Radiological Scoring Methods in Ankylosing Spondylitis. Reliability and Change Over 1 and 2 Years

ANNEKE SPOORENBERG, KURT de VLAM, SJEF van der LINDEN, MAXIME DOUGADOS, HERMAN MIELANTS, HILLE van de TEMPEL, and DÉSIRÉE van der HEIJDE

ABSTRACT. ***Objective.*** To compare reliability and change over time of radiological scoring methods in ankylosing spondylitis (AS).

Methods. Two trained observers scored 217 sets of radiographs from baseline and from one and 2 years' followup. Sacroiliac (SI) joints were grade 0–4 by the New York method and Stoke Ankylosing Spondylitis Spine Score (SASSS). Hips and cervical and lumbar spine were graded 0–4 by Bath Ankylosing Spondylitis Radiology Index (BASRI). BASRI spinal scores and New York SI are combined into BASRI-spine (score 2–12) and with the addition of BASRI-hips into BASRI-total (2–16). Cervical and lumbar spine were also scored in detail (SASSS, 0–36 each) and were combined into SASSS-total or "modified" SASSS (both range 0–72). To assess change a smallest detectable difference (SDD) was estimated for data on a quasi-interval scale.

Results. The SI scoring methods showed intra and interobserver kappa between 0.36 and 0.70. The BASRI-hip reached kappa between 0.59 and 0.84. Combined SASSS scores were most reliable, with intra and interobserver intraclass correlation coefficients (ICC) between 0.90 and 0.96. The ICC of the combined BASRI scores were also very good, ranging from 0.85 to 0.95. For SI New York, SI SASSS, and BASRI-hip, 0.3–1.2% of patients deteriorated 1 grade; 7.5% deteriorated 1 grade (6.3% of maximum score) in BASRI-spine and BASRI-total, and observers agreed in up to 48% of the cases that no change occurred. The SDD was lowest (7.5; 10% of maximum score) for "modified" SASSS. Only 0.8% of patients deteriorated more than the SDD and observers agreed in up to 92% of the cases that no change occurred.

Conclusion. Radiological scoring methods for AS are moderately to excellently reliable. Under the selected scoring conditions (concealed time order, average of 2 observers, SDD based on interobserver data, unselected patient population) there was too little change over 2 years to be detected reliably by the scoring methods. (J Rheumatol 2004;31:125–32)

Key Indexing Terms:
RADIOLOGY OUTCOME BATH ANKYLOSING SPONDYLITIS RADIOLOGY INDEX STOKE ANKYLOSING SPONDYLITIS SPINE SCORE ANKYLOSING SPONDYLITIS

Radiological damage is considered an important outcome in ankylosing spondylitis (AS)[1]. The evaluation of radiological change proves to be very difficult, for several reasons. Radiological sacroiliitis can easily be missed because of the complex anatomy of the sacroiliac (SI) joints. The undulating articular surfaces make it hard to image these joints on conventional radiographs. Squaring, erosions, and sclerosis appear in different stages of the disease[2] and syndesmophytes must be differentiated from osteophytes and disorders such as diffuse idiopathic skeletal hyperostosis (DISH). Usually AS is a slowly progressive disease and radiological change appears gradually: evaluation of radiographs with an interval of one year does not seem to be useful[3,4]. However, a detailed scoring method showed some change after a period of one year[5] and change after 2 years of followup could also be detected by a graded scoring method[3].

Changes of the SI joints are most frequently scored using the 5-grade New York criteria[6] or the nearly similar Stoke Ankylosing Spondylitis Spine Score (SASSS)[5,7,8]. To evaluate the spine in AS there are essentially 2 different scoring methods. The Bath Ankylosing Spondylitis Radiology Index (BASRI) is a global graded scoring method that is quick and easy to perform[9-11]. The first version of this BASRI was described in 1995[9] and a modified version was published in

From the University Hospital Maastricht, Maastricht, The Netherlands; University Hospital Gent, Gent, Belgium; Hôpital Cochin, Paris, France; Maasland Hospital, Sittard, The Netherlands; and Limburg University Center, Diepenbeek, Belgium.

