ORIGINAL ARTICLE

Hans-Joachim Wilke
Friederike Rohlmann
Cornelia Neidlinger-Wilke
Karin Werner
Lutz Claes
Annette Kettler

Validity and interobserver agreement of a new radiographic grading system for intervertebral disc degeneration: Part I. Lumbar spine

Abstract Many different radiographic grading systems for disc degeneration are described in literature. However, only a few of them are tested for interobserver agreement and none for validity. Furthermore, most of them are based on a subjective terminology. The aim of this study, therefore, is to combine these systems to a new one in which all subjective terms are replaced by more objective ones and to test this new system for validity and interobserver agreement. Since lumbar and cervical discs need to be graded differently, this study was divided into the present *Part I* for the lumbar and a *Part II* for the cervical spine. The new radiographic grading system covers the three variables "Height Loss", "Osteophyte Formation" and "Diffuse Sclerosis". On lateral and postero-anterior radiographs, each of these three variables first has to be graded individually. Then, the "Overall Degree of Degeneration" is assigned on a four-point scale from 0 (no degeneration) to 3 (severe degeneration). For validation, the radiographic degrees of degeneration of 44 lumbar discs were compared to the respective macroscopic ones, which were defined as "real" degrees of degeneration. The agreement between observers with different levels of experience was determined using the radiographs of 84 lumbar discs. Agreement was quantified using quadratic weighted Kappa coefficients (Kappa) with 95% confidence limits (95% CL). The validation of the new radiographic grading system revealed a substantial agreement between the radiographic and the "real" macroscopic overall degree of degeneration (Kappa = 0.714, 95% CL: 0.587–0.841). The radiographic grades, however, tended to be slightly lower than the "real" ones. The interobserver agreement was substantial for all the three variables and for the overall degree of degeneration (Kappa = 0.787, 95% CL: 0.702–0.872). However, the inexperienced observer tended to assign slightly lower degrees of degeneration than the experienced one. In conclusion, we believe that the new radiographic grading system is an almost objective, valid and reliable tool to quantify the degree of degeneration of individual lumbar intervertebral discs. However, the user should always remember that the "real" degree of degeneration tends to be underestimated and that slight differences between the ratings of observers with different levels of experience have to be expected.

Keywords Classification ·
Degeneration · Lumbar
intervertebral disc · Reliability ·
Validity

H.-J. Wilke (✉) · F. Rohlmann
C. Neidlinger-Wilke · K. Werner
L. Claes · A. Kettler
Institute for Orthopaedic Research and
Biomechanics, University of Ulm,
Helmholtzstrasse 14, 89081 Ulm, Germany
E-mail: hans-joachim.wilke@uni-ulm.de

Introduction

The morphology of intervertebral disc degeneration has often been described in literature [1, 9–11, 14, 18, 28, 33, 34]. Especially for research purposes, however, these changes need to be quantified. Therefore, in the past, many different grading systems have been developed especially for the lumbar and also for the cervical spine [21]. Some of these systems can only be used in vitro such as macroscopic or histologic grading systems. In contrast, some others are also applicable in clinical practice such as those based on plain radiographs, discography, computed tomography or magnetic resonance imaging. Out of these, magnetic resonance imaging has become increasingly popular since the intervertebral disc itself can be visualised and the procedure itself is not invasive. Nevertheless, grading systems based on plain radiographs still have several advantages. First, in contrast to discograms, they are less invasive. Second, in contrast to magnetic resonance imaging and computed tomography, they only require a standard X-ray machine and are much cheaper. And third, plain radiographs are often taken for diagnostic or follow-up purposes and, thus, are often already available.

A grading system has to fulfil certain requirements to become a valuable tool. First of all, the ratings should be the same irrespective of the experience of the observer. And second, they should be valid. Thus, they should reflect the "real" degree of degeneration. However, out of all the nine radiographic grading systems for lumbar or cervical disc degeneration found in literature [6, 17, 19, 20, 23–25, 31], only three are tested for interobserver agreement [19, 23, 24] and none for validity. Furthermore, most of them are based on terms such as "mild", "severe", "small" or "large" [6, 17, 19, 20, 23]. Since such terms are not well defined and tend to be subjective, the interobserver agreement of the respective grading systems is expected to be worse if more subjective terms were used.

Therefore, the first aim of this study was to combine the existing radiographic grading systems to a new one, in which all subjective terms were replaced by more objective ones. The second aim was to test this new grading system for validity and agreement between experienced and unexperienced observers. Due to the uncinate processes and the smaller dimensions of the cervical spine, lumbar and cervical discs need to be graded in a different way. In order to prevent confusion, this study was therefore divided into the present *Part I* for the lumbar and a *Part II* for the cervical spine.

