	0.66 (0.04)	0.75 (0.03)	*77.5*

All tests were two-sided and statistical significance was set at p £ 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 Croweclassification 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.

Table IV. Intra-observer agreement for the Crowe and Hartofilakidis classifications giving the unweighted and linearly-weighted kappa coefficient values and the SEM in parentheses

	Crowe classification			Hartofilakidis classification		
Observer pair	Unweighted kappa coefficient	Kappa with linear weighting	Observed agreement (%)	Unweighted kappa coefficient	Kappa with linear weighting	Observed agreement (%)
A/A	0.77 (0.03)	0.86 (0.02)	*83.6*	0.84 (0.03)	0.80 (0.02)	*90.0*
B/B	0.90 (0.02)	0.95 (0.01)	*93.4*	0.76 (0.03)	0.83 (0.02)	*84.7*
C/C	0.91 (0.02)	0.95 (0.01)	*94.2*	0.91 (0.02)	0.93 (0.01)	*95.3*

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.[6]

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.[6] 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.[12-18] Several factors may affect inter- and intra-observer reliability such as the experience of the raters,[19] 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.[20]

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 theacetabular 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 anatomical variation.[21]

In some studies it had been assumed that Crowe types I and II correspond to Hartofilakidis type I, Crowe type III to Hartofilakidis type II and Crowe type IV to Hartofilakidis type III.[22,23] This assumption is not always valid because the anatomy of the acetabulum is variable 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.[24] 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.[25] 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.[26,27] 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.[28] 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.[26] 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[11] proposed that a value > 0.80 represented excellent agreement between observers, but Fleiss[10] and Svanholm et al[28] 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.[12] Also, kappa depends on the

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 Hartofilakidis 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

1. **Crowe JF, Mani VJ, Ranawat CS.** Total hip replacement in congenital dislocation and dysplasia of the hip. *J Bone Joint Surg [Am]* 1979;61-A:15-23.
2. **Eftekhar NS.** *Total hip arthroplasty.* St. Louis: CV Mosby, 1993:925.
3. **Hartofilakidis G, Stamos K, Ioannidis TT.** Low friction arthroplasty for old untreated congenital dislocation of the hip. *J Bone Joint Surg [Br]* 1988;70-B:182-6.
4. **Hartofilakidis G, Stamos K, Karachalios T, Ioannidis TT, Zacharakis N.** Congenital hip disease in adults: classification of acetabular deficiencies and operative treatment with acetabuloplasty combined with total hip arthroplasty. *J Bone Joint Surg [Am]* 1996;78-A:683-92.
5. **Kerboul M, Mathieu M, Sauzieres P.** Total hip replacement for congenital dislocation of the hip. In: Postel M, Kerboul M, Evrard J, Courpied JP. *Total hip replacement.* New York: Springer, 1987.
6. **Chougle A, Hemmady MV, Hodgkinson JP.** Severity of hip dysplasia and loosening of the socket in cemented total hip replacement: a long-term follow-up. *J Bone Joint Surg [Br]* 2005;87-B:16-20.
7. **Chougle A, Hemmady MV, Hodgkinson JP.** Long-term survival of the acetabular component after total hip arthroplasty with cement in patients with developmental dysplasia of the hip. *J Bone Joint Surg [Am]* 2006;88-A:71-9.
8. **Eskelinen A, Helenius I, Remes V, et al.** Cementless total hip arthroplasty in patients with high congenital hip dislocation. *J Bone Joint Surg [Am]* 2006;88-A:80-91.
9. **Sim J, Wright CC.** The kappa statistics in reliability studies: use, interpretation and sample size requirement. *Phys Ther* 2005;85:257-68.
10. **Fleiss JL.** *Statistical methods for rates and proportions.* Second ed. New York: John Wiley & Sons, 1981:218.
11. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
12. **Blundell CM, Parker MJ, Pryor GA, Hopkinson-Woolley J, Bhonsle SS.** Assessment of the AO classification of intracapsular fractures of the proximal femur. *J Bone Joint Surg [Br]* 1998;80-B:679-83.
13. **Flinkkilä T, Nikkola-Sihto A, Kaarela O, Pääkkö E, Raatikainen T.** Poor interobserver reliability of AO classification of fractures of the distal radius: additional computed tomography is of minor value. *J Bone Joint Surg [Br]* 1998;80-B:670-2.
14. **Humphrey CA, Dirschl DR, Ellis TJ.** Interobserver reliability of a CT-based fracture classification system. *J Orthop Trauma* 2005;19:616-22.
15. **Wainwright AM, Williams JR, Carr AJ.** Interobserver and intraobserver variation in classification systems for fractures of the distal humerus. *J Bone Joint Surg [Br]* 2000;82-B:636-42.
16. **Malek IA, Machani B, Mevcha AM, Hyder NH.** Inter-observer reliability and intra-observer reproducibility of the Weber classification of ankle fractures. *J Bone Joint Surg [Br]* 2006;88-B:1204-6.
17. **Brorson S, Bagger J, Sylvest A, Hrøbjartsson A.** Improved interobserver variation after training of doctors in the Neer system: a randomised trial. *J Bone Joint Surg [Br]* 2002;84-B:950-4.
18. **Bhattacharya R, Vassan UT, Finn P, Port A.** Sanders classification of fractures of the os calcis: an analysis of inter- and intra-observer variability. *J Bone Joint Surg [Br]* 2005;87-B:205-8.
19. **Beaulé PE, Dorey FJ, Matta JM.** Letournel classification for acetabular fractures: assessment of interobserver and intraobserver reliability. *J Bone Joint Surg [Am]* 2003;85-A:1704-9.
20. **Doornberg J, Lindenhovius A, Kloen P, et al.** Two and three-dimensional computed tomography for the classification and management of distal humeral fractures: evaluation of reliability and diagnostic accuracy. *J Bone Joint Surg [Am]* 2006;88-A:1795-801.
21. **Hartofilakidis G, Karachalios T.** Total hip arthroplasty for congenital hip disease. *J Bone Joint Surg [Am]* 2004;86-A:242-50.
22. **Rozykdal Z, Janicek P, Smid Z.** Total hip replacement with the CLS expansion shell and a structural femoral head autograft for patients with congenital hip disease. *J Bone Joint Surg [Am]* 2005;87-A:801-7.
23. **Jarosynski G, Woodgate IG, Saleh KJ, Gross AE.** Total hip replacement for the dislocated hip. *J Bone Joint Surg [Am]* 2001;83-A:272.
24. **Decking R, Brunner A, Decking J, Puhl W, Günther KP.** Reliability of the Crowe and Hartofilakidis classifications used in the assessment of the adult dysplastic hip. *Skeletal Radiol* 2006;35:282-7.
25. **Garbuz DS, Masri BA, Esdaile J, Duncan CP.** Classification systems in orthopaedics. *J Am Acad Orthop Surg* 2002;10:290-7.
26. **Greenfield ML, Kuhn JE, Wojtys EM.** A statistics primer: validity and reliability. *Am J Sports Med* 1998;26:483-5.
27. **Kocher MS, Zurakowski D.** Clinical epidemiology and biostatistics: a primer for orthopaedic surgeons. *J Bone Joint Surg [Am]* 2004;86-A:607-20.
28. **Svanholm H, Starklint H, Gundersen HJ, et al.** Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. *APMIS* 1989;97:689-98.