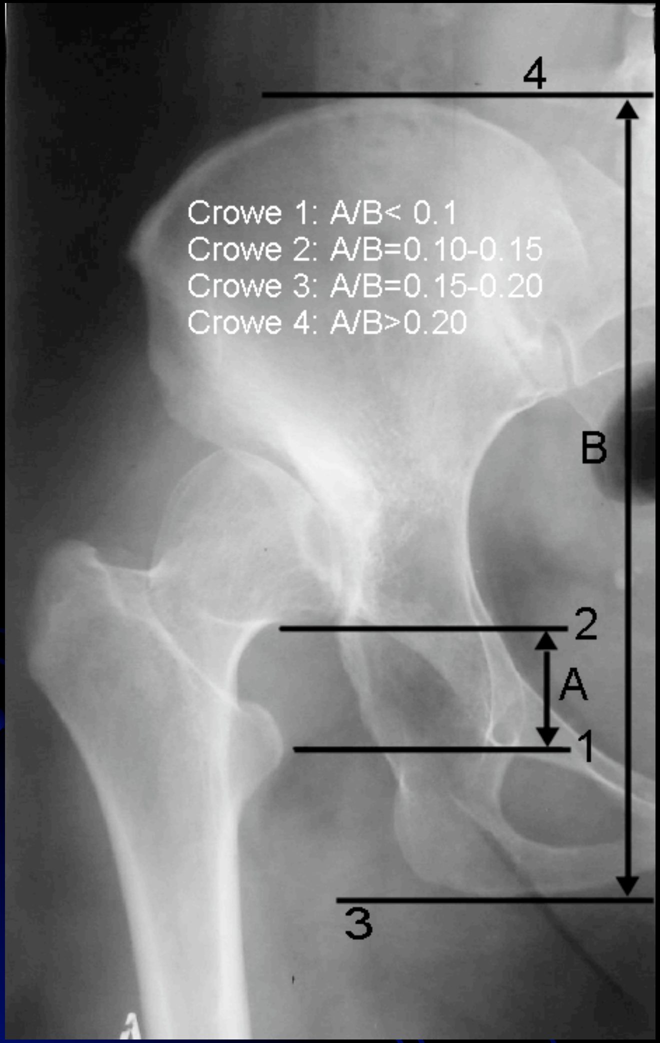
Inter- and intraobserver variability of the Crowe and Hartofilakidis classification systems for congenital hip disease in adults.

C.K. Yiannakopoulos, MD, A. Chougle, FRCS,
A. Eskelinen, MD, PhD, J.P. Hodgkinson, FRCS,
G. Hartofilakidis, MD, FACS

Orthopaedic Department, University of Athens, Athens, Greece,
Wrightington Hospital, Appley Bridge, England and

Department of Orthopaedics and Traumatology, Surgical Hospital, Helsinki
University Central Hospital, Helsinki, Finland

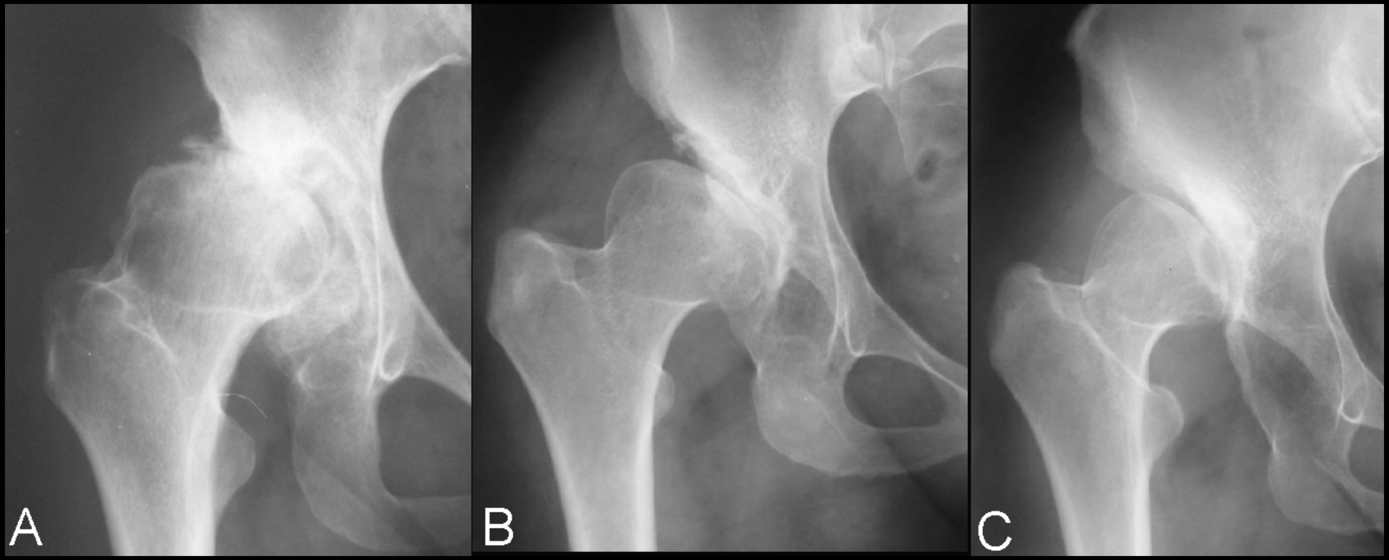
The Crowe et al. classification system.