A. Spoorenberg, MD, Rheumatologist; S. van der Linden, MD, PhD, Professor in Rheumatology, University Hospital Maastricht; K. de Vlam, MD, PhD, Rheumatologist; H. Mielants, MD, PhD, Professor in Rheumatology, University Hospital Gent; M. Dougados, MD, PhD, Professor in Rheumatology, Hôpital Cochin; H. van de Tempel, MD, Rheumatologist, Maasland Hospital; D.M.F.M. van der Heijde, MD, PhD, Professor in Rheumatology, University Hospital Maastricht, Limburg University Center.

Address reprint requests to Dr. D.M.F.M. van der Heijde, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ, Maastricht, The Netherlands. E-mail: dhe@sint.azm.nl

Submitted June 12, 2002; revision accepted June 10, 2003.

1998. Several BASRI scores are also combined in composite scores[10,11]. The SASSS for the spine is a more detailed scoring method assessing different features such as squaring, sclerosis, and erosions at various locations of each vertebra[5,7,8]. In an earlier study we compared the reliability and change over time over one year of these scoring methods. We concluded that both the SASSS method for the spine and BASRI achieved good reliability. All other scoring showed moderate reliability at best. No method showed change over a period of one year in a considerable number of patients[4]. At the time of our first study no scoring method was available to evaluate the hip in AS. Therefore the Larsen scoring method designed to score the hips in rheumatoid arthritis was used. Recently a new graded scoring method was developed to evaluate the hip in AS, the BASRI-hip[3]. In this second study we used the BASRI-hip.

The objective of this followup study was to compare all available AS radiological scoring methods for reliability and change over 2 years' time.

MATERIALS AND METHODS

Patients. A total of 217 consecutive outpatients who satisfied the modified New York criteria[12] were included in our study. Our study population comprises a cross sectional cohort of patients with AS, followed longitudinally. Sixty-nine percent of patients were male, a distribution usually found in AS populations. The median age at baseline was 42.2 years (range 18–78). There is a striking difference between median duration of complaints (17.0 yrs, range 0.3–54) and duration of disease since diagnosis (9.4 yrs, range 0.1–41), indicating patients with AS have complaints long before the diagnosis is made.

Scoring methods: SI joints. The SI joints were scored according to the New York method (0–4) and the SASSS (0–4). Both methods score the lower half of the SI joints and are graded from 0 to 4[5-7]. The main difference between these methods is grade 4 (complete ankylosis), where the New York method does not allow residual sclerosis. Both SI joints are scored separately, and thereafter the score is summed.

Scoring method: hips. The hips were scored according to BASRI-hips, graded 0 = normal, 1 = suspicious (possible focal joint space narrowing), 2 = minimal (definite narrowing, leaving a circumferential joint space > 2 mm), 3 = moderate (narrowing with circumferential joint space ≤ 2 mm or bone-on-bone apposition of < 2 cm), and 4 = severe (bone deformity or bone-on-bone apposition ≥ 2 cm or total hip replacement)[11]. Both hips are scored separately, and thereafter the score is averaged.

Scoring methods: spine. The BASRI was developed for the anteroposterior and lateral view of the lumbar spine and the lateral view of the cervical spine, and is graded 0 to 4 for each view[10]. The BASRI-spine is the composite score of the BASRI scored on the lumbar spine, the cervical spine, and the SI joints (New York method), with scores ranging from 2 to 12[10]. The highest score of the 2 views of the lumbar spine is applied in this BASRI-spine. In case of absence of one of the lumbar spine radiographs, the score of the available view was taken into account. The BASRI-total is a composite score of the BASRI-spine and the BASRI-hips, with scores ranging from 2 to 16[11].

The SASSS is scored from the lower border of the 12th thoracic vertebra down to and including the upper border of the sacrum. This scoring method is used on both the anterior and posterior site of the vertebrae, with a score ranging from 0 to 36 for each site, so the total score will range from 0 to 72[6,7].

The "modified" SASSS is scored from the lower border of the 2nd cervical vertebra to the upper border of the 1st thoracic vertebra and the lower border of the 12th thoracic vertebra to the upper border of the sacrum[8]. This scoring method is only used on the anterior site of the vertebrae, with a score ranging from 0 to 36 for the cervical spine and 0–36 for the lumbar spine. Therefore the total score of the modified version will also range from 0 to 72.

