How reliable are reliability studies of fracture classifications?

A systematic review of their methodologies

Laurent Audigé[1], Mohit Bhandari[2] and James Kellam[3]

[1]AO Clinical Investigation and Documentation, AO Center, Clavadelerstrasse, CH-7270 Davos Platz, Switzerland, [2]Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre, Hamilton, Canada, [3]Carolinas Medical Center, Charlotte N.C., USA
Correspondence LA: laurent.audige@aofoundation.org
Submitted 03-02-01. Accepted 03-07-01

ABSTRACT Two independent reviewers performed a search in MEDLINE and EMBASE for fracture classification reliability studies. Data were obtained on classifications, image modalities, fracture selection processes, sample sizes and their justification, type and number of raters, practical issues for the classification sessions, statistical methods, and results. A 10-item checklist was devised for quality assessment of methodologies.

44 studies assessing 32 fracture classification systems were included. We found a wide variation of methodologies. For instance, the median number of raters was 5 (2–36) and the median number of fractures was 50 (10–200). This selection was considered representative in 17/44 of the studies. The true distribution of classification categories was estimated in 9 studies. The kappa coefficient was mostly used (39/44) to quantify the raters' agreement.

Methodological issues are discussed. Given limitations in the use and interpretation of kappa coefficients, investigators should consider alternative methods that focus upon the accuracy of the classification systems. The development and adoption of a systematic methodological approach to the development and validation of fracture classification systems is needed.

A fracture classification system should be reliable and valid. It should also have prognostic value for patients to assist physicians in planning their management (Lindsjö 1985, Colton 1991, Burstein 1993, Martin and Marsh 1997). Numerous fracture classification systems have been proposed in orthopaedics (Rockwood et al. 1996, Bernstein et al. 1997, Browner et al. 1998); however, it remains unclear whether these classifications have been rigorously evaluated or validated. A classification is "validated" when several criteria are met (Bland and Altman 2002). First, the classification should "look good" on clinical grounds—i.e., it must have face and content validity. The second issue is related to how adequate the classification is in terms of its reliability and accuracy. Reliability means that various individuals should be able to produce the same results when using the classification, or that a single user should be able to obtain consistent results classifying the same fractures at different times. Assessing the reliability of a fracture classification system does not, however, guarantee accuracy. Accuracy measures how well an observation fits with the reality. The third issue is whether the classification has construct validity—that is, how closely are fracture categories associated with relevant patient outcomes in the context of specific fracture management plans.

There is growing concern that fracture classification systems should be formally validated before they are recommended for use in practice and research (Burstein 1993, Brady et al. 2000).

With increased appreciation of classification system reliability, there has been a rapid increase in the number of published reports evaluating

fracture classification systems. To our knowledge, there has been no systematic review of the methodologies utilized to evaluate such systems. Thus, given the paucity of information regarding methods to evaluate fracture classification systems, our purpose was threefold: 1) to conduct a systematic literature review of the methods used to assess the "reliability" of fracture classification systems, 2) to discuss determinants of quality among reliability studies and 3) determine the need for a methodological approach for the development and validation of these classifications.

Material and methods

Eligibility criteria

A systematic literature review of the methodological issues related to the validation of fracture classifications was done. We included studies (or indexed abstracts) reporting on the reliability (interobserver agreement) of classification systems for fractures.

Search strategy

One of us (LA) did the selection process, data extraction and analysis. This process was checked by a second independent reviewer (MB). The database MEDLINE was examined from 1969 to November 2002 using the keywords "agreement OR reliability" and "fracture". The titles and abstracts of reference hits were screened for eligibility, and complete selected articles obtained. The reference lists of all relevant articles were read until no further studies could be found. Finally, the database EMBASE was examined with the keyword "classification" added to the search terms mentioned above, and a similar screening process was used.

Data extraction

Data were extracted on the fracture type and classification being investigated, the image modalities used (i.e., radiographs, CT scan, 3D reconstruction), the selection process of the fractures (i.e., whether the sample of cases was representative of the population of cases), the sample size and its justification, the type and number of observers (i.e., orthopedic surgeons, radiologists, non-clinicians), practical issues for the classification sessions (e.g., the training of observers, the method of presentation of images, the time between two classification sessions, or methods to limit recording biases), the statistical methods used for the analyses (e.g., the calculation of raw agreement indices or specific measures of agreement such as the kappa coefficient), as well as the results. The kappa coefficient was described as (Po-Pe)/(1-Pe)—i.e., a "chance-corrected measure of agreement" with Po being the observed proportion of agreement and Pe the proportion of agreement expected by chance alone. It ranges from +1 (complete agreement) through 0 (agreement by chance alone) to less than 0 (less agreement than expected by chance).

Assessment of study quality

A validated checklist of recognized study quality criteria was sought to assess methodological issues. With the current lack of any such instrument, we agreed on a preliminary list of relevant items from diagnostic research (Pewsner et al. 2002). A description of these criteria is given in Table 1. In using this checklist, we gave a positive rating only when information was mentioned in the paper. Hence, lack of reporting was equated to poor quality. Two reviewers (LA & MB) assessed each study independently and discrepancies were resolved by consensus.

Data analysis

The statistical software Stata 7.0 (Stata Corporation, Texas, USA) was used for descriptive statistics.

Results

Study identification

With MEDLINE, 48 full-text articles were reviewed, which resulted in the inclusion of 37 studies. Four further studies were found in reference lists. A follow-up examination of EMBASE resulted in the review of 7 more full-text articles, and 3 additional studies. Thus, 44 studies fulfilled our criteria for this review (Figure).

