Radiographic classifications in Perthes disease

Interobserver agreement and association with femoral head sphericity at 5-year follow-up

Stefan HUHNSTOCK [1,5], Svein SVENNINGSEN [2], Else MERCKOLL [3], Anthony CATTERALL [4], Terje TERJESEN [1], and Ola WIIG [1]

[1] Department of Paediatric Orthopaedic Surgery, Oslo University Hospital, [2] Department of Orthopaedic Surgery, Sørlandet Hospital, Arendal, [3] Department of Radiology, Oslo University Hospital, Norway, [4] Royal National Orthopaedic Hospital, London, England, [5] Institute of Clinical Medicine, University of Oslo, Norway
Correspondence: stefan.huhnstock@oslo-universitetssykehus.no
Submitted 2017-02-21. Accepted 2017-05-12.

Background and purpose — Different radiographic classifications have been proposed for prediction of outcome in Perthes disease. We assessed whether the modified lateral pillar classification would provide more reliable interobserver agreement and prognostic value compared with the original lateral pillar classification and the Catterall classification.

Patients and methods — 42 patients (38 boys) with Perthes disease were included in the interobserver study. Their mean age at diagnosis was 6.5 (3–11) years. 5 observers classified the radiographs in 2 separate sessions according to the Catterall classification, the original and the modified lateral pillar classifications. Interobserver agreement was analysed using weighted kappa statistics. We assessed the associations between the classifications and femoral head sphericity at 5-year follow-up in 37 non-operatively treated patients in a crosstable analysis (Gamma statistics for ordinal variables, γ).

Results — The original lateral pillar and Catterall classifications showed moderate interobserver agreement (kappa 0.49 and 0.43, respectively) while the modified lateral pillar classification had fair agreement (kappa 0.40). The original lateral pillar classification was strongly associated with the 5-year radiographic outcome, with a mean γ correlation coefficient of 0.75 (95% CI: 0.61–0.95) among the 5 observers. The modified lateral pillar and Catterall classifications showed moderate associations (mean γ correlation coefficient 0.55 [95% CI: 0.38–0.66] and 0.64 [95% CI: 0.57–0.72], respectively).

Interpretation — The Catterall classification and the original lateral pillar classification had sufficient interobserver agreement and association to late radiographic outcome to be suitable for clinical use. Adding the borderline B/C group did not increase the interobserver agreement or prognostic value of the original lateral pillar classification.

Several prognostic indices have been proposed in Perthes disease. Major milestones were the introduction of the Catterall and lateral pillar classifications (Catterall 1971, Herring et al. 1992), attempting to predict the final radiographical outcome at an early stage of the disease. Catterall (1971) was the first to emphasize the relationship between the extent of femoral head involvement and final outcome. He defined 4 groups based on the site and extent of femoral head involvement, ranging from less than 25% in group I to a total head involvement in group IV. The classification was developed to be applied in the fragmentation phase. Limitation of the Catterall classification was a difficult and inaccurate initial assessment until the fragmentation phase. Grouping tended to change if the classification was applied too early (Van Dam et al. 1981). Another criticism has been the lack of sufficiently high levels of interobserver agreement (Hardcastle et al. 1980, Christensen et al. 1986, Simmons et al. 1990, Forster et al. 2006).

Herring et al. (1992) introduced a 3-group classification based on the height of the lateral portion of the femoral epiphysis (termed lateral pillar) compared with the unaffected side on AP radiographs. Group A hips showed no involvement of the lateral pillar. Group B hips had lucency and loss of height, but not exceeding 50%. Group C hips exhibited more lucency and > 50% loss of height. Reported limitations of this classification include difficulties to reliably classify hips in the initial stage (Lappin et al. 2002, Kuroda et al. 2009). Another limitation is the difficult use of the classification in bilateral cases since there is a lack of reference height to compare with.

The Herring group (2004a, 2004b) reviewed all the hips in the original study and identified a group of hips with radiographic findings that were more severe than those typical of group B but less severe than those seen in group C. Thus, they introduced a new group termed B/C borderline, transforming

DOI 10.1080/17453674.2017.1340040

their 3-group classification into a classification with 4 categories. The good to excellent interobserver results presented by the Herring group for the modified 4-group classification could not be confirmed by recently published results from the UK (Rajan et al. 2013). Thus, the first aim of our study was to assess the interobserver agreement of the modified lateral pillar classification compared with the Catterall and the original lateral pillar classifications.

Besides sufficient interobserver agreement, requirements of a good initial classification include a satisfactory ability to predict long-term outcome. Although the inventors of the modified lateral pillar classification reported good prognostic value (Herring et al. 2004a, 2004b), there seems to exist only 1 later study that has investigated this association (Froberg et al. 2011). Thus, our second aim was to assess the prognostic value of the modified lateral pillar classification and evaluate whether it was a better predictor compared with the Catterall and the original lateral pillar classifications.