Materials and methods

The new grading system covers the three main radiographic signs of disc degeneration: "Height Loss", "Osteophyte Formation" and "Diffuse Sclerosis" (Table 1). On lateral and postero-anterior radiographs each of these three variables first has to be graded individually on a scale from 0 to 3. Based on the sum of these three scores, the "Overall Degree of Degeneration" is
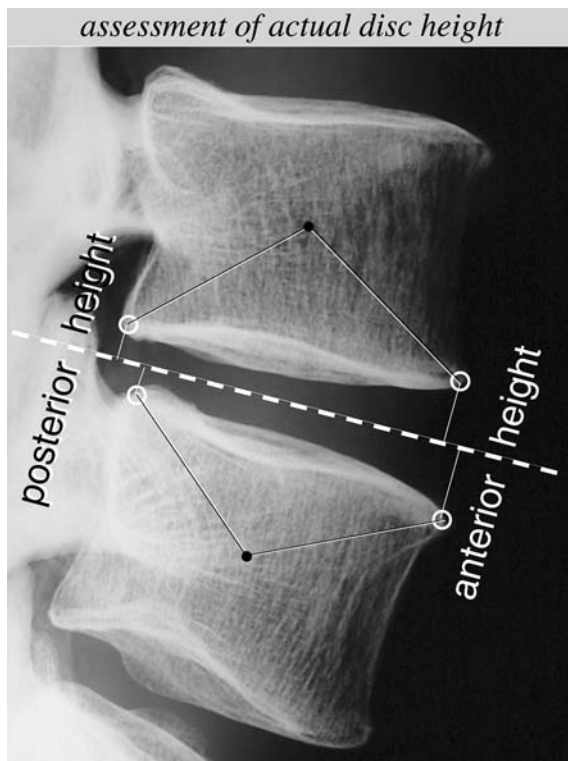
Table 1 New radiographic grading system for lumbar intervertebral disc degeneration modified according to the systems found in literature

Radiographic grading system for lumbar intervertebral disc degeneration (based on lateral and postero-anterior radiographs)

Height loss	Osteophyte formation	Diffuse sclerosis	Overall degree of degeneration
Anterior and posterior height loss with respect to the individual height before degeneration	Sum of points of eight edges No osteophytes: 0 points < 3 mm: 1 point ≥3 mm but <6 mm: 2 points ≥6 mm: 3 points	Sum of points of both adjacent vertebral bodies No sclerosis: 0 points 0.25 partially or completely affected: 1 point 0.5 partially or completely affected: 2 points >0.5 partially or completely affected: 3 points	Sum of points of "Height Loss", "Osteophyte Formation" and "Diffuse Sclerosis"
0 = 0% 1 = <33% 2 = ≥33 but <66% 3 = ≥66%	0 = 0 points 1 = 1–8 points 2 = 9–16 points 3 = 17–24 points	0 = 0 points 1 = 1–2 points 2 = 3–4 points 3 = 5–6 points	0 point = grade 0 (no degeneration) 1–3 points = grade 1 (mild degeneration) 4–6 points = grade 2 (moderate degeneration) 7–9 points = grade 3 (severe degeneration)

The three variables "Height Loss", "Osteophyte Formation" and "Diffuse Sclerosis" are first graded individually on a scale from 0 to 3. The "Overall Degree of Degeneration" is then assigned according to the sum of these three scores

assigned to each disc on a four-point scale from 0 (no degeneration) to 3 (severe degeneration).

"Height Loss" is defined as the average anterior and posterior (but not central) decrease in disc height referred to the respective height before degeneration. The anterior height before degeneration is estimated based on the normal values reported by Frobin et al. [15] (Fig. 1; Table 2). To account for interindividual differences, the ranges of normal disc height should be considered rather than their mean value. The posterior height before degeneration is estimated as being smaller or as high but not higher than the respective anterior height [12]. The central disc height was not included into the assessment of "Height Loss" since at this position the height increases in some cases of osteoporosis (fish-vertebra deformity).

Table 2 Normal values of anterior disc height normalised to the antero-posterior diameter of the cranial vertebral body (= 100%) (mean of male and female subjects according to Frobin et al. [15])

Normal values of anterior disc height (modified according to Frobin et al. [15])

	Mean (%)	Mean − 2 SD (%)	Mean + 2 SD (%)
T12-L1	24	18	30
L1-2	29	22	36
L2-3	33	26	41
L3-4	37	29	45
L4-5	42	32	51
L5-S1	41	31	51

Mean mean value, *SD* standard deviation

Fig. 1 To assess the degree of height loss, first, the actual disc height has to be determined. For this purpose, the anterior and posterior edges of the adjacent vertebral bodies (*small white circles*) are defined as those points having the largest distance to the centre of the vertebral body (*black points*). Then, the distance of each of these four edges to the midplane of the disc (*dashed line*) is measured. Finally, the sum of the two anterior distances is defined as actual anterior disc height, and the sum of the two posterior distances is defined as actual posterior disc height. This procedure is meant to support the estimation of actual disc height, but does not have to be carried out using drawings or digitisation. In a second step, this actual height is compared to the respective height before degeneration, which is estimated based on the normal values reported by Frobin et al. [15](Table 2)

The variable "Osteophyte Formation" is assessed in terms of the number and length of osteophytes growing at the two anterior, two posterior, two left lateral and two right lateral edges of the adjacent vertebral bodies (Fig. 2).

The variable "Diffuse Sclerosis" is graded in terms of the number of predefined regions that are affected by sclerosis (Fig. 3). A thickening of the bony endplates should not be counted if it is not diffuse.