Study characteristics

The 44 studies were classified according to 8 fracture localization groups (Table 2). 32 fracture

Table 1. Assessment of methodological quality of published reliability studies of fracture classification systems

Quality criteria	Number of studies	%
1. The classification system(s) was(were) clearly described	44	100
2. The study population was defined by clear inclusion and exclusion criteria	26	59
3. Selected cases were representative of the study population	17	39
4. The size of the sample was justified	0	0
5. The group of raters was representative of the intended users of the classification	4	9
6. The number of raters was appropriate	10	23
7. Raters classified all cases independently during classification sessions	31	70
8. Raters were blinded to patient clinical information	17	39
9. The true distribution of classification categories in the sample was estimated	9	20
10. Statistical methods used were adapted to study objectives	17	39

Definitions and comments related to quality criteria:

2. The study population was considered well-defined if authors provided a clear definition of the included fracture type, and if a procedure existed to ensure that all included fractures belonged to that targeted population (e.g. a panel of experts checked on the cases prior to the classification session or when cases were excluded when the majority or all raters agreed that they did not belong to the targeted population).
3. The sample was judged "representative" when authors reported that the selection was a "consecutive" series of patients, or patients randomly selected within a list. If the selection was based on the quality of the images, or the availability of specific image modalities (e.g. presence of a CT scan), the sample was judged "not representative".
5. This criterion would be met if the intended users of the classification were clearly defined and raters were clearly not selected on the basis of judgement. For instance, asking all intended users from one or several clinics to participate would be considered an appropriate method.
6. We classified raters as experienced surgeons, other surgeons and non-surgeons (e.g., radiologists) (see Table 1). The group of raters was judged adequate if at least 5 raters from any of these three groups participated. When the number of raters was not reported, it was not judged "adequate". If a study examined the effect of training, then at least 5 raters should be enrolled for each experience level defined a priori.
7. Raters classified cases independently if a classification meeting was organized or if it was mentioned clearly that this was conducted, e.g., "no discussion was allowed between raters", "no question could be asked during the classification session" or "raters were not aware of the other participants".
8. This was recorded as "blinded" when authors reported it specifically or "identifying marks on radiographs were covered with tape (patient ID covered)"
9. This is when a "gold standard" assessment was conducted, e.g. using consensus between raters, an assessment from an expert panel, or the intra-operative finding.
10. The use of only raw agreement indices was judged insufficient. The use of "chance-corrected" statistics (e.g., kappa, William's index, Sav) was considered adequate when assessing factors influencing reliability, but not to assess whether reliability was acceptable or when comparing reliability of two different classifications (see discussion section).

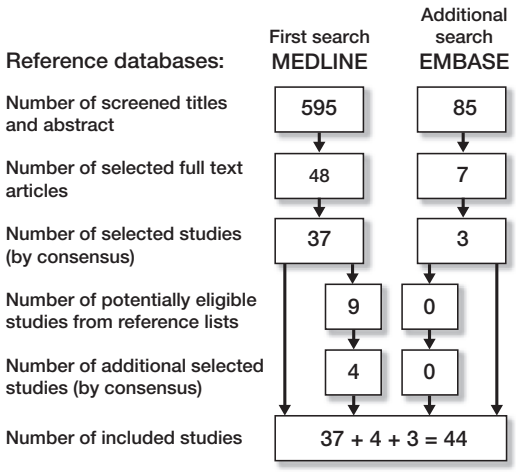
classification systems were also evaluated. All of these studies were cross-referenced with reports of previously-described classifications. Modifications or simplifications of these classifications had been reported and evaluated in 19 studies, although this had been stated in the aims of only 4 studies (Bernstein et al. 1996, Craig and Dirschl 1998, Dirschl and Adams 1997, Schipper et al. 2001). The median sample size of the 44 studies was 50 (10–200).

Methodological aspects of reliability studies

In the 44 studies included, we found considerable variations in the methodology (Tables 1 and 2).

These studies adequately referenced or described the categories of the classification systems evaluated. The "intent of use", however, was less clear. The study population was defined by clear-cut inclusion and exclusion criteria in only 26/44. The selection of cases was representative in 17/44 (i.e., consecutive series or random selection of cases), and had been selected "to represent most or all of the different patterns" in 7 more studies. One study selected cases that had been operated on (Brien et al. 1995). None of them justified the number of cases included.

We considered that participating raters were representative of the eventual users of the classification in only 4 studies. The type of raters, however, was reported in 40 studies, 20 of them included only orthopedic surgeons. Exclusively experienced surgeons (with or without non-surgeons) were involved in 9 studies, while others included



Study selection process for the review

surgeons with various levels of experience (i.e., fellows, registrars or residents participated). The experience level of raters with the targeted classification was reported in 21 studies. The median number of raters was 5 (2–36). 29 studies had between 4 and 6 raters. In 10/44, a group of at least 5 raters was involved in each evaluation.

It remained unclear how raters used the classifications. For instance, they could meet for a classification exercise (as we suspect in 10 studies), or classify cases at their own pace, but this was usually not stated. 12 studies had only one classification session and 32 had 2 sessions. An indication that raters classified each case independently of other raters was found in 31/44 studies. Moreover, the raters were blinded to clinical information about the patients in 17 studies.

The true distribution of classification categories in the sample (i.e., an attempt to use a "gold standard" classification process) was estimated in 9 studies from consensus between observers in 3 studies (Johnstone et al. 1993, Bernstein et al. 1996, Oskam et al. 2001), from an independent expert panel in 2 others (Lichtenhahn et al. 1991, Schipper et al. 2001), and 2 studies assessed intraoperative findings (Brady et al. 2000, Yacoubian et al. 2002). Silber et al. (2001) and Brorson et al. (2002) reported a distribution of cases according to fracture categories without mentioning how it was estimated.

16 studies reported raw agreement data, such as the observed percentage of agreement among raters, or the percentage of times that raters changed their ratings between various image modalities (e.g., Silber et al. 2001). The kappa coefficient was mostly used (39/44) to quantify agreement on intra- and interobserver reliability. Similar quantitative measures were used for the Sav statistic (Kreder et al. 1996, Swiontkowski et al. 1997) (2/44) and the Williams index (Swiontkowski et al. 1997) (1/44). Kappa coefficients for individual classification categories were reported in 3 studies (Thomsen et al. 1991, 1996, Swiontkowski et al. 1997). In addition, kappa values were used to investigate possible influencing factors, such as the type of observers (Bernstein et al. 1996, Illarramendi et al. 1998, Brady et al. 2000, Wainwright et al. 2000, Schipper et al. 2001), providing observer training (Rasmussen et al. 1993, Brorson et al. 2002), using binary questions (Craig and Dirschl 1998) or using different sets of diagnostic images (Bernstein et al. 1996, Chan et al. 1997, Brage et al. 1998, McAdams et al. 2002, Oner et al. 2002, Visutipol et al. 2000, Yacoubian et al. 2002). The statistical analyses, however, seemed adequate for the study objectives in only 17/44 studies.