Patients and methods

By a systematic search of the radiographic archive of our hospital, we identified 152 children who had been treated for Perthes disease between 1950 and 1984. 139 children had satisfying radiographic follow-up with good visual quality at least 5 years after diagnosis. We selected a random sample of 50 patients using a random-number generator. 5 patients with bilateral Perthes disease were excluded. We used for each patient true anteroposterior (AP) pelvis and frog-leg lateral radiographs at diagnosis, 1-year follow-up (mean interval 14 months) and at 5-year follow-up (mean interval 59 months). Radiographic staging according to Waldenström (1922) was applied. We excluded 3 patients due to advanced radiographic stage (reossification phase). Thus 42 patients (38 boys) with a mean age at diagnosis of 6.5 (3–11) years were included in the present study. At diagnosis, there were 36 patients in initial stage and 6 patients in fragmentation stage. 5 patients had been treated with femoral varus osteotomy and 37 patients had been treated non-operatively.

Observers

5 observers participated in the present study with the following professional background and individual contributions:

Observer SH: specialist in orthopedic surgery, senior pediatric orthopedic fellow.

Observer SS: consultant in orthopedic surgery with a great interest in pediatric orthopedic surgery. He received all radiographs stored on CDs but due to a hardware failure he was only able to retrieve images of 37 patients for the first session. A new set of CDs was sent for the second session and 40 patients could be assessed.

Observer EM: consultant in radiology, with special interest in pediatric orthopedics.

Observer AC: Professor emeritus of pediatric orthopedic surgery. He received all radiographs stored on CDs and found radiographs of 41 patients eligible for this study. No Stulberg classification was applied.

Observer OW: pediatric orthopedic consultant with special interest in Perthes disease.

All observers were familiar with the investigated classifications but nonetheless invited to a consensus-building meeting. All but 1 observer (AC) participated in the meeting before commencing the study. The 4 observers were provided with the original articles and a 20-minute tutorial, outlining the characteristics of each classification.

Radiographic assessment

The radiographs were assessed in 2 separate sessions. In the first session the original lateral pillar classification (Herring et al. 1992) and the Catterall classification (Catterall 1971) were applied, using the radiographs (at diagnosis or 1-year follow-up) that showed the greatest involvement of the femoral head at fragmentation. Radiographic outcome at 5-year follow-up was classified by 4 observers in the 37 non-operatively treated patients based on the shape of the femoral head. We modified the 5-group classification of Stulberg et al. (1981) into a simplified 3-group classification (Wiig et al. 2007), in which group A hips have spherical femoral head, group B have ovoid femoral head, and group C hips have flat femoral head. The second session was at least 1 month later and neither possible marks nor labelling from the first session could be traced on the radiographs. The observers were asked to classify the radiographs at fragmentation according to the modified lateral pillar classification (Herring et al. 2004a).

Interobserver analysis

We included all 42 patients and used an overall kappa statistic assessment of interobserver agreement by calculating the weighted kappa (Cohen 1968) for each pair of the 5 observers, yielding 10 kappa values for the lateral pillar classifications and the Catterall classification. Further, we calculated the weighted kappa for each pair of the 4 observers assessing the modified Stulberg classification, yielding 6 kappa values. Kappa statistics with linear weighting were used, defining the imputed relative distance between ordinal categories as 1 (Lowry 2015). The mean of kappa values for each classification was recorded as the overall kappa value (Light 1971) and they are presented with 95% confidence interval (CI). Possible values for kappa statistics range from –1 to 1, with 1 indicating perfect agreement and 0 indicating random agreement. As suggested by Landis and Koch (1977), we interpreted the weighted kappa values as follows: < 0.20 indicates poor agreement, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good agreement, and > 0.80 indicates excellent agreement.

Association to 5-year radiographic outcome

Only the non-operatively treated 37 were included in this part

Table 1. Interobserver agreement of the Catterall classification, the original and modified lateral pillar classifications and the modified Stulberg classification

Classification	Weighted kappa Mean	CI [a]	Range	Agreement [b]
Catterall	0.43	0.26–0.61	0–0.73	moderate
Original lateral pillar [c]	0.49	0.41–0.57	0.35–0.72	moderate
Modified lateral pillar [d]	0.40	0.29–0.51	0.15–0.59	fair
Modified Stulberg [c]	0.50	0.28–0.71	0.38–0.57	moderate

[a] CI is 95% confidence interval
[b] according to Landis and Koch (1977)
[c] 3 groups
[d] 4 groups

