Documentation of fracture severity with the AO classification of pediatric long-bone fractures

Theddy Slongo[1], Laurent Audigé[2], Nicolas Lutz[3], Steve Frick[4], Peter Schmittenbecher[5], James Hunter[6] and Jean-Michel Clavert[7]

[1]Deptartment of Pediatric Surgery, University Children's Hospital, CH-3010 Bern, [2]AO Clinical Investigation and Documentation, AO Foundation, Stettbachstrasse 10, CH-8600 Dübendorf, [3]Service de Chirurgie Pediatrique CHUV – HEL, CH-1011 Lausanne, Switzerland, [4]Department of Orthopaedic Surgery, Carolinas Medical Center, Charlotte, NC 28232, USA, [5]St. Hedwig's Hospital, Clinical Center Barmherzige Brüder, DE-93049 Regensburg, Germany, [6]Department of Paediatric Orthopaedic Surgery, Queen's Medical Centre, Nottingham University Hospital, NG7 2UH, UK, [7]Centre Hospitalier Hautepierre, Service de Chirurgie Infantile, FR-67098 Strasbourg Cedex, France

Correspondence LA: laurent.audige@aofoundation.org

Submitted 06-. Accepted 06-09-01

Background **The AO comprehensive pediatric long-bone fracture classification system describes the localization and morphology of fractures, and considers severity in 3 categories: (1) simple, (2) wedge, and (3) complex. We evaluated the reliability and accuracy of surgeons in using this rating system.**

Material and methods **In a first validation phase, 5 experienced pediatric (orthopedic) surgeons reviewed radiographs of 267 prospectively collected pediatric fractures (agreement study A). In a second study (B), 70 surgeons of various levels of experience in 15 clinics classified 275 fractures via internet. Simple fractures comprised about 90%, 99% and 100% of diaphyseal (D), metaphyseal (M), and epiphyseal (E) fractures, respectively.**

Results **Kappa coefficients for severity coding in D fractures were 0.82 and 0.51 in studies A and B, respectively. The median accuracy of surgeons in classifying simple fractures was above 97% in both studies but was lower, 85% (46–100), for wedge or complex D fractures.**

Interpretation **While reliability and accuracy estimates were satisfactory as a whole, the ratings of some individual surgeons were inadequate. Our findings suggest that the classification of fracture severity in children should be done in only two categories that distinguish between simple and wedge/complex fractures.**

■

The AO Pediatric Expert Group (PAEG) in cooperation with AO Investigation and Documentation (AOCID) and the International Working Group for Paediatric Traumatology (iAGKT) have developed the first comprehensive classification of long-bone fractures in children (Slongo et al. 2006), based on the Müller AO classification for adults (Müller and Narzarian 1990).

Audigé et al. (2005) recommended that three research phases should be successively completed before a classification can be considered as validated. The first two phases involve series of agreement studies to evaluate the reliability and accuracy of the classification—initially by experienced and trained clinicians, and then more pragmatically by surgeons of different levels of experience (Audigé et al. 2004b, Slongo et al. 2007). The classification system should be clinically relevant, reliable, and accurate (Burstein 1993, Martin and Marsh 1997, Garbuz et al. 2002, Audigé et al. 2004a, Audigé et al. 2005). Only then can it be used for documentation and evaluation of treatment options and outcomes in a third phase of validation.

The first two phases have been completed and presented for the classification of pediatric long-bone fractures according to the type (epiphyseal, metaphyseal, or diaphyseal) and the pattern-specific child code, and showed satisfactory results (Audigé et al. 2004b, Slongo et al. 2006, 2007).

DOI 10.1080/17453670710013753

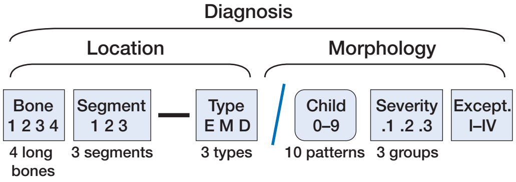


Figure 1. Overall structure of the pediatric fracture classification. From Slongo et al. (2006b).

The objective of our study was to estimate the inter-rater reliability and individual accuracy of surgeons, specifically regarding the documentation of severity of supracondylar, radial and tibial fractures (as documented by the number of fracture fragments).

