Acta Orthop. Belg., 2010, **76**, 48-57

ORIGINAL STUDY

Reliability and reproducibility of classification systems for Legg-Calvé-Perthes disease : A systematic review of the literature

Dhirendra MAHADEVA, Mark CHONG, David J. LANGTON, Anthony M. TURNER

From University Hospitals Coventry and Warwickshire and New Cross Hospital, Wolverhampton, United Kingdom

Several classification systems are in use for Legg-Calvé-Perthes disease. Three of them : Catterall, Salter Thompson and Herring (Lateral Pillar) are most commonly used. There has been debate on which is most reliable. The purpose of this paper was to systematically analyse the literature when the classifications were compared. The Ovid (Medline) Database was used and the MeSH terms Perthes Classification and Reliability were inserted. Eleven studies were retrieved but only five were suitable for analysis as they attempted to compare the classifications. Most studies used kappa agreements as the principal outcome measure, although intraclass coefficients and percentage agreements were also used. Only four studies assessed for both intraobserver reproducibility and interobserver reliability. A further study from the references appendages was found to be suitable, and was included in the analysis. Kappa ranged from poor to fair (Salter Thompson), fair to moderate (Catterall) and moderate to good (Herring). The outcome from Legg-Calvé-Perthes disease is extremely variable. Inconsistent interpretation of the plain films may explain this, although it is likely this is multifactorial. The papers in this study show that on balance, the Lateral Pillar classification was most reliable, probably secondary to ease of use. A persistent theme was that the subchondral fracture line in the Salter Thompson system was difficult to interpret and not always present. It also showed that whilst reliability and reproducibility tended to improve with experience, disagreement was not always restricted to more junior personnel. Each classification has its merits but reliability and reproducibility remains unsatisfactory. Digital technology in the future may help delineate the lesions better and improve agreement.

Keywords : Perthes ; classification ; reliability ; reproducibility.

INTRODUCTION

Various controversies exist in Legg-Calvé-Perthes disease (LCP), from aetiology to management (*17*). The challenge for the orthopaedic surgeon remains the ability to discriminate between patients who are likely to have good outcome

■ Dhirendra Mahadeva, BMBS MRCSEd, Registrar, Trauma and Orthopaedics.
■ Mark Chong, BMBS MRCS, Speciality Registrar, Trauma and Orthopaedics.
■ David J. Langton, MBBS MRCS, Research Registrar, Trauma and Orthopaedics.
University Hospitals Coventry and Warwickshire, UK.
■ Anthony M Turner, FRCS, Consultant Paediatric Orthopaedic Surgeon.
Department of Trauma and Orthopaedic Surgery, New Cross Hospital, Wolverhampton, UK.
Correspondence : Dhirendra Mahadeva, Flat 4, 37 Portland Rd, Birmingham, B16 9HS.
E-mail : mahadeva501@yahoo.co.uk

