Eur Spine J (2002) 11 : 235–245
DOI 10.1007/s00586-001-0364-8

ORIGINAL ARTICLE

F. C. Oner
L. M. P. Ramos
R. K. J. Simmermacher
P. T. D. Kingma
C. H. Diekerhof
W. J. A. Dhert
A. J. Verbout

Classification of thoracic and lumbar spine fractures: problems of reproducibility

A study of 53 patients using CT and MRI

Received: 22 June 2001
Revised: 10 October 2001
Accepted: 23 October 2001
Published online: 29 January 2002

F.C. Oner (✉)
University Hospital Utrecht, G.05.228,
PO Box 85500, 3508 GA Utrecht,
The Netherlands
e-mail: f.c.oner@chir.azu.nl,
Tel.: +31-30-2506971,
Fax: +31-30-2541944

F.C. Oner · P.T.D. Kingma
C.H. Diekerhof · W.J.A. Dhert
A.J. Verbout
Department of Orthopedics,
University Medical Center, Utrecht,
The Netherlands

L.M.P. Ramos
Department of Radiology,
University Medical Center Utrecht,
The Netherlands

R.K.J. Simmermacher
Department of General Surgery,
University Medical Center Utrecht,
The Netherlands

Abstract Reproducibility of fracture classification systems in general has been a matter of controversy. The reproducibility of spinal fracture classifications has not been sufficiently studied. We studied the inter-observer and intra-observer reproducibility of the Magerl (AO) classification using radiograms, CTs and MRIs of 53 patients. We compared this classification with the older and simpler Denis classification. Five observers classified the fractures, first using the radiograms and CTs and, 6 weeks later, with radiograms and MRIs. Three of the observers repeated the readings after 3 months. Three observers also classified the fractures according to Denis. Agreement was measured using Cohen's κ test. The type (A, B, C) classification of the AO system was fairly reproducible with CTs. With MRI this was only moderate. Group subclassification of the types yielded higher κ values, corresponding to substantial agreement. The agreement was, in general, better with the Denis classification, but the variance was higher due to the difficulty of finding proper categories for some injury patterns. Although the AO classification allows proper registration of all kinds of injury, the reproducibility, especially at the type level, is problematic. Use of MRI and better definition of the distinctive properties of the three different types may enhance the reproducibility of the scheme.

Keywords Spine · Thoracolumbar · Spine fractures · Classification · MRI

Introduction

Fracture classification schemes are considered necessary tools as a conceptual framework for diagnosis and treatment. They are also systems for communication about the relative severity of injuries and the result of different treatment options. However, classification schemes used for the peripheral skeleton have been shown to have poor to moderate inter-observer and intra-observer reproducibility [1, 2, 3, 6, 10, 11, 13, 20, 21, 27, 28, 29, 31]. This raises questions about the usefulness of any classification scheme about fractures, which represent a continuum of different injuries resulting from the chaotic processes of trauma.

Injuries of the thoracolumbar spine pose an even greater challenge for classification attempts, due to the involvement of soft-tissue structures aside from different bone-fracture patterns. Thoracic and lumbar spine fractures represent complex injuries of a structure composed of parts with different susceptibility to injury and different healing potentials. This complexity was already recognized by

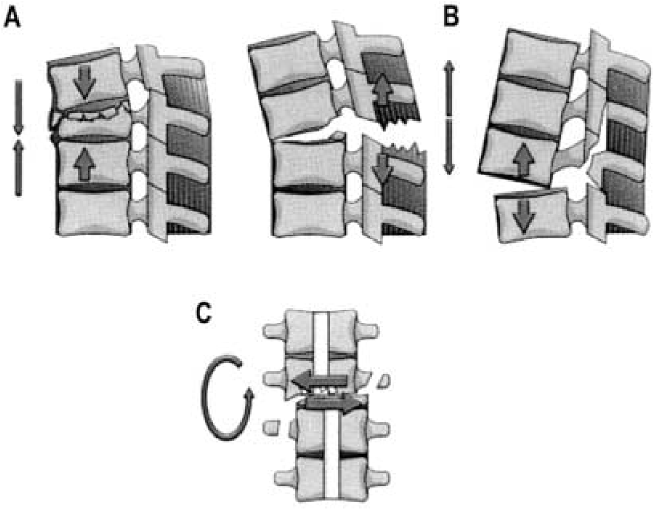

Fig. 1 Essential characteristics of the three injury types according to Magerl et al. [14]. Type A: compression injury of the anterior column; type B: two column injury with transverse disruption; type C: two column injury with superimposed rotation

Böhler, who devised the first schematic classification of these fractures [5]. Subsequent concepts tried to capture the various injury patterns using architectonic abstractions, such as columns. The two-column concept of Holdsworth [12] was followed by the three column concepts of Louis [18] and Denis [9]. The main concern of these authors was the relation of different fracture patterns to immediate and long-term mechanical and neurologic stability. These patterns were identified on radiograms and, later, in the case of Denis, also transverse CT images. Although these schemes were used for a long time in the literature, it appears that no studies have systematically questioned the reproducibility of these classifications.

In 1994 a new and comprehensive classification scheme was proposed by the AO group as a result of a review of 1,445 patients over a 10-year period [19]. This scheme takes into account the morphologic appearance on radiographs (including extent of soft-tissue involvement), the mechanisms of injury, and the increasing severity of the injury. Three main types of injury are defined by common morphologic characteristics and a common injury-producing force (Fig. 1). Extent and direction of soft-tissue injury are the main determinants of these types. Each type is further divided into groups and subgroups, using the common AO 3–3-3 grid (Fig. 2). An experimental study showed good relation between the type categorization of the scheme and the resulting mechanical instability in a cadaveric fracture model [17].

The type (A, B, C) classification depends mainly on the question of the mechanical integrity of the posterior column. Injury to the posterior column means allocation of the injury from the type A to the more severe types B or C. Although the authors emphasize that the involvement of soft tissues in transverse plane is the key determinant in type level of classification, the integrity of the posterior column was indirectly deduced from radiograms and CT scans in the original series. It can be expected that this main distinction, based on judgments of a predominantly soft-tissue injury, would prove to be difficult because soft-tissue injury patterns associated with spinal fractures cannot be sufficiently depicted based only on radiograms and CT. The authors did not mention MRI findings but, theoretically, addition of MRI can potentially increase the reliability of this level of the classification. MRI has been shown to be capable of detecting ligamentary injuries associated with thoracolumbar spine fractures in experimental and clinical studies [15, 22, 24,30]. It has been suggested that future classifications should include MRI findings of soft-tissue injuries [23,25].

