RESEARCH ARTICLE

Interobserver and intraobserver reliability of Salter-Harris classification of physeal injuries

Tzavellas AN, Kenanidis E, Potoupnis M, Pellios S, Tsiridis E, Sayegh F

3rd Academic Orthopaedic Unit, Medical School, Aristotle University of Thessaloniki, "Papageorgiou" General Hospital, Thessaloniki, Greece

Abstract

Background: Prognostic value of Salter-Harris (SH) classification is well established. Its reliability, however, can be questioned. We aim to evaluate the interobserver and intraobserver reliability of SH classification and to correlate the level of rater's experience with the correct scoring for each SH subclass.

Methods: Twenty-eight independent raters stratified in three levels of seniority evaluated 50 randomly selected radiographs of physeal injuries. The interval for intraobserver reliability was 12 weeks. The overall agreement between raters was assessed using kappa statistics. Student's t-test and Spearman correlation coefficient used to compare results between groups.

Results: Overall kappa for interobserver reliability was 0.45. The mean kappa difference between specialists and residents was significant ($p < 0.001$). The mean kappa difference was also significant between senior and junior residents ($p < 0.001$), favoring senior residents. Intraobserver kappa differs between specialists (0.55) and residents (0.49), but this did not reach statistical significance ($p = 0.34$). SH type II and III demonstrated the highest category-specific kappa coefficient. Seniority was correlated significantly with the number of correct answers (Spearman rho $= 0.6$ $p = 0.001$).

Conclusions: Moderate interobserver reliability that was improved with greater rater's experience was found. Type II and III are the best scored regardless rater's experience. Type I, IV, and V when in doubt, require additional imaging. Hippokratia 2016, 20(3): 222-226

Keywords: Reliability, Salter-Harris classification, fractures, growth plate

Corresponding Author: Anastasios-Nektarios Tzavellas, MD, MSc, 3 Spirou Loui str., PS 55534, Pylaia, Greece, tel: +302311280314, mob: +306973520878, e-mail: tzav_a@hotmail.com

Introduction

The management of physeal injuries is based on patient's history, physical findings, and radiographic evaluation[1]. The radiological assessment of specific physeal injuries is challenging due to occasional obscurity between the growth plate and fracture lines[2-4]. Several classification systems have been suggested for the evaluation of physeal fractures[4-7]. Among them, the Salter-Harris (SH) classification, introduced half a century ago, is widely accepted[1,4]. Specific studies, however, challenged its prognostic value[3,5]. A few studies tried to evaluate the reliability of certain anatomic regions or fracture patterns[1,8,9]. Though, to the best of our knowledge, the interobserver and intraobserver agreement on SH classification among orthopedic surgeons of stratified seniority has never been studied in English literature.

Aim of this study was to evaluate the reliability of SH classification. Our objectives were firstly to estimate the interobserver and intraobserver agreement and secondly to correlate the level of training with the correct scoring for each SH subclass.

Material and Methods

This is a reliability study that was conducted by the 3rd Academic Orthopaedic Unit of "Papageorgiou" hospital. Even though ethical approval did not deem necessary for a reliability study as radiographs were from patients that received the standard care and their data was protected, the study was initially approved by the institution's Scientific and Ethical Committee and this approval was finalised when data analysis was completed (2016). It was conducted between January and December 2013, in accordance with the World Medical Association Declaration of Helsinki of 1964 as revised in 1983 and 2000.

Radiographic selection criteria

We included patients aged between 2-16 years who had suffered physeal injuries. They were identified using ICD-10 coding in our trauma-radiology database (IMPAX 6.4 Clinician, Agfa Healthcare, Mortsel, Belgium).

