Reliability of the AO/ASIF classification for pertrochanteric femoral fractures

Inger B Schipper[1,3], Ewout W Steyerberg[2], Rene M Castelein[1] and Arie B van Vugt[3]

[1]Isala Clinics, Weezenlanden Hospital, Departments of General and Orthopedic Surgery, Zwolle, The Netherlands, [2]Erasmus University Rotterdam, Department of Public Health, Rotterdam, The Netherlands, [3]University Hospital Rotterdam Dijkzigt, Department of Traumatology, Rotterdam, The Netherlands. Correspondence: Dr. I.B. Schipper, Department of Traumatology, University Hospital Rotterdam, Dr. Molewaterplein 40, NL-3015 GD Rotterdam, The Netherlands. Tel +31 10 4639222. E-mail: ibschip@xs4all.nl
Submitted 00-04-30. Accepted 00-08-07

ABSTRACT – 20 radiographs of pertrochanteric femoral fractures were classified as to fracture "group" and "subgroup" according to the AO/ASIF Fracture Classification (type 31A) by 15 observers. 3 months later, the same radiographs were reviewed by the same observers. Mean agreement of the observers with the final consensus ranged from 53% (with subgroup classification) to 81% (without subgroup). The mean kappa value for interobserver reliability was 0.33 and 0.34 for classification with subgroup in both observer sessions, respectively. Omission of the subgroup classification resulted in better mean kappa values (0.67 and 0.63, respectively). Mean intraobserver reliability was 0.48 in the fracture "subgroup" and 0.78 in the "group" classification.

In conclusion, the results show that the AO/ASIF classification for pertrochanteric fractures is reliable for fracture subgroups 31A1, A2 or A3. The group classification should be used to compare scientific data and determine the best treatment. Further classification of fracture subgroups leads to poor reproducibility of results.

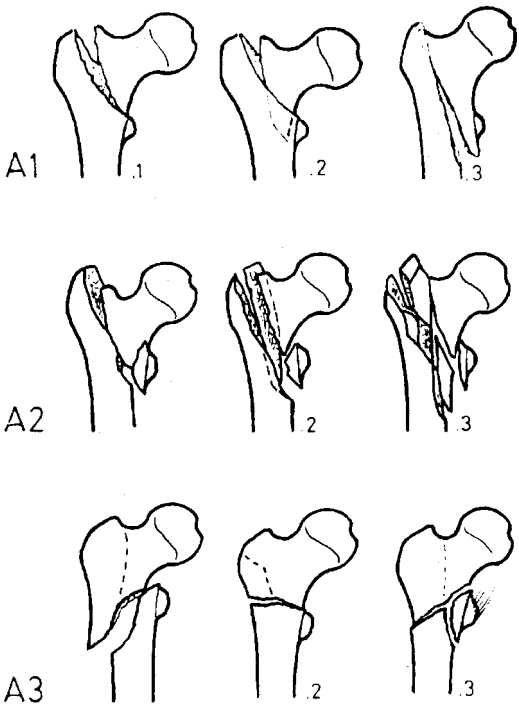
■

Nowadays, the commonest classification is that of Fractures of the Long Bones introduced by the AO/ASIF group (Müller et al. 1990). This classification is organized into hierarchical triads. For every bone segment (e.g., femur, tibia or humerus), 3 "types" of possible fractures exist (A, B, C), each of which can be divided into 3 "groups" (e.g., for pertrochanteric fracture groups A1, A2, A3). The 3 groups are each divided into 3 "subgroups" according to increasing fracture severity, indicating a greater difficulty in operative treatment, a higher likelihood of complications, and a poorer prognosis (Figure). Despite its common use and wide acceptance, its reliability and reproducibility have been questioned for a small number of specific fracture types (Johnstone et al. 1993, Siebenrock and Gerber 1993, Kreder et al. 1996, Martin et al. 1997, Craig and Dirschl 1998). Systematic classification of a pertrochanteric fracture may reduce problems related to interpretation of the fracture and facilitate the choice of the appropriate method of treatment. To assess clinical studies on various types of pertrochanteric hip fractures, a reproducible classification is mandatory.