To validate the new radiographic grading system, first, the radiographic degrees of degeneration of 44 intervertebral discs from 16 fresh frozen mono or polysegmental human osteoligamentous lumbar spine specimens were determined. The age of the donors ranged between 16 and 91 years (mean 66 years) and none of them had a known history of trauma or spinal disease. These radiographic degrees of degeneration were then compared to the respective macroscopic ones, which were defined as "real" degrees of degeneration. For this purpose, the specimens were first x-rayed in the lateral and postero-anterior direction (43805 X-Ray System, Faxitron Series, Hewlett Packard, USA; film to source distance 61 cm) using a tube voltage of 45–50 kV and an exposure time of 5 min. Then, still being frozen, they were cut in the mid-sagittal plane. The cutting surfaces were photographed and stored for evaluation. To be able to directly compare the radiographic with the macroscopic degrees of degeneration, the macroscopic grading system also covered the three variables "Height Loss", "Osteophyte Formation" and "Diffuse Sclerosis" (Tables 1 and 3). However, macroscopically, the three variables "Nucleus Pulposus", "Annulus Fibrosus" and "Endplate Cartilage" were added to reflect the "real" degree of degeneration as closely as possible (modified according to Thompson et al. [32]).
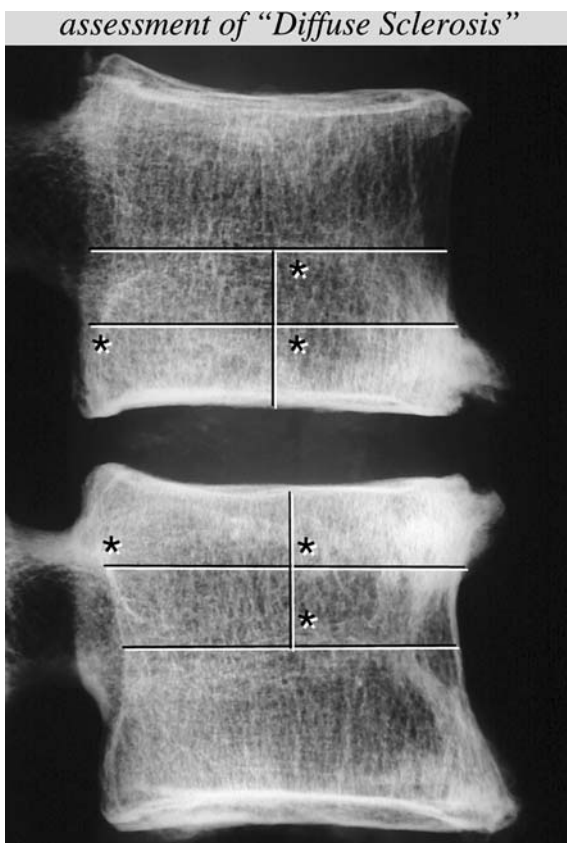
Using this modified macroscopic grading system, the "real" degree of degeneration of the 44 lumbar discs was determined by two observers independently. Both of them were familiar with disc degeneration and had

Fig. 2 To assess the variable "Ostophyte Formation", the two anterior (e1, e2), two posterior (e3, e4), two right lateral (e5, e6) and two left lateral edges (e7, e8) of the adjacent vertebral bodies are screened for osteophytes. Their number is counted and their length is measured along their long axis beginning at the former border of the vertebral body and ending at their tips (*white lines* in the edges e1, e2, e5, e6, e7 and e8)
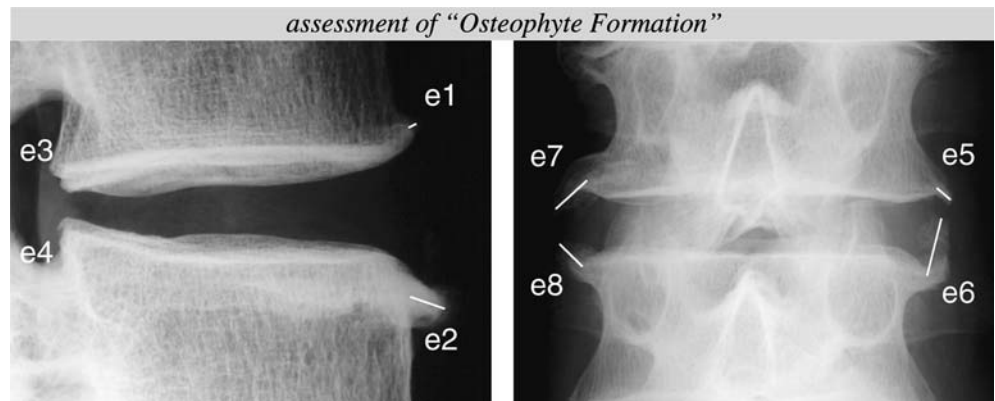
several years of experience in spinal research. The "real" degree of degeneration was then defined as the mean value of the results of both the observers. Then, the 44 discs were additionally graded radiographically by one

Fig. 3 The variable "Diffuse Sclerosis" is assessed on the lateral radiographs only. The lower half of the upper vertebral body and the upper half of the lower vertebral body are each divided into four regions. Then, the number of regions is counted, which are covered by sclerosis. Note that a partially covered region is counted as if it was completely covered. In this example, the number of affected regions (*asterisk*) would be three for the upper and three for the lower vertebral body

of these two observers. In order to ensure that this observer was not biased by the evaluation of the macroscopic slices carried out a few days before, the radiographs were blinded and put in a randomised order. The postero-anterior radiographs of four discs could not be evaluated due to poor quality. These four discs could therefore only be included for the variables "Height Loss" and "Diffuse Sclerosis". The remaining 40 discs, however, could be evaluated completely. To statistically assess the agreement between the radiographic and the "real", macroscopic degree of degeneration weighted Kappa coefficients (quadratic weights) with 95% confidence limits (95% CL) were calculated according to Fleiss and Cohen [13] using the software SAS 8.2 [30]. These calculations were carried out under the assumption of independency of the observation of each intervertebral disc.