Reported agreement

A survey of published kappa coefficients is given in Table 3. When considering reported kappa coefficients (whether an overall kappa, or a mean or mid-range of observer pair-wise kappa values) for interobserver reliability of classifications at the first classification session, only 4 of 86 coefficients were above 0.80, 17 between 0.60–0.80, 32 between 0.40–0.60 and 33 < 0.40. Almost all authors (35/39) used one of several proposed guidelines for the interpretation of the kappa coefficient or a modification (Table 4).

Discussion

Most of the existing classifications were developed to meet face and content validity criteria and were used in practice without having passed the test of reliability and accuracy. Our review showed that the majority of reliability studies were done in the last decade on the basis of well-accepted and commonly used classifications.

Table 2. Description of the methodological designs of published reliability studies of the fracture clasification systems

Reference	Fracture	Classification	Selection of fractures and images				Type and experience of observers					Classification session			
			A	B	C	D	E	F	G	H	I	J	K	L	M
Long bones															
Lichtenhahn et al. 1992	Long bones	Müller-AO	58	1		2	-				-	1	1	-	2
Johnstone et al. 1993	Long bones	Müller-AO	10	-		-	18	x	x	o	2	3	1		1
Humerus															
Kristiansen et al. 1988	Proximal humerus	Neer	100	-		-	4	x	x	o	1	2	1	-	-
Sidor et al. 1993	Proximal humerus	Neer	50	2		1	5	x	x	x	1	2	2	6	-
Siebenrock and Gerber 1993		Proximal humerusNeer; Müller-AO (groups & types)	95	1		-	5	x	o	o	1	2		2	2
Brien et al. 1995	Proximal humerus	Neer (reduced to 3 categories)	28	1		1	4	x	o	x	1	-		-	-
Bernstein et al. 1996	Proximal humerus	Neer (complete & simplified)	20	1	2	1	4	x	x	o	1	-	2	2	1
Sjöden et al. 1997	Proximal humerus	Müller-AO (groups); Neer (simplified)	26	1	1	-	10	x	o	x	-	-	-	2	-
Sjöden et al. 1999	Proximal humerus	Müller-AO (groups); Neer (simplified)	24	-	6	-	7	x	o	x	-	-	-	2	-
Wainwright et al. 2000	Distal humerus	Müller-AO (subgroups, groups & types); Riseborough and Radin; Jupiter and Mehne	33	1		-	9	o	x	x	-	-	1	3	-
Barton et al. 2001	Supracond. humerus	Modified Gartland	50	1		-	5	x	x	x	-	-	1	0.5	-
Brorson et al. 2002	Proximal humerus	Neer	42	1		2	14	x	x	o	-	3	2	0.5	4
Radius															
Andersen et al. 1991	Colles' fracture	Older	185	1		-	4	x	x	o	1	3	1	1	-
Andersen et al. 1996	Distal radius	Frykman; Melone; Mayo; Müller-AO	55	1		1	4	x	o	x	-	-	1	0.75	-
Kreder et al. 1996	Distal radius	Müller-AO (subgroups, groups & types)	30	2		-	36	x	x	x	2	3	2	0.5	-
Morgan et al. 1997	Radial head	Mason	25	-		-	20			o	-	-	1	0.75	-
Flinkkilä et al. 1998	Colles' fracture	Müller-AO modified & simplified	30	1	3	-	5	x	x	x	1	3	1	2	-
Illarramendi et al. 1998	Distal radius	Frykman (simplified); Müller-AO (simplified)	200	-		1	6	x	x	o	1	2	1	2	-
Oskam et al. 2001	Distal radius	Müller-AO (type A, B or C)	124	1		1	2	x	o	o	1	1	1	-	1
Femur															
Frandsen et al. 1988	Femoral neck	Garden's classification	100	1		-	8	o	x	x	1	3	1	-	-
Andersen et al. 1990	Trochanter	Evans (complete & simplified)	49	-		-	6	o	x	o	1	-	1	1	-
Gehrchen et al. 1993	Trochanter	Evans (complete & simplified)	52	1		-	4	o	x	o	1	-	-	1.5	-
Thomsen et al. 1996	Femoral neck	Garden (4 & 2 groups)	96	1		2	6	x	x	x	-	2	1	3	-
Gehrchen et al. 1997	Subtrochanteric	Seinsheimer (complete & simplified)	50	1		1	4	x	x	o	1	2	1	1.5	-
Blundell et al. 1998	Intracapsular prox.	Müller-AO (groups & types)	71	1		-	5	x	x	o	-	-	1	2	-
Brady et al. 2000	After THR	Vancouver	40	1		1	6	x	x	o	-	-	1	0.5	3
Schipper et al. 2001	Pertrochanteric	Müller-AO 31A (subgroups & groups)	20	2		2	15	x	x	x	3	3	2	3	2
Bjorgul and Reikeras 2002	Femoral neck	Garden	32	3		2	6	x	o	o	1	1	-	-	-
Pervez et al. 2002	Trochanteric	Jensen (modification of Evans); Müller-AO 31A (subgroups & groups)	88	2			5	x	x	o	-	-	-	3	-
Tibia															
Nielsen et al. 1990	Malleolar	Lauge-Hansen (complete & simplified)	118	1		-	4	o	x	o	1	-	1	1.5	-
Thomsen et al. 1991	Ankle	Lauge-Hansen; Weber	94	1		2	4	o	x	x	-	3	1	3	-
Rasmussen et al. 1993	Ankle	Lauge-Hansen (random group with training)	100	1		2	8	x	x	x	-	3	2	1	-
Martin et al. 1997	Distal tibia	Müller-AO 43 (groups & types); Rüedi-Allgöwer	43	2	2	-	6	x	x	x	2	2	1	2	-
Swiontkowski et al. 1997	Pilon (AO 43)	AO/OTA (subgroups, groups & types)	84	1		-	5	x	x	x	-	1	-	-	-
Chan et al. 1997	Tibia plateau	Schatzker	21	1	2	1	6	x	x	x	1	-	2	2	-
Yacoubian et al. 2002	Tibia plateau	Müller-AO; Schatzker	52	-	1*	-	3	x	o	o	-	-	1	-	3
Dirschl and Adams 1997	Tibia plafond	Rüedi-Allgöwer (original & with binary questions)	25	2		1	6	x	x	o	-	2	1	3	-
Brage et al. 1998	Ankle	Lauge-Hansen; Danis-Weber	99	1		1	4	x	x	x	-	3	1	0.5	-
Craig and Dirschl 1998	Malleolar	Müller-AO 44 (groups & types, binary system)	50	2		1	6	x	x	o	-	2	2	-	-
Spine															
Blauth et al. 1999	Thoracolombar spine	Magerl (subgroups, groups & types)	14	1	1	-	21				-	1	1	-	-
Oner et al. 2002	Thoracolombar spine	Müller-AO; Denis	53	1	1*	1	5	x	x	x	1	2	1	1.5	-
Pelvis															
Visutipol et al. 2000	Acetabular	Letournel and Judet	20	-	5	-	5			o	-	2	-	2	-
Silber et al. 2001	Pelvic	Torode and Zeig	62	1	2	-	3	x	o	o	-	-	1	-	4
Other															
McAdams et al. 2002	Scapular neck	No name given	21	1	4	-	4	x	x	x	-	1	2	1	-