Our goals in this study were to determine the inter-observer reliability and intra-observer reproducibility of the AO classification scheme, and to test the hypothesis that

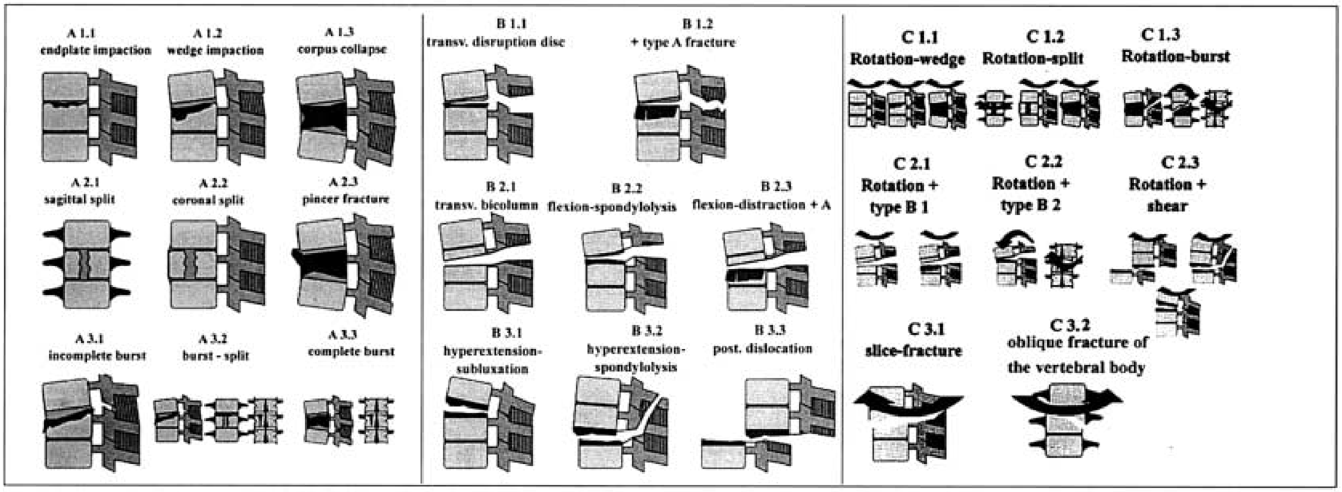

Fig. 2 Group and subgroup divisions of type A, B and C fractures

MRI would result in a better agreement about the type categorization of the fractures. We also studied the same issues for the older and simpler Denis classification. Denis classification was based on radiograms and transverse CT images. Novel imaging technology, such as CT-MPR (multi planar reconstruction) and MRI may show the flaws of this classification, which is, in our view, an overt simplification of complex injuries.

Materials and methods

Since 1994 we have obtained MRIs of all patients with a thoracic or lumbar spine fracture admitted to our hospital. T1-weighted (TR 578; TE 25) and T2-weighted (TR 2000; TE 100) images were obtained during the first week after admission. MRI was not possible in cases of polytrauma necessitating long periods of assisted ventilation or emergency intervention before imaging. MRI was therefore not performed for 13 patients. MRIs were obtained for 78 patients in the period from September 1994 to September 1997. Fifty-three of these patients also had adequate CTs, with multiplanar 2D reconstructions. Standard AP and lateral radiograms, CT scans and MRI of these 53 patients were collected and filed in an anonymous fashion, blinded for all patient data. Five observers participated in the study: one orthopedic spine surgeon, one general trauma surgeon, one neuroradiologist and two orthopedic residents in their fifth (resident 1) and third (resident 2) year of training, respectively. In our hospital, a spinal-injury work group, consisting of the orthopedic spine surgeon, the general trauma surgeon, a neurosurgeon and the neuroradiologist, meet weekly to discuss all patients with spinal injury. Orthopedic residents also attend these meetings. We have been using the AO classification in this work group since 1995, so each participant was acquainted with the scheme. Prior to the start of the study each participant read the original article by Magerl et al. [19], describing the basic concepts of the scheme. Each participant was provided with a visual representation of the classification, with a short description of the classification at the first three levels of the scheme (i.e., type, group, and subgroup, such as A 1.1 or B 2.3) (Figs. 1, 2). Observers were asked to note every fracture seen and to complete a separate form for each of the fractures. Subsequently, all five observers rated the files, first only with radiograms and CTs and then, 6–8 weeks after the first rating, with radiograms and MRIs. These ratings were used for inter-observer agreement between the five observers and intra-observer agreement between CT and MRI readings. Three months after the first rating the orthopedic spine surgeon and the two orthopedic residents rated all the files again in the same manner. These ratings were used to determine the intra-observer agreement between the first and second CT and MRI readings. To compare these results with an older and simpler scheme, three observers (the orthopedic spine surgeon and the two residents) rated the injuries in the same manner according to the Denis classification, 6 months after the last readings.

Cohen's κ test was used for inter-observer and intra-observer agreement. The guidelines proposed by Landis and Koch [16] were used to categorize κ values: 0.00–0.20, slight reliability; 0.21–0.40, fair reliability; 0.41–0.60, moderate reliability; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement. Agreement on the presence and levels of observed fractures were first determined. For inter-observer measurements only, the cases were included if a fracture was reported at the same level in both of the readings.

According to the basic foundation of the AO classification scheme, the distinction between type A and the other two types is an essential feature concerning posterior column involvement. The crucial distinction at the type level is whether the injury belongs to the common and more stable type A, or to the potentially more unstable type B or C category. This distinction depends largely on the recognition of soft-tissue involvement in transverse plane, which is expected to be more difficult with radiograms and CTs and, according to our hypothesis, would be better established using MRI. For this reason, we first measured the agreements for type A and non-type A (type B or C) distinction. Thus, the A/non-A distinction reflects essentially the judgment of the observer on the integrity of the posterior column. The agreements were measured on the separate types (A, B, C) as the second level. The basic subdivision of the types follows largely the subdivision of the type A, therefore agreement on groups and subgroups was measured for type A fractures only and was reported in cases when both of the readings reported a type A fracture. Finally, agreement in all three levels was measured.

From the ten readings (one with CT and one with MRI for each participant) inter-observer agreement was measured separately for the CT and MRI readings. Agreement between the CT and MRI readings of each participant was also measured as intra-observer agreement between CT and MRI. Intra-observer agreement was measured between the two CT readings and the two MRI readings of the three participants who did the entire procedure for the second time. As a summary measure for the κ coefficients, generalized κ's were used for the inter-observer agreement and mean κ's for intra-observer agreement.