Out of 513 recordings in our database, an initial sample of 200 suitable cases was selected during a 12 consecutive month period (2013). The selection was based on a consensus assessment between three orthopedic surgeons and a musculoskeletal radiologist who did not participate further as raters. Consensus assessment excluded cases of other fracture patterns (i.e., buckle fractures, greenstick or complete metaphy-

seal fractures). Radiographs of poor quality or not consisting of a set of anteroposterior and lateral views were excluded too. Correct scoring was established by the same group of specialists based on the full medical records, magnetic resonance imaging (MRI) or computed tomography (CT), additional imaging, and surgical intervention notes.

Finally, 50 cases were randomly selected using Statistical Package for the Social Sciences (SPSS) statistical software, version 21 (IBM Corp., Armonk, NY, USA). Table 1 shows the distribution of cases in relation to the anatomic sites of injury and types of SH classification.

Radiograph processing

Radiographs were processed by Photoshop CS5 Extended, Version 12.0x32 (Adobe, California, USA). All radiographs had personal data protected, and health information masked out. To eliminate repetition bias, all radiographs were placed in different files. Each file was comprised from the same radiographs arranged in different sequence. Six such files were created and used by the raters.

Radiographs' evaluation

Twenty-eight blinded independent raters participated in the current study. Three levels of seniority were employed, namely specialists orthopedic surgeons, senior and junior orthopedic residents. Senior residents had 3-6 years in training, while junior residents had a maximum of two years in training. One of the six files was randomly assigned to each rater. Before evaluating the radiographs, all raters were encouraged to study SH classification.

The interobserver reliability was estimated according to the first evaluation of radiographs to prevent recall bias. Twelve weeks later, each rater reevaluated a different file (same radiographs in a different order) to determine intraobserver reliability.

Statistics

Kappa statistics were used for evaluating reliability[10]. Fleiss' kappa (κ) was used to estimate agreement among raters[11] and category-specific kappa coefficient to evaluate the interobserver reliability among different fracture types[11]. Intraobserver reliability was evaluated using the mean Cohen's kappa measurement[12-14]. Kappa values were interpreted based on Landis and Koch's guidelines[15].

Standard statistical methods were used for descriptive statistics. The normality of data distribution was tested according to the Kolmogorov-Smirnov test. Student's t-test was used for comparisons between two groups of raters. Spearman correlation coefficient was used to determine the correlation between the time training and the number of correct answers. All statistical tests were two-tailed. Analyses were performed with SPSS statistical software.

Sample size estimation

The number of radiographs rated was determined based on the number of independent raters, the classification subclasses as well as the expected kappa value[11,16-19]. The kappa difference for interobserver reliability between groups of different experience ranges between 0.1-0.2 in other reliability studies[14,18-20]. To estimate the final sample size, we took into consideration that clinically important reliability means a value of kappa ≥0.4, sufficient power is 0.8, and α value is 0.05[16,19]. The statistical analysis showed that, in order to find a difference between two individual raters of a value between 0.1-0.2, at least 130 radiographs had to be included in the study[14,18-20]. According to Altaye et al, duplication of raters limits the need of subjects almost to the half[16]. Thus in this study, we increased the number of raters to 28, so a sample of 50 radiographs was considered appropriate to reach firm conclusions.

Results

Fourteen specialists and 14 resident orthopedists participated in the study. Residents were divided into two levels of in training seniority; there were seven junior residents and seven senior residents (mean five years in training). Twenty-eight doctors (100 %) responded to the first evaluation set of radiographs while 24 (response 85.7 %) returned the second set of radiographs.

Interobserver agreement

Overall kappa was 0.45 ($p < 0.001$), representing an

Table 1: Distribution of the 50 randomly selected radiographs of physeal injuries in relation to the anatomic sites of injury and the types of Salter-Harris classification.