Table 2. Association between the prognostic classifications and the femoral head sphericity at 5-year follow-up assessed by the modified 3-group Stulberg classification. Initials are observer

	Femoral head sphericity Correlation coefficient [a]	p-value
Original lateral pillar classification		
AC	0.86	< 0.001
SH	0.61	0.02
OEM	0.95	0.001
SS	0.73	0.006
OW	0.61	0.02
Mean	0.75	
Catterall classification		
AC	0.61	< 0.001
SH	0.57	0.006
EM	0.72	0.001
SS	0.71	0.002
OW	0.61	0.02
Mean	0.64	
Modified lateral pillar classification		
AC	0.54	0.02
SH	0.66	0.003
EM	0.38	0.1
SS	0.63	0.02
OW	0.53	0.02
Mean	0.55	

[a] Gamma statistics for ordinal variables.

of the study. 4 observers classified the radiographic outcome at 5-year follow-up according to the femoral head sphericity using the 3-group modification of the Stulberg classification (round, ovoid, flat femoral head). We termed the category as "true" if there was more than 50% consensus among the observers. If there was ≤ 50% consensus, the radiographs were reassessed by 2 observers (SH and OW). Loss of height within 2 mm of a concentric circle on AP and frog-leg projection was defined as round and more than 2 mm as ovoid. The associations of the Catterall and the lateral pillar classifications were assessed in a cross-table analysis with "true" Stulberg 3-group classification as outcome variable. Gamma statistics for ordinal variables were used (Goodman and Kruskal 1954, 1959), calculating γ correlation coefficients, which were interpreted as follows: values < 0.24 indicate no association, 0.25–0.49 means weak association, 0.50–0.74 moderate association and values > 0.74 indicate strong association. Statistical analysis was done using SPSS® statistics version 21 (IBM, Armonk, NY, USA).

Results

Interobserver analysis

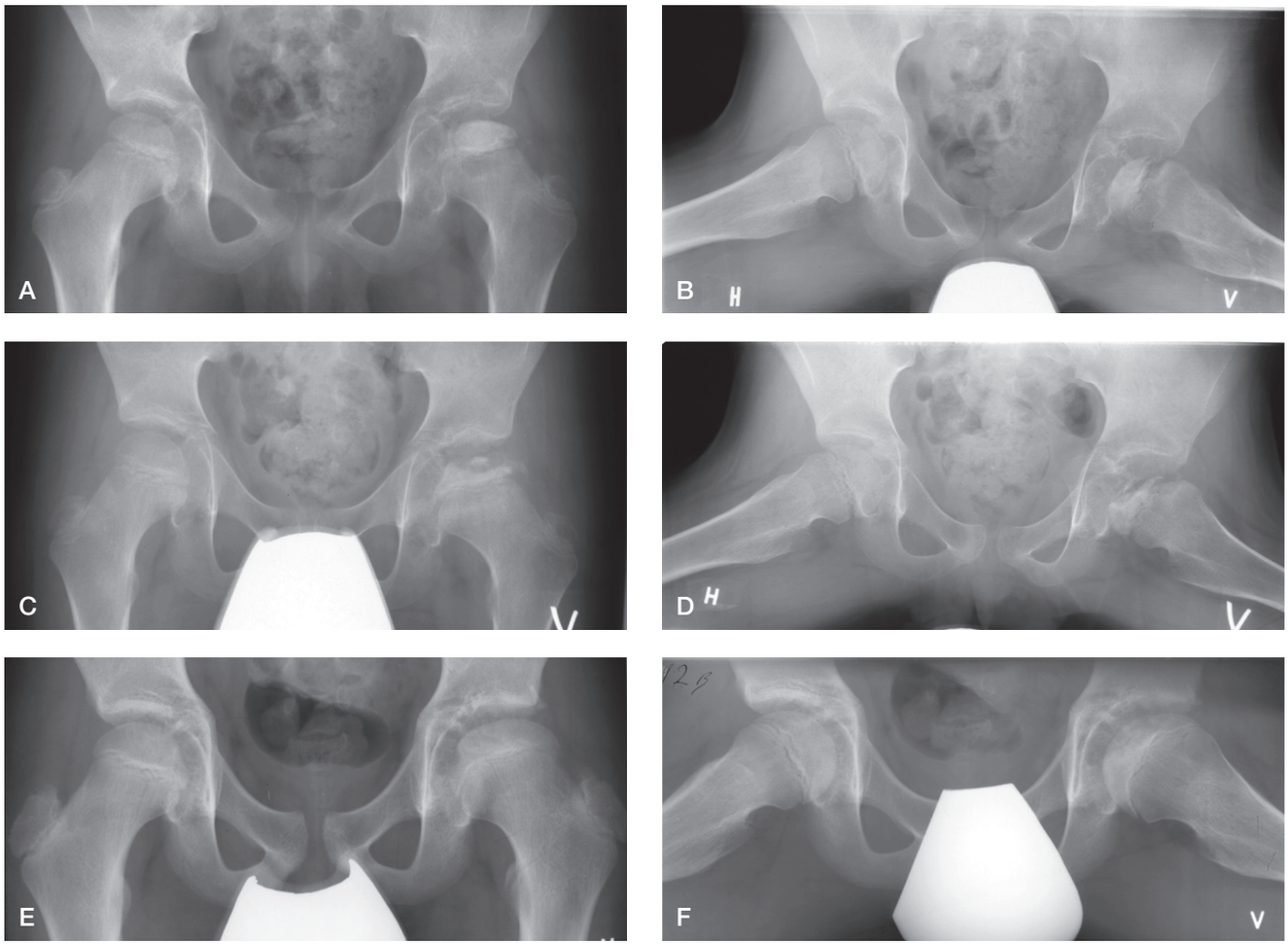
The kappa analysis (Table 1) revealed that the original lateral pillar classifications had an overall moderate interobserver agreement (mean weighted kappa 0.49, CI: 0.41–0.57). An overall moderate interobserver agreement was also found for the Catterall classification (mean weighted kappa 0.43, CI: 0.26–0.61), with a broader variation for individual kappa values. The modified lateral pillar classification scored lowest with fair overall interobserver agreement (mean weighted kappa 0.40, CI: 0.29–0.51) and individual kappa values ranging from 0.15 to 0.59. The 3-group modification of the Stulberg classification had an overall moderate interobserver agreement with a mean weighted kappa value of 0.50 (CI: 0.28–0.71).