For the Denis classification [9], the inter-observer agreement was first measured at the type level (compression, burst, seat-belt and fracture dislocations) and then the whole classification (subgroups of the four basic types as described by Denis) with CT and MRI readings of the three observers (spine surgeon, resident 1 and resident 2). Intra-observer agreement was also measured between the CT and MRI ratings of the three observers.

Statistics were performed using SPSS/PC + version 5.0.1.

Results

The classifications according to the AO scheme provided by the observers on the CT and MRI readings are shown in Table 1. Multiple fractures in the same patient are reported under the same patient number with different levels. Seventy-six fractures were reported at least once. Sixty fractures were reported by every observer in every reading. The frequencies of different classes reported by the five observers on the CT and MRI readings are shown in Table 2. The highest frequency of non-A class report was by the spine surgeon, the lowest by resident 1. Twenty of the total possible 27 categories were reported at least once.

Considering the inter-observer agreement of the number and level of the fractures, the mean κ value was 0.65 (0.53–0.94) for the CT readings and 0.62 (0.43–0.95) for MRI readings. For the same issue, the mean κ between the CT and MRI readings of each participant was 0.77 (0.62–0.90).

The simple κ values concerning type A/non-type A distinction, type categorization, type A fractures group and subgroup, and agreement on all three levels (complete classification) are shown in Tables 3, 4, 5, 6 and 7. Generalized κ's and mean κ's, with the ranges, are summarized in Table 8. The distinction between type A/non-type A inter-observer agreement was better with MRI, but reached only moderate levels. Agreement on complete type classification and classification in all three levels were fair in both

Table 1 Fractures seen and classified by the five observers on CT and MRI readings

Patient	Level	Trauma surgeon		Radiologist		Spine surgeon		Resident 1		Resident 2	
		CT	MRI	CT	MRI	CT	MRI	CT	MRI	CT	MRI
1	T7	B 2.3	A 3.1	A 2.3	A 2.3	B 2.3	B 2.3	A 3.3	A 3.3	B 2.3	B 2.3
2	L3	B 1.2	A 3.1	A 3.3	A 3.3	A 3.3	A 3.3	A 3.2	A 3.3	A 3.3	A 3.3
3	L1	A 3.1	A 1.2	A 2.3	A 2.3	A 3.3	A 3.3	A 3.1	A 1.2	A 3.3	A 3.3
4	L1	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3
5	L1	B 2.3	A 3.1	B 2.3	A 3.3	B 1.2	A 3.2	A 3.1	A 3.1	B 1.2	A 3.3
6	L1	A 3.3	A 3.3	A 3.3	A 3.2	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3
7	L3	B 2.3	A 3.1	A 3.3	A 3.3	A 3.2	A 3.3	A 3.1	A 3.1	B 2.3	A 3.1
8	L1	B 2.3	B 1.2	A 3.3	A 3.3	B 1.2	B 1.2	A 3.2	B 1.2	B 1.2	B 1.2
9	L 1	A 1.2	A 1.2	A 1.3	A 1.3	B 1.2	B 1.2	A 1.2	A 1.2	A 1.2	B 1.2
10	T12	A 3.2	A 3.1	A 2.3	A 3.2	A 3.2	A 3.1	A 3.2	A 3.1	A 3.2	A 3.2
11	L1	A 2.1	A 3.3	A 3.1	A 3.1	A 3.1	A 3.1	A 3.1	A 3.1	A 3.3	A 3.3
12	L1	A 3.1	A 3.1	A 3.1	B 2.3	A 3.1	A 3.1	A 3.2	A 3.1	A 3.1	A 3.1
13	T12	A 1.2									
13	L1	A 3.1	A 1.2	A 1.3	A 1.3	A 3.1	A 3.1	A 3.1	A 3.1	A 1.3	A 1.3
14	T12	A 3.1			A 1.2		A 1.1				
14	L2	B 2.3	A 3.3	A 3.2	A 3.1	A 3.2	A 1.2	A 1.2	A 3.1	A 3.2	A 3.2
14	L3	B 2.3	A 1.2	A 3.3	A 3.3	A 3.3	A 3.2	A 3.2	A 3.3	B 2.3	B 2.3
15	T8	A 1.2	A 3.1	A 2.3	A 3.1	A 3.1	A 3.1	A 1.2	A 1.3	A 3.1	A 3.1
15	T11				A 1.1		A 1.1				
16	T12	A 3.2	A 3.2	A 2.2	A 3.3	A 3.2	A 3.2			A 3.2	
16	L3	B 2.3	A 3.3	A 3.3	A 2.3	A 3.3	A 3.3	A 3.3	A 3.1	A 3.3	B 2.2
17	L3	B 2.3	A 2.2	A 3.3	A 3.3	B 2.3	B 1.2	A 1.2	A 2.3	B 2.3	B 2.1
18	L1	A 1.2	A 1.2	B 1.2	B 3.2	B 1.2	B 1.2	A 1.2	A 2.3	B 1.2	B 1.2
19	T8	A 3.3	A 1.2	A 1.1	A 1.3	A 1.2	A 3.1				A 1.3
19	T11	A 1.2	A 1.2	A 1.1	A 1.2	A 1.2	A 1.1	A 1.2	A 1.2	A 1.2	A 1.2
19	L3	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2	A 1.1	A 1.2	A 1.2	A 1.2	A 1.2
20	T12	A 3.1	A 1.2	A 1.2	A 1.2	A 1.2	A 3.2	A 1.2	A 1.2	A 1.2	A 1.2
21	T12	A 3.1	A 3.1	A 2.3	A 2.3	A 3.1	A 3.2	A 1.2	A 3.1	A 1.2	A 1.2
22	T12	A 1.1	A 1.1							A 1.1	
22	L1	A 3.2	B 1.2	B 1.2	A 1.3	A 3.2	A 3.1	A 2.3	A 3.3	A 3.2	A 3.2
23	L1		A 1.1	A 1.3	A 1.1	A 1.3	A 1.1	A 1.2	A 1.2		A 1.3
23	L2		A 1.2	A 1.3	A 1.1	A 1.3	A 1.1	A 1.2	A 1.2	A 1.3	A 1.3
23	L3		A 1.1	A 1.3	A 1.1	A 1.3	A 1.1	A 1.2	A 1.2	A 1.3	A 1.3
24	L 1	B 2.3	A 3.2	A 3.2	A 3.2	A 3.3	A 3.3	A 3.3	A 3.3	B 2.3	B 2.3
25	L1	A 3.2	A 3.1	A 3.2	A 3.2	A 3.2	B 2.3	A 3.1	A 3.1	A 3.3	A 3.3
26	L2	A 3.1	B 1.2	A 1.3	A 1.3	A 1.2	A 1.2	A 1.2	A 1.3	A 1.2	A 1.2
27	T12	B 2.1	A 3.1	A 3.3	A 3.3	B 2.3	B 1.2	A 3.2	A 3.1	B 2.1	B 2.2
28	L1	A 3.2	A 2.3	A 3.3	A 3.3	A 3.3	C 1.3	A 3.2	A 3.3	A 3.3	A 3.3
29	T12	A 1.2	B 1.2	B 1.2	B 1.2	B 1.2	B 1.2	A 1.2	B 1.2	B 1.2	B 1.2
30	L1	A 3.1	B 1.2	A 2.3	B 1.2	A 3.1	B 1.2	A 1.2	A 1.2	C 2.2	C 2.2
31	T12	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3
32	L2	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	B 2.3	A 3.3	A 3.3	A 3.3	A 3.3
33	L1	A 1.2	A 3.1	A 3.1	A 3.1	A 3.1	A 3.1	A 1.2	A 1.2	A 3.1	A 3.1
34	L1	A 1.2	A 1.1	A 1.2	A 1.2	A 1.1	A 1.1	A 2.1	A 1.2	A 1.2	A 1.2
34	L2	A 1.2	A 1.2	A 1.3	A 1.2	A 1.2	A 1.1	A 1.2	A 1.2	A 1.2	A 1.2
34	L4	A 1.2	A 1.1					A 1.2			
34	L5	A 3.2	A 3.3	A 3.3	A 3.3	A 3.3	A 3.1	A 3.2	A 3.3	A 3.3	A 3.3
35	L1	A 1.2	B 2.1	A 1.2	B 1.2	A 1.2	B 1.2	A 1.2	B 1.2	A 1.2	A 1.2
36	T3	A 3.2	A 3.3	A 3.3	A 3.3	B 2.3	B 2.3	A 3.3	A 3.3	A 2.3	A 3.3
36	T4			A 3.2	A 3.2	B 1.2	B 2.3	A 3.3	A 3.3		A 3.3
37	L1	A 3.3	B 1.2	A 3.3	A 3.3	A 3.2	A 3.1	A 3.1	A 3.1	A 3.3	A 3.3
38	L1	A 3.3	A 3.1	A 3.2	A 3.2	A 3.2	A 3.1	A 3.2	A 1.2	A 3.3	A 3.3
38	L5	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 1.2	A 3.3	A 3.2	A 3.2
39	L1	A 1.2	A 1.2	A 1.2	A 1.2	B 1.2	A 1.2	A 1.2	A 1.2	A 3.2	A 3.2