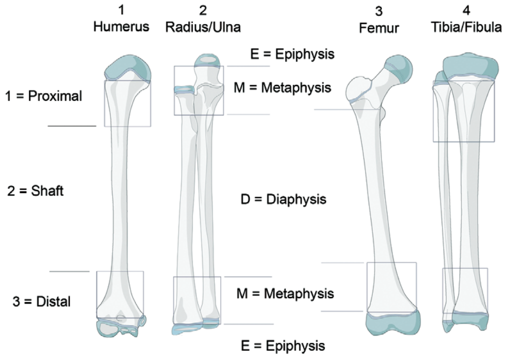


Figure 2. Coding of pediatric long bones and localization of fractures.

Material and methods

Fracture classification system

The AO comprehensive classification system for pediatric long-bone fractures (Slongo et al. 2006) includes several dimensions related to localization and morphology (Figure 1). Briefly, the anatomy is related to the 4 long bones and their 3 segments, defined as proximal (1), shaft (2) and distal (3). It is further described by the fracture type, recorded as epiphyseal (E), metaphyseal (M), or diaphyseal (D), whereby proximal and distal fractures are classified as E or M and shaft fractures are always D (Figure 2). The distinction between metaphyseal and diaphyseal fractures is achieved by localizing the center of fracture lines relative to a square drawn over the respective growth plates. The morphology of the fracture is documented by a type-specific child pattern code (Table 1), a severity code, and also an additional code for displacement of specific fractures such as supracondylar fractures and radial heads. The classification process requires observers trained to read standard radiographic images.

In the present evaluation, we concentrated on the classification of fracture severity, as defined by the number of fracture fragments. Within the overall fracture classification code for pediatric long bones, the severity code is given after the child pattern classification (Figure 1). When surgeons classify fractures as part of agreement studies (see the next section), this code distinguishes between

Table 1. Specific patterns of pediatric fractures (child code)

Epiphysis	Metaphysis	Diaphysis
/1 Salter - Harris I		/1 Bowing fracture
/2 Salter - Harris II	/2 Buckle or greenstick	/2 Greenstick fracture
/3 Salter – Harris III	/3 Complete fracture	
/4 Salter – Harris IV		/4 Transverse fracture < 30°
/5 Two-plane fracture		/5 Oblique / spiral fracture > 30°
/6 Tri-plane fracture		/6 Monteggia lesion
/7 Ligament avulsion	/7 Ligament avulsion	/7 Galeazzi lesion
/8 Flake fracture		
/9 Other fractures	/9 Other fractures	/9 Other fractures, incl. toddler fracture [a]

[a] Toddler fractures were initially given the code /3, but according to Slongo et al. (2006a) these fractures were not reliably documented. The code /9 should be given to any fractures not fitting into one of the other anticipated patterns.

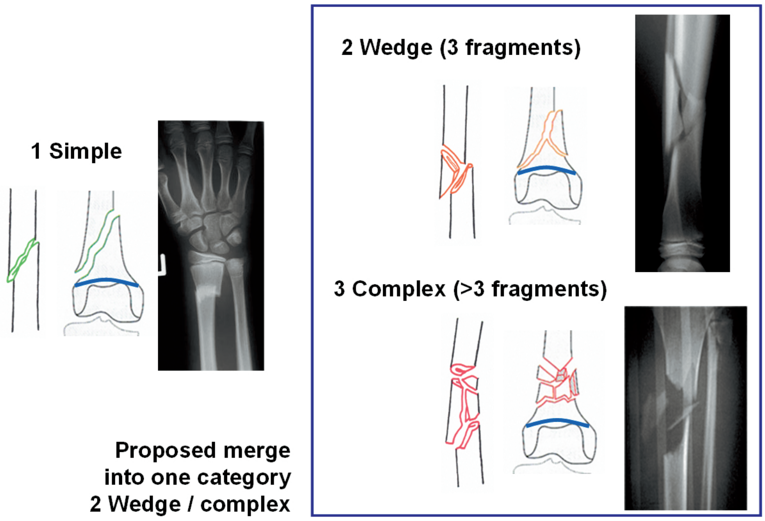


Figure 3. Definition of fracture severity code and proposed simplification. The classification of fracture severity was developed and evaluated with 3 categories. Our results suggest that the two more severe categories should be combined.