In order to show whether the grade assigned to a disc depends on the degree of experience of the observer, the agreement between one experienced and one inexperienced observer was determined. Both observers graded the lateral and postero-anterior radiographs of 27 osteoligamentous mono or polysegmental spine specimens with an overall of 84 lumbar intervertebral discs. The age of the donors ranged between 16 and 92 years (mean 67 years) and none of them had a known history of trauma or spinal disease. The experienced observer was the one who also evaluated the macroscopic slices and the radiographs for validation. In contrast, the inexperienced observer, being a mechanical engineer without any medical training, had no experience in reading radiographs and was not familiar with disc degeneration. However, he was trained before grading the discs: the grading system was explained using some training radiographs. Furthermore, the radiographic appearance of the most common spinal diseases such as osteoporosis, osteoporotic fractures, fish-vertebra deformities, spondylolysis, Bechterew's disease or spinal metastases was demonstrated in a 30 min session. Then, written instructions were handed over, in which the assessment of the three variables was explained again

Table 3 Macroscopic grading system for lumbar intervertebral disc degeneration used as the "gold standard" to define the "real" degree of degeneration (modified according to the systems found in literature)

Macroscopic grading system for lumbar intervertebral disc degeneration (based on mid-sagittal sections)

Nucleus pulposus	Annulus fibrosus	Endplate cartilage
Appearance	Appearance	Sum of points of both cartilaginous endplates
		Normal: 0 points
		Thickness irregular: 1 point
		Focal defect(s): 2 points
		(Almost) complete destruction: 3 points
0 = Bulging gel	0 = Discrete fibrous lamellas	0 = 0 points
1 = Fibrous tissue; loss of annular–nuclear demarcation	1 = Mucinous infiltration	1 = 1–2 points
2 = Focal clefts	2 = Focal disruptions	2 = 3–4 points
3 = Complete disruption of nucleus or complete transformation into tissue other than nucleus tissue	3 = Complete disruption of anterior plus posterior annulus or complete transformation into tissue other than annulus tissue	3 = 5–6 points

Height loss	Osteophyte formation	Diffuse sclerosis	Overall degree of degeneration
Anterior and posterior height loss with respect to the individual height before degeneration	Sum of points of four edges	Sum of points of both adjacent vertebral bodies	Sum of points of "Nucleus Pulposus", "Annulus Fibrosus", "Endplate Cartilage", "Height Loss", "Osteophyte Formation", and "Diffuse Sclerosis"
	No osteophytes: 0 points	No sclerosis: 0 points	
	<3 mm: 1 point	0.25 partially or completely affected: 1 point	
	≥3 mm but <6 mm: 2 points	0.5 partially or completely affected: 2 points	
	≥6 mm: 3 points	>0.5 partially or completely affected: 3 points	
0 = 0%	0 = 0 points	0 = 0 points	0 point = grade 0 (no degeneration)
1 = <33%	1 = 1–4 points	1 = 1–2 points	1–6 points = grade 1 (mild degeneration)
2 = ≥33 but <66%	2 = 5–8 points	2 = 3–4 points	7–12 points = grade 2 (moderate degeneration)
3 = ≥66%	3 = 9–12 points	3 = 5–6 points	13–18 points = grade 3 (severe degeneration)

The six variables "Nucleus Pulposus", "Annulus Fibrosus", "Endplate Cartilage", "Height Loss", "Osteophyte Formation" and "Diffuse Sclerosis" are first graded individually on a scale from 0 to 3. The "Overall Degree of Degeneration" is then assigned according to the sum of these six scores. Note that the macroscopic and the radiographic grading system are almost identical for the three variables, namely, "Height Loss", "Osteophyte Formation" and "Diffuse Sclerosis"

and the normal values of anterior lumbar disc height were listed similar to the Figs. 1, 2, 3 and to Table 2. Besides these instructions, the inexperienced observer did not get any further help during grading.

Statistically, the agreement between the ratings of the experienced and the inexperienced observer was evaluated using the same type of weighted Kappa coefficient as for validation [13]. For both, the validation and the assessment of the interobserver agreement, a Kappa of < 0.00 was interpreted as poor agreement, 0.00–0.20 as slight agreement, 0.21–0.40 as fair agreement, 0.41–0.60 as moderate agreement, 0.61–0.80 as substantial agreement and > 0.81 as almost perfect agreement [22].

Results

The agreement between the macroscopic ratings of the two experienced observers was almost perfect (Kappa between 0.874 and 0.920) for the overall degree of degeneration and the variables "Height Loss", "Nucleus Pulposus", "Annulus Fibrosus" and "Endplate Cartilage" (Table 4). For the variables "Osteophyte Formation" and "Diffuse Sclerosis" the agreement was somewhat lower, but still substantial (Kappa 0.675, respectively 0.707). These good agreements would almost have allowed to define the rating of only one observer as a "real" degree of degeneration. To further increase objectivity, however, the average ratings of both were used instead.