A, vertical distance between the reference interteardrop line (line 1) and the head-neck junction (line 2).

B, the vertical distance between the line connecting the ischial tuberosities (line 3) and the line connecting the iliac crests (line 4).

The Hartofilakidis et al. system for the classification of CHD in adults



3 types of increasing severity of the deformity

Dysplasia, Low and High dislocation

Purpose

This study was designed to assess the reproducibility (inter- and intraobserver agreement) of the Crowe et al. and the Hartofilakidis et al. CHD classification systems.

Materials-Methods

- 145 AP pelvis radiographs
- 209 adult patients with OA hips secondary to CHD
- Randomly assigned a number between 1 and 145
- Evaluated twice by three surgeons from different European countries (England, Finland and Greece)
- All observers were familiar with the two CHD classification systems
- Prior to embarking on the study the reviewers were provided with the same classification descriptions and diagrams showing clearly the distinguishing features of each type and with a DVD containing all radiographs.

Materials-Methods

Independent rating of all affected hips using both systems
Second evaluation 2 months later.

Interobserver variability = comparison of the ratings of all observers at each time

Intraobserver reliability = comparison of the two assessments of each observer

Sample Size Estimation

Sample size calculation was based on the primary outcome with the aim of showing a reliability that was at least substantial (Kappa >0.7) and the power was set to 90%.

A sample size of n=three observers and k=20 x-rays per observer was calculated for each type of hip deformity (dysplasia, low and high dislocation).

Statistical Analysis

Assessment of inter- and intraobserver consistency was accomplished by the use of two parameters: the proportion of agreement and the weighted kappa coefficient as proposed by Fleiss.

Weighted kappa coefficients were calculated using quadratic weights. The weighted kappa coefficient involves adjustment of the observed proportion of agreement by correction for the proportion of agreement which arises due to chance.

Intraobserver comparison was assessed by calculating weighted kappa coefficients.

Statistical Analysis

Interobserver agreement was assessed by calculating weighted kappa coefficients for every possible pair of observers.

The observed proportion of agreement is the percentage of instances in which the observers agreed.

The weighted kappa coefficient involves adjustment of the observed proportion of agreement by correction for the proportion of agreement which arises due to chance.

The kappa value may vary between +1 (complete agreement), through 0 (agreement by chance) to -1 (complete disagreement).

For interobserver agreement the multirater kappa was also calculated.

According to Landis and Koch

agreement was graded as slight (K=0–0.2), fair (K=0.21–0.40), moderate (K=0.41–0.60), sub-stantial (K=0.61–0.80), and almost perfect (K=0.81–1).

The level of significance was p=0.05.

The statistical analysis was performed using the SPSS version 13.00 statistical package.

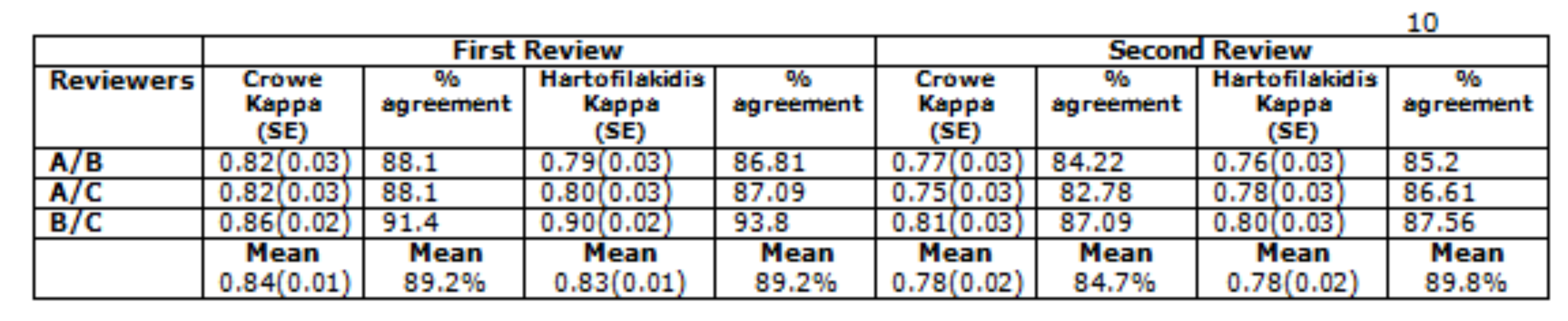
There was no failure to classify any hip by the reviewers.