3 categories (Figure 3), defined as:

“simple” (1), with only 2 main fracture fragments

“wedge” (2), with 2 main fracture fragments and a third fully separated intermediate fragment. These fractures were considered to be partially unstable

“complex” (3), with 2 main fracture fragments and at least 2 fully separated intermediate fragments. These fractures were considered to be totally unstable.

Case collection

A prospective collection of pediatric humerus supracondylar fractures, forearm fractures and tibia fractures (patients < 16 years old, open physis) with single long-bone fractures was conducted at the University Children’s Hospital, Bern, Switzerland. The selection of cases was independent of the perceived quality of the radiographs. In this case collection, there was only 1 fracture per child. Anteroposterior and lateral standard preoperative radiographs were saved as digital images and presented in random order to surgeons during agreement studies.

Agreement studies on fracture classification

As part of a validation process involving different fracture classification systems (Audigé et al. 2005), two phases of validation were conducted. In the first phase, the classification system was developed and adjusted successively following a series of 4 pilot agreement studies involving experienced surgeons. In the present paper, the first study (A) was the fourth and last of this initial series; it was was conducted during the summer of 2003 and included 267 single fractures (Audigé et al. 2004b, Slongo et al. 2006). 5 fully trained and experienced pediatric orthopedic surgeons from one clinic classified each case independently. Radiographic images were sent on a CD to be viewed on a personal computer. Data were recorded electronically by the surgeons using a Microsoft Excel data form and stored centrally for the analyses.

In a second validation phase, the classification system was assessed among a large number of surgeons with various levels of experience. This second study (B) was conducted as a web-based multicenter study involving 70 surgeons in 15 clinics and 5 countries. Collectively, the surgeons had a wide range of experience (pediatric orthopedic surgeons, pediatric surgeons, trauma surgeons) (Slongo et al. 2007). Training was provided in each clinic prior to the classification exercise. Between August 2004 and July 2005, participants classified 275 fractures at their own pace using the internet after going through a training module with 15 cases. The fracture diagnosis was made following the hierarchy of the classification system, using

Table 2. Raters' pairwise kappa for classification of severity codes

	Epiphyseal		Metaphyseal		Diaphyseal	
	n (%)	kappa	n (%)	kappa	n (%)	kappa
Study A						
Number of ratings per case	5		5		4–5	
Full agreement (100% of raters)	45 (85)		106 (97)		97 (94)	
1 – simple	55	–	107	0.79	95	0.82
2 – wedge	–		1	–	7	0.82
3 – complex	–		1	–	1	–
Overall kappa						0.82
Study B						
Number of ratings per case	35–69		41–70		42–70	
Full agreement (100% of raters)	14 (31)		58 (54)		55 (47)	
1 – simple	42	0.13	106	0.34	105	0.61
2 – wedge	–		1		8	0.37
3 – complex	–		–		4	–
Overall kappa						0.51

both the clinical terminology and the corresponding codes.

We analyzed the classification of fracture severity separately for each fracture type (E, M, and D) in both studies. Inter-rater reliability (agreement between surgeons) was evaluated via overall, category-specific, and surgeons' pairwise kappa coefficients using the statistical software Intercooled Stata version 9.1 (Stata Corporation, College Station, TX). Kappa coefficients were reported for groups of at least 5 cases. Classification accuracy (agreement of surgeons' rating with the truth) for fracture severity was estimated from studies A and B by estimating the most likely distribution of "true" fracture categories in the samples.

With 70 surgeons involved in study B, we defined the "true" fracture categories by considering the ratings given by the majority of surgeons (Slongo et al. 2007). Because only 5 surgeons were involved in study A, we applied an alternative approach called latent class analysis (Audigé et al. 2004b) using the software Latent GOLD version 3.0.1 (Statistical Innovations Inc., Belmont, MA); this technique is based on the hypothesis that each fracture belongs to one of several real clinically relevant categories (or classes). Although these categories can be theoretically defined, however, they may not be directly observable in practice, hence they are said to be "latent". The analysis aims at identifying the most likely number of these latent classes in the population, given the selected sample of fractures and the agreement data collected among the various surgeons. The modeling process assesses how many fracture classes can be reasonably identified in the sample, estimates for each class the surgeons' accuracy of classification, and allocates each case to the most probable "latent class". The analysis therefore provides an estimate of the "true" fracture distribution in the sample.