Therefore, we assessed the interobserver and intraobserver reliability of the AO/ASIF classification system for pertrochanteric femoral fractures. Interobserver reliability was assessed for fracture "group" classification and for "subgroup" classification during 2 radiograph sessions. We also evaluated interobserver reliability among 3 different groups of observers (surgeons, surgical residents and radiologists).

Material and methods

The preoperative anteroposterior and lateral radiographs of 20 patients admitted to our hospital in 1998 with pertrochanteric femoral fractures were selected from a trauma database. No special



A simplified version of the pertrochanteric femoral fracture classification 31A made by the AO/ASIF group was given to each observer (From Müller ME et al.: The Comprehensive Classification of Fractures of the Long Bones. Springer-Verlag Berlin/Heidelberg 1990. Reprinted with permission of the publisher).

criteria were set as to the quality of the radiographs, except that they had been accepted to form the basis of treatment. The radiographs were assessed by an expert panel consisting of the senior authors and 2 consultant radiologists from our clinics, to ensure representation of the full spectrum of pertrochanteric hip fractures, classified according to segment 31 type A of the AO/ASIF classification. Each fracture was then classified by consensus of the panel. Fractures were defined as pertrochanteric when the fracture lines went through the major or minor trochanter. A fracture was considered to have subtrochanteric extension (A3) when the fracture lines extended distally from either the major or minor trochanter, to a maximum of 3 cm below the minor trochanter.

The radiographs were then reviewed by 15 observers: 5 surgeons involved in trauma-care, 5 surgical residents with special interest in orthopedic trauma and 5 radiologists. None of the observers had previous experience with the AO/ASIF classification. Fracture classification sessions were conducted by one of the authors (I.B.S) in a standardized fashion. An explanation of the AO/ASIF classification segment 31 type A, its division into groups and subgroups and a copy of the original AO/ASIF classification (Figure) were given as reference to each observer separately. Each observer then classified each of the 20 selected fractures as to group and subgroup (9 possible fracture classifications). Observers were not given any feedback after the first session, nor were radiographs available to observers between the first and second classification session. 3 months later, the same observers under the same conditions classified the same radiographs in a different order.

Statistics

We determined the interobserver reliability by comparing the classification results assigned by the 15 observers. Kappa values were calculated for interobserver reliability with and without subgroup classification of the fracture in the first and second sessions.

Intraobserver reproducibility was assessed by comparing the classifications with subgroup and without subgroup by each observer in the two classification sessions. The kappa coefficient of reliability provides a pairwise proportion of agreement between or among observers, corrected for chance. Kappa values can vary from $-p_e/1-p_e$ (complete disagreement) through 0 (chance agreement) to +1 (complete agreement).

Interobserver kappa values were calculated for each possible pair of the 15 observers for both classification sessions. Intraobserver kappa values were calculated comparing classification scores of each observer on the two classification sessions.

An average kappa value was calculated to reflect the overall agreement between the observers. The uncertainty associated with this estimate could not be calculated with standard statistical methods, since the dependency between kappas from the same observation had to be taken into account. We therefore used a bootstrap resampling procedure (Efron and Tibshirani 1993). One set of observers was replaced by another set of observers in the analysis. Note that if kappa values (rather than observers) had been used, the dependence between kappas would have been ignored. The

Table 1. Radiograph classification of each fracture in both readings by 15 observers. Consensus classification is given separately