Missing scoring sites for the SASSS were handled as follows: when up to 3 scoring sites for each view could not be scored, the mean of the other scoring sites was applied; when more than 3 scoring sites could not be scored the whole SASSS score for that particular view was scored "missing."

In all spine-scoring methods, syndesmophytes were differentiated from osteophytes using the following description: an osteophyte was defined as a bony deformity on the edge of the vertebra projecting > 0.5 cm horizontally. Osteophytes were not included in the analyses. Scoring methods for the SI joint and spine are described in detail in our first study[4].

Inter and intraobserver reliability. To obtain inter and intraobserver reliability of the scoring method for the hips (BASRI-hip), 2 experienced observers (AS and KV) scored 30 randomly selected baseline radiographs from the 217 consecutive outpatients with AS. The University Hospital Maastricht, the University Hospital Gent, and the Hôpital Cochin in Paris each provided 10 blinded radiographs of anteroposterior views of the pelvis to score the hips. The 2 observers had a training session to gain experience with the scoring method. All abnormalities present on the radiographs were discussed in detail. After this training session the observers scored a set of radiographs independently and discussed the results with each other, and this session was followed by a consensus meeting with the 2 observers and 2 other experts in AS. The study on BASRI-hip reliability was started when few (≤ 5) discrepancies existed between the 2 observers. Two different sets of 30 radiographs were used for training, and again a different set of 30 radiographs was used for assessment of reliability of BASRI-hip. For BASRI-hip interobserver reliability was calculated, and in addition, intraobserver reliability based on the scores of the radiographs was scored a second time after 2 weeks.

For all other scoring methods of the SI joints and the spine inter and intraobserver reliability was assessed in our first study[4]. Baseline and one-year radiographs were scored during our first study and again in this present study. Intraobserver reliability could also be computed using data from baseline and at one year of our first study and this present study, except for BASRI-total, because the BASRI-hip scoring method (which is part of BASRI-total) was not available at the time of our first study.

This second intraobserver reliability was based on an interval of 2 years between the first and the second reading. Interobserver reliability could also be calculated for baseline and one year and 2 year data of this study.

Change over time over 2 years. We included 217 consecutive outpatients who satisfied the modified New York criteria for AS[12]; 137 patients from the University Hospital Maastricht and Maasland Hospital Sittard (The Netherlands), 55 patients from the Hôpital Cochin, Paris (France), and 25 patients from the University Hospital Gent (Belgium). These hospitals are secondary and tertiary referral centers. From these 217 patients, we studied 3 sets of radiographs taken with an interval of one year. All 3 sets of radiographs were scored viewing the radiographs simultaneously (paired), without knowledge of the chronology of the radiographs, in a random order by the same 2 experienced observers independently (AS and KV). The scoring methods were also scored in random order. As a result, SI joints were scored separately from the hip joints. We used the scoring methods as described in the previous section.

Statistics. Inter and intraobserver agreement of the different scoring methods was analyzed for categorical data by the linear weighted kappa statistic and for continuous data by the random effects, average measure intraclass correlation coefficient (ICC, type 3.1) with observer as fixed facet[13]. Joint pairs (hips and SI joints) were regarded as independent units, i.e., their possible correlation was ignored. To visualize the observer agreement we plotted the continuous data using the Bland and Altman method[14]. Change over time of the scoring methods was assessed by a cutoff value

based on interobserver reliability of change. For grading scales such as all BASRI methods and the New York and SASSS methods for the SI joints, a change of one grade was defined as the minimum detectable difference. For data on a semicontinuous scale, such as the SASSS methods for the spine, a smallest detectable difference (SDD) was estimated in the situation of 2 fixed observers, yielding a mean change[15]. The SDD is the smallest change that can be detected apart from measurement error. In the BASRI scores and the SI scores, these are categorical scales, which do not allow calculation of a SDD.

RESULTS

Inter and intraobserver reliability of BASRI-hip (single radiographs). BASRI-hip (0–4) scores of 30 baseline radiographs showed good to very good reliability with intraobserver kappa of 0.73 and 0.84 and interobserver kappa 0.63 and 0.66.