Table 1 continued

Patient	Level	Trauma surgeon		Radiologist		Spine surgeon		Resident 1		Resident 2	
		CT	MRI	CT	MRI	CT	MRI	CT	MRI	CT	MRI
40	L1	A 3.3	A 3.3	A 3.3	A 3.3	C 1.3	A 3.3	A 3.3	A 3.1	A 3.3	A 1.2
41	T12	B 1.2	B 1.2	A 3.3	B 1.2	B 2.3	B 1.2	A 3.1	B 2.2	B 2.3	A 1.1
41	L1	A 1.1		A 1.1	A 1.1	A 1.2	A 1.1			A 1.1	A 1.1
42	L3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.2	A 3.2	A 3.3	A 3.3	A 3.3
43	L1	C 1.3	C 1.3	C 3.2	C 3.2	C 1.3	C 2.1	C 1.3	B 1.2	C 1.3	C 1.3
44	T9	A 1.2	A 1.2	A 3.3	A 3.3	C 1.3	C 1.2	B 2.2	A 3.3	C 1.3	C 1.3
45	L1	A 3.1	A 3.1	A 2.3	A 3.3	A 3.1	A 3.1	A 3.1	A 3.1	A 3.3	A 3.3
46	L1	A 1.2	A 1.2	A 1.1	A 1.1	A 1.3	A 1.2	A 1.2	A 3.1	A 1.2	A 1.2
47	L1	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2	A 1.1	A 3.2	A 1.1	A 1.1	A 1.1
47	L2	A 1.2	A 1.2	A 3.1	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2
47	L3	A 3.3	A 3.1	A 3.3	A 3.3	A 3.2	A 3.1	A 1.2	A 3.1	A 3.3	A 3.3
48	T12	A 1.2	B 1.2	A 1.2	B 1.2	B 1.2	B 1.2	A 1.2	A 1.2	A 1.2	B 1.2
48	L2	A 1.2		A 1.1	A 1.1		A 1.1				A 1.1
48	L3			A 1.1	A 1.1		A 1.1				A 1.1
48	L4						A 1.1				
49	L1	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2	A 1.2
50	L1	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3	A 3.3
51	L1	A 1.2	A 1.1	A 1.2	A 1.3	A 1.2	A 3.1	A 1.2	A 1.2		A 1.3
51	L2	A 1.2	A 1.2	A 1.2	A 1.3	A 3.1	A 1.2	A 1.1	A 1.2	A 1.1	A 1.3
51	L3	A 1.1	A 1.2	A 1.2	A 1.3	A 1.2	A 1.2	A 1.2	A 1.2	A 1.1	A 1.3
52	L1	A 1.2	A 1.2	A 1.2	A 1.2	A 1.3	A 1.2	A 1.2	A 1.3	A 1.2	A 1.2
53	L1	A 3.3	A 3.1	A 3.3	A 3.3	A 3.2	B 1.2	A 3.2	A 3.1	A 3.3	A 3.3

Table 2 Frequencies of different fracture classes reported by the five observers on CT and MRI readings

	Trauma surgeon		Radiologist		Spine surgeon		Resident 1		Resident 2	
	CT	MRI	CT	MRI	CT	MRI	CT	MRI	CT	MRI
A 1.1	3	6	6	8	1	14	1	1	5	5
A 1.2	23	20	12	11	13	9	29	20	14	13
A 1.3			7	8	5			3	3	8
A 2.1	1						1			
A 2.2		1	1					2		
A 2.3		1	7	4			1	17	1	
A 3.1	9	15	4	4	9	14	9		3	4
A 3.2	7	2	5	6	10	6	12	17	6	5
A 3.3	13	13	23	23	14	11	11		19	20
B 1.1								4		
B 1.2	2	9	3	5	8	11			4	5
B 1.3										
B 2.1	1								1	1
B 2.2							1	1		2
B 2.3	9		1	1	5	5			6	3
B 3.1										
B 3.2				1						
B 3.3										
C 1.1										
C 1.2						1				
C 1.3	1	1			3	1	1		2	2
C 2.1						1				
C 2.2									1	1
C 2.3										
C 3.1										
C 3.2			1	1						