Legends to Table 2.

A: Number of fractures
B: Selection process; 1= Random selection or consecutive series; 2= Represents "most" or all of the different patterns or chosen to provide a wide range of injury type and fracture severity; 3= Fractures belonging to one particular fracture category; "-"=Not reported
C: Use of CT scans and/or 3D images; 1=CT together with X-rays in each case; 2=comparison without and then added to Xray (same exercise); 3=comparison without and then added to Xray at second exercise; 4=Xray & CT separately & then combined in 3 exercises; 5=Xray & 3D separately; 6=CT and 3D together with Xrays in each case - *Use of MRI together with radiographs in one classification session
D: Quality of images; 1=Exclusions due to poor quality of images (Acceptability of the images assessed by independent orthopaedic surgeon(s) or fractures found "classifiable" by the authors); 2=Use of all cases independent of their quality; 3=Analysis of "good" quality images separately; ?=not reported
E: Number of observers; "-"=Not reported
F: Experienced surgeons (surgeons described such as "Experienced surgical specialist for the fracture type (experts)" or "Adult orthopaedic reconstructive surgeons (7–8 years exp)" or "Experienced surgeons (traumatologist)" or "attending surgeons" or "General orthopaedic surgeons (15-20 years exp)" or "Senior house officer") x = yes; o = no
G: Other surgeons (surgeons described such as "Orthopedic fellows / consultants / residents" or "Junior orthopedic resident" or "Junior house officer" or "Chief orthopedic resident" or "Senior / career registrar" or "registrar" or "trainee"); x = yes; o = no
H: Non-surgeons; x = yes; o = no (assessor such as "non-surgeon clinicians" or "radiologists" or "non-clinicians, e.g. Research coordinator, nurses")
I: Experience with classification; 1=Each assessor was familiar with the classification; 2=Some of the assessors were familiar with the classification; 3=No previous experience with the classification; "-"=Not reported
J: Instruction of observers prior to classification sessions; 1=No specific training; 2=Description of the classification given to observers; 3=Active reinstruction provided (e.g. presentation, pilot testing); "-"=Not reported
K: Type of classification session; 1=Observers classified fractures independently at their own pace; 2=Meeting of observers; "-"=not clearly reported
L: Minimum time between two sessions (month); "-"=only one session was conducted
M: "Gold standard" assessment; 1=Using consensus agreement between observers; 2=Using independent expert panel; 3=Using observations during operation; 4=distribution estimated but method not provided; "-"=Not reported

Methodological quality of reliability studies

For this review we short-listed some methodological items we believed were important from current knowledge regarding diagnostic studies (Pewsner et al. 2002) but recognize that such a scale should be further developed and validated. To associate lack of reporting with poor quality may underestimate the actual scientific value of the selected studies, and thus call for better reporting from authors in future studies.

Full descriptions of classification categories were available for these studies. In assessing a classification, however, we feel it is important to describe the intended classification process—i.e., on which basis, by whom, when and how the classification should be used. Omitting it would be equivalent to using a laboratory test without a protocol. Although not stated by the authors, all the studies included seemed to be essentially pragmatic—i.e., in choosing the study design, investigators may have reflected on the most intuitive applications in clinical practice.

For these studies, an unbiased selection of cases, independent of the quality of the diagnostic images, fracture pattern or treatment used, is needed. The spectrum of fractures should reflect the spectrum of the setting for which the classification is intended. Nonrandom selection of fractures will lead to spectrum bias (Pewsner et al. 2002). Some authors excluded cases that had "incomplete diagnostic images" (11 studies) or "images with inadequate quality" (13 studies). To limit (and assess) selection bias, it is more appropriate to include all cases, irrespective of the quality of the images, to agree on quality criteria and to evaluate their effects on the results.

The classification sessions were poorly described, so it was difficult to assess possible recording biases. In pragmatic studies, the classification exercise should relate as much as possible to the "intent of use" in practice, while logistical and practical issues will call for compromises in the study design. Raters, however, should remain "blinded" or unaware of prior patient information to limit their interpretation of diagnostic images.