Association to radiographic outcome

There was consensus on the femoral head shape in 32 of 37 patients. In the remaining 5 patients no primary consensus was reached since 2 observers chose "round " and 2 chose "ovoid". These 5 patients were reassessed by 2 of the observers and the "true" category was agreed upon (Figure). Thus the "true" 3-group Stulberg category was round femoral head in 10 patients, ovoid femoral head in 22 patients and a flat femoral head in 5 patients. The original lateral pillar classification was moderately to strongly associated with 5-year radiographic outcome, with mean (range) γ correlation coefficient of 0.75 (0.61–0.95) among the 5 observers (Table 2). The modified lateral pillar had a weak to moderate association with radiographic outcome, with mean γ correlation coefficient of 0.55 (0.38–0.66), and the Catterall classification showed moderate association with mean γ correlation coefficient 0.64 (0.57–0.72).

Discussion

Our results revealed moderate interobserver agreement for the Catterall and the original lateral pillar classifications and fair agreement for the modified lateral pillar classification. The original lateral pillar classification applied at fragmentation was strongly associated with the final radiographic outcome assessed by femoral head shape. The introduction of the borderline B/C group did not increase the interobserver agreement or association to late radiographic outcome of the lateral pillar classification system.



Radiographs of a boy with Perthes disease of the left hip.
A and B. AP and lateral projections at diagnosis (age 8 years) of Perthes disease.
C and D. Radiographs taken 8 months after diagnosis show fragmentation of the femoral head. The observers classified the radiographs with the following categories: Catterall group 3 (4 observers), Catterall group 2 (1 observer); original lateral pillar type B (4 observers), type C (1 observer); modified lateral pillar type B/C (3 observers), type B (2 observers).
E and F. AP and frog-leg radiographs taken 4 years and 7 months after diagnosis, at an age of 13 years. Both projections show healing and were classified according to the modified Stulberg classification as follows: round femoral head (2 observers) and ovoid femoral head (2 observers). 2 observers reassessed the radiographs. These observers agreed upon round femoral head as "true" modified Stulberg classification.

Before discussing the clinical and scientific implications of these findings, it is important to address the limitations of our study. We did not perform a prior power calculation to identify the minimal sample size required for the interobserver analysis. However, the number of patients was similar to that of previous studies (Tables 3–5). Park et al. (2012) performed a structured approach and determined the need of 36 patients, similar to the number of patients in our interobserver evaluation. 2 observers classified no hips as lateral pillar group A, while the other observers identified only 1 or 2 hips as belonging to group A. It is known from the literature that group A hips are truly underrepresented (< 5%) in comparison with group B and group C in the Perthes population (Herring et al. 2004b, Terjesen et al. 2010). This prevalence problem may cause kappa values to be unrepresentatively low (Byrt et al. 1993). In the evaluation of the prognostic value of the classifications, we included patients who had been treated with non-weightbearing and/or physiotherapy, since none of these methods have been proven to have any effect on the natural history of Perthes disease (Wiig et al. 2008). A limitation with the prognostic evaluation was the relatively small number of patients in this analysis compared with other reports on the natural history (Norlin et al. 1991, Joseph et al. 2003, Terjesen et al. 2010). However, the radiographs in these studies were mainly classified by 1 of the authors alone, which poses uncertainty regarding the reliability of the classification applied. We tried to reduce this uncertainty by multiple readings of the prognostic classifications and by establishing a consensus of the final radiographic outcome. This approach requires a substantial amount of ratings per radiograph, which is only