Table 3 κ Values concerning the A/non-A distinction

		Trauma surgeon		Radiologist		Spine surgeon		Resident 1		Resident 2	
		CT	MRI	CT	MRI	CT	MRI	CT	MRI	CT	MRI
Trauma surgeon	**CT**	xxx	**0.11**	0.12	0.05	0.36	0.21	0.08	0.16	0.71	0.48
	MRI		xxx	0.33	0.61	0.27	0.38	0.12	0.58	0.28	0.29
Radiologist	**CT**			xxx	**0.41**	0.30	0.15	0.25	0.40	0.34	0.23
	MRI				xxx	0.31	0.43	0.16	0.63	0.35	0.36
Spine surgeon	**CT**					xxx	**0.66**	0.19	0.26	0.60	0.57
	MRI						xxx	0.15	0.31	0.50	0.57
Resident 1	**CT**							xxx	**0.30**	0.20	0.21
	MRI								xxx	0.28	0.15
Resident 2	**CT**									xxx	**0.76**
	MRI										xxx

Table 4 κ Values concerning the type categorization

		Trauma surgeon		Radiologist		Spine surgeon		Resident 1		Resident 2	
		CT	MRI	CT	MRI	CT	MRI	CT	MRI	CT	MRI
Trauma surgeon.	CT	xxx	**0.12**	0.14	0.06	0.38	0.23	0.10	0.10	0.72	0.50
	MRI		xxx	0.34	0.62	0.29	0.40	0.13	0.49	0.26	0.26
Radiologist	CT			xxx	**0.42**	0.32	0.17	0.26	0.28	0.36	0.25
	MRI				xxx	0.32	0.44	0.17	0.54	0.31	0.32
Spine surgeon	CT					xxx	**0.67**	0.14	0.20	0.62	0.59
	MRI						xxx	0.11	0.26	0.48	0.55
Resident 1	CT							xxx	**0.14**	0.15	0.15
	MRI								xxx	0.22	0.09
Resident 2	CT									xxx	**0.77**
	MRI										xxx

Table 5 κ Values concerning the type A group distinction

		Trauma surgeon		Radiologist		Spine surgeon		Resident 1		Resident 2	
		CT	MRI	CT	MRI	CT	MRI	CT	MRI	CT	MRI
Trauma surgeon	CT	xxx	**0.70**	0.47	0.50	0.61	0.65	0.56	0.59	0.66	0.61
	MRI		xxx	0.59	0.71	0.77	0.68	0.54	0.65	0.78	0.78
Radiologist	CT			xxx	**0.82**	0.67	0.59	0.53	0.58	0.66	0.68
	MRI				xxx	0.80	0.71	0.56	0.62	0.79	0.81
Spine surgeon	CT					xxx	**0.77**	0.56	0.66	0.86	0.87
	MRI						xxx	0.58	0.66	0.75	0.66
Resident 1	CT							xxx	**0.61**	0.56	0.54
	MRI								xxx	0.58	0.58
Resident 2	CT									xxx	**0.91**
	MRI										xxx

readings. Agreement over the groups of type A reached a level of substantial agreement for both readings. There was also fair agreement in both readings of subgroups of type A. The intra-observer agreement between the CT and MRI readings of the five observers were higher for all items, but followed the same pattern.

In 30 fractures at least one of the observers reported a non-A fracture in one of his readings. For only one fracture was there a non-A categorization in all of the ten readings. For the CT readings, out of the 60 fractures reported by every observer, in 23 at least one of the observers reported a non-A fracture. Only in one case was there

Table 6 κ Values concerning the type A subgroup distinction

		Trauma surgeon		Radiologist		Spine surgeon		Resident 1		Resident 2	
		CT	MRI	CT	MRI	CT	MRI	CT	MRI	CT	MRI
Trauma surgeon	CT	xxx	**0.45**	0.36	0.33	0.38	0.23	0.41	0.29	0.51	0.36
	MRI		xxx	0.28	0.30	0.42	0.43	0.30	0.41	0.39	0.31
Radiologist	CT			xxx	**0.64**	0.46	0.33	0.18	0.30	0.43	0.42
	MRI				xxx	0.35	0.29	0.16	0.24	0.42	0.52
Spine surgeon	CT					xxx	**0.33**	0.32	0.39	0.50	0.46
	MRI						xxx	0.33	0.30	0.32	0.27
Resident 1	CT							xxx	**0.39**	0.28	0.22
	MRI								xxx	0.22	0.21
Resident 2	CT									xxx	**0.88**
	MRI										xxx

Table 7 κ Values concerning the all three levels of the classification scheme

		Trauma surgeon		Radiologist		Spine surgeon		Resident 1		Resident 2	
		CT	MRI	CT	MRI	CT	MRI	CT	MRI	CT	MRI
Trauma surgeon	CT	xxx	**0.33**	0.28	0.24	0.29	0.16	0.32	0.21	0.50	0.32
	RI		xxx	0.26	0.32	0.34	0.36	0.25	0.36	0.31	0.28
Radiologist	CT			xxx	**0.57**	0.36	0.25	0.16	0.28	0.34	0.33
	MRI				xxx	0.27	0.28	0.14	0.25	0.30	0.41
Spine surgeon	CT					xxx	**0.33**	0.26	0.31	0.50	0.44
	MRI						xxx	0.22	0.24	0.26	0.26
Resident 1	CT							xxx	**0.35**	0.23	0.18
	MRI								xxx	0.18	0.17
Resident 2	CT									xxx	**0.79**
	MRI										xxx

Table 8 Summary of the ranges of κ values. Generalized κ values for inter-observer and mean κ values for the intra-observer measurements are shown in brackets

	Inter-observer CT	Inter-observer MRI	Intra-observer CT/MRI
A/non-A	0.08–0.71 (0.34)	0.15–0.63 (0.42)	0.11–0.76 (0.45)
Type	0.10–0.72 (0.35)	0.09–0.62 (0.39)	0.12–0.77 (0.41)
A group	0.47–0.86 (0.61)	0.58–0.81 (0.73)	0.61–0.91 (0.76)
A subgroup	0.18–0.51 (0.37)	0.21–0.52 (0.34)	0.33–0.88 (0.54)
Complete	0.16–0.50 (0.31)	0.17–0.41 (0.28)	0.33–0.79 (0.47)

agreement among all observers on the non-A classification. For the MRI readings, 63 fractures were reported by all observers. In 26 of these at least one of the observers reported a non-A fracture. In two cases there was agreement between all observers on non-A categorization.

κ Values for intra-observer agreement of the first and second CT and MRI of the three observers are shown in Table 9. These values were, as expected, higher than the inter-observer agreements.