No.	Position	Session	Fracture																		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Consensus (expert panel)			3.3	3.3	3.1	1.2	2.2	1.1	2.2	1.2	3.3	3.1	3.3	3.3	3.3	2.3	3.1	3.1	1.2	1.2	2.2	1.2
1	surgeon	1	2.2	2.3	3.3	1.1	2.2	1.1	2.1	1.2	2.3	3.1	2.3	2.3	3.3	2.2	3.1	1.1	1.2	1.2	2.1	1.2
2	surgeon	1	2.1	3.3	3.1	1.1	2.2	1.2	2.2	1.2	3.3	3.3	3.1	3.3	3.3	2.3	3.1	3.1	1.2	1.2	2.1	1.2
3	surgeon	1	3.3	3.3	3.3	1.2	1.3	1.1	2.2	3.3	1.3	3.1	3.3	3.3	3.3	2.2	3.1	3.1	1.2	1.2	2.2	1.2
4	surgeon	1	2.3	3.3	3.1	1.2	2.2	1.1	2.1	3.2	3.2	3.1	3.1	2.3	3.3	2.2	3.1	1.2	1.1	1.2	2.1	1.3
5	surgeon	1	2.2	3.3	3.3	2.1	2.1	1.1	2.1	1.2	3.2	3.1	3.3	3.3	3.3	2.3	3.1	1.2	1.2	1.2	2.2	1.2
6	resident	1	2.2	2.3	3.1	1.2	2.1	1.2	2.1	1.2	3.3	3.3	3.3	2.3	3.3	2.3	3.1	3.2	1.2	2.1	2.3	1.2
7	resident	1	2.1	2.3	3.1	2.1	2.1	1.2	2.1	1.2	2.3	3.3	3.3	2.2	3.3	2.3	3.3	3.1	2.1	1.2	2.1	1.2
8	resident	1	2.3	3.3	3.3	1.1	2.2	1.1	2.2	1.1	3.3	3.2	3.3	1.3	3.3	2.3	3.1	3.2	1.1	1.1	2.1	1.2
9	resident	1	2.3	3.3	3.1	1.2	2.1	1.1	2.2	1.2	3.3	3.3	3.3	3.2	3.3	2.3	3.1	1.2	1.2	1.2	2.2	1.2
10	resident	1	2.1	3.3	3.1	1.2	2.1	1.1	1.1	1.2	3.3	3.1	3.3	2.3	3.1	2.2	3.3	3.2	3.2	1.2	2.1	1.2
11	radiologist	1	2.3	3.3	3.1	1.1	2.1	1.1	2.1	2.1	2.1	3.1	3.3	3.3	3.3	2.3	3.1	1.3	1.1	1.1	2.2	1.1
12	radiologist	1	2.1	2.3	3.1	3.2	2.1	1.1	2.1	1.1	3.3	3.2	3.3	2.3	3.3	2.3	3.1	3.2	1.2	1.1	2.1	2.2
13	radiologist	1	2.2	2.3	3.1	1.1	2.1	1.1	2.2	2.3	2.1	3.1	3.3	2.3	3.3	2.2	3.1	2.2	1.1	1.2	2.2	1.2
14	radiologist	1	2.2	3.3	3.1	1.3	1.2	1.3	2.1	1.3	3.3	3.1	3.3	1.3	3.3	2.3	3.1	3.1	1.2	2.1	2.3	1.1
15	radiologist	1	2.2	2.3	3.3	2.2	2.2	1.1	2.3	2.2	3.2	3.3	3.3	3.3	3.3	2.3	3.3	3.3	1.2	1.2	2.3	1.2
1	surgeon	2	2.2	3.3	3.3	1.1	2.2	1.1	2.2	1.1	3.3	3.3	3.3	2.3	3.3	2.3	3.1	3.1	1.2	1.2	2.2	1.2
2	surgeon	2	3.1	3.3	3.3	1.1	1.2	1.2	1.2	1.1	2.3	3.1	3.3	3.1	3.3	2.3	3.1	1.2	2.1	1.2	2.2	1.2
3	surgeon	2	3.3	3.3	3.3	1.2	2.2	1.2	2.2	1.2	3.3	3.3	3.3	3.3	3.3	1.3	3.1	3.1	1.2	1.2	2.2	1.2
4	surgeon	2	2.3	2.3	3.1	1.2	2.1	1.1	2.1	1.1	1.3	3.1	1.3	2.3	3.3	2.3	3.1	3.1	1.2	1.1	2.2	1.2
5	surgeon	2	3.1	3.3	3.3	1.2	2.2	1.1	2.1	1.2	3.2	3.1	3.3	3.3	3.3	2.3	3.1	1.2	2.1	1.2	2.3	1.1
6	resident	2	2.1	3.3	3.1	1.1	2.1	1.2	1.1	1.2	1.3	3.1	3.3	2.3	3.3	2.3	3.1	3.2	1.2	1.2	2.3	1.2
7	resident	2	2.1	2.3	3.3	1.2	2.1	1.1	1.2	1.2	3.3	3.1	3.3	2.1	3.3	2.1	3.3	3.1	1.2	1.2	2.1	1.2
8	resident	2	2.3	3.3	3.3	1.2	2.1	1.2	2.3	1.3	3.3	3.3	3.3	3.1	3.3	2.2	3.1	3.2	1.1	1.1	2.1	1.2
9	resident	2	2.2	3.3	3.1	1.2	2.2	1.1	2.2	1.2	2.2	3.3	3.3	3.3	3.3	2.3	3.1	1.2	1.2	2.1	2.1	1.2
10	resident	2	2.1	2.3	3.3	1.1	2.1	1.1	2.1	1.1	2.1	3.1	3.3	2.2	3.3	2.3	3.1	3.1	1.1	1.2	2.1	1.2
11	radiologist	2	2.3	3.3	3.3	1.2	3.2	1.1	1.2	1.1	2.2	3.3	3.3	3.3	3.3	3.3	3.1	3.2	1.1	1.2	2.2	1.2
12	radiologist	2	2.2	2.3	3.3	1.1	2.2	1.1	2.2	1.1	2.1	3.3	3.3	2.3	3.3	2.3	3.1	3.2	2.1	1.1	2.1	1.2
13	radiologist	2	2.3	2.3	3.3	1.2	2.3	1.1	2.3	1.3	3.2	3.1	3.3	2.3	3.3	2.3	3.1	3.1	1.2	2.1	2.2	1.1
14	radiologist	2	2.3	3.1	3.3	1.3	1.3	1.2	1.1	2.1	2.1	3.1	3.3	3.1	3.3	2.3	3.1	3.1	1.2	2.1	2.2	1.2
15	radiologist	2	2.2	2.3	3.3	2.2	2.3	2.3	2.2	2.2	3.2	3.1	3.1	2.3	3.3	2.3	3.1	3.1	1.3	2.2	2.3	1.2