In the first evaluation of the radiographs by the three reviewers paired comparisons showed a mean interobserver weighted kappa coefficient of 0.84(0.01) with a mean interobserver agreement of 89.2% for Crowe's et al. classification and 0.83(0.01) with a mean interobserver agreement of 89.2% for the Hartofilakidis's et al. classification.

In the second evaluation of the radiographs by the three reviewers paired comparisons showed a mean interobserver weighted kappa coefficient of 0.78(0.02) with a mean interobserver agreement of 84.7% for Crowe’s et al. classification and 0.78(0.02) with a mean interobserver agreement of 89.8% for the Hartofilakidis’s et al. classification.

For Crowe’s classification the interobserver agreement among the three pairs showed a minimum weighted kappa coefficient of 0.75 for Observers 1 and 3 and a maximum weighted kappa coefficient of 0.86 for Observers 2 and 3. For Hartofilakidis’s classification the minimum weighted kappa value was 0.76 for Observers 1 and 2 and a maximum value of 0.90 for Observers 2 and 3.

Paired comparisons of interobserver agreement among the three observers for Crowe and Hartofilakidis classifications.



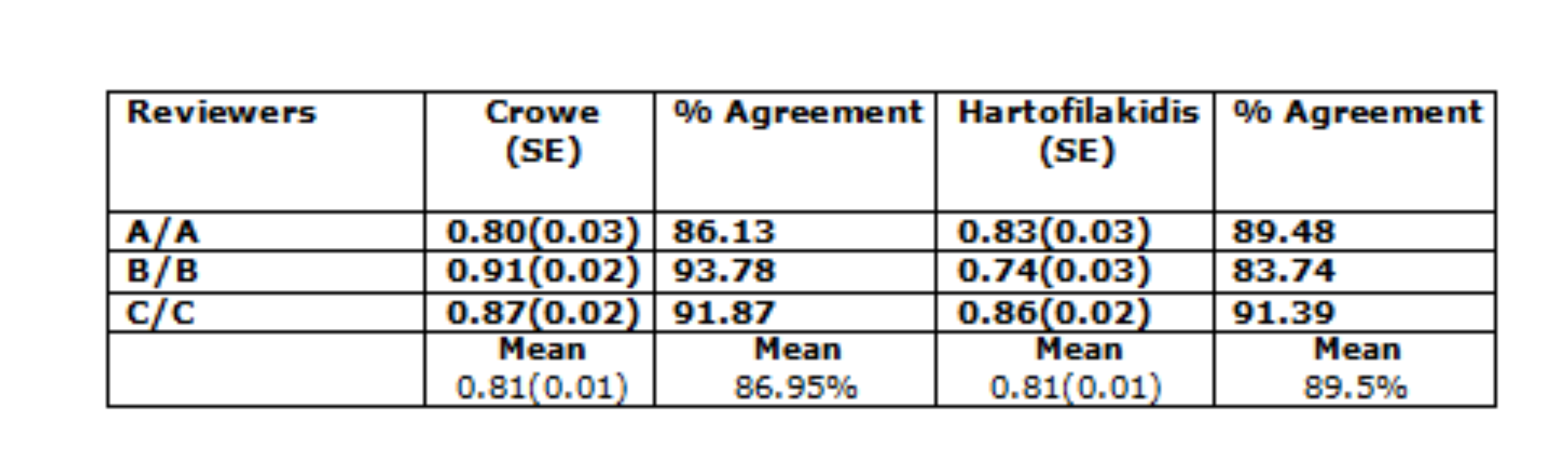
	First Review				Second Review			
Reviewers	Crowe Kappa (SE)	% agreement	Hartofilakidis Kappa (SE)	% agreement	Crowe Kappa (SE)	% agreement	Hartofilakidis Kappa (SE)	% agreement
A/B	0.82(0.03)	88.1	0.79(0.03)	86.81	0.77(0.03)	84.22	0.76(0.03)	85.2
A/C	0.82(0.03)	88.1	0.80(0.03)	87.09	0.75(0.03)	82.78	0.78(0.03)	86.61
B/C	0.86(0.02)	91.4	0.90(0.02)	93.8	0.81(0.03)	87.09	0.80(0.03)	87.56
	Mean 0.84(0.01)	Mean 89.2%	Mean 0.83(0.01)	Mean 89.2%	Mean 0.78(0.02)	Mean 84.7%	Mean 0.78(0.02)	Mean 89.8%

Intraobserver agreement for the two classification systems, i.e. the comparison between the first and the second evaluation.