Anatomic site	Type I	Type II	Type III	Type IV	Type V	Overall
Distal radius	2	6	2	0	2	12
Metacarpals and fingers	1	7	2	1	1	12
Distal tibia	1	2	3	1	1	8
Distal fibula	1	1	0	0	1	3
Metatarsals and toes	0	3	2	1	0	6
Proximal humerus	2	2	0	0	0	4
Distal humerus	0	0	1	1	0	2
Distal femur	0	2	1	0	0	3
Overall	7	23	11	4	5	50

overall agreement of 60.6 %. The kappa value for specialists was 0.53 (p <0.001), but it was only 0.39 (p <0.001) for the residents. An agreement of 67.2 % was achieved between specialists and 55.4 % between residents (Table 2). The mean kappa value difference between specialists and residents as well as between specialists and senior residents was statistically significant (p <0.001). Kappa value for junior residents was 0.3 and for senior residents 0.44 (p <0.001).

Intraobserver agreement

The mean kappa for intraobserver reliability was 0.52, demonstrating 65.3 % of agreement (Table 3). The overall kappa achieved by specialists was higher than residents, but not significantly (0.55 versus 0.49, p =0.346). Kappa value for intraobserver reliability did not differ between specialists and senior residents (p =0.84); there was a significant difference however between the two groups of the in training seniority (p =0.037) (Table 3).

Category-specific agreement

SH type II fractures, followed by type III, demonstrated the highest category-specific kappa coefficient (Table 4). Specialists demonstrated higher kappa coefficient for each category of SH classification than residents (Table 4).

Correct scoring

At first assessment, the accuracy of the evaluations of the radiographs reached 68.7 %. Specialists gave significantly more correct answers than residents (73.2 % versus 64.1 %, p =0.002) (Table 5). Overall, seniority correlated significantly with the number of correct answers (Spearman rho =0.6, p =0.001).

Discussion

A modern fracture classification should be clinically relevant in deciding the correct treatment, follow-up strategy, and prognosis. Our study demonstrated moder-

Table 2: Interobserver reliability between the groups of different experience for the Salter-Harris classification.

	Fleiss' kappa	proportion of agreement	p value
Specialists	0.53	0.672	<0.001
Residents (overall)	0.39	0.554	<0.001
Senior residents	0.44	0.596	<0.001
Junior residents	0.30	0.485	<0.001
Overall	0.45	0.606	<0.001

Table 3: Intraobserver reliability between the groups of different experience for the Salter-Harris classification.

	Fleiss' kappa	proportion of agreement	p value
Specialists	0.55	0.686	-
Residents(overall)	0.49	0.627	0.346*
Senior residents	0.56	0.687	0.84*
Junior residents	0.37	0.532	0.037**
Overall	0.52	0.653	-

*: Comparison of mean Fleiss' kappa between specialists and residents (senior or overall). Tests performed using Students t-test, **: Comparison of mean Fleiss' kappa between senior and junior residents. Tests performed using Students t-test.

Table 4: Category-specific reliability of Salter-Harris classification for each of the different experience groups of raters for each Salter-Harris type.

	Specialists	Residents	Senior residents	Junior residents	Total	p*
Type I	0.5	0.34	0.36	0.28	0.41	0.018
Type II	0.69	0.59	0.64	0.52	0.62	0.35
Type III	0.61	0.41	0.53	0.24	0.5	0.001
Type IV	0.32	0.16	0.21	0.03	0.22	0.005
Type V	0.19	0.14	0.14	0.1	0.17	0.32

*: Comparison between category-specific reliability of specialists and residents for each Salter-Harris type. Tests performed using Students t-test.

Table 5: Percentage of correct answers for each of the different experience groups of raters.

	Correct answers (%)	p value
Specialists	73.2	-
Residents (overall)	64.1	0.002*
Senior residents	66.7	0.009*
Junior residents	60.6	0.235**
Overall	68.7	-

*: Comparison of the percentage of correct answers between specialists and residents (senior or overall). Tests performed using Students t-test, **: Comparison of the percentage of correct answers between senior and junior residents. Tests performed using Students t-test.

ate interobserver reliability for SH classification with an overall agreement reaching 60.6 %.