The validation of the radiographic grading system revealed an almost perfect agreement with the macroscopic, "real" degree of degeneration for the variable "Height Loss" (Kappa 0.862) and a slightly lower but still substantial agreement for "Osteophyte Formation" (Kappa 0.613) (Table 5). For the overall degree of degeneration the agreement also was substantial, the radiographic grades, however, tended to be lower than

Table 4 Agreement between the macroscopic ratings of the two experienced observers (weighted Kappa coefficients with 95% CL)

Interobserver agreement macroscopy (n = 44 lumbar intervertebral discs)

	Kappa	95% CL	
		Lower	Upper
Height loss	0.905	0.848	0.962
Osteophyte formation	0.675	0.537	0.814
Diffuse sclerosis	0.707	0.504	0.910
Nucleus pulposus	0.879	0.796	0.963
Annulus fibrosus	0.874	0.787	0.961
Endplate cartilage	0.913	0.852	0.973
Overall grade	0.920	0.865	0.975

CL confidence limits

Table 5 Agreement between the radiographic and the macroscopic "real" degrees of degeneration of 40 and 44 lumbar intervertebral discs, respectively (weighted Kappa coefficients with 95% CL)

Agreement between radiography and macroscopy (n = 44; n = 40[a] lumbar intervertebral discs)

	Kappa	95% CL	
		Lower	Upper
Height loss	0.862	0.783	0.941
Osteophyte formation	0.613[a]	0.463[a]	0.763[a]
Diffuse sclerosis	0.343	0.121	0.565
Overall grade[b]	0.714[a]	0.587[a]	0.841[a]

CL confidence limits
[b]Note that the overall degree of degeneration covers only three variables in the radiographic grading system ("Height Loss", "Osteophyte Formation" and "Diffuse Sclerosis"), but six variables in the macroscopic grading system (additionally "Nucleus Pulposus", "Annulus Fibrosus" and "Endplate Cartilage")

the macroscopic ones: in 20 out of 40 discs the "real" overall degree of degeneration was underestimated, but in only three it was overestimated (Fig. 4). As to the variable "Diffuse Sclerosis", Kappa was 0.343 reflecting an only fair agreement. In this case, much fewer sclerotic areas were detected radiographically than macroscopically.

The agreement between the radiographic ratings of the experienced and the inexperienced observer was substantial (Kappa between 0.681 and 0.798) for all the three variables as well as for the overall degree of degeneration (Table 6). However, the inexperienced observer generally tended to assign lower degrees of degeneration than the experienced one (Fig. 5). For example, concerning the overall degree of degeneration, 15 discs were rated 1° lower by the inexperienced observer but only one disc was rated 1° higher. Nevertheless, most ratings were identical: the same degree of "Height Loss" was assigned by both observers to 65% of all the discs, the same degree of "Osteophyte Formation" to 79%, the same degree of "Diffuse Sclerosis" to 80% and the same overall degree of degeneration to 81% of all the discs. The differences in grade assignment were never higher than 1° except for one disc concerning the variable "Osteophyte Formation" and two discs concerning the variable "Diffuse Sclerosis", where the difference was 2°. Differences of more than 2° did not occur.

Discussion

In this study, the radiographic grading systems for lumbar intervertebral disc degeneration available from literature were combined to a new one, in which undefined and subjective terms were replaced by better

Fig. 4 Agreement between the radiographic and the macroscopic "real" degree of degeneration of 40 and 44 lumbar intervertebral discs, respectively. Each field contains the number of discs rated with 0, 1, 2 or 3 points radiographically (rating of one experienced observer) and with 0, 0.5, 1, 1.5, 2, 2.5 or 3 points macroscopically (mean value of the ratings of two experienced observers)

HEIGHT LOSS — radiography

pts.:	0	1	2	3	total
0	7	3	0	0	10
0.5	2	3	0	0	5
1	0	5	0	0	5
1.5	0	4	0	0	4
2	0	4	4	0	8
2.5	0	1	1	0	2
3	0	0	2	8	10
total	9	20	7	8	44

(macroscopy)

OSTEOPHYTE FORMATION — radiography

pts.:	0	1	2	3	total
0	3	4	0	0	7
0.5	2	6	0	0	8
1	1	6	6	0	13
1.5	0	1	5	0	6
2	0	2	1	1	4
2.5	0	0	1	1	2
3	0	0	0	0	0
total	6	19	13	2	40

(macroscopy)

DIFFUSE SCLEROSIS — radiography

pts.:	0	1	2	3	total
0	14	0	0	0	14
0.5	2	0	1	0	3
1	2	2	1	0	5
1.5	8	0	0	1	9
2	4	0	0	0	4
2.5	2	0	1	0	3
3	1	2	1	2	6
total	33	4	4	3	44

(macroscopy)

OVERALL GRADE — radiography

pts.:	0	1	2	3	total
0	2	2	0	0	4
0.5	1	1	0	0	2
1	0	9	0	0	9
1.5	0	2	0	0	2
2	0	8	2	0	10
2.5	0	1	3	0	4
3	0	0	5	4	9
total	3	23	10	4	40

(macroscopy)

defined and more objective ones. Finally, similar to the grading system of Mimura et al. [25], the height loss of the disc was estimated as the percentage decrease in height referred to the height before degeneration. Osteophytes were assessed in terms of their number and absolute length and the degree of sclerosis was determined according to the number of predefined areas that were affected.

Despite these attempts to create a more objective grading system than those known from literature, a certain degree of subjectivity still remained. In the assessment of the variable "Height Loss", for example, the initial disc height still needs to be estimated. In vivo, this estimation becomes even more difficult due to the diurnal changes in disc height [2, 5, 29]. But even in vitro, its assessment is difficult due to the large spread of normal values [15]. Thus, a wide variety of different estimations are possible. Therefore, it is not surprising that the ratings of the two observers were not equal: for the inexperienced observer the height loss of the discs often seemed to be less severe than for the experienced one.