bootstrapping process duplicated the procedure used by other observers who had made the classification and therefore gave insight into the variability of the estimated average kappa. We took 500 bootstrap samples and calculated the standard error of the estimated kappas in these samples. Based on the suggestion of a reviewer, we also used a SAS macro (Fleiss 1981). This resulted in identical estimates of the overall kappa value and somewhat smaller estimates of the standard error (results available from the authors).

We also compared the agreement of separate consultant groups (surgeons, surgical residents, radiologists). Average kappas were calculated for each consultant group, with its standard error indicated by the bootstrapping procedure (500 replications).

Results

20 radiographs were reviewed twice by 15 observers (Table 1). Of the 600 (15 × 20 × 2) classifications obtained, correspondence with the final consensus ranged from 0% to 100%. Mean correspondence was 53%. A substantial improvement in agreement was found when fractures were classified only according to main groups, rather than according to subgroups as well. Classification without subgroups resulted in an increase of mean agreement with the final consensus to 81%.

Interobserver reliability (Table 2)

The mean kappa coefficient was 0.33 in the first classification session, in the second session, the kappa value was similar (0.34). Interobserver

Table 2. Kappa values for interobserver and intraobserver reliability (SE)

Kappa values	First session	Second session
Interobserver		
With subgroup classification	0.33 (0.01)	0.34 (0.01)
Without subgroup classific.	0.67 (0.01)	0.63 (0.01)
residents	0.69 (0.04)	0.51 (0.05)
surgeons	0.62 (0.03)	0.64 (0.05)
radiologists	0.65 (0.03)	0.69 (0.03)
Intraobserver		
With subgroup classification	0.48 [a]	
Without subgroup classific.	0.72 (0.02)	
residents	0.70 (0.05)	
surgeons	0.73 (0.02)	
radiologists	0.72 (0.05)	

[a] Intraobserver reproducibility of subgroup classification was calculated only for 4 observers (kappa values 0.26, 0.48, 0.54, 0.64).

agreement improved significantly if subgroup classifications were left out: the mean kappa coefficient was 0.67 in the first session, and 0.63 in the second.