Interpretation of kappa coefficients is difficult

It is tempting to interpret kappa values reported by different studies using one of the guidelines mentioned previously (Table 4) and compare the reliability of various classification systems for similar locations of fractures. In the following text we focus on the results obtained with reference to the Neer classification of proximal humerus fractures. A simplified version of this classification (using 3–6 categories) was investigated in 6 studies (Kristiansen et al. 1988, Siebenrock and Gerber 1993, Bernstein et al. 1996, Sjödén et al. 1997,

Table 3. Selected results [a] of reliability studies of fracture classification systems

References	Fractures	Clasification system	Number of categories	Percentage of observers' agreement	Inter-observer reliability (Kappa) Overall kappa coefficient (95%CI)	Mean (range) of pair-wise kappa coefficients
Humerus						
Kristiansen et al. 1988	Proximal	Neer (simplified)	5	24–59 [b]		0.30 (0.07–0.48)
			3	49–75 [b]		0.29 (0.03–0.47)
Sidor et al. 1993	Proximal	Neer	16	32	0.48	(0.43–0.58)
Siebenrock & Gerber 1993	Proximal	Neer (simplified)	4	26		0.40 (0.25–0.51)
		Müller-AO 11 (groups)	9	15		0.42 (0.36–0.49)
		Müller-AO 11 (types)	3	38		0.53 (0.50–0.58)
Brien et al. 1995	Proximal	Neer (simplified)	3	57–71 [b]		(0.37–0.75)
Bernstein et al. 1996	Proximal	Neer	16	15	0.52 (CR); 0.56 (only CT); 0.50 (CR & CT)	
		Neer (simplified)	6		0.54 (CR & CT)	
Sjöden et al. 1997	Proximal	Müller-AO 11 (groups)	9			0.31
		Neer (simplified)	6			0.42
Sjöden et al. 1999	Proximal	Müller-AO 11 (groups)	9			0.32
		Neer (simplified)	6			0.44
Wainwright et al. 2000	Distal	Riseborough and Radin	4		0.51	
		Müller-AO 13 (groups)	9		0.52	
		Müller-AO 13 (types)	3		0.66	
		Jupiter and Mehne	19		0.29	
Barton et al. 2001	Supracond.	Modified Gartland	3	64		0.77 (0.58–0.88)
Brorson et al. 2002	Proximal	Neer	6			0.27 (baseline) 0.62 (after teach.)
Radius						
Andersen et al. 1991	Colle's	Older	4	49		(0.60–0.75)
Andersen et al. 1996	Distal radius	Frykman	8	47	0.36	
		Melone	6	48	0.34	
		Mayo	4	59	0.43	
Morgan et al. 1997	Radial head	System of Mason	3	16		0.54 [d] (0.14–0.82)
Flinkkilä et al. 1998	Colle's	Müller-AO 23 modified	5	23–50 (CR) [b]	0.23 (CR); 0.25 (CR & CT)	
		Müller-AO 23 modified	2	60–87 (CR) [b]	0.48 (CR); 0.78 (CR & CT)	
Illarramendi et al. 1998	Distal radius	Frykman (simplified)	4	51–90 [b]		0.43 (0.36–0.84)
		Müller-AO 23 (simplified)	5	51–68 [b]		0.37 (0.25–0.48)
Oskam et al. 2001	Distal radius	Müller-AO 23 (type A, B or C)	3	80	0.65	
Femur						
Andersen et al. 1990	Trochanter	Evans	5	18		(0.38–0.69)
		Evans (stable/unstable)	2	57		
Gehrchen et al. 1993	Trochanter	Evans	5	44		
		Evans (stable/unstable)	2	65		(0.41–0.77)
Thomsen et al. 1996	Neck	Garden	4	56	0.39 (0.36–0.42)	(0.17–0.55)
		Garden (2 groups)	2	86	0.68 (0.63–0.72)	
Gehrchen et al. 1997	Subtroch.	Seinsheimer	5	42–60 [b]		
		Seinsheimer (3A yes/no)	2	72–84 [b]		(0.20–0.57)
Blundell et al. 1998	Intracaps. prox.	Müller-AO	9			0.30 (0.21–0.51)
		Müller-AO (revised)	3			0.85 (0.77–0.96)
Brady et al. 2000	After THR	Vancouver	5		0.61 (SE 0.09)	
Schipper et al. 2001	Pertrochanteric	Müller-AO 31A (subgroups)	9			0.33
		Müller-AO 31A (groups)	3			0.67
Bjorgul and Reikeras 2002	Neck	Garden	4			0.35 (0–0.56) [d]
Pervez et al. 2002	Trochanteric	Jensen (modification of Evans)	5			0.34 (0.17–0.38)
		Müller-AO 31A (subgroups)	9			0.33 (0.14–0.48)
		Müller-AO 31A (groups)	3			0.62 (0.50–0.71)
Tibia						
Craig and Dirschl 1998	Malleolar	Müller-AO 44 (groups)	9			0.61 (SD 0.07) [e]
	Malleolar	Müller-AO 44 (types)	3			0.77 (SD 0.08) [e]
Thomsen et al. 1991	Ankle	Lauge-Hansen	13	67	0.49	
		Weber	3	74	0.58	
Rasmussen et al. 1993	Ankle	Lauge-Hansen	15	63	0.51 (0.45–0.57) [c]	
Martin et al. 1997	Distal tibia	Müller-AO 43 (groups)	9		0.39 (0.35–0.43) (CR)	
		Müller-AO 43 (types)	3		0.59 (0.52–0.56) (CR)	
		Rüedi-Allgöwer	3		0.51 (0.45–0.57) (CR)	
Chan et al. 1997	Tibia plateau	Schatzker	6		(CR): 0.62 (0.31 –0.83); (CR & CT): 0.61 (0.37–0.83)	
Yacoubian et al. 2002	Tibia plateau	Müller-AO; Schatzker (Results not presented for a specific classification)	-		(CR): 0.68 (0.56–0.81) (CR+CT): 0.73 (0.68–0.77) (CR+MRI): 0.85 (0.82–0.89)	
Dirschl and Adams 1997	Tibia plafond	Rüedi-Allgöwer	3			0.43 (0.07–0.79) [e]
Brage et al. 1998	Ankle	Lauge-Hansen	15		0.58 (3 CR views)	
		Danis-Weber	3		0.69 (3 CR views)	

[a] Inter-observer reliability estimated by the Kappa coefficient following the only or first classification session

Legends to Table 3

Note = For simplification and illustration of interpretation issues, we extracted one or two typical kappa coefficients for inter-rater reliability from each study of long bone fractures. Given the number of fractures included in most samples, we excluded from this table the assessment of the Müller-AO classification at the sub-group level (usually including 27 potential categories), but considered results at the group and type levels. Data from abstracts and kappa coefficients for intra-rater reliability were ignored.
CR = Conventional radiography; CT = Computed tomography; SD = Standard deviation; SE = Standard error
[a] Results that were published as abstracts are not presented in this table
[b] Range for pairs of observers
[c] Group of raters with no training provided
[d] Median value
[e] Without use of binary questions

Sjödén et al. 1999, Brorson et al. 2002). The raw percentages of agreement were reported in 2 studies. Siebenrock and Gerber (1993) found a 26% full agreement among 5 raters, while Kristiansen et al. (1988) gave a range of pair-wise percentages of agreement of 24%–59%. Since the percentage of full agreement among all raters will decrease as the number of raters increases, we prefer the report of pair-wise percentages of agreement. It would also be useful to know whether the (dis)agreement concerns only particular classification categories.