Results

In study A, while all 55 epiphyseal (E) fractures were classified as simple, only 2 of 109 metaphyseal (M) fractures and 8 of 103 diaphyseal (D) fractures were identified by the 5 raters as either wedge or complex (Table 2). With between 35 and 70 ratings per case in study B, 1 of 107 M fractures and 12 of 117 D fractures were identified as wedge or complex. Consequently, we analyzed the data also by combining grades 2 and 3. In study B, the proportion of surgeons classifying each fracture as wedge or complex ranged from 0% (when all surgeons agreed about simple fractures) to 100% (when all surgeons agreed about wedge/complex fractures) (Figure 4). In this analysis, we considered fractures to be simple when more than 50% of surgeons reported them as being simple, but still 69%, 43% and 47% of simple E, M and D fractures were incorrectly classified as wedge or complex by at least 1 surgeon.

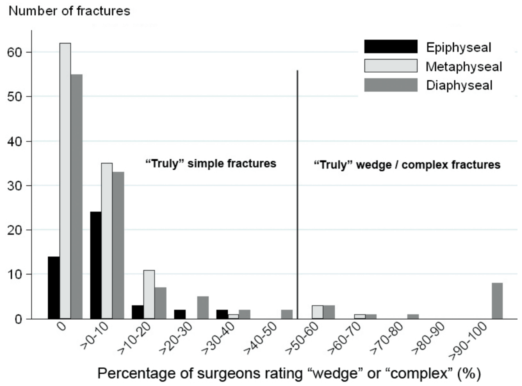


Figure 4. Distribution of the number of fractures in relation to the percentage of surgeons classifying epiphyseal, metaphyseal, and diaphyseal fractures as wedge or complex.

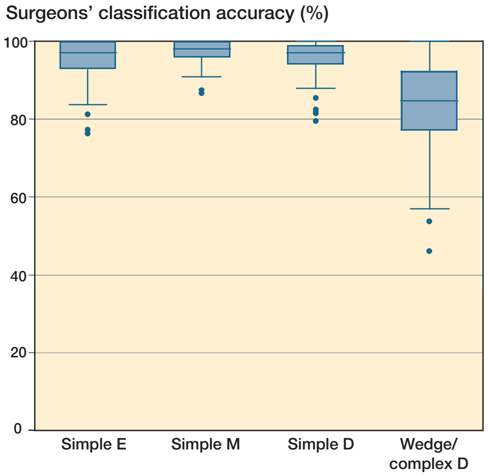


Figure 5. Surgeons' median and ranges of accuracy of classifying the severity (simple versus wedge/complex) of epiphyseal, metaphyseal, and diaphyseal fractures.

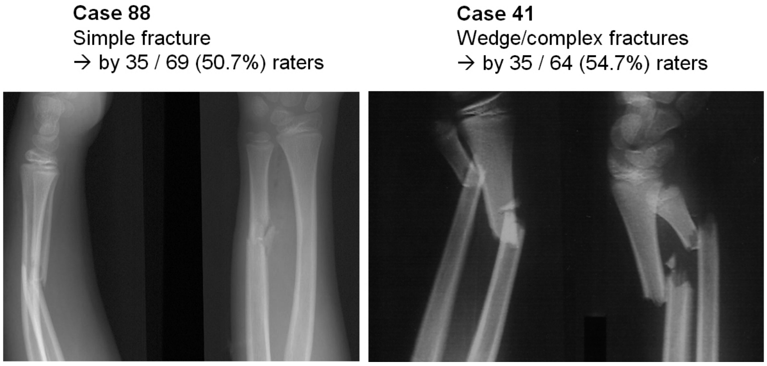
In study A, kappa coefficients of agreement in the identification of simple M and D fractures (vs. wedge and complex combined) were 0.79 and 0.82, respectively. The median raters' pairwise kappa for M and D fractures were 0.80 (0.49–1.00) and 0.84 (0.64–1.00), respectively. In study B, an overall kappa coefficient of 0.51 was estimated for D fractures. Kappa coefficients of agreement in the identification of simple E, M, and D fractures were 0.13, 0.34, and 0.61, respectively.