Concerning the Denis classification, the κ values are shown in Tables 10 and 11. With the CT evaluation at the Denis type level, the highest agreement was achieved between the two residents, and the lowest between the spine surgeon and resident 1. The same pattern was observed for the whole classification. The highest agreement at both levels with the MRI evaluation also was achieved between the two residents, and the lowest between the spine surgeon and resident 1. The highest agreement was achieved between the CT and MRI readings by resident 1, and the lowest by the spine surgeon. The common observation of all three observers was that many MRI, but also CT, reconstruction findings could not be incorporated into the Denis scheme. It was not possible to make a clear distinction between the compression-type fractures with or without posterior column soft-tissue involvement. The same was true of the burst-type fractures. In many cases, the fractures would be classified as Denis type 1 (compression) or 2 (burst) on the basis of radiograms and transverse CT im-

Table 9 κ Values of intra-observer agreement of the three observers between the first and second CT and MRI readings

	A/non-A		Type		A group		A subgroup		Complete	
	CT	MRI	CT	MRI	CT	MRI	CT	MRI	CT	MRI
Spine surgeon	0.65	0.80	0.56	0.41	0.76	0.95	0.35	0.60	0.35	0.61
Resident 1	0.29	0.30	0.28	0.32	0.62	0.77	0.40	0.40	0.60	0.35
Resident 2	0.72	0.70	0.66	0.65	0.90	0.95	0.88	0.90	0.77	0.72

Table 10 κ Values concerning the Denis classification

	CT type	CT whole	MRI type	MRI whole
Spine surgeon – resident 1	0.44	0.28	0.37	0.28
Spine surgeon – resident 2	0.63	0.39	0.55	0.37
Resident 1 – resident 2	0.72	0.67	0.65	0.53
Mean	0.60	0.45	0.52	0.39

Table 11 κ Values concerning the intra-observer between CT-MRI readings of the Denis classification

	Type	Whole
Spine surgeon	0.66	0.55
Resident 1	0.93	0.91
Resident 2	0.88	0.87

ages. But where CT–MPR or MRI findings suggest posterior-column involvement, problems arise because these injury patterns do not correspond with the Denis type 3 (seatbelt injuries) or Denis type 4 (fracture dislocations), but are also not the same type of injuries as simple compression or burst fractures. The spine surgeon particularly tended to categorize these injuries under Denis type-3, while the residents classified them as Denis type 1 or 2, which explains the higher agreement between the two residents.

Discussion

Fracture classification systems are useful conceptual tools for understanding the basic mechanisms involved. A classification system is based upon a presumption about an underlying common characteristic of the subsets of a domain. In the case of a fracture classification system, this is based upon the presumption that the interaction of various forces with the parts of a living organism involved create some basic observable patterns. The main difficulty of all fracture classification schemes lies in the innumerable variables involved in a traumatic lesion. The classification has to presuppose an "all or none" result of some of the interactions. A classification scheme tries to compress the available information into reproducible categories without loss of information content. It is inevitable that two kinds of problems arise with categorization schemes. Either there is a loss of information content in favor of simplicity and thus higher reproducibility; or loss of simplicity and reproducibility in favor of higher fidelity to the information content. Changes in the information content, for example as a result of novel technology, may have different effects on these two strategies.

The Magerl (AO) scheme seems to choose fidelity to the information content, by providing categories for all kinds of possible injury patterns. This leads, inevitably, to an increase in the complexity of the scheme, but also provides means for classification in accordance with increasing information content following novel technology. Although the AO scheme recognizes the difference in injuries with or without transverse plane soft-tissue involvement, the means to make this distinction reliably have not been sufficiently explored. The classification presupposes that the posterior ligamentary complex is either injured or not, although the authors recognize that transient forms do exist. However mechanically sound [17] this distinction may be, in reality we observed varying *degrees* of involvement of the posterior ligamentary complex, corresponding to the transient forms mentioned by the designers of the scheme. We observed different changes in the posterior ligamentary complex, varying from slight edema to complete ruptures (Figs 3, 4, 5, 6). Our operative findings

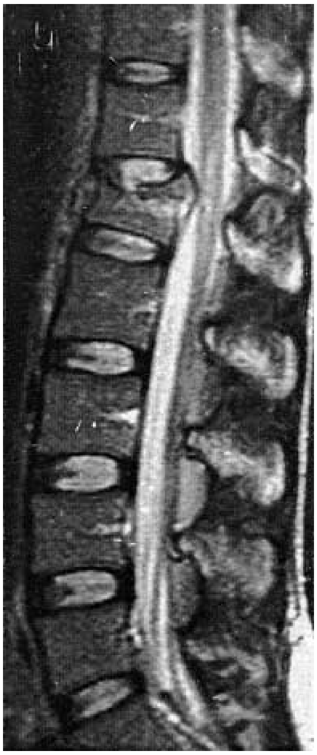

Fig. 3 MRI of a fracture that was classified as type A by all observers on both CT and MRI readings

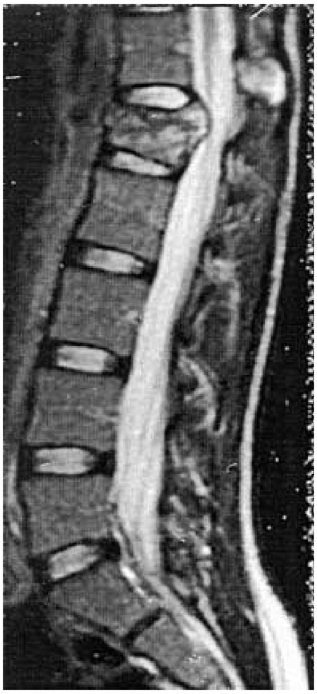

Fig. 4 This fracture is classified as type A by all observers on CT reading, and as type B by two observers on MRI reading

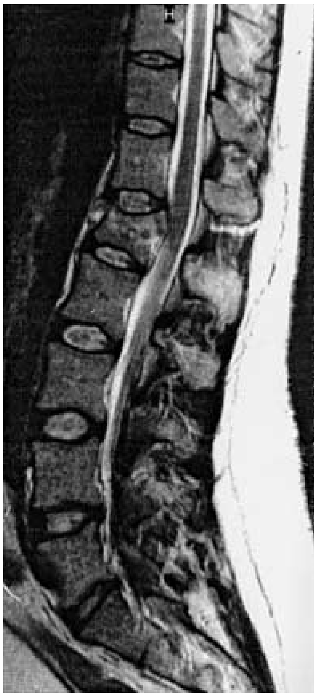

Fig. 6 One observer classified this fracture as type B on CT reading, and two observers on MR reading. Both of the residents classified this fracture as "burst" according to Denis, while the spine surgeon classified it as "seat-belt type"

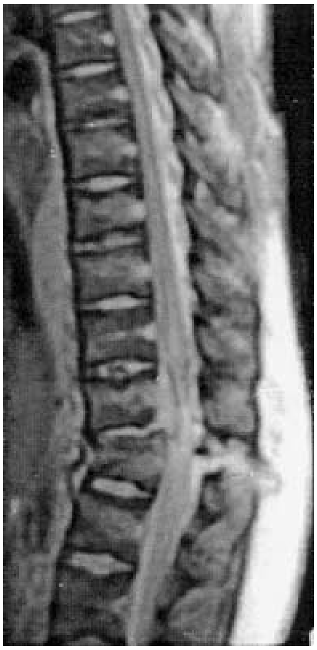

Fig. 5 This fracture is classified as type B by three observers on CT reading and by all observers on MRI reading

were consistent with MRI as reported in other studies [23, 24, 30].