Table 3. Interobserver agreement of the Catterall 4-group classification in 6 previous studies and the present study. Statistics in all studies are weighted kappa

Study	n	Observers	Mean (range) [a]	Agreement [b]
Pietrzak et al. (2004)	63	3	0.39 (0.28–0.42)	fair
Present study	42	5	0.43 (0–0.73)	moderate
Nathan Sambandam et al. (2006)	44	2	0.44	moderate
de Billy et al. (2002) [c]	19	9	0.54 (0.36–0.77)	moderate
Wiig et al. (2002)	63–158	3	(0.49–0.62)	moderate to good
Simmons et al. (1990)	40	15	0.55 (0.49–0.64)	moderate
Christensen et al. (1986)	100	4	0.62 (0.50–0.67)	good

[a] weighted kappa
[b] interpretation of kappa values (Landis and Koch 1977): < 0.2 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 good agreement;
[c] recalculated with data given in the original article.

Table 4. Interobserver agreement of Herring's original lateral pillar (3-group) classification in 6 previous studies and the present study

Study	n	Observers	Statistics	Mean (range)	Agreement [a]
Present study	42	5	Weighted kappa	0.49 (0.35–0.72)	moderate
Podeszwa et al. (2000)	33	5	Cohen´s kappa	0.51 (0.43–0.62)	moderate
Herring et al. (1992)	32	16	Kappa, unspecified	0.52	moderate
Akgun et al. (2004)	50	3	Kappa, unspecified	0.53 (0.53–0.54)	moderate
Wiig et al. (2002)	63–158	3	Weighted kappa	0.56–0.70	moderate to good
Pietrzak et al. (2004)	63	3	Weighted kappa	0.65 (0.61–0.70)	good
Nathan Sambandam et al. (2006)	44	2	Weighted kappa	0.72	good

[a] interpretation of kappa values (Landis and Koch 1977): < 0.2 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 good agreement.

Table 5. Interobserver agreement of Herring's modified lateral pillar classification (4 groups) in 2 previous studies and the present study

Study	n	Observers	Statistics	Mean (range)	Agreement[a]
Rajan et al. (2013)	35	6	Weighted kappa	0.39 (0.05–0.56)	fair
Present study	42	5	Weighted kappa	0.40 (0.15–0.59)	fair
Herring et al. (2004)	20	6	Modified weighted kappa	0.71 (0.49–0.89)	good

[a] interpretation of kappa values (Landis and Koch 1977): < 0.2 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 good agreement.

feasible for a limited number of patients. Thus, the chosen approach posed both a limitation and a strength of our study.

Interobserver analysis

In studies on interobserver agreement, it is crucial to specify which statistic was used to compute agreement, i.e. Cohen's kappa (1960), Fleiss kappa (1971), and intra-class correlation (ICC) (McGraw and Wong 1996) and which variant of the statistics was computed (Siegel and Castellan 1988, McGraw and Wong 1996). The different statistical variants can substantially influence the interpretation of interobserver estimates as shown in the following example: we reassessed a study examining the interobserver agreement of the Catterall 4-group classification (de Billy et al. 2002). The authors presented excellent interobserver agreement using ICC statistics (ICC = 0.94), without clearly stating the variant that was used (missing unit and effect of ICC). We reanalysed the given raw data using linear weighted kappa statistics, yielding 36 pairs of observations and an average kappa value of 0.54 (moderate agreement). Norman and Streiner (2008) showed that ICC (2-way, mixed, single-measures, consistency) is identical to a weighted kappa with quadratic weighting, which tends to result in higher kappa values than when using linear weighting.

Herring et al. (2004a) provided each observer with a 20-minute tutorial before rating the radiographs with the modified lateral pillar classification. Although all participating

observers in our study were experienced in evaluating radiographs of hips with Perthes disease, they were nevertheless provided with a tutorial. Despite a detailed review of the modified 4 group lateral pillar classification with special attention to the borderline B/C group, we were not able to achieve similar results to those of Herring et al. (2004b). Previous studies have highlighted an increasing reproducibility of the classifications when assessed by experienced observers (Simmons et al. 1990, Podeszwa et al. 2000, Wiig et al. 2002, Kalenderer et al. 2005), but similar interobserver levels could not be reproduced in our study. Many reports assessing the interobserver agreement have been performed at 1 institution only (Nathan Sambandam et al. 2006, Park et al. 2012, Rajan et al. 2013), whilst the present study included 3 different hospitals. It is our belief that the present approach gives a more realistic estimate of interobserver agreement as compared with studies performed at a single institution.