No benefits or funds were received in support of this study

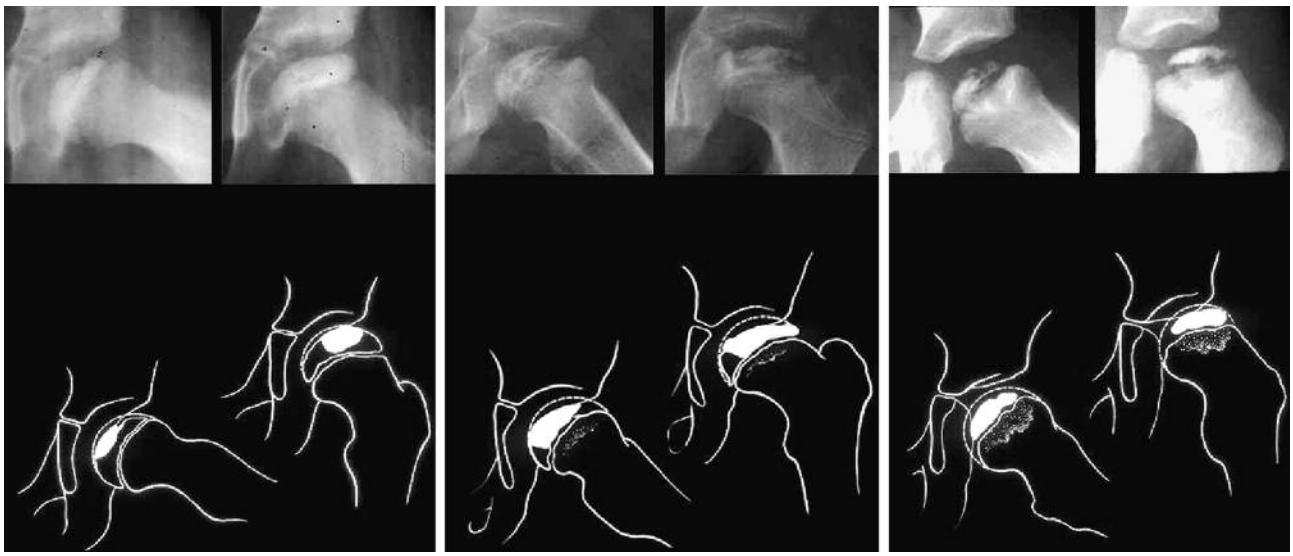

Fig. 1. — Catterall's classification for LCP disease (type 2 to 4, excluding stage 1 because of limited femoral head involvement)

compared to those with poorer prognosis. Various orthopaedic diseases have radiological classification systems, which were introduced to aid treatment decisions. LCP is no exception. Three main classification systems have been described (*3,7,15*). The Catterall (*3*) classification, first to be widely used, described four groups based on the amount of the femoral head that was involved (fig 1). Catterall suggested that groups 1 and 2 were benign, requiring symptomatic treatment whilst groups 3 and 4 had more extensive head involvement with less favourable outcome. In addition he described four "at risk signs" that indicated poor prognosis (Gage sign, calcification lateral to the epiphysis, lateral subluxation and the angle of the epiphyseal line).

The Salter and Thompson (*15*) classification is based on the extent of the subchondral fracture line, which appears early in the course of LCP and also in the resorptive phase. A fracture line involving less than half of the femoral head was associated with a good prognosis (fig 2a), whilst if more than half the head was involved, the prognosis was less favourable (fig 2b) (*15*).

Herring *et al* proposed a classification in relation to the height of the lateral pillar of the femoral head on antero-posterior (AP) radiographs. Hips are classified during the fragmentation stage into three groups (fig 3) (*9*).

The question as to which of these classifications is optimal to be utilised in executing management decisions is contentious. Three factors can be analysed to determine which is most useful ; validity of its prognostic value, ease of use and finally reproducibility and reliability. Each factor is important. However, reproducibility and reliability are critical. Inconsistency could allow different treatment regimes to be instigated for the same stage of the disease. This may explain the highly variable outcome in the literature with respect to LCP disease (*6,10,18*).

The purpose of this paper was to carry out a systematic review of all studies in the literature that have attempted to compare the above classifications directly for inter-observer and/or intra-observer agreement and deliver a weighted conclusion based on this.

MATERIALS AND METHODS

All articles assessing the reliability and/or reproducibility of the LCP classifications were considered eligible. The principal inclusion criteria was that the