Others have also observed that MRIs of fractures classified as compression fractures show signs of posterior column involvement in almost 50% of cases [23, 24, 25, 30]. In an experimental study it has been shown that MRI is capable of detecting ligamentary injuries associated with a fracture [15]. However, attempts in another study to redefine the AO classification based on MRI led to difficulties because it was not clear which kind of soft-tissue involvement should be considered indicative of non-A injury [23].

Our study group is not an unselected population, because of the fact that a number of patients with probably mainly type B and C patterns were excluded due to the difficulties of advanced imaging within a week after trauma, or before intervention. Inclusion of these patients would possibly result in higher κ values in the distinction between A/non-A. But this does not explain the fact that, in almost half of the fractures detected by every observer on both readings, at least one observer, at least once, doubted this major distinction. The designers of the scheme recognize these difficulties when they state, "it is quite natural that injuries occur which constitute transient forms between types... (a) type A injury can become type B when the degree of flexion exceeds the point beyond which the posterior ligament complex definitely fails" [19]. There is, however, no clue about how to define a "definitive failure" of this complex. The designers' solution to this problem is "transient forms may either be allocated to the lesser or more severe category, depending on which characteristics predominate." Although this might be the best strategy for an individual surgeon to decide over the treatment modalities, this ambiguity renders the scheme less reliable for comparison of patient populations from different locations. The designers also recognize that " some type B injuries ... were missed and classified as type A injuries when only standard radiographs are available" [19].

Our findings concur with the results of a study reported by Blauth et al. in which the radiograms and CTs of 14 cases were classified by 22 clinics specialized in the treatment of spinal injuries [4]. In this study also there was a high agreement over simple A3 cases, but a high disagreement in cases of complex injuries. Blauth et al. also point to the difficulties with the definition of posterior injury, and recommend the use of MRI.

In a multi-layer classification scheme it is expected that the agreement rates decrease in subsequent levels, as observed for the AO classification of peripheral fractures [1, 2, 3, 5, 10, 11, 13, 14, 20, 21, 27, 29,31]. This does not seem to be the case in the classification scheme we studied. The agreement on the group classification of the common type A fractures was higher than the agreement on type categorization or A/non-A distinction. Subgroup classification, however, dropped to lower values, as expected. This is another indication that the type categorization of the scheme is problematic. Although the inter-observer agreement on type A/non-type A distinction was higher with MRI readings in our study, it reached only moderate levels. Inclusion of MRI as a diagnostic tool, as proposed by Blauth et al. [4], may thus enhance the depiction of ligament injuries. However, MRI findings should first be described in a reproducible manner and should be integrated into the scheme [23]. The κ values obtained with CT and MRI were comparable for all other parameters. Considering the potential of MRI to provide a better agreement on A/non-A distinction, and no further advantage of CT, we conclude that, as far as this classification is concerned, MRI can replace CT as the diagnostic tool of choice for thoracic and lumbar spine fractures, as suggested by others [25].

Although fracture classification systems of the peripheral skeleton based on morphological appearances provide valuable information on the severity of the injury, one should be careful in the application of this principle to the spine. A fracture with exactly the same morphological appearance would have different mechanical consequences in the thoracic spine, thoracolumbar junction or lumbar spine. The level of the injury should always be included in the scheme.

In its present form the type categorization of the AO scheme is not sufficiently reproducible to be used for comparison of different patient series. The inter-observer and intra-observer agreement on group and subgroup levels of the common type A fractures are comparable with reports in the literature of some common peripheral fractures. The highest agreement was achieved on the type A group classification. This is practically the same as the distinction between the "wedge-compression" and "burst" fractures of the Denis scheme [9], which is why we decided to test the Denis scheme with the same available information.

The Denis classification was based on the, for that time, novel technology of transverse CT images and represents a strategy of simplification. Refinement of the imaging technology, in the form of CT-MPR and MRI proved that much of the information from these new modalities is difficult to integrate into this scheme. Much of the ligamentary involvement of the posterior column cannot be accounted for in the scheme. As a result, either injury patterns with or without posterior ligamentary complex involvement are grouped together into categories based upon the patterns of bony involvement, or injury patterns with posterior ligamentary complex involvement are assigned to higher categories, constituting an over-estimation of the severity. In our study, the more experienced observer more often assigned these injuries to higher categories, leading to marked variance in the results. This type of confusion may have contributed in the past to the widely different results of conservative treatment strategies reported in the literature.

Although intuitively one would think that inter-observer agreement between experienced observers and intra-observer agreement of more experienced observers would be better, earlier studies showed that this is not the case [10, 20, 27,31]. In our study, the highest inter-observer agreement was between the orthopedic spine surgeon and one of the residents. One of the residents achieved also the highest intra-observer consistency.

A fundamental discussion about parameter definition in clinical orthopedic research is necessary. κ Values for inter-observer and intra-observer agreements varying from 0.38 to 0.77 have been reported for peripheral fracture classification systems. There is no consensus about the level of κ values that should be considered acceptable for fracture-classification systems [20]. We used the distinction proposed by Landis and Koch [16], as did many of the other studies. However, in an editorial, Sanders has suggested that fracture classification systems should have an inter-observer reproducibility level exceeding a κ value of 0.55 [26]. This would introduce very stringent criteria, which is probably not achievable in traumatology. From a skeptical point of view it can be argued that fractures are not reliably classifiable in any meaningful way. It can also be argued, though, that any degree of agreement higher than chance distribution can forward our common understanding of the patterns involved. "So is 1 per cent vision better than total blindness" [8]. In that case, fracture classification schemes should be seen as evolvable entities of pattern recognition, which should be subject to a continuous process of assessment, reassessment, and refinement. A potentially serious complication of such an evolving process, however, can occur in the increasingly popular instrument in clinical orthopedic research: the meta-analysis. For the sake of a possible future meta-analysis, authors are asked to convey their data according to schemes accepted in the literature. However, without a proper appreciation of the inherent uncertainties of these schemes, there is the danger of these meta-analyses leading to meta-errors.

