J Child Orthop (2007) 1:101–105
DOI 10.1007/s11832-007-0020-x

ORIGINAL CLINICAL ARTICLE

Inter-observer reliability of the Stulberg classification in the assessment of Perthes disease

Ola Wiig · Terje Terjesen · Svein Svenningsen

Received: 12 December 2006 / Accepted: 22 March 2007 / Published online: 28 April 2007

Abstract

Purpose Accurate and reliable radiographic classifications are of great importance as a basis of treatment decisions and prognosis in Perthes disease. The classification of Stulberg is widely used as a predictor of long-term outcome. The aim of the present study was to determine whether the Stulberg classification is sufficiently reliable for routine clinical use in the assessment of Perthes disease.

Methods We used this classification to assess the radiographs of 101 hips in two separate sessions (55 and 46 hips, respectively), interfered by an educational intervention in which the classification algorithm was discussed and clarified.

Results We obtained good agreement between experienced examiners (weighted kappa 0.65) and a percentage agreement of 71%. We obtained weighted kappa values of 0.51 and 0.57 (moderate agreement) and percentage agreements of 62% and 65% between the least experienced observer and the two experienced examiners. Combining Stulberg class I and II, and IV and V into a simpler three-group classification gave better agreement between all observers. The agreement between the two experienced observers was improved to 81%.

Conclusions We conclude that the reliability of the Stulberg classification is acceptable when the radiographic assessment is carried out by experienced examiners. A simpler three-group classification based on the shape of the femoral head (spherical, ovoid and flat) gave better agreement and is, therefore, recommended for routine clinical use.

Keywords Stulberg classification · Inter-observer agreement · Perthes disease

O. Wiig (✉)
Orthopaedic Centre, Ullevål University Hospital,
0407 Oslo, Norway
e-mail: ola.wiig@ulleval.no

T. Terjesen
Orthopaedic Department, Rikshospitalet University Hospital,
Oslo, Norway

S. Svenningsen
Orthopaedic Department, Sørlandet Sykehus,
Arendal, Norway

Introduction

Decisions regarding treatment and prognosis in Perthes disease are mainly based upon the assessment of radiographs, and accurate and reliable radiographic classifications are of great importance. Stulberg et al. [1] proposed a five-class classification to predict long-term prognosis and the onset of degenerative joint disease after Perthes disease. They identified three types of congruency between the femoral head and the acetabulum: spherical congruency (classes I and II), aspherical congruency (classes III and IV) and aspherical incongruency (class V). They suggested that each class was associated with a predictable clinical and radiographic course. Although the classification of Stulberg is widely used, only a few studies have evaluated the inter-observer reliability of this system. Agus et al. [2] and Farsetti et al. [3] have reported excellent inter- and intra-observer reliability, whereas Neyt et al. [4] found that the inter-observer variance was marked and they called into question the usefulness and reliability of the Stulberg classification.

In the present study, the inter-observer reliability of the Stulberg classification system with the use of the consensus algorithm by Neyt et al. [4] was evaluated. The aim was to determine whether the Stulberg classification is sufficiently reliable for routine clinical use in the assessment of Perthes disease.

Patients and methods

During the 5-year period 1996–2000, a nationwide prospective study on Perthes disease, initiated by the Norwegian Paediatric Orthopaedic Society, was carried out. Four hundred and twenty-five (425) patients were registered and they were followed for 5 years. We evaluated the radiographs of 101 hips in 90 patients (11 had bilateral affection). They were patients from the first 3 years of inclusion (1996–1998) with available 5-year follow-up radiographs when we started this study. The radiographs were assessed by an orthopaedic surgeon with great experience in examining the radiographs of children (SS), an experienced paediatric orthopaedic surgeon (TT) and an orthopaedic surgeon in the training of paediatric orthopaedic practice (OW). We used the Stulberg classification as modified by the algorithm of Neyt et al. [4]. Class I hips are spherical with normal femoral head, neck and acetabulum. Class II hips have spherical femoral head with either coxa magna, short neck or steep acetabulum. Class III hips have ovoid (not flat) femoral head contour, in the sense that they do not fall within 2 mm of the Mose [5] concentric circles in both anteroposterior and Lauenstein projections. Class IV hips have flat outline of the femoral head (at least 1/3rd of the contour of the femoral head resembles a straight line on at least one projection) and there is congruency between the femoral head and the acetabulum. Class V hips have flat femoral head and normal acetabulum (aspherical incongruency).

We evaluated the radiographs in two separate sessions (55 and 46 hips, respectively), interfered by an educational intervention in which the classification algorithm was discussed and clarified. When two or three observers agreed on the classification, the result was termed as the ‘‘true’’ Stulberg class.