Mean pair-wise kappa coefficients ranged from 0.27 to 0.62. Bernstein et al. (1996) noted an overall kappa of 0.54. All these coefficients were far from 1, and thus their reliability could be judged as unacceptable. The lowest kappa value of 0.27 was reported by using radiographs alone without additional training (Brorson et al. 2002). Kristiansen et al. (1988) found a similar value with radiographs alone and reported a lower kappa than other studies using CT scans (Bernstein et al. 1996, Sjödén et al. 1997, 1999), which suggests at first sight higher reliability of the classification with CT imaging. Using radiographs alone, Siebenrock and Gerber (1993) reported a kappa similar to those of Sjöden et al. (1997, 1999), but only experienced surgeons participated, while some studies included other less-experienced surgeons. Moreover, Siebenrock and Gerber (1993) assessed a simplified classification with 4 categories instead of 6, and probably organized a meeting of raters to obtain the data. Considering the variations among methodologies between studies, it appears difficult to interpret kappa values between studies and to find ways of improving reliability. Moreover, the utility of kappa is a matter of controversy, because its values depend strongly on the distribution of cases between the various classification categories within a sample (Gjørup 1988, Cook 1998)—often referred to as the "base rate problem" for binary classifications (Shrout 1998). Any differences observed may actually be caused simply by variations in the study samples (hence the need to have an estimate of this distribution). Therefore comparison of kappa coefficients is indicated only when the distribution of true categories is fairly similar among study samples or when the same sample is used on several occasions. The 2 studies of Sjöden et al. (1997, 1999) used the same study methods (except image modalities) and might have used a similar

Table 4. Guidelines used for the interpretation of the kappa coefficient

<table>
<tr><th rowspan="2">Guidelines proposed by</th><th colspan="9">Kappa coefficient scale</th><th rowspan="2">No. of studies</th><th rowspan="2">%</th></tr>
<tr><th><0</th><th>0.00</th><th>0.20</th><th>0.40</th><th>0.50</th><th>0.60</th><th>0.75</th><th>0.80</th><th>1.00</th></tr>
<tr><td>Landis and Koch 1977</td><td>Poor</td><td>Slight</td><td>Fair</td><td colspan="2">Moderate</td><td colspan="2">Substantial</td><td>Excellent [a]</td><td></td><td>26</td><td>59</td></tr>
<tr><td>Altman 1990</td><td colspan="2">Poor</td><td>Fair</td><td colspan="2">Moderate</td><td colspan="2">Good</td><td>Very good</td><td rowspan="5">Perfect</td><td>1</td><td>2</td></tr>
<tr><td>Fleiss 1981</td><td colspan="3">Poor</td><td colspan="3">Fair to Good</td><td colspan="2">Excellent</td><td>3</td><td>7</td></tr>
<tr><td>Svanholm et al. 1989</td><td colspan="4">Poor</td><td colspan="2">Fair</td><td colspan="2">Good or excell.</td><td>4</td><td>9</td></tr>
<tr><td>Martin et al. 1997</td><td colspan="4">Poor</td><td colspan="2">Fair to Good</td><td colspan="2">Excellent</td><td>1</td><td>2</td></tr>
<tr><td>Brage et al. 1998</td><td colspan="4">Poor</td><td colspan="2">Good</td><td colspan="2">Excellent</td><td>1</td><td>2</td></tr>
</table>

[a] or "Almost perfect"

sample of cases. Kappa values could therefore be compared to conclude that 3D images added little more than CT scans to the reliability of the simplified Neer classification. In a recent study, Thomsen et al. (2002) ensured that the distribution of "true" categories should be similar between samples so that kappa values could be compared.

Methodological standards are needed

In essence, reliability studies were done to evaluate indirectly the validity (or accuracy) of a classification—i.e., how close are reported categories to the true fracture trait in question. High reliability, however, does not mean high accuracy. For the validation of classification systems, however, the frequent lack of an adequate "gold standard" classification process is a major concern. This explains why, in our review, the investigation of classification accuracy was unusual.

The appropriate research design depends on the aims of the study and the clinical questions to be answered. For instance, assessing the use of a classification in real-life clinical practice from unrepresentative samples of cases and observers is inappropriate and can be misleading. However, this might be acceptable for small targeted pilot studies during the development phase. More detailed practical recommendations should be developed. The kappa coefficient must be used with care, and estimates should be reported separately for each category in the classification. Moreover, we encourage reporting the probable distribution of "true" categories (even if only derived from consensus agreement) to help in the interpretation of kappa coefficients. The kappa coefficient is useful for studying factors influencing reliability, but we believe existing guidelines for interpretation (Table 4) should not be used to draw conclusions about the reliability of a classification. These guidelines rely on categories that are arbitrary and liable to subjective interpretation (Uebersax 2002a). The result is that no agreement exists as to what constitutes an "acceptable" kappa coefficient. Indeed, the kappa statistic should not be used to answer all important questions in the evaluation of classification systems. Alternative methods need to be examined (Uebersax 2002b), in particular, for the assessment of the accuracy of classifications.

Most reliability studies have concluded that classification systems are unreliable. This review showed, however, that results and conclusions should be interpreted in the light of their methodological strength. As the flexibility for corrective measures in existing classifications is limited, we believe the validation process should be done before classification systems are recommended for use. The development and adoption of a systematic methodological approach for the development and validation of fracture classifications is needed.