A limited number of studies have tried to evaluate the reliability of SH classification[1,8,9]. The majority of them employed a small number of radiographs and raters that were usually specialists in pediatric orthopedics. At the Emergency Department, however, initial assessment of such fractures relies on the judgment of residents. To overcome the latter limitation, in our study we stratified the assessment in three different levels of seniority. Also, we provided a strong power calculation to validate our results. The raters' response of more than 85 % satisfied the precalculated power analysis necessary to reach firm conclusions. The large number of raters and radiographs employed in the current study strengthen our outcomes compared with similar studies in the literature[12,13,21,22].

The inclusion of physeal fractures regardless the anatomic site of injury is an important divergence of our study. It could be a weakness, as it enlarges the study's field of interest. Both clinical and radiological evaluations of the elbow joint in a growing skeleton are considered to be problematic. The number, sequence, morphology, and complexity of the ossification centers of the elbow joint can vary significantly[23]. On the other hand, our aim was to evaluate SH classification as a whole in order to reveal such limitations of the classification. A greater number of radiographs per anatomic region is undoubtedly needed to reach reliable results on each joint.

A natural limitation, as it appears in similar studies in the literature[24], was the non-equal number of cases at each SH subclass. There was a greater number of SH II cases compared with the other subclasses, especially the fourth and fifth (Table 1). This was expected as we used a random sample from our trauma-radiology database. Type II is reported to be the most common type of physeal fractures whereas type V the rarest[2,3,25].

Raters were encouraged to study the classification before evaluating the radiographs. That could potentially be considered a limitation. Cicchetti however, considers that observers' training prior to rating is of major importance as it increases the statistical power of a reliability study[17].

The moderate interobserver reliability found could be explained by the following parameters. The most important are the division of SH classification in five subgroups and the varying level of experience of the raters. The ensuing are the obscurity of radiographic appearance of the physeal plate in the various stages of skeletal development and the expected "blindness" of raters towards the mechanism of injury[26].

The greater the number of subclasses, the smaller the expected reliability[24,26]. One of the highest interobserver reliability (0.74) has been reported by Barton et al[8]. It concerns the Gartland classification of supracondylar humeral fractures that has only three subclasses[8]. On the contrary, Flikkila et al reported poor interobserver reliability (0.18) of the "Arbeitsgemeinschaft fur Osteosynthesefragen" (AO) fracture classification of the distal radius when using the original 27 subclasses[27]. In this study, limiting the original AO subclasses from 27 to only two, improved interobserver reliability (0.18 to 0.48). Compared to the latter two studies, in our setting, 24 raters evaluated 50 radiographs addressing to five subgroups showing an overall interobserver kappa of 0.45 and an intraobserver kappa of 0.52 in a 12-weeks interval. Therefore, our results coincide to the recorded experience in the published literature regarding physeal and skeletal injuries[8,26,27].

In our study, specialists demonstrated significantly greater agreement and gave considerably more correct answers than the residents. The effect of experience on reliability has been previously described in other classification systems[14,20,26]. Randsborg et al reported variation in interobserver reliability among raters of different experience that evaluated distal non-physeal radius fractures in children[14]. In this study, agreement among specialists was greater than senior and junior residents[14]. Classifying physeal fractures, however, is more challenging compared to the above anatomical classification. The latter explains low agreement found in our setting and simultaneously identifies the limitations of SH classification. Thus, SH classification has been proved to be resilient to time testing, having however certain limitations.

In addition, the moderate intraobserver agreement (k =0.52) found in our study was not affected by the time interval; Kottner et al supported that time interval among the two evaluations should be enough to avoid recall bias, but also small to prevent a significant change of rater's experience[28]. The time interval of 12 weeks used in our study and other studies[14] is considered satisfactory.