Inter and intraobserver reliability of all scoring methods (baseline, 1 and 2 years). SI scoring methods (New York and SASSS, 0–4) showed moderate to good intraobserver reliability with kappa ranging from 0.36 to 0.67. For both scoring methods interobserver reliability was good with kappa between 0.66 and 0.70 (Table 1).

The BASRI grading scores of the various parts of the spine (0–4) showed moderate to good reliability. For BASRI scores of the lumbar spine, the intraobserver kappa with 2-year interval between the scoring sessions ranged from 0.61 to 0.65 and interobserver kappa from 0.58 to 0.78. BASRI of the cervical spine showed intraobserver kappa ranging from 0.41 to 0.56 and interobserver kappa from 0.61 to 0.62. BASRI-hip (0–4) scores again showed good interobserver reliability, kappa ranging from 0.59 to 0.60 (Table 1). Intraobserver reliability could not be calculated because at the time our first study was performed the scoring method for BASRI-hip was not available.

The combined BASRI scores showed good to excellent reliability. For the BASRI-spine (2–12) the intraobserver ICC ranged from 0.85 to 0.90 and interobserver ICC from 0.92 to 0.94. This was even slightly better for BASRI-total (0–16) with ICC ranging from 0.94 to 0.96 for interobserver reliability (Table 1). Intraobserver reliability could not be computed because the BASRI-hip scoring method that is part of BASRI-total was not available at the time of our first study.

The SASSS also showed excellent reliability. The SASSS scored on the anterior and posterior site of the lateral view of the lumbar spine (both 0–36) showed intraobserver ICC 0.94–0.95 and interobserver ICC 0.94–0.98.

The combined score, SASSS-total (0–72), showed intra and interobserver ICC of 0.92–0.96 and 0.98, respectively (Table 1). The SASSS applied on the anterior site of the lateral view of the cervical spine showed intra and interobserver ICC of 0.92–0.96 and 0.95–0.96, respectively. The combined score of the anterior sites of both the lateral view of the lumbar and cervical spine (modified SASSS, 0–72) showed good intra and interobserver ICC of 0.95–0.96 and 0.97–0.98, respectively (Table 1).

Table 2 shows the concordance rates of the 2 observers at baseline and at 1 year and 2 year followup for individual and combined scoring methods. For each instance perfect concordance rates are low for all scoring methods. The concordance rates for the combined SASSS methods are

Table 1. Inter and intraobserver reliability of all scoring methods (baseline, 1 year, 2 year) (lower and upper border of the 95% confidence interval).

Method	Interobserver Agreement			Intraobserver Agreement (radiographs scored with a 2 year interval)			
	Observer 1 and 2			Observer 1		Observer 2	
	T0	T12	T24	T0	T12	T0	T12
SI joints New York, 0–4	k = 0.68 0.61–0.72 n = 400	k = 0.69 0.62–0.74 n = 388	k = 0.66 0.62–0.72 n = 365	k = 0.63 0.56–0.69 n = 348	k = 0.55 0.47–0.63 n = 342	k = 0.40 0.32–0.48 n = 348	k = 0.40 0.32–0.48 n = 339
SI joints SASSS, 0–4	k = 0.68 0.63–0.72 n = 400	k = 0.69 0.64–0.74 n = 388	k = 0.70 0.64–0.74 n = 365	k = 0.67 0.63–0.74 n = 352	k = 0.62 0.56–0.64 n = 344	k = 0.42 0.34–0.49 n = 348	k = 0.36 0.28–0.43 n = 339
BASRI-hip, 0–4	k = 0.60 0.54–0.67 n = 30	k = 0.59 0.52–0.66 n = 30	k = 0.60 0.52–0.67 n = 30				
BASRI-spine, 2–12	ICC = 0.92 0.90–0.94 n = 192	ICC = 0.94 0.93–0.96 n = 190	ICC = 0.93 0.91–0.95 n = 176	ICC = 0.89 0.84–0.92 n = 154	ICC = 0.90 0.86–0.92 n = 163	ICC = 0.85 0.79–0.89 n = 144	ICC = 0.86 0.80–0.89 n = 161
BASRI-total, 2–16	ICC = 0.94 0.92–0.95 n = 192	ICC = 0.96 0.94–0.97 n = 190	ICC = 0.95 0.93–0.96 n = 176				
SASSS modified, 0–72	ICC = 0.98 0.97–0.98 n = 162	ICC = 0.97 0.96–0.98 n = 172	ICC = 0.97 0.96–0.98 n = 153	ICC = 0.96 0.95–0.97 n = 154	ICC = 0.96 0.95–0.97 n = 164	ICC = 0.96 0.95–0.97 n = 136	ICC = 0.95 0.94–0.97 n = 155
SASSS total, 0–72	ICC = 0.98 0.97–0.98 n = 161	ICC = 0.98 0.97–0.98 n = 163	ICC = 0.98 0.97–0.98 n = 157	ICC = 0.93 0.92–0.94 n = 152	ICC = 0.96 0.95–0.97 n = 162	ICC = 0.92 0.91–0.93 n = 140	ICC = 0.95 0.94–0.96 n = 156