Apart from the complexity of statistical methods and interpretation, studies on interobserver agreement in Perthes disease differ substantially in number of observers and radiographs analyzed. Hence direct comparison of presented results should be undertaken cautiously. Nevertheless, we summarized the results of the most relevant studies assessing the Catterall and lateral pillar classifications using kappa statistics and compared them with our results. We found moderate interobserver agreement in 5 out of 7 studies assessing the Catterall classification (Table 3) and in 5 of 7 studies assessing the original lateral pillar classification (Table 4). The introduction of the borderline B/C group decreased the reproducibility of the lateral pillar classification system in our study (Table 5), which is in accordance with the results of Rajan et al. (2013). They found a fair interobserver agreement (mean kappa = 0.39), similar to our mean kappa value of 0.40. The Herring group (2004b) found in average good interobserver agreement (kappa = 0.71) using a modified weighted kappa analysis (quadratic weighting). The authors attributed only half as much importance in misclassifying the borderline group B/C to its adjacent groups as they attributed misclassifying between groups A, B and C. Since quadratic weighting may increase kappa values artificially if an extra category is introduced within a classification system (Brenner and Kliebsch 1996), this may have led to an unrepresentatively high level of agreement.

Our results showed that the Catterall and the original lateral pillar classifications had moderate interobserver agreement. There is no established common understanding of which degree of interobserver agreement may be necessary or appropriate to define a classification system as satisfactory in clinical practice. Some authors abandoned the use of classifications on the basis of moderate agreement (Christensen et al. 1986) while others redefined them as acceptable (Akgun et al. 2004). However, the fact that the introduction of the borderline B/C group even reduced the reproducibility of the lateral pillar classification raises concerns about the usefulness of this modification.

Association to radiographic outcome

We assessed the radiographic outcome on the basis of femoral head sphericity because this is strongly associated with long-term outcome (Mose 1980, Stulberg et al. 1981). Our study revealed significant associations between the 3 classification systems at fragmentation and the femoral head sphericity at 5-year follow up. The original 3-group lateral pillar classification had the strongest association, which is in accordance with previous studies (Herring et al. 1992, Ritterbusch et al. 1993, Farsetti et al. 1995, Ismail and Macnicol 1998, Lappin et al. 2002, Terjesen et al. 2010). All hips that had a flat femoral head shape at 5-year follow up had been classified as group C at the fragmentation phase. The Catterall classification as predictor of radiographic outcome is controversial. It did not correlate well with the final radiographic outcome in some studies (Weinstein 1985, Ismail and Macnicol 1998, Gigante et al. 2002), while others confirmed it as a prognostic factor (Dickens and Menelaus 1978, Meurer et al. 1999, Terjesen et al. 2010). Especially when modified into a 2-group classification, distinguishing between more and less than 50% of femoral head necrosis, the Catterall grouping system was a strong predictor of radiographic outcome (Wiig et al. 2008). Our results confirm a significant association between the Catterall 4-group classification and the femoral head sphericity at 5-year follow up, but the association seemed to be somewhat weaker than that of the original lateral pillar classification.

The Herring group reviewed the hips in their original study and identified a group of hips with radiographic findings more severe than those typical of group B but less severe than those in group C (Herring et al. 2004b). Because of difficulties in defining the borderlines between the groups, a new classification group (borderline B/C) was introduced. The authors found that the modified classification was a strong prognostic factor. Our results and the findings of other authors (Froberg et al. 2011) confirm a significant association between the modified 4-group lateral pillar classification and modified Stulberg as outcome variable. To our knowledge the present study is the first to compare the value of the modified lateral pillar classification with the original classification as predictors. Our findings suggest that the introduction of the new borderline B/C group did not improve the association of the lateral pillar system to the femoral head sphericity at 5-year follow-up.

Summary

Our results underline that each of the classifications has its limitations; none is perfect. We think that the original lateral pillar system (3 groups) is the most suitable classification in the early radiographic stages of Perthes disease. It is easier to apply (needs only AP radiographs) and was somewhat better associated with the final radiographic outcome as compared with the Catterall 4-group classification. The introduction of the borderline B/C group increased neither the reproducibility nor the prognostic value of the lateral pillar system, which raises concerns about its usefulness in clinical practice.