In our study, SH IV and V demonstrated the lowest reliability among the raters. Salter & Harris pointed out these difficulties, especially for type V fracture[4]. Mann et al confirmed that the initial diagnosis of type V fractures is rare[2]. Our study aimed to identify those limitations and to alert musculoskeletal physicians of variable experience to raise concerns regarding certain SH subclasses that may require further evaluation. CT, MRI or both are compulsory when there is doubt about diagnosis and decision making of physeal injuries. Lippert et al strongly recommend the use

of MRI or CT scan on every SH III fracture of the distal femur[29]. MRI visualizes better the soft tissues, especially physeal plate structures[30,31]. It can be used successfully for early detection of physeal fractures' complications[30,32,33]. However, it cannot be recommended routinely.

The clinical importance of our findings is that type I and V fractures can be easily diagnostic outliers when evaluated by the junior personnel in the Emergency Department. Injuries close to physes especially those of the early childhood should be reevaluated by a specialist pediatric orthopedic surgeon. Plain radiographs are still considered the diagnostic cornerstone of evaluation, classification, and decision of treatment of pediatric injuries[34]. Power analysis and the two levels of randomization of our design increases our confidence to support that, when in doubt, even in experienced hands further imaging including MRI or CT scan should be advised.

Conflict of interest

None of the authors has any conflicts of interest to declare.