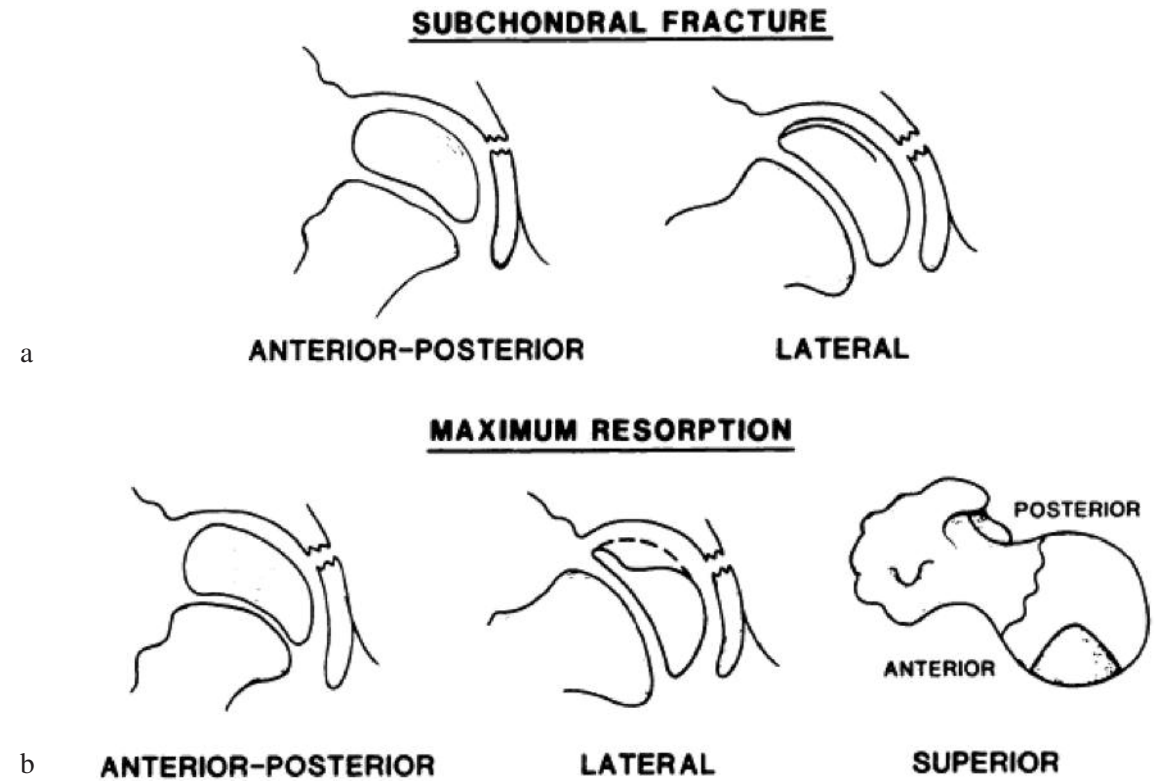

Fig. 2a & 2b. — Salter Thompson Classification for LCP disease with 2a showing a subchondral fracture of < 50% and 2b of >50% with subsequent appearance/s in the resorption phase.

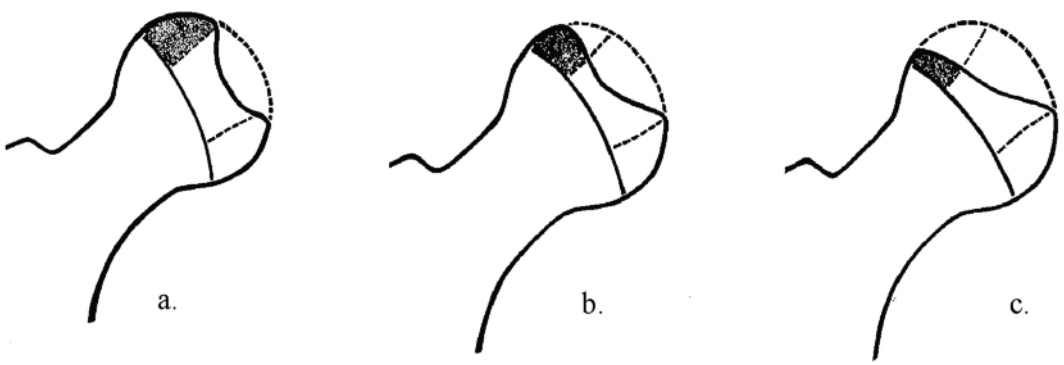

Fig. 3. — Herring Lateral Pillar classification for LCP disease. a. normal height of the lateral pillar maintained, b. Over 50% of the height of the lateral pillar maintained, c. Less than 50% of the lateral pillar maintained.

studies must compare the classification systems. Published studies were identified from the Medline (OVID) medical database by using a combination of the Mesh terms Perthes (LCP) and reliability. The two terms were chosen as keyword/s and the mesh term Perthes (LCP) was exploded on the subheading of classification. A combination search was then carried out. No expert opinion for identification of any further articles was obtained and no additional searches from major orthopaedic proceedings were performed. The papers

were retrieved by a single author *(DM*) and agreed to be suitable for analysis only after discussion and agreement with all co authors *(MC, DJL, AMT*). Studies that only compared two classification systems instead of all three were also considered suitable for analysis. Only papers published in English were included.

RESULTS

Only five of the eleven studies analysed inter- and intra-observer reliability between various LCP classifications (*1,11,14,16,20*). One further abstract from the reference appendages of most of the retrieved articles was reviewed, and deemed suitable for inclusion (*17*). No direct comparison was made between articles but instead the results are expanded in the analysis below under each classification subheading. The data could not be pooled because study designs varied. Among the study variations were that some used more observers than cases, some vice versa. Some used paediatric orthopaedic specialists only whilst others used a varied spectrum of inexperienced to experienced doctors. Two studies did not carry out any intra-observer assessment. Those that did generally did so with a suitable time gap between tests. Also the earlier studies tended to use a solitary set of radiographs whilst the latter studies found it of value to assess them serially over time. With regards to outcome measures, some used percentage agreement, whilst others used intraclass coefficients for statistical analysis to compare the inter-observer reliability of the classifications. Most, however, tended to use weighted Kappa statistics. The information pertaining to the sample population, study design, intervention and results and conclusions drawn from the articles are discussed, and table I presents a summary of this.