Table 2. Concordance rate (%), observer 1 and 2.

	SI SASSS n = 434 Range 0–4	SI New York n = 434 Range 0–4	BASRI-hip n = 434 Range 0–4	BASRI-spine n = 217 Range 2–12	BASRI-total n = 217 Range 2–16	SASSS Total 1 wk ant and post n = 217 Range 0–72	Modified SASSS 1 wk ant and cwk ant n = 217 Range 0–72
Baseline							
Perfect agreement*	70	74	66	32	38	35	22
< 2 grades difference				69	72		
< 6 points difference						78	78
1 year							
Perfect agreement*	71	76	64	35	36	33	23
< 2 grades difference				74	75		
< 6 points difference						80	77
2 years							
Perfect agreement*	72	76	67	31	35	27	21
< 2 grades difference				68	68		
< 6 points difference						77	71

* Perfect agreement: < 1 grade/point difference between observers.

between 71% and 80%, accepting less than 6 points difference between the 2 observers on a scale of 0–72. Accepting less than 2 grades difference, the concordance rates of BASRI-total (68–75%) are comparable with those of the combined SASSS methods accepting less than 6 points difference. One grade change in BASRI-total represents 6.3% of the maximum scoring range. One grade in the BASRI-total can be compared to 5 points in the combined SASSS methods, which represents 4.9% of the maximum range.

To illustrate observer agreement over the complete range of observed scores, Figures 1 and 2 show Bland and Altman plots of baseline data and progression data of the modified SASSS. Progression data are based on the difference between baseline data and 2 year followup data. The Bland and Altman plot of baseline data of the modified SASSS shows a maximum difference of 26 points between the 2 observers on a scoring range of 0–72; the 95% limits of agreement of the difference between the 2 observers is 1.96 times the SD (4.4) (Figure 1). Observer 1 scores systematically somewhat higher than Observer 2. The Bland and Altman plots of 1 and 2 year data are very similar to this baseline plot (data not shown). Figure 2, showing the progression data of the 2 observers over 2 years, shows a