Classification accuracy was estimated for the classification of simple vs. wedge/complex diaphyseal fractures. In study A, the median accuracies from 5 surgeons of classifying simple and non-simple D fractures were 100% (96–100) and 91% (91–92), respectively. In study B with 70 surgeons, the median accuracies of classifying simple E, M, and D fractures were 97% (76–100), 98% (87–100), and 97% (79–100), respectively (Figure 5). The median accuracy of classifying wedge or complex D fractures was lower at 85% (46–100).

Discussion

The current AO pediatric long-bone classification has been developed and evaluated through the first two phases as recommended Audigé et al. (2005), before it is further assessed in the context of prospective studies. The coding system is hierarchical, with a succession of diagnoses regarding fracture localization and morphology. Satisfactory results related to the higher level of the hierarchy, i.e. localization and child pattern, have been presented and discussed (Audigé et al. 2004b, Slongo et al. 2006, 2007). However, the more detailed part of the classification system—including the fracture severity—is no less important to consider. Yet the clinical importance of the severity of fractures in childhood should be assessed together with the chosen methods of treatment. The evaluation of the severity coding should not be conducted together with the rest of the code, as was often done for the evaluation of the Müller-AO long-bone classification (Audigé et al. 2004a). By conducting separate analyses for each feature and clinically relevant fracture subgroup, appropriate recommendations to improve the classification system can be made. We suspected that coding of fracture severity would differ between fracture types (E, M or D), but not between bones, because of the differing pediatric fracture patterns.

The number of cases available within each classification category should be sufficient (ideally at least 10 (Audigé et al. 2004b)) to provide reliable results. Our fracture sample was fairly large, and sufficiently large for assessment of the most frequent categories but inadequate for rare categories. Our results therefore apply mainly to the coding of simple fractures for all fracture types, and the coding of wedge fractures in diaphyseal fractures. The rarity of wedge or complex E and M fractures



Case 88
Simple fracture
→ by 35 / 69 (50.7%) raters

Case 41
Wedge/complex fractures
→ by 35 / 64 (54.7%) raters

Figure 6. Two cases for which surgeons disagreed in identification of the severity.

observed in our sample is consistent with previous observations (Slongo et al. 1995), where in a pilot documentation study only 16% and 1% of fractures were B and C fractures, respectively—considered to be unstable according the Müller-AO classification. An alternative sampling strategy would be required to properly assess the validity of classifying wedge or complex pediatric long-bone fractures.

The two studies (A and B) were conducted under different settings, but case samples were similar; hence kappa coefficients can be compared. The kappa coefficient is a useful indicator of classification reliability, but it is dependent on the distribution of fracture categories in the sample (Audigé et al. 2004a). We considered only inter-rater reliability in these studies, as the most relevant indicator. Documenting intra-rater reliability in addition requires more resources and previous research has shown that it is almost always better than inter-rater reliability (Audigé et al. 2004a). The estimation of classification accuracy is an important validation step, but relies on the quality of the reference classification used. We are aware that both methods used to derive the true fracture status are not perfect; for instance, there were 2 cases that can be considered misclassified for fracture severity by the majority of surgeons (Figure 6). Given the lack of a gold standard classification, however, this approach was the best available reference standard for these studies.

In the original classification system proposal (Slongo et al. 2006), a grade of fracture severity was considered, not so much because of its influence on healing, as in adults, but because of the need to investigate the indications for various methods of osteosynthesis. We recognize that the terminology "severity" may be related to many aspects of the injury. Within this classification system, however, "severity" is defined by the fragmentation of the fracture, which leads to some interpretation and judgment about the stability of the fracture after reduction and therefore can support treatment decision. The current evaluation shows that all wedge or complex fractures were identified mostly in the diaphysis (around 10% of cases) and much less frequently in the metaphysis (around 1% of cases), but they remained rare. Multifragmentary epiphyseal fractures are extremely rare; thus, no conclusion can be drawn from our study regarding their diagnosis.