Salter Thompson

The Salter Thompson Classification was compared in 5 of the 6 studies. Only the Ritterbusch *et al* (*14*) study did not perform an analysis on it. In the various study designs around the same theme which may have differed on the number of observers or number of radiographs, average agreement values tended to be from poor to fair at best.

The Simmons *et al* (*17*) study offered better results (table IIa and IIb). In fact, their principal finding was that experience improved inter-observer agreement in both classifications, being superior for the Salter-Thompson system. They deduce that its simplicity (two groups instead of four) and its use earlier in the course of treatment (the subchondral fracture line is observed earlier) are advantageous. However, the observers in this study analyzed the radiographs at one point only. The radiographs here were of children who presented to a tertiary unit within two years. The radiograph inclusion criteria were not commented upon, but they were of 'good quality' possibly inferring a selection bias. Therefore, it is likely that the selection of cases for this analysis was such that the fracture line existed, suggesting this may have been a skewed group. The subchondral fracture is difficult to identify if the child presents late, which in reality is common.

Compare this to the Wiig *et al* (*20*) paper, where only in 30% of their series could a subchondral fracture line be seen. For the Salter-Thompson classification, moderate to good Kappa agreements at the initial phase became very low on the one year follow-up radiograph due to a high drop out rate. The subchondral fracture is difficult to identify if the child presents late, which in reality is common. This makes it difficult to apply the classification. Complicating this, in their study, there was more disagreement between observers when the fracture line was actually present ! (table IIIa/IIIb).

This is further highlighted by the Kalender *et al* (*1,11*) study, where classifications were made over a series of radiographs on the same patient. On the first review of radiographs (set one) the Salter-Thompson classification system was found to be most reliable in the pre-treatment series (although this was only 'fair') and was significantly better than the other two classification systems ('poor'). On second review the Salter-Thompson classification was not discussed, presumably because the subchondral fracture line was no longer distinguishable in the fragmentation phase. If a classification system cannot be employed for the second set of radiographs (during the fragmentation stage), a genuine comparison cannot be made. Finally, in the fragmentation stage (set 2) the intra- and inter-

Table I. — Summary of finding from the six studies retrieved

Journal	No	Title	Study Design	Sample size	Outcome measure	Results	Conclusion	Level of Evidence
Journal of Bone and Joint Surgery(Br)	1	Interobserver variability in grading Perthes disease	Observational (Cross sectional analysis)	40	Interobserver agreement via Kappa statistics	1)Kappa for Catterall = fair to good(across all grades) 2)Kappa for Salter Thompson = poor (residents) to excellent (seniors)	Salter Thompson is simpler to use and yields higher degree of interobserver agreement amongst senior personel	Level 4
Journal of Pediatric Orthop	2	Comparison of Lateral Pillar Classification and Catterall Classification of LCP disease	Observational (Cross sectional analysis)	78	1)Interobserver agreement 2)Correlating classification with final Stulberg outcome	Interobserver reliability for Herring classification was significantly better on both.	Herring classification more predictive of outcome and allows for better communication about disease	Level 4
Journal of Pediatric Orthop	3	Intraobserver and interobserver reliability of Catterall, Herring, Salter-Thompson and Stulberg classification systems in Perthes Disease	Observational (Cross sectional analysis)	10	Inter/intra observer agreement via intraclass correlation coefficient	ST and Catterall had better agreement results pre treatment but all systems faired poorly in set 2 radiographs.	Catterall and ST are preferred prior to treatment. No suitable classification system is ideal during treatment	Level 4
Journal of Pediatric Orthop	4	The importance of Surgeon's Experience on Intraobserver and Interobserver Reliability of Classifications Used for Perthes Disease	Observational (Cross sectional analysis)	10	Inter/intra observer agreement via intraclass correlation coeeficient	Agrement levels improved with experience from poor to fair/good in ST/Catterrall and fair to excellent with Herring	Reliability of classification systems improve with experience , but there is still unacceptable high error.	Level 4
Acta Orthopaedica	5	Interobserver reliability of radiographic classifications and measurements in the assessment of Perthes disease	Observational (Cross sectional analysis)	Variable (63-158)	Interobserver agreement via kappa statistics	At primary, Kappa for Catterall = moderate to good Herring = moderate to good ST = fair to good All were poor at follow up	1) Catterall classification not well agreed by less experienced doctors. 2) Salter Thompson not easy to use(subchondral fracture line only present in 30%) 3) Herring classification is simple to use and is useful in routine clinical work	Level 4
Journal of Pediatric Orthop	6	Reliability of radiological classifications used in LCP disease	Observational (Cross sectional analysis)	44	Inter and intraobserver agreement via kappa statistics	ST = 0.16/0.3,0.7 Herring=0.72/0.71,0..66 Catterall=0.43/ 0.38,0.58	Herring classification has the best agreement. Catterall classification can be improved if certain radiological parameters are optimized	Level 4