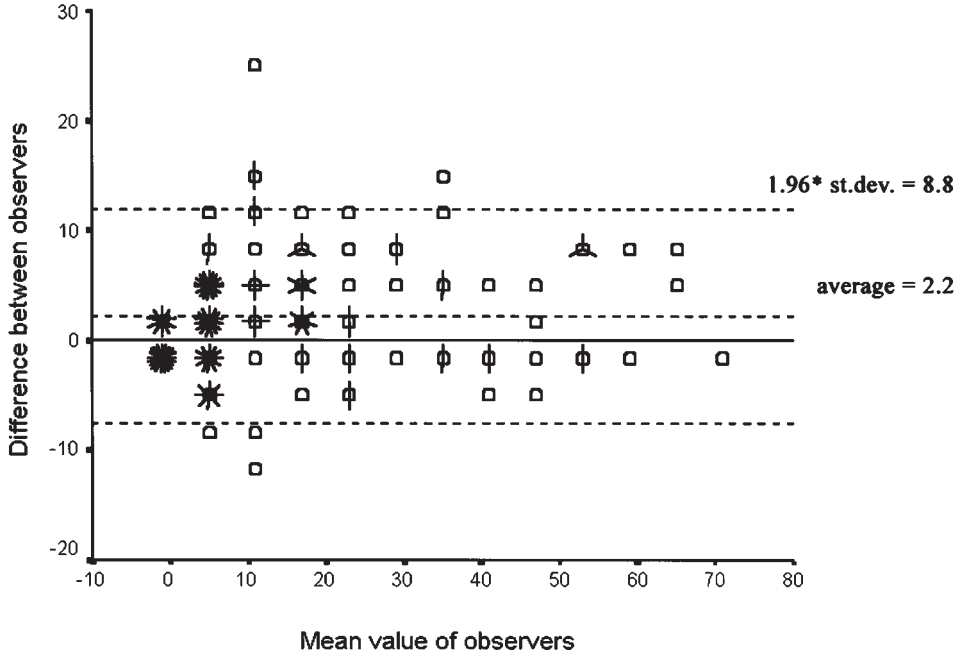

Figure 1. Bland and Altman plot: mean versus difference of 2 observers at baseline; SASSS-modified (SASSS lumbar anterior and cervical anterior). ○ = 1 patient, every dash represents an extra patient.

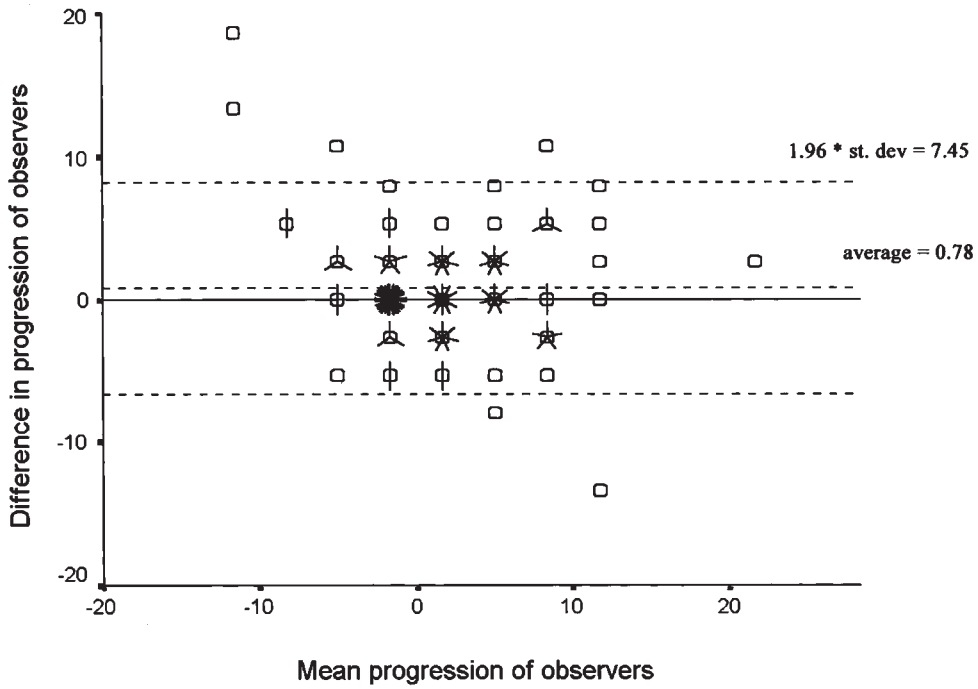

Figure 2. Bland and Altman plot: mean versus difference of the progression scores of the 2 observers over 2 years; SASSS-modified (SASSS lumbar anterior and cervical anterior). ○ = 1 patient, every dash represents an extra patient.

maximum difference of 18 points between the observers, with 95% limit of agreement of 7.45. Because the BASRI scores are on categorical scales, it is not possible to calculate SDD and Bland and Altman plots.

Change over time. Overall, we found little change over the course of 2 years. No difference was found in mean, median, and SD of the entire group for SI New York, SI SASSS, BASRI-hip score, BASRI-spine, and BASRI-total (Table 3). By SASSS (spine) there was little difference in mean, median, and SD at baseline and 1 and 2 years (Table 3). These differences did not reach statistical significance. The distribution of (dis)agreement based on the minimum detectable difference of 1 grade for data on graded scales or the SDD for data on a semicontinuous scale is shown in Table 4. If a patient deteriorated or improved more than the SDD or 1 grade, the change was judged as real. This is

Table 3. Summary statistics on the group level at baseline and after 1 year and 2 year followup.