In order to test whether a simpler classification would give better agreement, we combined Stulberg class I and II, and IV and V, creating a three-group classification based on the shape of the femoral head (group A with spherical femoral head, group B with ovoid femoral head and group C with flat femoral head).

The data were analysed using the weighted kappa statistics [6]. We also calculated the percentage agreement between the observers. The kappa has a maximum of 1.00 when agreement is perfect, whereas a value of 0.00 indicates no agreement better than chance, and negative values show worse than chance agreement. As suggested by Altmann [6], we applied the following interpretations of kappa values: below 0.20 as poor agreement, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as good and over 0.80 as very good agreement.

Results

There were 67 boys and 23 girls, and 11 patients had bilateral disease. The mean age at diagnosis was 5.4 years (SD=1.9) and the mean age at the 5-year follow-up was 10.4 years (SD=1.9). The Stulberg distribution for all examiners is shown in Fig. 1. All observers agreed on 49 hips, two observers on 52 hips and there were no hips where all observers disagreed. The ‘‘true’’ Stulberg distribution was 23 hips (22.8%) in class I, 29 hips (28.7%) in class II, 41 hips (40.6%) in class III, 8 hips (7.9%) in class IV and 0 hips in class V.

We obtained a weighted kappa value of 0.65 (good agreement) and a percentage agreement of 71% for all hips (Table 1) between the experienced observers (TT and SS).

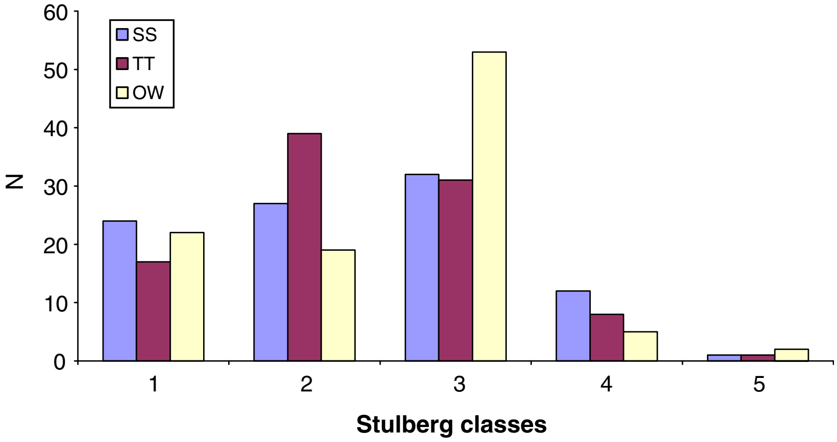

Fig. 1 Distribution of the Stulberg classes among the three observers (SS, TT and OW)

Table 1 Inter-observer agreement, given as the number of hips according to the Stulberg classification

	Agreement								Kappa$_{(W)}$
	N_1	Stulberg classes					N_2	%	
		1	2	3	4	5			
SS/TT	101	17	26	24	5	0	72	71	0.65
SS/OW	101	19	14	30	3	0	66	65	0.57
TT/OW	101	15	15	29	4	0	63	62	0.51

SS, TT and OW are the paediatric orthopaedic surgeons, N_1 is the number of hips examined, N_2 is the number of patients agreed upon, % is the percentage agreement and Kappa$_{(W)}$ is the weighted kappa

We obtained a weighted kappa value of 0.57 (moderate agreement) and a percentage agreement of 65% between OW and SS, and a weighted kappa of 0.51 (moderate agreement) and a percentage agreement of 62% between OW and TT.

Prior to the consensus meeting, the experienced observers obtained good agreement (weighted kappa of 0.66 and percentage agreement of 76%). After the meeting, the agreement was slightly lower (weighted kappa of 0.60 and percentage agreement of 63%).

Combining Stulberg class I and II (group A), maintaining Stulberg class III (group B) and combining Stulberg class IV and V (group C) gave better agreement between all observers (Table 2). The weighed kappa was 0.70 among the experienced observers and 0.56–0.60 between the less experienced observer and the more experienced observers.

No hip was considered flat by one observer and spherical by other observers. Using the modified Stulberg classification, one or both experienced observers considered the femoral head to be spherical in 62 hips. The observers agreed that the femoral head was spherical in 51 of 62 hips (82%). In the remaining 11 hips, one observer thought that the head was spherical and the other thought it was oval in eight hips, whereas the opposite occurred in three hips. One or both of the experienced observers thought that the femoral head was ovoid in 39 hips. The observers agreed that the femoral head was ovoid in 31 hips (79%). In the eight hips with discrepancy, one of the observers proposed ovoid head and the other proposed flat head in six hips, whereas the opposite occurred in two hips.