The mean kappa values for the different observer groups did not differ (p = 0.35) in the first classification session. Residents showed significantly worse interobserver reliability (p = 0.04) than the surgeons and radiologists during the second reading.

Intraobserver reliability

The mean kappa coefficient for intraobserver reliability for classification of fracture groups with their subgroups could not be calculated for 11 observers, since not all classifications used in the first reading were used in the second. For example, observer 1 classified fractures number 7 and 19 as 2.1 in the first reading and no fractures as such in the second (Table 1). The mean kappa coefficient was 0.48 among 4 observers for whom an intraobserver kappa could be calculated. Mean intraobserver reliability for groups was substantially better, with a kappa coefficient of 0.72 (Table 2). Intraobserver agreement values among observer groups did not differ (p = 0.09) for the surgeons, surgical residents and radiologists.

Discussion

A valid fracture classification should meet 4 criteria (Burstein 1993, Martin and Marsh 1997). It should provide: 1) guidelines for treatment, 2) be a method which we can report, compare and assess results of treatment of similar fractures, 3) should provide a reliable language of communication, and 4) be reasonably reliable and reproducible. Many authors have studied various fractures and fracture classification systems regarding their reliability and reproducibility (Table 3). Some classifications are based strictly on a specific location, whereas the AO/ASIF classification provides systematic guidelines for classifying of all fracture locations in the long bones. Both the AO classification system and the non-AO classifications have wide ranges of kappa values. The former system requires 3 sequential decisions about fracture classification. Each step in categorizing fracture type, group and subgroup adds a risk of error to the previous classification step. Due to this cumulative error risk, interobserver and intraobserver disagreement increases. In our study, interobserver reliability was poor for fracture subgroup classification (kappa value 0.33) according to the scales of strength of agreement proposed by Fleiss (Seigel et al. 1992). These results are consistent with those of previous AO/ASIF classification investigations (Table 3). Fracture group classification, however, was good, and with a kappa value of 0.67 even better than in other reports (Siebenrock and Gerber 1993, Kreder et al. 1996, Martin and Marsh 1997, Martin et al. 1997). The relatively low interobserver agreement among the residents confirms that experience with classification of fractures and their treatment improves the reliability of using a classification system (Johnstone et al. 1993, Kreder et al. 1996, Dirschl and Adams 1997, Martin et al. 1997). Intraobserver reliability in groups showed a kappa value of 0.72 and was also better than most results reported in the literature.

The main difficulty of a classification for pertrochanteric fractures lies in the variety of fracture patterns, the possible involvement of the greater and lesser trochanters and the differentiation from lateral collum fractures and subtrochanteric fractures. Pertrochanteric fractures, extending to the

Table 3. Mean interobserver kappa values for 5 fracture classification systems. The kappa values of AO/ASIF classifications are given separately for classification of fracture type, fracture group and subgroup

Author	Fracture classification	Mean interobserver Kappa value	
Horn and Rettig 1993	Gustillo-Andersen (open fractures)	0.53	
Siebenrock et al. 1993	Neer (shoulder)	0.30	
Kristiansen et al. 1988	Neer (shoulder)	0.30	
Dirschl and Adams 1997	Ruedi-Algower (ankle, distal tibia)	0.48	
Martin et al. 1997	Ruedi-Algower (ankle, distal tibia)	0.46	
Thomsen et al. 1991	Lauge-Hansen (ankle)	0.55	
	Weber (ankle)	0.57	
Siebenrock et al. 1993	proximal humerus; AO/ASIF segment 20	0.53	type
		0.42	group
Kreder et al. 1996	distal radius; AO/ASIF segment 23	0.68	type
		0.48	group
Martin et al. 1997	distal tibia; AO/ASIF segment 43	0.60	type
		0.38	group
Craig et al. 1998	ankle; AO/ASIF segment 44	0.77	type
		0.61	group