References

1. Thawrani D, Kuester V, Gabos PG, Kruse RW, Littleton AG, Rogers KJ, et al. Reliability and necessity of computerized tomography in distal tibial physeal injuries. J Pediatr Orthop. 2011; 31: 745-750.
2. Mann DC, Rajmaira S. Distribution of physeal and nonphyseal fractures in 2,650 long-bone fractures in children aged 0-16 years. J Pediatr Orthop. 1990; 10: 713-716.
3. Mizuta T, Benson WM, Foster BK, Paterson DC, Morris LL. Statistical analysis of the incidence of physeal injuries. J Pediatr Orthop. 1987; 7: 518-523.
4. Cepela DJ, Tartaglione JP, Dooley TP, Pate PN. Classifications In Brief: Salter-Harris Classification of Pediatric Physeal Fractures. Clin Orthop Relat Res. 2016; 474: 2531-2537.
5. Chadwick CJ, Bentley G. The classification and prognosis of epiphyseal injuries. Injury. 1987; 18: 157-168.
6. Joeris A, Lutz N, Blumenthal A, Slongo T, Audigé L. The AO Pediatric Comprehensive Classification of Long Bone Fractures (PCCF). Acta Orthop. 2017; 88: 123-128.
7. Sferopoulos NK. Classification of distal radius physeal fractures not included in the salter-harris system. Open Orthop J. 2014; 8: 219-224.
8. Barton KL, Kaminsky CK, Green DW, Shean CJ, Kautz SM, Skaggs DL. Reliability of a modified Gartland classification of supracondylar humerus fractures. J Pediatr Orthop. 2001; 21: 27-30.
9. Slongo T, Audigé L, Clavert JM, Lutz N, Frick S, Hunter J. The AO comprehensive classification of pediatric long-bone fractures: a web-based multicenter agreement study. J Pediatr Orthop. 2007; 27: 171-180.
10. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006; 59: 1033-1039.
11. Roberts C. Modelling patterns of agreement for nominal scales. Stat Med. 2008; 27: 810-830.
12. Maripuri SN, Rao P, Manoj-Thomas A, Mohanty K. The classification systems for tibial plateau fractures: how reliable are they? Injury. 2008; 39: 1216-1221.
13. Niemeyer T, Wolf A, Kluba S, Halm HF, Dietz K, Kluba T. Interobserver and intraobserver agreement of Lenke and King classifications for idiopathic scoliosis and the influence of level of professional training. Spine (Phila Pa 1976). 2006; 31: 2103-2107; discussion 2108.
14. Randsborg PH, Sivertsen EA. Classification of distal radius fractures in children: good inter- and intraobserver reliability, which improves with clinical experience. BMC Musculoskelet Disord. 2012; 13: 6.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33: 159-174.
16. Altaye M, Donner A, Eliasziw M. A general goodness-of-fit approach for inference procedures concerning the kappa statistic. Stat Med. 2001; 20: 2479-2488.
17. Cicchetti DV. Sample size requirements for increasing the precision of reliability estimates: problems and proposed solutions. J Clin Exp Neuropsychol. 1999; 21: 567-570.
18. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Stat Med. 1992; 11: 1511-1519.
19. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005; 85: 257-268.
20. Lindenhovius A, Karanicolas PJ, Bhandari M, Ring D; COAST Collaborative. Radiographic arthrosis after elbow trauma: interobserver reliability. J Hand Surg Am. 2012; 37: 755-759.
21. Elzinga M, Segers M, Siebenga J, Heilbron E, de Lange-de Klerk ES, Bakker F. Inter- and intraobserver agreement on the Load Sharing Classification of thoracolumbar spine fractures. Injury. 2012; 43: 416-422.
22. Ramappa M, Bajwa A, Singh A, Mackenney P, Hui A, Port A. Interobserver and intraobserver variations in tibial pilon fracture classification systems. Foot (Edinb). 2010; 20: 61-63.
23. Beaty JH, Kasser JR. The elbow region: General concepts in the pediatric patient. In: Beaty JH, Kasser JR (eds). Rockwood and Wilkins' Fracture in children. 6th edition, Lippincott Williams & Wilkins Publishers, Philadelphia, 2010, 475-486.
24. Schneidmüller D, Röder C, Kraus R, Marzi I, Kaiser M, Dietrich D, et al. Development and validation of a paediatric long-bone fracture classification. A prospective multicentre study in 13 European paediatric trauma centres. BMC Musculoskelet Disord. 2011; 12: 89.
25. Joeris A, Lutz N, Blumenthal A, Slongo T, Audigé L. The AO Pediatric Comprehensive Classification of Long Bone Fractures (PCCF). Acta Orthop. 2017; 88: 129-132.
26. Andersen DJ, Blair WF, Steyers CM Jr, Adams BD, el-Khouri GY, Brandser EA. Classification of distal radius fractures: an analysis of interobserver reliability and intraobserver reproducibility. J Hand Surg Am. 1996; 21: 574-582.
27. Flikkilä T, Nikkola-Sihto A, Kaarela O, Pääkkö E, Raatikainen T. Poor interobserver reliability of AO classification of fractures of the distal radius. Additional computed tomography is of minor value. J Bone Joint Surg Br. 1998; 80: 670-672.
28. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011; 64: 96-106.
29. Lippert WC, Owens RF, Wall EJ. Salter-Harris type III fractures of the distal femur: plain radiographs can be deceptive. J Pediatr Orthop. 2010; 30: 598-605.
30. Jaramillo D, Hoffer FA, Shapiro F, Rand F. MR imaging of fractures of the growth plate. AJR Am J Roentgenol. 1990; 155: 1261-1265.
31. White PG, Mah JY, Friedman L. Magnetic resonance imaging in acute physeal injuries. Skeletal Radiol. 1994; 23: 627-631.
32. Lurie B, Koff MF, Shah P, Feldmann EJ, Amacker N, Downey-Zayas T, et al. Three-dimensional magnetic resonance imaging of physeal injury: reliability and clinical utility. J Pediatr Orthop. 2014; 34: 239-245.
33. Shi DP, Zhu SC, Li Y, Zheng J. Epiphyseal and physeal injury: comparison of conventional radiography and magnetic resonance imaging. Clin Imaging. 2009; 33: 379-383.
34. Lohman M, Kivisaari A, Kallio P, Puntila J, Vehmas T, Kivisaari L. Acute paediatric ankle trauma: MRI versus plain radiography. Skeletal Radiol. 2001; 30: 504-511.