In contrast, the variable "Osteophyte Formation" could be defined much more objectively . Nevertheless, the inexperienced observer tended to see fewer osteophytes, thus, for example, the inexperienced observer tended to define pointed edges as normal, whereas the experienced one tended to define them as osteophytes. Thus, even though the terms used in the new grading system are more objective than those used in the systems known from literature, interobserver differences still have to be expected. The tendencies seen in this study, however, should not be generalised to all experienced and inexperienced observers since the ratings of only one experienced and one inexperienced observer were compared with each other and also since the quality of the radiographs is not always and everywhere the same. For example, the alignment of the patient during X-raying may be more difficult than the alignment of a spine specimen.

However, despite these tendencies, the differences between the ratings of the two observers were only little: for the three variables and the overall degree of degeneration they did not differ by more than 1° in all except for three cases where the difference was 2°. Furthermore, the interobserver agreement was substantial with Kappa

Table 6 Agreement between the radiographic ratings of one experienced and one inexperienced observer (weighted Kappa coefficients with 95% CL)

Interobserver agreement radiography ($n = 84$ lumbar intervertebral discs)

	Kappa	95% CL	
		Lower	Upper
Height loss	0.798	0.713	0.884
Osteophyte formation	0.687	0.559	0.814
Diffuse sclerosis	0.681	0.490	0.872
Overall grade	0.787	0.702	0.872

CL confidence limits

Fig. 5 Agreement between the radiographic ratings of one experienced and one inexperienced observer. Each field contains the number of lumbar intervertebral discs rated with the respective scores

HEIGHT LOSS

experienced obs.

pts.:	unexperienced obs.				total
	0	1	2	3	
0	35	0	0	0	35
1	19	15	0	0	34
2	0	6	0	1	7
3	0	0	3	5	8
total	54	21	3	6	84

OSTEOPHYTE FORMATION

experienced obs.

pts.:	unexperienced obs.				total
	0	1	2	3	
0	18	0	0	0	18
1	4	45	1	0	50
2	0	11	3	0	14
3	0	1	1	0	2
total	22	57	5	0	84

DIFFUSE SCLEROSIS

experienced obs.

pts.:	unexperienced obs.				total
	0	1	2	3	
0	65	6	1	0	72
1	4	0	1	0	5
2	1	1	2	0	4
3	0	0	3	0	3
total	70	7	7	0	84

OVERALL GRADE

experienced obs.

pts.:	unexperienced obs.				total
	0	1	2	3	
0	14	0	0	0	14
1	7	47	1	0	55
2	0	4	7	0	11
3	0	0	4	0	4
total	21	51	12	0	84

coefficients of 0.798 for "Height Loss", 0.687 for "Osteophyte Formation", 0.681 for "Diffuse Sclerosis" and 0.787 for the overall degree of degeneration. Even though, these values reflect the agreement between the two observers with different levels of experience, they were not much lower than those reported by Lane et al. [23] for three observers with similar experience. The Kappa coefficients of Lane et al. were 0.95 for "Narrowing", 0.91 for "Osteophytes" and 0.93 for the "Summary Grades". For "Sclerosis", however, Lane et al. reported a Kappa of only 0.55. Thus, even though the three observers of Lane et al. were all experienced, their interobserver agreement for this variable was significantly lower than the respective agreement of the new system. In contrast to Lane et al., but similar to the present study, Madan et al. [24] reported the agreement between five observers with different levels of experience. Their interobserver Kappa coefficients varied between 0.351 and 0.673 for the overall disc grade and thus, were lower than the respective value of the new system. These results indicate that the use of undefined terms may work with experienced observers for the variables "Height Loss" and "Osteophyte Formation", but does not work for the variable "Sclerosis" and not with inexperienced observers.

Similar to the work of Madan et al., the agreement between observers with different degrees of experience was also reported by Pfirrmann et al., who developed a grading system based on magnetic resonance images [27]. The Kappa coefficients reported by this group ranged between 0.74 and 0.81 for the overall disc grade. This range covers the respective value for the new radiographic system (0.787), but is higher than the range reported by Madan et al. (0.351–0.673) [24]. These differences between Pfirrmann et al. and Madan et al. indicate that the assessment of signal intensity and homogeneity on magnetic resonance images may per se be more objective than the assessment of bony structures and densities on radiographs. This would probably also be the case for Modic's classification of vertebral body marrow changes [26]. According to Modic et al., these changes are associated with degenerative disc disease and, thus, are often used to quantify disc degeneration.

Due to their small number, each of the three variables "Height Loss", "Diffuse Sclerosis" and "Osteophyte Formation" strongly influences the overall degree of degeneration. Thus, discs of one and the same overall degree of degeneration may have completely different appearances. Depending on the purpose of the study, it might therefore be advantageous to report the three variables individually instead of the overall degree of degeneration only. Another possibility to reduce the weight of each variable would be to include further variables such as "Listhesis" or "Disc Calcification" into the grading system. These variables are assumed to be associated with disc degeneration and can be seen on X-rays [7, 8, 35]. An objective grading, however, is difficult. The degree of listhesis seen on a radiograph, for example, strongly depends on the loading of the spine during X-raying. For instance, the degree of listhesis of one and the same patient may be completely different in a lying position when compared to a standing or sitting position. And whether calcifications can be seen on radiographs or not strongly depend on the quality of the radiograph and the voltage used. Therefore, these two variables were not included into the new grading system.