Table 2 Inter-observer agreement, given as the number of hips according to the three-group classification

	Agreement						Kappa$_{(W)}$
	N_1	Three-group classification			N_2	%	
		A	B	C			
SS/TT	101	51	24	7	82	81	0.70
SS/OW	101	38	30	6	74	73	0.60
TT/OW	101	37	29	7	73	72	0.56

SS, TT and OW are the paediatric orthopaedic surgeons, N_1 is the number of hips examined, N_2 is the number of patients agreed upon, % is the percentage agreement and Kappa$_{(W)}$ is the weighted kappa

Discussion

Most of the hips examined in this study were skeletally immature. The mean age at follow-up was 10.4 years. Ten percent (10%) had closure of one or both of the triradiate cartilages and 9% had one or both femoral head physes closed at the 5-year follow-up. However, all hips were completely healed at the final follow-up. In theory, a spherical femoral head (class I or II) at the time of healing can become slightly oval during the remaining growth until skeletal maturity if there exists a partial physeal arrest and, hence, move the hip from class I or II to class III. However, we believe that such a physeal arrest occurs early in the course of the disease; consequently, the femoral head will be deformed before healing. We find it unlikely that an oval or flat femoral head (class III–V) at healing will become spherical (class I or II) during remaining growth after healing. Consequently, as Cooperman and Stulberg [9] and Martinez et al. [10] concluded, we think that we can reliably assign each patient to a Stulberg class, even if the hips are not completely skeletally mature.

Stulberg et al. [1] and Neyt et al. [4] classified 20% and 4%, respectively, in class I, 12% and 18% in class II and the majority of the hips in class III (17% and 36%) and class IV (32% and 39%). They classified 18% and 4%, respectively, of the hips in class V. In our study, we found that the majority of the hips were class II or III (29% and 41% hips, respectively) and no hip was classified as class V. This may indicate that the hips in our series have better prognosis, which is most likely due to a relatively low mean age at diagnosis (5.4 years) in our study compared to Stulberg et al. [1] (7.5 years).

Neyt et al. [4] assessed the inter- and intra-rater reliability of the Stulberg classification and found that, although the intra-rater reliability was marginally acceptable, the inter-observer variance was marked (reliability coefficients ranging from 0.603 to 0.732). Unfortunately, the authors used generalisability coefficients on their categorical data instead of kappa statistics, making comparison with other studies difficult. They called into question the usefulness and reliability of the Stulberg classification. However, Agus et al. [2] and Farsetti et al. [3] have reported excellent inter-observer reliability using the Stulberg system, but certain objections could be raised on both studies. Agus et al. [2] used intra-class correlation coefficients (ICC), which is a widely used method for measuring the reliability of quantitative scales [7]. They had agreement in 50% using the original Stulberg classification and 78% using a simplified Stulberg classification, which is similar to our results. Farsetti et al. [3] obtained an inter-observer agreement of 98%, but no further information or comments that could explain this exceptionally good accordance was provided. In our study, we obtained good agreement between the experienced examiners (weighted kappa=0.65). The agreement was less when the radiographs were assessed by a less experienced examiner, which accords with Kalenderer et al. [8]. Neyt et al. [4], however, reported no uniform effect of experience on reliability.

Herring et al. [11] investigated the inter-observer reliability of a redefined Stulberg classification. Because they found that the use of the Mose [5] template of concentric circles was a source of error, they used a new method of measuring the circle fit of the femoral head. In addition, they defined class IV as a femoral head with more than 1 cm of flattening in a weight-bearing area on either anteroposterior or the Lauenstein projection, and they omitted class V hips. They found 91% agreement and a weighted kappa of 0.82, and concluded that the redefined Stulberg classification was sufficiently reliable and accurate for use in studies of Perthes disease.

The least experienced observer had more class III hips compared with the other observers, and the major source of disagreement was deciding whether the femoral head was round or slightly ovoid (contour within 2 mm or 4 mm of the Mose circles). Mose [5] used a transparent plate equipped with a series of 28 concentric circles with a difference of 2 mm in radius between the circles to decide whether the femoral head was spherical or not. A spherical femoral head had to follow the same circle on the template within a variation of 2 mm. As with Herring et al. [11], we found it difficult to see the outline of the femoral head clearly in some low-contrast radiographs. This was especially hard when the radiographs showed small-scale images of the hip. In our experience, a correct placement of the centre of the femoral head is essential to decide whether the shape of the surface of the epiphysis is spherical or ovoid. As observed by Herring et al. [11] and Neyt et al. [4], we found no effect of an educational intervention; indeed, the agreement between the observers was slightly less after this session.

In Stulberg et al's original work from 1981 [1], the class V hips were described as distinct from class IV, as class V hips had flat femoral heads in normal acetabuli. Furthermore, he stated that severe osteoarthritis developed before the age of 50 years in the class V hips, in contrast to the class III and IV hips, where mild to moderate arthritis developed in late adulthood. Ippolito et al. [12] reported that 37% of Stulberg class III hips and 70% of class IV hips had osteoarthritis at the ages of 30–40 years. Weinstein [13] stated that class V hips deteriorate at the end of the fourth decade of life, whereas patients with class III and IV hips undergo significant functional deterioration one decade later. Based on this, we think that it is reasonable to consider that class IV and V hips for all practical purposes have quite similar prognosis, as these hips deteriorate functionally when the patients are in their thirties and forties.