Table IIa. — Interobserver agreement using the Catterall system

Observer	Weighted Kappa	Standard Error	95% CI %	%Agreement
Staff	0.64	0.035	0.58 to 0.71	71
Fellows	0.51	0.041	0.43 to 0.59	66
Residents	0.49	0.044	0.41 to 0.58	64

Table IIb. — Interobserver agreement using the Salter Thompson system

Observers	Weighted Kappa	Standard Error	95% CI	%Agreement
Staff	0.99	0.002	0.99 to 1	93
Fellows	0.58	0.054	0.47 to 0.69	80
Residents	0.49	0.056	0.34 to 0.56	68

Interobserver agreement results from Simmons *et al* (*J Bone Joint Surg* 1990 ; 72-B : 202-204. Redrawn with permission).

Table IIIa. — Interobserver agreement, Catterall Classification

		Agreement						$Kappa_{(w)}$
		Catterall grouping						
	N_1	1	2	3	4	N_2	%	
Primary								
O/SS	158	8	11	38	43	100	63	0.49
TT/SS	76	2	5	35	15	57	75	0.62
Follow-up								
O/SS	115	2	1	12	51	66	57	0.28
TT/SS	63	–	–	20	26	46	73	–

Table IIIb. — Interobserver agreement, Salter- Thompson Classification

		Agreement				$Kappa_{(w)}$
		Salter-Thompson groups				
	N_1	A	B	N_2	%	
Primary						
O/SS	149	18	110	128	86	0.54
TT/SS	73	9	56	65	89	0.63
Follow-up						
O/SS	91	2	81	83	91	0.29
TT/SS	63	1	55	56	89	0.18

O local orthopaedic surgeons, SS and TT Paediatric Orthopaedic surgeons, Primary radiographs at time of Diagnosis, Follow-up radiographs at time of diagnosis N_1 number of patients examined, N_2 number of patients agreed upon, % percentage agreement, $Kappa_{(w)}$ weighted Kappa.

Interobserver agreement results from Wiig *et al.* (*Acta Orthop Scand* 2002 ; 73 : 523-530. Reprinted with permission).

observer agreements did not always improve with experience. It is this stage where treatments may have to be altered to improve outcome (i.e. decision to intervene with surgery). Difficulty in interpretation in this stage (set 2) offers an explanation to why there has been variability in the outcome from this condition.

Herring (Lateral Pillar)

Ritterbusch *et al* (*14*) found that the inter-observer reliability for the Herring classification (56 of 78 hips) was significantly ($p < 0.01$) better than the Catterall method (32 of 78 hips). This is shown below in table IV.

Their study only compared the two classifications above. They did not perform any intra-observer analysis, thereby restricting internal validity to their results. They commented that the majority of radiographs in their study were either Herring B/C at the initial stage, as is commonly the case in a tertiary referral centre. This may have naturally meant that the sets of radiographs may have been more complex. The three observers here included a medical student, an orthopaedic resident and a paediatric orthopaedic specialist. They

Table IVa. — Interobserver reliability of the classification systems

Classification system	Average ICC	95% CI	Rating
Catterall (set 1)	0.6203	0.4843-0.7205	Good
Salter-Thompson (set 1)	0.6037	0.4622-0.7079	Good
Herring (set 1)	0.5955	0.4521-0.7014	Good
Catterall (set 2)	0.5782	0.4286-0.5481	Good
Herring (set 2)	0.3878	0.1708-0.5481	Poor
Stulberg (set 3)	0.7912	0.7192-0.8448	Excellent

ICC, intraclass correlation coefficient ; CI, confidence interval. Set 1, before treatment ; Set 2, 6-12 months after the initiation of treatment ; Set 3, at least 5 years after the end of the treatment at skeletal maturity.