The reliability of coding wedge fractures in the diaphysis was poor, with a kappa value of 0.37 in study B, which may be related to definition or imaging. We defined a "wedge" fracture as a fracture with a free-floating bone fragment. In many images, fragments may be perceived as still being attached to one of the main fragments, especially if the number or quality of the radiographs is inadequate. This would also explain why after combining the "wedge" and "complex" categories, the kappa coefficient remained poor (0.61) in the internet-based study (B). In clinical settings, we believe that this classification is likely to be more reliable and accurate when used after treatment, when additional relevant information (e.g. visualization of fragments during fracture reduction) is available. We considered the cut-off of 50% of surgeons for the identification of "true" simple fractures, which may have an influence on the classification accuracy but not the classification reliability. Our estimates of classification accuracy of simple fractures are high, above 80–90% for most surgeons, and all 3 fracture types E, M and D, but they fall below 50% for wedge/complex fractures. We realize that the proposed definitions of “wedge” and “complex” include the consideration of stability, which remains subjective, thus opening the way for disagreement. There should be more detailed investigation of whether increased training—and also obtaining codes by group consensus with experienced observers, such as recommended by Slongo et al. (2007)—can increase the accuracy of classifying fracture severity.

Our investigation further supports the need to implement a validation process for fracture classification systems before they are fully put into practice. The initial proposal was based on the Müller-AO classification system for adults, where severity is determined in a series of triads, including the same 3 categories “simple”, “wedge” and “complex”, such as types A, B, and C for diaphyseal fractures and subtypes 1, 2, and 3 for extraarticular fractures. However, our validation results highlight some deficiencies in adopting a similar severity coding for pediatric fractures and support the suggestion that a severity code in 3 categories is not relevant. Surgeons did not agree on this diagnosis to the extent that a distinction between wedge and complex fractures would be reliable in practice. Thus, we recommend that these two categories be combined for routine classification (Figure 3). Further evaluations are required regarding the diagnosis and relevance of complex epiphyseal fractures, and there is a need for post-reduction data for severity coding. Possible measures for improvements in current practice are the use of additional post-treatment information, and the classification of fractures by consensus in clinics.

Contributions of authors

TS and LA: contributed equally to this work and wrote the manuscript. TS: initiator of the project, conducted the case collection, developed the study protocol, participated in the classification sessions, and provided the clinical expertise as principal investigator. LA: developed the study protocol, coordinated the project, and provided the methodological expertise, including data analysis. The other 5 authors are members of the AO Pediatric Classification Group and were involved in the development of the classification and the clasification sessions as experienced surgeons. All authors approved the current version of the manuscript.

The authors are grateful to all surgeons who participated in the successive agreement studies for their fruitful support in the development and validation of this pediatric fracture classification system.

No competing interests declared.

Audigé L, Bhandari M, Kellam J. How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. Acta Orthop Scand 2004a; 75: 184-94.

Audigé L, Hunter J, Weinberg A, Magidson J, Slongo T. Development and evaluation process of a paediatric long-bone fracture classification proposal. Eur J Trauma 2004b; 30: 248-54.

Audigé L, Bhandari M, Hanson B, Kellman J. A concept for the validation of fracture classifications. J Orthop Trauma 2005; 19: 404-9.

Burstein A H. Fracture classification systems: do they work and are they useful? J Bone Joint Surg (Am) 1993; 75: 1743-4.

Garbuz D S, Masri B A, Esdaile J, Duncan C P. Classification systems in orthopaedics. J Am Acad Orthop Surg 2002; 10: 290-7.

Martin J S, Marsh J L. Current classification of fractures. Rationale and utility. Radiol Clin North Am 1997; 35: 491-506.

Müller M, Narzarian S. The comprehensive classification for fractures of long bones. Springer, Berlin, Heidelberg, New York 1990.

Slongo T, Schaerli, A F, Koch P, Buehler M. Klassifikation und Dokumentation der Frakturen im Kindesalter-Pilotstudie der internationalen Arbeitsgemeinschaft für Kindertraumatologie. Zentralbl Kinderchi 1995; 4: 157-63.

Slongo T, Audigé L, Schlickewei W, Clavert J M, Hunter J. Development and validation of the AO pediatric comprehensive classification of long bone fractures by the Pediatric Expert Group of the AO Foundation in collaboration with AO Clinical Investigation and Documentation and the International Association for Pediatric Traumatology. J Pediatr Orthop 2006; 26: 43-9.

Slongo T, Audigé L, Clavert J M, Nicolas L, Frick S, Hunter J. The AO paediatric comprehensive Classification of long- bone Fractures: a web- based multicenter agreement study. J Pediatr Orthop 2007; in press.