To validate the radiographic grading system, the macroscopic degree of degeneration was defined as being "real". This definition was used since macroscopic slices directly reflect the changes within the disc, whereas

radiographs only depict the surrounding bony structures. Histologically, disc degeneration shows regional variations within one and the same disc [4]. The macroscopic grading system used here, however, does not account for these differences. This also applies for the radiographic grading systems since the disc itself cannot be depicted. Thus, in both the grading systems, the macroscopic and the radiographic one, the disc is assessed as "average".

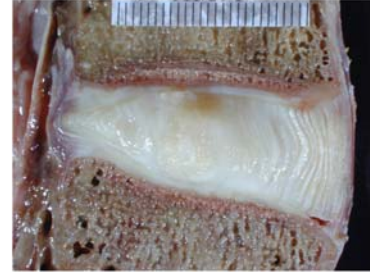**Grade 0 (no degeneration)** (L4-5, male, 37 years):
postero-anterior radiograph   *lateral radiograph*   *midsagittal frozen cut*

Grade 1 (mild degeneration) (L4-5, female, 49 years):
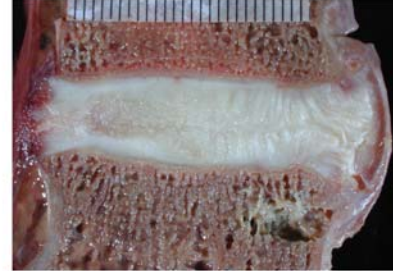*postero-anterior radiograph*   *lateral radiograph*   *midsagittal frozen cut*

Grade 2 (moderate degeneration) (L4-5, male, 88 years):
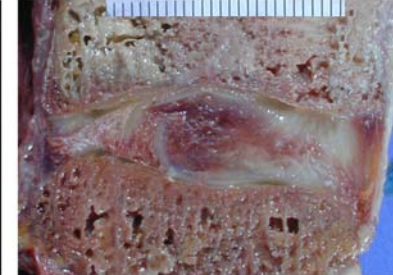*postero-anterior radiograph*   *lateral radiograph*   *midsagittal frozen cut*

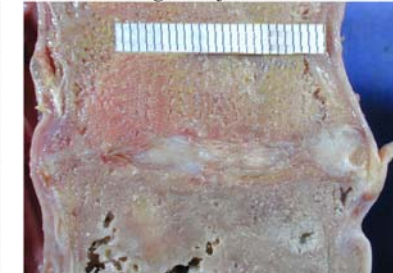**Grade 3 (severe degeneration)** (L3-4, female, 89 years):
postero-anterior radiograph   *lateral radiograph*   *midsagittal frozen cut*

Fig. 6 Examples of the four degrees of degeneration

Compared to the macroscopic "real" degrees of degeneration, the radiographic degrees of degeneration tended to be lower. This underestimation has two reasons: first, radiographically, the loss of intervertebral height, the formation of osteophytes and endplate sclerosis are indirect signs of degeneration, while changes within the disc itself cannot be seen directly. Thus, early degenerative changes, such as a discolouring of the nucleus cannot be detected on radiographs. Similarly Frobin et al. could show that signal loss within the intervertebral disc is possible without the radiographic loss of height [16]. In such cases, the disc may radiographically have grade 0, and macroscopically, however, grade 1. Thus, in the detection of early degenerative changes within the disc, magnetic resonance imaging may have certain advantages compared to plain radiography. However, according to Benneker et al., a magnetic resonance imaging score does not necessarily have to correlate better with morphology than a radiographic score [3].Also, the variable "Diffuse Sclerosis" is easily underestimated on radiographs. This underestimation may become even more pronounced if radiographs of patients instead of osteoligamentous specimens have to be rated, since on the radiographs of patients much more tissue surrounds the spine and influences the x-ray transparency around the vertebral bodies. The only variable, where radiography revealed a higher degree of degeneration than macroscopy was "Osteophyte Formation" since macroscopically, the assessment of osteophytes was restricted to the mid-sagittal plane.

Despite these discrepancies between the radiographic and the macroscopic "real" degrees of degeneration, however, the agreement for the overall degree of degeneration still was substantial (Fig. 6). Thus, the overall validity of the new radiographic grading system is deemed to be good.

Conclusions

In conclusion, we believe that the new radiographic grading system is an almost objective, valid and reliable tool if the degree of degeneration of individual lumbar intervertebral discs has to be quantified. However, the user should always remember that the radiographic degree of degeneration tends to be lower than the "real" macroscopic one and that slight differences between the ratings of observers with different degrees of experience have to be expected.

This study was focused on the agreement between one experienced and one inexperienced observer to evaluate the objectivity of the new system. Other parameters such as the intraobserver agreement, the agreement between observers with similar degrees of experience, the agreement between whole institutions or the effect of the quality of the radiographs on the ratings need to be investigated in future studies.