Table IVb. — Intraobserver reliability of the classification systems

Classification system	Average ICC	95% CI	Rating
Catterall (set 1)	0.6862	-0.7407 to 0.9180	Good
Salter-Thompson (set 1)	0.5758	-0.6667 to 1.0000	Good
Herring (set 1)	0.4946	-0.8989 to 1.0000	Fair
Catterall (set 2)	0.3864	-0.9756 to 0.9000	Poor
Herring (set 2)	0.1133	-0.8333 to 0.7240	Poor
Stulberg (set 3)	0.7733	-0.2881 to 0.9472	Excellent

ICC, intraclass correlation coefficient ; CI, confidence interval. Set 1, before treatment ; Set 2, 6-12 months after the initiation of treatment ; Set 3, at least 5 years after the end of the treatment at skeletal maturity.

Inter/intraobserver agreement results from Agus *et al* (*J Pediatr Orthop* 2004 ; 13 : 166-169. Reprinted with permission).

concluded that the Herring classification can be practised by junior personnel more reliably but this cannot be supported in the display of their results as observers 1 to 3 are not correlated with their level of experience.

Both Wiig *et al* (*20*) and Sambandan *et al* (*16*) demonstrated that the lateral pillar classification showed the best interrater agreement levels within their observers, and they attribute this to its ease of use (only an AP film is required and measurements can be quantified). Both these studies however mainly used experienced paediatric observer consultants, which may explain their superior results. Sambandan *et al* (*16*) incorporated a simple magnifying glass and ruler to aid quantitative "height" measurements which is commended. In the Wiig *et al* (*20*) study, all three observers categorised the radiographs at presentation and at 1 year follow-up, however, the most experienced surgeon only reviewed every other case. The Lateral Pillar Classification results here showed moderate to good agreement which persisted from the initial radiograph to the one year review, however, Kappa could not always be calculated because of missing data.

Experienced observers, who will be involved in the final decision making, are obvious choices to test a classification system. However, we feel that trainees and general orthopaedic surgeons are required to have a grasp of the classifications to aid discussion between colleagues and allow for suitable referral. A quantitative assessment of less experienced surgeons could have added weight to both studies.

In the two Kalenderer *et al* papers (*1,11*) which incorporated raw data from one study, 10 patients had three sets of radiographs (AP and frog lateral) before treatment (set 1), 6-12 months after initiation of treatment (set 2) and finally at five years following the end of treatment (set 3). Twelve orthopaedic surgeons (of 3-30 years experience) and 6 residents acted as observers. In set 1, the Herring classification was generally poor, this is despite the radiographs

Table V. — Intra and Interobserver reliability of classification systems among observers from Kalenderer *et al.* (*J Pediatr Orthop* 2005 ; 25 460-464. Reprinted with permission).

	Residents		Senior Surgeons		Paediatric Orthopaedists	
	Interobserver	Intraobserver	Interobserver	Intraobserver	Interobserver	Intraobserver
Catterall (set 1)	0.2625	0.6299	0.7326	0.6706	0.7410	0.7412
95% CI	-0.2520-0.5656	-0.8863-0.9202	0.5620-0.8492	-0.5176-0.9698	0.5664-0.8453	-0.3983-0.9796
Salter-Thompson (set 1)	0.4089	0.4788	0.3858	0.5557	0.7717	0.7177
95% CI	0.0013-0.6501	-0.5960-0.9656	-0.0888-0.6536	-0.8296-0.9346	0.6176-0.8636	-0.8788-1.0000
Herring (set 1)	0.1863	0.4663	0.6372	0.6454	0.7586	0.7631
95% CI	-0.3684-0.5162	-0.7686-0.9630	0.3607-0.7941	-0.8657-0.9317	0.5958-0.8558	-0.4312-1.0000
Catterall (set 2)	0.4763	0.4300	0.7427	0.6097	0.5644	0.5691
95% CI	0.1193-0.6886	-0.6610-0.9591	0.5465-0.8540	-0.6824-0.9153	0.2707-0.7398	-0.7327-0.9752
Herring (set 2)	0.0747	0.3573	0.4717	0.5777	0.4952	0.3582
95% CI	-0.5561-0.4498	-0.8120-0.9314	0.0690-0.7002	-0.7614-0.9290	0.1548-0.6984	-0.8541-0.7705
Stulberg (set 3)	0.7991	0.8088	0.8135	0.8165	0.7645	0.7906
95% CI	0.6621-0.8805	-0.6767-0.9869	0.6878-0.8886	-0.4176-0.9863	0.6022-0.8606	-0.4817-0.9836

Set 1, initial phase ; set 2, fragmentation phase ; set 3, healed phase at skeletal maturity.
Data are given as, average intraclass correlation coefficient and 95% confidence interval (CI).