There are very few class V hips reported: Stulberg et al. [1] had 18 hips, Ippolito et al. [12] had one hip and Martinez et al. [10] had two hips. In our opinion, it is rather unlikely that the acetabulum remains normal after having been influenced for several years by a flat femoral head, even though the ability of the acetabulum to remodel declines after 8 years of age [14].

When using a classification like the Stulberg grading, we find it to be hardly adequate to omit one of the five classes, as was done by Herring et al. [11], who omitted class V. Although class V was proposed four times (two hips by one observer and one hip by each of the other two observers), there were no ‘‘true’’ Stulberg class V hips in the present study. This confirms the results of Neyt et al. [4] that class V is very rare and difficult to distinguish from class IV, which is the reason why we, in our modified Stulberg grading, have combined classes IV and V.

Hips classified as Stulberg class I and II have the same good prognosis [1] and the misclassification of I as II or the reverse would be of no clinical importance [11]. In addition, a hip in which femoral varus osteotomy has been performed will be classified as a class II hip, even with a perfectly spherical head due to the altered anatomy of the proximal femur.

We evaluated a simpler three-group classification, where group A (Stulberg class I and II) constitutes hips with a spherical femoral head, group B (Stulberg class III) is hips with an ovoid femoral head and group C (Stulberg class IV and V) comprises hips with a flat femoral head. This simple classification gave a somewhat better agreement among the observers. The agreement between the experienced observers improved from 71% to 81%.

Because the original Stulberg classes I and II have similar good long-term prognosis and because of the scarcity of class V hips and the difficulty in distinguishing between class IV and class V, we advocate our simpler three-group modification to be used in routine clinical practice. According to the present results, the classification is more reliable when experienced examiners evaluate the radiographs.

References

1. Stulberg D, Cooperman DR, Wallenstein R (1981) The natural history of Legg-Calvé-Perthes disease. J Bone Joint Surg Am 63(7):1095–1108
2. Agus H, Kalenderer O, Eryanlmaz G, Ozcalabi IT (2004) Intraobserver and interobserver reliability of Catterall, Herring, Salter-Thompson and Stulberg classification systems in Perthes disease. J Pediatr Orthop B 13(3):166–169
3. Farsetti P, Tudisco C, Caternini R, Potenza V, Ippolito E (1995) The Herring lateral pillar classification for prognosis in Perthes disease. Late results in 49 patients treated conservatively. J Bone Joint Surg Br 77(5):739–742
4. Neyt JG, Weinstein SL, Spratt KF, Dolan L, Morcuende J, Dietz FR, Guyton G, Hart R, Kraut MS, Lervick G, Pardubsky P, Saterbak A (1999) Stulberg classification system for evaluation of Legg-Calvé-Perthes disease: intra-rater and inter-rater reliability. J Bone Joint Surg Am 81(9):1209–16
5. Mose K (1980) Methods of measuring in Legg-Calvé-Perthes disease with special regard to the prognosis. Clin Ortop Relat Res 150:103–109
6. Altmann DB (1991) Practical statistics for medical research. Chapman and Hall, London
7. Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educat Psychol Meas 33:613–619
8. Kalenderer O, Agus H, Ozcalabi IT, Ozluk S (2005) The importance of surgeons' experience on intraobserver and interobserver reliability of classifications used for Perthes disease. J Pediatr Orthop 25(4):460–464
9. Cooperman DR, Stulberg D (1986) Ambulatory containment treatment in Perthes' disease. Clin Orthop Relat Res 203:289–300
10. Martinez AG, Weinstein SL, Dietz FR (1992) The weight-bearing abduction brace for the treatment of Legg-Perthes disease. J Bone Joint Surg Am 74(1):12–21
11. Herring JA, Kim HT, Browne R (2004) Legg-Calvé-Perthes disease. Part I: classification of radiographs with use of the modified lateral pillar and Stulberg classifications. J Bone Joint Surg Am 86(10):2103–2120
12. Ippolito E, Tudisco C, Farsetti P (1987) The long-term prognosis of unilateral Perthes' disease. J Bone Joint Surg Br 69(2):243–250
13. Weinstein SL (1997) Natural history and treatment outcomes of childhood hip disorders. Clin Orthop Relat Res 344:227–242
14. Lindstøm JR, Ponseti IV, Wenger DR (1979) Acetabular development after reduction in congenital dislocation of the hip. J Bone Joint Surg Am 61(1):112–118