References

1. Adams MA, Dolan P, Hutton WC (1986) The stages of disc degeneration as revealed by discograms. J Bone Joint Surg Br 68(1):36–41
2. Althoff I, Brinckmann P, Frobin W, Sandover J, Burton K (1992) An improved method of stature measurement for quantitative determination of spinal loading. Application to sitting postures and whole body vibration. Spine 17(6):682–693
3. Benneker LM, Heini PF, Anderson SE, Alini M, Ito K (2005) Correlation of radiographic and MRI parameters to morphological and biochemical assessment of intervertebral disc degeneration. Eur Spine J 14(1):27–35
4. Boos N, Weissbach S, Rohrbach H, Weiler C, Spratt KF, Nerlich AG (2002) Classification of age-related changes in lumbar intervertebral discs: 2002 Volvo Award in basic science. Spine 27(23):2631–2644
5. Botsford DJ, Esses SI, Ogilvie-Harris D J (1994) In vivo diurnal variation in intervertebral disc volume and morphology. Spine 19(8):935–940
6. Brooker AE, Barter RW (1965) Cervical spondylosis. A clinical study with comparative radiology. Brain 88(5):925–936
7. Chanchairujira K, Chung CB, Kim JY et al (2004) Intervertebral disc calcification of the spine in an elderly population: radiographic prevalence, location, and distribution and correlation with spinal degeneration. Radiology 230(2):499–503
8. Cheng XG, Brys P, Nijs J et al (1996) Radiological prevalence of lumbar intervertebral disc calcification in the elderly: an autopsy study. Skeletal Radiol 25(3):231–235
9. Collins DH (1949) The pathology of articular and spinal diseases. Edward Arnold & Co., CO
10. Coventry MB (1945) The intervertebral disc: its macroscopic anatomy and pathology: Part III. Pathological changes in the intervertebral disc lesion. J Bone Joint Surg (Am) 27:460–473
11. Coventry MB (1945) The intervertebral disc: its microscopic anatomy and pathology: Part II. Changes in the intervertebral disc concomitant with age. J Bone Joint Surg (Am) 27:233–247
12. Farfan HF (1973) Mechanical disorders of the low back. Lea & Febiger, Philadelphia
13. Fleiss J, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 33:613–619
14. Friberg S, Hirsch C (1949) Anatomical and clinical studies on lumbar disc degeneration. Acta Orthop Scand 19:222–242

15. Frobin W, Brinckmann P, Biggemann M, Tillotson M, Burton K (1997) Precision measurement of disc height, vertebral height and sagittal plane displacement from lateral radiographic views of the lumbar spine. Clin Biomech (Bristol, Avon) 12(Suppl. 1):S1–S63

16. Frobin W, Brinckmann P, Kramer M, Hartwig E (2001) Height of lumbar discs measured from radiographs compared with degeneration and height classified from MR images. Eur Radiol 11(2):263–269

17. Gordon SJ, Yang KH, Mayer PJ, Mace AH Jr, Kish VL, Radin EL (1991) Mechanism of disc rupture. A preliminary report. Spine 16(4):450–456

18. Hirsch C (1972) Some morphological changes in the cervical spine during ageing. In: Hirsch C, Zotterman Y (eds) Cervical pain. Pergamon Press, Oxford, pp 21–32

19. Kellgren JH, Jeffrey MR, Ball J (1963) In: The epidemiology of chronic rheumatism. Vol. II: Atlas of standard radiographs of arthritis. Blackwell Scientific Publications, Oxford, pp 14–19

20. Kellgren JH, Lawrence JS (1952) Rheumatism in miners. II. X-ray study. Br J Ind Med 9(3):197–207

21. Kettler A, Wilke H-J (accepted) Review of existing grading systems for cervical and lumbar disc and facet joint degeneration. Eur Spine J

22. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174

23. Lane NE, Nevitt MC, Genant HK, Hochberg MC (1993) Reliability of new indices of radiographic osteoarthritis of the hand and hip and lumbar disc degeneration. J Rheumatol 20(11):1911–1918

24. Madan SS, Rai A, Harley JM (2003) Interobserver error in interpretation of the radiographs for degeneration of the lumbar spine. Iowa Orthop J 23:51–56

25. Mimura M, Panjabi MM, Oxland TR, Crisco JJ, Yamamoto I, Vasavada A (1994) Disc degeneration affects the multidirectional flexibility of the lumbar spine. Spine 19(12):1371–1380

26. Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR (1988) Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 166(1 Pt 1):193–199

27. Pfirrmann CW, Metzdorf A, Zanetti M, Hodler J, Boos N (2001) Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine 26(17):1873–1878

28. Resnick D (1985) Degenerative diseases of the vertebral column. Radiology 156(1):3–14

29. Roberts N, Hogg D, Whitehouse GH, Dangerfield P (1998) Quantitative analysis of diurnal variation in volume and water content of lumbar intervertebral discs. Clin Anat 11(1):1–8

30. SAS (1999) SAS Institute Inc., Cary, NC, USA

31. Silberstein CE (1965) The evolution of degenerative changes in the cervical spine and an investigation into the "Joints of Luschka". Clin Orthop 40:184–204

32. Thompson JP, Pearce RH, Schechter MT, Adams ME, Tsang IK, Bishop PB (1990) Preliminary evaluation of a scheme for grading the gross morphology of the human intervertebral disc. Spine 15(5):411–415

33. Töndury G (1972) The behaviour of the cervical discs during life. In: Hirsch C, Zotterman Y (eds) Cervical pain. Pergamon Press, Oxford, pp 59–66

34. Vernon-Roberts B, Pirie C J (1977) Degenerative changes in the intervertebral discs of the lumbar spine and their sequelae. Rheumatol Rehabil 16(1):13–21

35. Vogt MT, Rubin DA, San Valentin R et al (1999) Degenerative lumbar listhesis and bone mineral density in elderly women. The study of osteoporotic fractures. Spine 24(23):2536–2541