In conclusion, we recommend the use of the Magerl (AO) classification because it allows categorization of injuries to all relevant parts of the spine. However, the classification scheme should be revised based on MRI, so that it will be clear which kinds of soft-tissue injuries should be considered indicative of various types and subgroups.

Acknowledgements The authors thank Dr Ir J.A.J. Faber of the Centrum voor Biostatistiek at the University of Utrecht, for his kind assistance in performing the statistical analysis.

References

1. Andersen DJ, Blair WF, Steyers CM JR, Adams BD, el-Khouri GY, Brandser EA (1996) Classification of distal radius fractures: an analysis of interobserver reliability and intraobserver reproducibility. J Hand Surg Am 21:574–582
2. Andersen E, Jorgensen LG, Hededam LT (1990) Evans' classification of trochanteric fractures: an assessment of the interobserver and intraobserver reliability. Injury 21:377–378
3. Andersen GR, Rasmussen JB, Dahl B, Solgaard S (1991) Older's classification of Colles' fractures. Good intraobserver and interobserver reproducibility in 185 cases. Acta Orthop Scand 62: 463–464
4. Blauth M, Bastian L, Knop C, Lange U, Tusch G. (1999) Interobserverreliabilität bei der Klassifikation von thorakolumbalen Wirbelsäulenverletzungen. Orthopäde 28:662–81
5. Böhler L (1943) Die Technik der Knochenbruchbehandelung. 9.-11 Auflage. Verlag von Wilhelm Maudrich, Wien
6. Brumback RJ, Jones AL (1994) Interobserver agreement in the classification of open fractures of the tibia. J Bone Jt Surg Br Vol 76-A:1162–1165
7. Burstein AH (1993) Fracture classification systems: do they work and are they useful? J Bone Jt Surg Am Vol 75-A:1743–1744
8. Dawkins R (1991) The blind watchmaker. Penguin Books, Middlesex, England, p. 81
9. Denis F (1983) The three column spine and its significance in the classification of acute thoracolumbar spinal injuries. Spine 8:817–831
10. Dirschl DR, Adams GL (1997) A critical assessment of factors influencing reliability in the classification of fractures, using fractures of the tibial plafond as a model. J Orth Trauma 11:471–476
11. Gehrchen PM, Nielsen JO, Olesen B (1993) Poor reproducibility of Evans' classification of the trochanteric fracture. Assessment of 4 observers in 52 cases. Acta Orthop Scand 64:71–72
12. Holdsworth FW (1963) Fractures, dislocations and fracture-dislocations of the spine. J Bone Jt Surg Br Vol 45B:6–20
13. Horn BD, Rettig ME (1993) Interobserver reliability in the Gustilo and Anderson classification of open fractures. J Orthop Trauma 7:357–360
14. Johnstone DJ, Radford WJ, Parnell EJ (1993) Interobserver variation using the AO/ASIF classification of long bone fractures. Injury 24:163–165
15. Kliewer MA, Gray L, Paver J, Richardson WD, Vogler JB, McElhaney JH, Myers BS (1993) Acute spinal ligament disruption: MR imaging with anatomic correlation. J Magn Reson Imaging 3:855–861
16. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33: 159–174
17. Lange U, Bastian L, Knop C, Blauth M (1998) Stabilitätsverhalten verschiedener verletzungen der thoracolumbalen wirbelsäule. Eine biomechanische studie (abstract). Ned Tijd Traumatol 110 Suppl:110
18. Louis R (1977) Les theories de l'instabilite. Rev Chir Orthop 63:423–425
19. Magerl F, Aebi M, Gertzbein SD, Harms J, Nazarian S (1994) A comprehensive classification of thoracic and lumbar injuries. Eur Spine J 3:184–201
20. Martin JS, Marsh JL, Bonar SK, DeCoster TA, Found EM, Brandser EA (1997) Assesment of the AO/ASIF fracture classification for the distal tibia. J Orthop Trauma 11:477–483
21. Nielsen JO, Dons-Jensen H, Sorensen HT (1990) Lauge-Hansen classification of malleolar fractures. An assessment of the reproducibility in 118 cases. Acta Orthop Scand 61:385–387
22. Oner FC, vd Rijt RHH, Ramos LMP, Groen GJ, Dhert WJA, Verbout AJ (1999) Correlation of MR images of disc injuries with anatomic sections in experimental thoracolumbar spine fractures. Eur Spine J 8:194–198
23. Oner FC, van Gils AP, Dhert WJ, Verbout AJ (1999) MRI findings of thoracolumbar spine fractures: a categorisation based on MRI examinations of 100 fractures. Skeletal Radiol 28:433–43
24. Petersilge CA, Pathria MN, Emery SE, Masaryk TJ (1995) Thoracolumbar burst fractures: Evaluation with MR imaging. Radiology 194:49–54
25. Saifuddin A, Noordeen H, Taylor BA, Bayley I (1996) The role of imaging in the diagnosis and management of thoracolumbar burst fractures: current concepts and a review of the literature. Skeletal Radiol 25:603–613
26. Sanders RW (1997) Editorial. The problem with apples and oranges. J Orthop Trauma 465–466
27. Sidor ML, Zuckerman JD, Lyon T, Koval K, Cuomo F, Schoenberg N (1993) The Neer classification system for proximal humeral fractures. An assessment of interobserver reliability and intraobserver reproducibility. J Bone Jt Surg Am Vol 75-A:1745–1750
28. Siebenrock KA, Gerber C (1993) The reproducibility of classification of fractures of the proximal end of the humerus. J Bone Jt Surg Am Vol 75-A:1751–1755
29. Swiontkowski MF, Sands AK, Agel J, Diab M, Schwappach JR, Kreder HJ (1997) Interobserver variation in the AO/OTA fracture classification system for pilon fractures: Is there a problem? J Orthop Trauma 11:467–470
30. Terk MR, Hume-Neal M, Fraipont M, Ahmadi J, Colletti PM (1997): Injury of the posterior ligament complex in patients with acute spinal trauma: evaluation by MR imaging. Am J Roentgen 168:1481–1486
31. Thomsen NO, Overgaard S, Olsen LH, Hansen H, Nielsen ST (1991) Observer variation in the radiographic classification of ankle fractures. J Bone Jt Surg Br Vol 73-B:676–678