As regards the intraobserver reliability the weighted kappa coefficients between the 2 evaluations of the same observer ranged for the Crowe classification from 0.80 to 0.91, while the respective values for the Hartofilakidis classification were 0.74 and 0.86.

The mean weighted kappa coefficient for Crowe’s et al. classification was 0.81(0.01) with a mean agreement of 86.95% and for the Hartofilakidis’s et al. classification the respective figures were 0.81(0.01) with a mean agreement of 89.5%.

Intraobserver agreement for the two classification systems

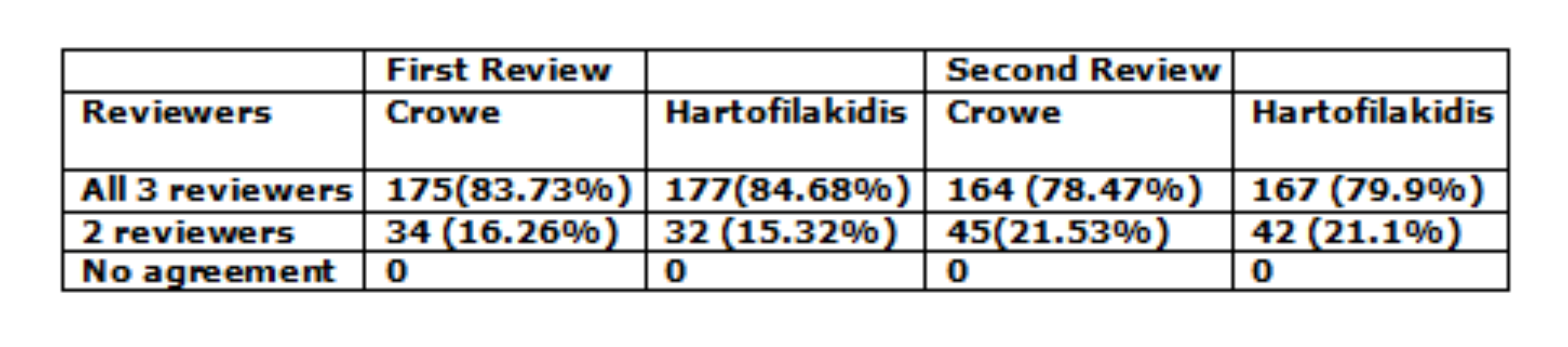


Reviewers	Crowe (SE)	% Agreement	Hartofilakidis (SE)	% Agreement
A/A	0.80(0.03)	86.13	0.83(0.03)	89.48
B/B	0.91(0.02)	93.78	0.74(0.03)	83.74
C/C	0.87(0.02)	91.87	0.86(0.02)	91.39
	Mean 0.81(0.01)	Mean 86.95%	Mean 0.81(0.01)	Mean 89.5%

Agreement between the observers for both classification systems

For Crowe classification all the observers agreed on the characterization of 175 of the 209 hip at the first review and for 164 of the 209 hip at the second review.

For the Hartofilakidis classification all observers agreed on the characterization of 177 hips at the first review and 167 hips at the second.



	First Review		Second Review	
Reviewers	Crowe	Hartofilakidis	Crowe	Hartofilakidis
All 3 reviewers	175(83.73%)	177(84.68%)	164 (78.47%)	167 (79.9%)
2 reviewers	34 (16.26%)	32 (15.32%)	45(21.53%)	42 (21.1%)
No agreement	0	0	0	0

Classifications in orthopaedic surgery are usually performed employing an order of increasing severity based on morphologic, radiographic or survival criteria.

A classification system that aspires to universal adoption should be as simple as possible and easy to use in clinical practice, reliable and reproducible.

To evaluate a classification system sufficient reliability between different observers or the same observer in different time points is essential.

Several studies revealed increased variability and low reproducibility with radiographic classification systems used in orthopaedic surgery.

The limitations of the Crowe et al. classification are the need of a radiograph including the whole pelvis, and the variability of locating the femoral head-neck junction, depending on limb rotation.

The limitations of the Hartofilakidis et al. classification include the difficulty in classifying borderline cases of dysplasia and low dislocation and low and high dislocation.

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