The authors thank Dr Beate Hanson (Director of AOCID, Davos, CH) for her on-going support and suggestions in the preparation of this manuscript.

No competing interests declared.

Altman D G. Inter-rater agreement. In: Practical statistics for medical research. Chapman & Hall / CRC London 1990: 403-9.

Andersen D J, Blair W F, Steyers C M, Jr., Adams B D, el Khouri G Y, Brandser E A. Classification of distal radius fractures: an analysis of interobserver reliability and intraobserver reproducibility. J Hand Surg (Am) 1996; 21 (4): 574-82.

Andersen E, Jorgensen L G, Hededam L T. Evans' classification of trochanteric fractures: an assessment of the interobserver and intraobserver reliability. Injury 1990; 21 (6): 377-8.

Andersen G R, Rasmussen J B, Dahl B, Solgaard S. Older's classification of Colles' fractures. Good intraobserver and interobserver reproducibility in 185 cases. Acta Orthop Scand 1991; 62 (5): 463-4.

Barton K L, Kaminsky C K, Green D W, Shean C J, Kautz S M, Skaggs D L. Reliability of a modified Gartland classification of supracondylar humerus fractures. J Pediatr Orthop 2001; 21 (1): 27-30.

Bernstein J, Adler L M, Blank J E, Dalsey R M, Williams G R, Iannotti J P. Evaluation of the Neer system of classification of proximal humeral fractures with computerized tomographic scans and plain radiographs. J Bone Joint Surg (Am) 1996; 78 (9): 1371-5.

Bernstein J, Monaghan B A, Silber J S, DeLong W G. Taxonomy and treatment–a classification of fracture classifications. J Bone Joint Surg (Br) 1997; 79 (5): 706-7.

Bjorgul K, Reikeraos O. Low interobserver reliability of radiographic signs predicting healing disturbance in displaced intracapsular fracture of the femoral neck. Acta Orthop Scand 2002; 73 (3): 307-10.

Bland J M, Altman D G. Statistics notes: Validating scales and indexes. BMJ 2002; 324 (7337): 606-7.

Blauth M, Bastian L, Knop C, Lange U, Tusch G. [Interobserver reliability in the classification of thoraco-lumbar spinal injuries]. Orthopäde 1999; 28 (8): 662-81.

Blundell C M, Parker M J, Pryor G A, Hopkinson-Woolley J, Bhonsle S S. Assessment of the AO classification of intracapsular fractures of the proximal femur. J Bone Joint Surg (Br) 1998; 80 (4): 679-83.

Brady O H, Garbuz D S, Masri B A, Duncan C P . The reliability and validity of the Vancouver classification of femoral fractures after hip replacement. J Arthroplasty 2000; 15 (1): 59-62.

Brage M E, Rockett M, Vraney R, Anderson R, Toledano A. Ankle fracture classification: a comparison of reliability of three X-ray views versus two. Foot Ankle Int 1998; 19 (8): 555-62.

Brien H, Noftall F, MacMaster S, Cummings T, Landells C, Rockwood P. Neer's classification system: a critical appraisal. J Trauma 1995; 38 (2): 257-60.

Brorson S, Bagger J, Sylvest A, Hobjartsson A. Improved interobserver variation after training of doctors in the Neer system. A randomised trial. J Bone Joint Surg (Br) 2002; 84 (7): 950-4.

Browner B D, Jupiter J B, Levine A M, Trafton P G. Skeletal trauma - fractures, dislocations, ligamentous injuries. W.B. Saunders Company, A Division of Harcourt Brace & Company Philadelphia, London, Toronto, Montreal, Sydney, Tokyo 1998.

Burstein A H. Fracture classification systems: do they work and are they useful? J Bone Joint Surg (Am) 1993; 75 (12): 1743-4.

Chan P S, Klimkiewicz J J, Luchetti W T, Esterhai J L, Kneeland J B, Dalinka M K, Heppenstall R B. Impact of CT scan on treatment plan and fracture classification of tibial plateau fractures. J Orthop Trauma 1997; 11 (7): 484-9.

Colton C L. Telling the bones. J Bone Joint Surg (Br) 1991; 73 (3): 362-4.

Cook R J. Kappa and its dependence on marginal rates. In: Encyclopedia of biostatistics (Eds. Armitage P and Colton T). Wiley New York 1998: 2166-8.

Craig W L, III, Dirschl D R. Effects of binary decision making on the classification of fractures of the ankle. J Orthop Trauma 1998; 12 (4): 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 (7): 471-6.

Fleiss J L. Statistical methods for rates and proportions, 2nd ed. In: John Wiley & Sons New York 1981: 217-8.

Flinkkilä T, Nikkola-Sihto A, Kaarela O, Paakko 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 (4): 670-2.

Frandsen P A, Andersen E, Madsen F, Skjodt T. Garden's classification of femoral neck fractures. An assessment of inter-observer variation. J Bone Joint Surg (Br) 1988; 70 (4): 588-90.

Gehrchen P M, Nielsen J O, Olesen B. Poor reproducibility of Evans' classification of the trochanteric fracture. Assessment of 4 observers in 52 cases. Acta Orthop Scand 1993; 64 (1): 71-2.

Gehrchen P M, Nielsen J O, Olesen B, Andresen B K. Seinsheimer's classification of subtrochanteric fractures. Poor reproducibility of 4 observers' evaluation of 50 cases. Acta Orthop Scand 1997; 68 (6): 524-6.

Gjørup T. The kappa coefficient and the prevalence of a diagnosis. Methods Inf Med 1988; 27 (4): 184-6.

Illarramendi A, Gonzalez D, V, Segal E, De Carli P, Maignon G, Gallucci G. Evaluation of simplified Frykman and AO classifications of fractures of the distal radius. Assessment of interobserver and intraobserver agreement. Int Orthop 1998; 22 (2): 111-5.

Johnstone D J, Radford W J, Parnell E J. Interobserver variation using the AO/ASIF classification of long bone fractures. Injury 1993; 24 (3): 163-5.

Kreder H J, Hanel D P, McKee M, Jupiter J, McGillivary G, Swiontkowski M F. Consistency of AO fracture classification for the distal radius. J Bone Joint Surg (Br) 1996; 78 (5): 726-31.