being copied, digitized and visualized on a monitor. This is demonstrated in table Va/b. In set 2, they comment that there was 'good' agreement in the Herring classification. When the results are analysed by experience of observer, there was excellent agreement and both intra and inter-observer reliability in the Herring classification systems amongst the paediatric orthopaedic specialists in the first review. In set 2 however, the Herring classification systems mainly had fair agreement levels across all levels (there were two good agreement levels found and they were actually in the residents for inter-observer Herring classification and in the senior surgeons for inter-observer Catterall classification), demonstrated in table VI. This suggests that in the fragmentation stage (set 2) the intra- and inter-observer agreements for the Herring classification did not always improve with experience.

Table VI. — Interobserver reliability from Ritterbusch *et al.* (*J Pediatr Orthop 1993* ; 13 :200-202. Redrawn with permission).

Observers	Catterall Classification	LP Classification
Total	78	78
All 3 agreed	32	56
Obs 1 and 2	41	74
Obs 2 and 3	36	57
Obs 1 and 3	53	58

Catterall

The bulk of the classification systems agreement results for Catterall has been demonstrated in various tables above when compared. Of note, for the Catterall classification, moderate to good agreement obtained at the initial phases decreased to poor in the fragmentation phase. Secondly, the Catterall classification required increased experience to obtain better Kappa statistics and general use may pose difficulties. Returning to Rittersbuch *et al* (*14*), they commented that the majority of radiographs in their study were either Catterall 3 or 4 at the initial stage. The more complex appearance of these radiographs made their classification less

consistent as displayed in table VI (when compared with Herring).

The Sambandan *et al* (*16*) study is of interest because it offers some explanation to why the Catterall classification consistently produces poor inter-observer reliability. No previous paper had analysed this classification in this way. For the Catterall classification, radiological parameters such as identification of a sequestrum, posterior remodelling and anterior extent of the involved epiphysis had fair reliability but identification of the subchondral fracture, metaphyseal reaction and junction of involved to uninvolved region showed poor reliability. The authors of the paper suggest that improving these parameters could increase the reliability of the system, although they do not offer any substantial suggestions.

DISCUSSION

The purpose of any classification system is to guide decision making and indicate prognosis. There are numerous studies that have attempted to test inter-observer agreement for the classification systems of LCP, but often in isolation (*2,4,9,15*). This review includes for the first time all English language papers that have attempted to compare classifications. It shows inconsistent inter- intra-observer agreement levels. However, there are some salient points to be drawn from this analysis.

Firstly, none of the studies above attempted a multicentre analysis of the inter- intra-observer agreement. Without such data, it is difficult to get the confidence intervals to obtain a better estimate of the agreement levels. This may not be possible as kappa values tend to change with the prevalence of a condition (*5,12*), and therefore meaningful comparisons cannot be made between different studies. If however, the same series of radiographs were presented from one centre to another with observers of the appropriate or similar experience taking part, with suitable blinding for reproducibility and reliability, this could be circumvented. However, we acknowledge that this would be difficult to conduct.

Agreement between various levels and at multiple centres is perhaps not necessary with final decision making. This was suggested in the Sambandan *et al* (*16*) study. After all, this is a complex condition in which treatment decisions are most appropriately carried out by the surgeons with the most knowledge and experience to execute definitive treatment. They should have the best agreement on classification. But it was interesting to note that this was not the case with respect to certain classifications (Catterall in set 2 radiographs (*11*) and Salter-Thompson). However, management in tertiary units may only be practical in a well funded healthcare system and may not be possible in the developing world. In these countries it would be expected that a local orthopaedic surgeon will manage LCP disease. Therefore, if the classification system is to be truly global, there must be better agreement across all experience levels.

A common criticism in the selection of radiographs for all the above studies was that a comment was often made that they were of good quality or chosen because they were in the fragmentation stage. This in facts infers selection bias. This however is not always borne out in clinical practice. Only in the Wiig *et al* study (*20*) was there comment that, at the time of diagnosis, 45% of the hips were in the initial phase, 51% in the fragmentation phase whilst another 4% were in the reossification phase. Although they do not state who determined this, the spread series of radiographs appear to be random, reflecting a spectrum of cases mirroring day to day practice.

Generally though, most of the studies demonstrated that experience brought a better level of agreement. However, there was no analysis found to quantify improvement as an individual surgeon becomes more experienced. Analysis of kappa agreement amongst trainees at three different points in their training may have been useful. This allows informed discussions between junior staff, senior general orthopaedic surgeons and paediatric specialists at tertiary referral centres.