SH: contributed to the design of the study, selected appropriate conventional radiographs according to the study requirements, monitored the digitalization process, performed statistical analysis and wrote the manuscript. SS, AC, EM: classified radiographs according to the Catterall, Stulberg and lateral pillar classifications. TT: initiated and contributed to the design of the study. He identified and selected radiographs of Perthes patients from the hospital radiographic archive and participated in the writing process of the manuscript. OW: identified and selected radiographs of Perthes patients from the hospital radiographic archives, contributed to the study design, and classified radiographs according to the Catterall, Stulberg and lateral pillar classifications.

We thank Are Hugo Pripp, Department of Biostatistics, Epidemiology and Health, Oslo University Hospital for valuable support and assistance in refining the statistical analysis and Heidi-Karin Lundlie, secretary at the Department of Radiology, Oslo University Hospital, who helped in digitalizing conventional radiographs and provided observer AC and SS with radiographs on CDs.

Akgun R, Yazici M, Aksoy M C, Cil A, Alpaslan A M, Tumer Y. The accuracy and reliability of estimation of lateral pillar height in determining the herring grade in Legg–Calve–Perthes disease. J Pediatr Orthop 2004; 24 (6): 651-3.

Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 1996; 7 (2): 199-202.

Byrt T, Bishop J, Carlin J B. Bias, prevalence and kappa. J Clin Epidemiol 1993; 46 (5): 423-9.

Catterall A. The natural history of Perthes' disease. J Bone Joint Surg Br 1971; 53 (1): 37-53.

Christensen F, Soballe K, Ejsted R, Luxhoj T. The Catterall classification of Perthes' disease: an assessment of reliability. J Bone Joint Surg Br 1986; 68 (4): 614-15.

Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20 (1): 37-46.

Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 1968; 70 (4): 213-20.

de Billy B, Viel J F, Monnet E, Garnier E, Aubert D. Interobserver reliability in the interpretation of radiologic signs in Legg–Calve–Perthes disease. J Pediatr Orthop B 2002; 11 (1): 10-14.

Dickens D R, Menelaus M B. The assessment of prognosis in Perthes' disease. J Bone Joint Surg Br 1978; 60-b (2): 189-94.

Farsetti P, Tudisco C, Caterini R, Potenza V, Ippolito E. The Herring lateral pillar classification for prognosis in Perthes disease: Late results in 49 patients treated conservatively. J Bone Joint Surg Br 1995; 77 (5): 739-42.

Fleiss J L. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76 (5): 378-82.

Forster M C, Kumar S, Rajan R A, Atherton W G, Asirvatham R, Thava V R. Head-at-risk signs in Legg–Calve–Perthes disease: poor inter- and intraobserver reliability. Acta Orthop 2006; 77 (3): 413-17.

Froberg L, Christensen F, Pedersen N W, Overgaard S. Long-term follow-up of a patient cohort with Legg–Calve–Perthes disease. J Pediatr Orthop B 2011; 20 (5): 273-7.

Gigante C, Frizziero P, Turra S. Prognostic value of Catterall and Herring classification in Legg–Calve–Perthes disease: Follow-up to skeletal maturity of 32 patients. J Pediatr Orthop 2002; 22 (3): 345-9.

Goodman L A, Kruskal W H. Measures of association for cross classifications. J Am Stat Assoc 1954; 49 (268): 732-64.

Goodman L A, Kruskal W H. Measures of Association for cross classifications. II: Further discussion and references. J Am Stat Assoc 1959; 54 (285): 123-63.

Hardcastle P H, Ross R, Hamalainen M, Mata A. Catterall grouping of Perthes' disease: An assessment of observer error and prognosis using the Catterall classification. J Bone Joint Surg Br 1980; 62-B (4): 428-31.

Herring J A, Neustadt J B, Williams J J, Early J S, Browne R H. The lateral pillar classification of Legg–Calve–Perthes disease. J Pediatr Orthop 1992; 12 (2): 143-50.

Herring J A, Kim H T, Browne R. Legg–Calve–Perthes disease. Part I: Classification of radiographs with use of the modified lateral pillar and Stulberg classifications. J Bone Joint Surg Am 2004a; 86-A (10): 2103-20.

Herring J A, Kim H T, Browne R. Legg–Calve–Perthes disease. Part II: Prospective multicenter study of the effect of treatment on outcome. J Bone Joint Surg Am 2004b; 86-A (10): 2121-34.

Ismail A M, Macnicol M F. Prognosis in Perthes' disease: A comparison of radiological predictors. J Bone Joint Surg Br 1998; 80 (2): 310-14.

Joseph B, Varghese G, Mulpuri K, Narasimha Rao K, Nair N S. Natural evolution of Perthes disease: A study of 610 children under 12 years of age at disease onset. J Pediatr Orthop 2003; 23 (5): 590-600.