Kristiansen B, Andersen U L, Olsen C A, Varmarken J E. The Neer classification of fractures of the proximal humerus. An assessment of interobserver variation. Skeletal Radiol 1988; 17 (6): 420-2.

Landis J R, Koch G G. The measurement of observer agreement for categorical data. Biometrics 1977; 33 (1): 159-74.

Lichtenhahn P, Fernandez D L, Schatzker J. [Analysis of the "user friendliness" of the AO classification of fractures]. Helv Chir Acta 1992; 58 (6): 919-24.

Lindsjö U. Classification of ankle fractures: the Lauge-Hansen or AO system? Clin Orthop 1985; (199): 12-6.

Martin J S, Marsh J L. Current classification of fractures. Rationale and utility. Radiol Clin North Am 1997; 35 (3): 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 (7): 477-83.

McAdams T R, Blevins F T, Martin T P, DeCoster T A. The role of plain films and computed tomography in the evaluation of scapular neck fractures. J Orthop Trauma 2002; 16 (1): 7-11.

Morgan S J, Groshen S L, Itamura J M, Shankwiler J, Brien W W, Kuschner S H. Reliability evaluation of classifying radial head fractures by the system of Mason. Bull Hosp Jt Dis 1997; 56 (2): 95-8.

Nielsen J O, Dons-Jensen H, Sorensen H T. Lauge-Hansen classification of malleolar fractures. An assessment of the reproducibility in 118 cases. Acta Orthop Scand 1990; 61 (5): 385-7.

Oner F C, Ramos L M, Simmermacher R K, Kingma P T, Diekerhof C H, Dhert W J, Verbout A J. Classification of thoracic and lumbar spine fractures: problems of reproducibility. A study of 53 patients using CT and MRI. Eur Spine J 2002; 11 (3): 235-45.

Oskam J, Kingma J, Klasen H J. Interrater reliability for the basic categories of the AO/ASIF's system as a frame of reference for classifying distal radial fractures. Percept Mot Skills 2001; 92 (2): 589-94.

Pervez H, Parker M, Pryor G, Lutchman L, Chirodian N. Classification of trochanteric fracture of the proximal femur: a study of the reliability of current systems. Injury 2002; 33 (8): 713-5.

Pewsner D, Battaglia M, Bucher H, Minder C, Grossenbacher F, Egger M. The Bayes Library of Diagnostic Studies and Reviews. Basel Institute for Clinical Epidemiology, Institute for Social and Preventive Medicine University of Berne Basel, Berne 2002.

Rasmussen S, Madsen PV, Bennicke K. Observer variation in the Lauge-Hansen classification of ankle fractures. Precision improved by instruction. Acta Orthop Scand 1993; 64 (6): 693-4.

Rockwood C A, Green D P, Bucholz R W, Heckman J D. Rockwood and Green's Fractures in Adults. Lippincott-Raven Philadelphia, New York 1996.

Schipper I B, Steyerberg E W, Castelein R M, van Vugt A B. Reliability of the AO/ASIF classification for pertrochanteric femoral fractures. Acta Orthop Scand 2001; 72 (1): 36-41.

Shrout P E. Measurement reliability and agreement in psychiatry. Stat Methods Med Res 1998; 7 (3): 301-17.

Sidor M L, Zuckerman J D, Lyon T, Koval K, Cuomo F, Schoenberg N. The Neer classification system for proximal humeral fractures. An assessment of interobserver reliability and intraobserver reproducibility. J Bone Joint Surg (Am) 1993; 75 (12): 1745-50.

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 (12): 1751-5.

Silber J S, Flynn J M, Katz M A, Ganley T J, Koffler K M, Drummond D S. Role of computed tomography in the classification and management of pediatric pelvic fractures. J Pediatr Orthop 2001; 21 (2): 148-51.

Sjödén G O, Movin T, Guntner P, Aspelin P , Ahrengart L, Ersmark H, Sperber A. Poor reproducibility of classification of proximal humeral fractures. Additional CT of minor value. Acta Orthop Scand 1997; 68 (3): 239-42.

Sjödén G O, Movin T, Aspelin P, Guntner P , Shalabi A. 3D-radiographic analysis does not improve the Neer and AO classifications of proximal humeral fractures. Acta Orthop Scand 1999; 70 (4): 325-8.

Svanholm H, Starklint H, Gundersen H J, Fabricius J, Barlebo H, Olsen S. Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. APMIS 1989; 97 (8): 689-98.

Swiontkowski M F, Sands A K, Agel J, Diab M, Schwappach J R, Kreder H J. Interobserver variation in the AO/OTA fracture classification system for pilon fractures: is there a problem? J Orthop Trauma 1997; 11 (7): 467-70.

Thomsen N O, 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 (4): 676-8.

Thomsen N O, Jensen C M, Skovgaard N, Pedersen M S, Pallesen P, Soe-Nielsen N H, Rosenklint A. Observer variation in the radiographic classification of fractures of the neck of the femur using Garden's system. Int Orthop 1996; 20 (5): 326-9.

Thomsen N O, Olsen L H, Nielsen S T. Kappa statistic in the assessment of observer variation: the significance of multiple observers classifying ankle fractures. J Orthop Sci 2002; 7 (2): 163-6.

Uebersax J. Kappa Coefficients. http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm 2002a.

Uebersax J. Statistical Methods for Rater Agreement. http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm 2002b.

Visutipol B, Chobtangsin P, Ketmalasiri B, Pattarabanjird N, Varodompun N. Evaluation of Letournel and Judet classification of acetabular fracture with plain radiographs and three-dimensional computerized tomographic scan. J Orthop Surg 2000; 8 (1): 33-7.

Wainwright A M, Williams J R, Carr A J. Interobserver and intraobserver variation in classification systems for fractures of the distal humerus. J Bone Joint Surg (Br) 2000; 82 (5): 636-42.

Yacoubian S V, Nevins R T, Sallis J G, Potter H G, Lorich D G. Impact of MRI on treatment plan and fracture classification of tibial plateau fractures. J Orthop Trauma 2002; 16 (9): 632-7.