In clinical decision making, not all categories in the various Perthes classification carry equal weight. In a child under six years, it is likely that symptomatic supportive therapy will produce a satisfactory outcome, even if an agreement on a Herring A or B cannot be made. However, if the disagreement is whether there is a subchondral

fracture, there could be a delay in diagnosis and treatment, with a potentially adverse outcome.

The ability to predict prognosis is the most important aspect of any classification system. This ability needs to be reliable and reproducible. The classification systems in use for LCP disease provide a framework to guide the practising orthopaedic surgeon. All three classifications have disadvantages. On balance the Herring classification is the most reliable because of its ease of use. The Salter-Thompson classification is best avoided by less experienced surgeons, mainly because the subchondral fracture is difficult to delineate on plain radiographs (particularly if presentation is delayed beyond the fragmentation phase). Finally, it is possible that improved digital radiographic technology may contribute to better agreement levels in the future.

REFERENCES

1. **Agus H, Kalenderer O, Eryanlmaz G, Ozcalabi IT.** Intraobserver and interobserver reliability of Caterall, Herring, Salter-Thompson and Stulberg classification systems in Perthes disease. *J Pediatr Orthop* 2004 ; 13 : 166-169.
2. **Catterall A.** The natural history of Legg-Calve-Perthes Disease. *J Bone Joint Surg* 1971 ; 53-B : 37-53.
3. **Catterall A.** Natural history, classification, and X-ray signs in Legg-Calve-Perthes disease. *Acta Orthop Belg* 1980 ; 46 : 346-351.
4. **Christiansen F, Soballe K, Ejsted R, Luxhoj T.** The Catterall classification of Perthes disease : an assessment of reliability. *J Bone Joint Surg* 1986 ; 68-B : 614-615.
5. **Cohen J.** Weighted kappa : nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968 ; 70 : 213.
6. **Gigante C, Frizzerio 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 : 345-349.
7. **Hardcastle PH, 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* 1980 ; 62-B : 428-431.
8. **Herring JA.** *Legg-Calve-Perthes Disease*. American Academy of Orthopaedic Surgeons (Monograph series).
9. **Herring JA, Neustadt JB, Williams JJ, Early JS, Browne RH.** The lateral pillar classification of Legg-Calve-Perthes disease. *J Pediatr Orthop* 1992 ; 12 : 143-150.
10. **Ismail AM, Macnicol MF.** Prognosis in Perthes disease A comparison of radiological predictors. *J Bone Joint Surg* 1998 ; 80-B : 310-314.
11. **Kalenderer O, Agus H, Ozcalabi IT, Ozluk S.** The importance of surgeons' experience on intraobserver and interobserver reliability of classifications used for Perthes disease. *J Pediatr Orthop* 2005 ; 25 : 460-464.
12. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. *Biometrics* 1977 ; 33 : 159-174.
13. **Podeszwa DA, Stanitski CL, Stanitski DF, Woo R, Mendelow MJ.** 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 :562-565.
14. **Ritterbusch JF, Shantaram SS, Gelinas C.** Comparison of laterall pillar classification and Catterall classification of Legg-Calve-Perthes disease. *J Pediatr Orthop* 1993 ; 13 : 200-202.
15. **Salter RB, Thompson GH.** Legg-Calve-Perthes Disease. The prognostic significance of the subchondral fractures and a two-group classification of the femoral head involvement. *J Bone Joint Surg* 1984 ; 66-A : 479-489.
16. **Sambandan NS, Gul A, Shankar R, Goni V.** Reliability of radiological classifications used in Legg-Calve-Perthes disease. *J Pediatr Orthop* 2006 ; 15 267-270.
17. **Simmons ED, Graham HK, Szalai JP.** Interobserver variability in grading Perthes disease. *J Bone Joint Surg* 1990 ; 72-B : 202-204.
18. **Stulberg SD, Cooperman DR, Wallensten R.** The natural history of Legg-Calve-Perthes disease. *J Bone Joint Surg* 1981 ; 63-A :1095 – 1108.
19. **Van Dam BE, Crider RJ, Noyes JD, Larsen LJ.** Determination of Catterall classification in Legg-Calve-Perthes disease. *J Bone Joint Surg* 1981 ; 63-A : 906-914.
20. **Wiig O, Terjesen T, Svenningsen S.** Interobserver reliability of radiographic classifications and measurements in the assessment of Perthes disease. *Acta Orthop Scand* 2002 ; 73 : 523-530.