subtrochanteric region, are difficult to categorize with the AO classification, since the AO/ASIF classification guidelines take no account of specific classification of subtrochanteric fractures. The complexity of pertrochanteric and especially subtrochanteric fractures may prevent further improvement of reliability of their classification. Applying a classification system for a complex fracture in a standardized manner does not necessarily mean improvement in reliability. It is therefore recommended that classification of each fracture should be done by consensus, in order to teach and encourage colleagues to discuss and determine specific characteristics of each fracture. Guidelines for treatment may be based on the same systematic classification of fracture groups and, if classified by consensus of an expert team, of fracture subgroups. Using this classification, all fractures classified as 31.A.1 are treated with a Dynamic Hip Screw in our clinics. Patients with fractures classified as 31.A.2 and 31.A.3 are, because of their unstable fracture characteristics, treated by implantation of a Gamma-Nail or a Proximal Femoral Nail. However, the optimal treatment of pertrochanteric femoral fractures, particularly of types A.1.3, A.2.1 and A.3.3 remains debatable despite use of a valid fracture-group classification system. These examples emphasize the need for a reliable subgroup classification and the clinical importance of valid further subdivision of stable and unstable pertrochanteric femoral fractures. Other and simpler subgroup classifications may be used for this purpose.

In our opinion, the AO/ASIF classification of pertrochanteric femoral fractures (AO/ASIF 31A) meets the above-mentioned criteria for a valid classification (Burstein 1993, Martin and Marsh 1997).

Our study also shows that fracture classification systems have limitations. Poor interobserver reliability of subgroup classification raises the question whether subdivision into fracture subgroups should be encouraged.

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

Craig W L, Dirschl D R. Effects of binary decision making on the classification of fractures of the ankle. J Orthop Trauma 1998; 12: 280-3.

Dirschl D R, Adams G L. A critical assessment of factors influencing reliability in the classification of fractures, using fractures of the tibial plafond as a model. J Orthop Trauma 1997; 11: 471-6.

Efron B, Tibshirani R. An introduction to the bootstrap. Monographs on statistics and applied probability. Chapman and Hall, New York 1993; xvi: 436.

Fleiss J L. Statistical methods for rates and proportions, second edition. John Wiley & Sons Inc., New York 1981.

Horn B D, Rettig M E. Interobserver reliability in the Gustilo and Anderson classification of open fractures. J Orthop Trauma 1993; 7: 357-60.

Johnstone J, Radford W J P, Parnell EJ. Interobserver variation using the AO/ASIF classification of long bone fractures. Injury 1993; 24: 163-5.

Kreder K J, Hanel D P, McKee M, Jupiter J, McGillivary G, Swiontkowski MF. Consistency of AO fracture classification for the distal radius.J Bone Joint Surg (Br) 1996; 78: 726-31.

Kristiansen B, Andersen U L S, Olsen CA , Varmarken J E. The Neer classification of fractures of the proximal humerus. Skeletal Radiol 1988; 17: 420-2

Martin J S, Marsh J L. Current classification of fractures. Imag Orthop Trauma 1997; 35: 491-506.

Martin J S, Marsh J L, Bonar S K, DeCoster T A, Found E M, Brandser E A. Assessment of the AO/ASIF Fracture Classification for the distal tibia. J Orthop Trauma 1997; 11: 477-83.

Müller M E, Nazarian S, Koch P, Schatzker J. The comprehensive classification of fractures of the long bones. Springer-Verlag, Berlin/Heidelberg 1990.

Seigel D G, Podgor M J, Remaley N A. Acceptable values of Kappa for comparison of two groups. Am J Epidemiol 1992; 135: 571-8.

Siebenrock K A, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. J Bone Joint Surg (Am) 1993; 75: 1751-5.

Thomsen N O B, Overgaard S, Olsen L H, Hansen H, Nielsen S T. Observer variation in the radiographic classification of ankle fractures. J Bone Joint Surg (Br) 1991; 73: 676-8.