Kalenderer O, Agus H, Ozcalabi I T, Ozluk S. The importance of surgeons' experience on intraobserver and interobserver reliability of classifications used for Perthes disease. J Pediatr Orthop 2005; 25 (4): 460-4.

Kuroda T, Mitani S, Sugimoto Y, Asaumi K, Endo H, Akazawa H, et al. Changes in the lateral pillar classification in Perthes' disease. J Pediatr Orthop B 2009; 18 (3): 116-19.

Landis J R, Koch G G. The measurement of observer agreement for categorical data. Biometrics 1977; 33 (1): 159-74.

Lappin K, Kealey D, Cosgrove A. Herring classification: How useful is the initial radiograph? J Pediatr Orthop 2002; 22 (4): 479-82.

Light R J. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychol Bull 1971; 76 (5): 365-77.

Lowry R. VassarStats: Website for statistical computation. VassarStats: Website for statistical computation. Vassar College; 2015.

McGraw K O, Wong S P. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1996; 1 (1): 30-46.

Meurer A, Schwitalle M, Humke T, Rosendahl T, Heine J. [Comparison of the prognostic value of the Catterall and Herring classification in patients with Perthes disease]. Z Orthop Ihre Grenzgeb 1999; 137 (2): 168-72.

Mose K. Methods of measuring in Legg–Calve–Perthes disease with special regard to the prognosis. Clin Orthop Relat Res 1980; (150): 103-9.

Nathan Sambandam S, Gul A, Shankar R, Goni V. Reliability of radiological classifications used in Legg–Calve–Perthes disease. J Pediatr Orthop B 2006; 15 (4): 267-70.

Norlin R, Hammerby S, Tkaczuk H. The natural history of Perthes' disease. Int Orthop 1991; 15 (1): 13-6.

Norman G R, Streiner D L. Biostatistics: The bare essentials. B C Decker 2008.

Park M S, Chung C Y, Lee K M, Kim T W, Sung K H. Reliability and stability of three common classifications for Legg–Calve–Perthes disease. Clin Orthop Relat Res 2012; 470 (9): 2376-82.

Pietrzak S, Napiontek M, Tomaszewski M. Catterall and Herring classifications in assessing Perthes disease: Inter- and intra-observer study. Ortop Traumatol Rehabil 2004; 6 (5): 561-6.

Podeszwa D A, Stanitski C L, Stanitski D F, Woo R, Mendelow M J. The effect of pediatric orthopaedic experience on interobserver and intraobserver reliability of the Herring lateral pillar classification of Perthes disease. J Pediatr Orthop 2000; 20 (5): 562-5.

Rajan R, Chandrasenan J, Price K, Konstantoulakis C, Metcalfe J, Jones S. Legg–Calve–Perthes: Interobserver and intraobserver reliability of the modified Herring lateral pillar classification. J Pediatr Orthop 2013; 33 (2): 120-3.

Ritterbusch J F, Shantharam S S, Gelinas C. Comparison of lateral pillar classification and Catterall classification of Legg–Calve–Perthes' disease. J Pediatr Orthop 1993; 13 (2): 200-2.

Siegel S, Castellan N J. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill 1988.

Simmons E D, Graham H K, Szalai J P. Interobserver variability in grading Perthes' disease. J Bone Joint Surg Br 1990; 72 (2): 202-4.

Stulberg S D, Cooperman D R, Wallensten R. The natural history of Legg–Calve–Perthes disease. J Bone Joint Surg Am 1981; 63 (7): 1095-108.

Terjesen T, Wiig O, Svenningsen S. The natural history of Perthes' disease. Acta Orthop 2010; 81 (6): 708-14.

Van Dam B E, Crider R J, Noyes J D, Larsen L J. Determination of the Catterall classification in Legg–Calve–Perthes disease. J Bone Joint Surg Am 1981; 63 (6): 906-14.

Waldenström H. The definite form of coxa plana. Acta Radiologica 1922; 1: 384.

Weinstein S L. Legg–Calve–Perthes disease: Results of long-term follow-up. Hip 1985: 28-37.

Wiig O, Terjesen T, Svenningsen S. Inter-observer reliability of radiographic classifications and measurements in the assessment of Perthes' disease. Acta Orthop Scand 2002; 73 (5): 523-30.

Wiig O, Terjesen T, Svenningsen S. Inter-observer reliability of the Stulberg classification in the assessment of Perthes disease. J Child Orthop 2007; 1 (2): 101-5.

Wiig O, Terjesen T, Svenningsen S. Prognostic factors and outcome of treatment in Perthes' disease: A prospective study of 368 patients with five-year follow-up. J Bone Joint Surg Br 2008; 90 (10): 1364-71.