Scoring Method (range)	Average of the 2 Observers* mean, SD Median, minimum–maximum		
	Baseline	1 Year	2 Year
SI SASSS (0–4)	3.1, 0.8	3.1, 0.8	3.1, 0.8
	3.0, 0–4	3.0, 0–4	3.0, 0–4
SI New York (0–4)	3.1, 0.8	3.1, 0.7	3.1, 0.8
	3.0, 0–4	3.0, 0–4	3.0, 0–4
BASRI-hip (0–4)	0.7, 1.0	0.7, 1.0	0.6, 1.0
	0.0, 0–4	0.0, 0–4	0.0, 0–4
BASRI-spine (2–12)	6.9, 2.8	6.8, 2.7	6.9, 2.8
	6.6, 0–12	6.5, 0–12	7.0, 0–12
BASRI-total (2–16)	7.6, 3.4	7.5, 3.3	7.5, 3.3
	7.1, 0–16	7.1, 0–16	7.0, 0–16
SASSS total, lumbar anterior and posterior (0–72)	11.5, 17.5	12.2, 17.9	12.3, 18.0
	4.0, 0–72	5.0, 0–72	4.5, 0–72
SASSS modified, lumbar and cervical anterior (0–72)	13.8, 16.4	13.7, 16.3	15.0, 16.8
	6.7, 0–72	6.7, 0–72	7.6, 0–72

* Average score of the 2 observers = (score observer 1 + score observer 2) divided by 2.

Table 4. Sensitivity to change of AS scoring methods [values are mean ± standard deviation of the difference (SDD)].

Method	Interobserver Agreement[1] of Change[2]		
	0–1 year	1–2 year	0–2 year
SI joints New York (left and right, 0–4)	0.04 ± 0.32 P0 85.6% ≥ 1 grade change: P– 0.3%, P+ 0.3%, P(–) 7.2%, P(+) 6.6%	0.02 ± 0.24 P0 88.5% ≥ 1 grade change: P– 0.3%, P+ 0.3%, P(–) 5.2%, P(+) 5.7%	0.03 ± 0.29 P0 88.8% ≥ 1 grade change: P– 1.2%, P+ 0.6%, P(–) 7.5%, P(+) 3.7%
SI joints SASSS (left and right, 0–4)	0.07 ± 0.17 P0 89.7% ≥ 1 grade change: P– 0%, P+ 0%, P(–) 5.6%, P(+) 4.5%	0.05 ± 0.14 P0 91.7% ≥ 1 grade change: P– 0%, P+ 0%, P(–) 4.6%, P(+) 3.7%	0.01 ± 0.17 P0 90.5% ≥ 1 grade change: P– 0.3%, P+ 0%, P(–) 5.2%, P(+) 4.0%
BASRI-hip (left and right, 0–4)	0.02 ± 0.24 P0 84.6% ≥ 1 grade change: P– 0.3%, P+ 0.0%, P(–) 8.2%, P(+) 1.6%	0.02 ± 0.26 P0 84.6% ≥ 1 grade change: P– 0.9%, P+ 0.0%, P(–) 7.4%, P(+) 6.8%	0.0 ± 0.25 P0 85.3% ≥ 1 grade change: P– 0.3%, P+ 0.0%, P(–) 7.3%, P(+) 7.3%
BASRI-spine Lumbar-cervical spine, SI New York (2–12)	0.03 ± 1.28 P0 48.6% ≥ 1 grade change: P– 5.6%, P+ 2.8%, P(–) 22.5%, P(+) 8.5%	0.14 ± 1.0 P0 51.9% ≥ 1 grade change: P– 1.8%, P+ 1.2%, P(–) 28.7%, P(+) 17.4%	0.15 ± 1.22 P0 48.6% ≥ 1 grade change: P– 7.5%, P+ 1.9%, P(–) 32.2%, P(+) 18.7%
BASRI-total Lumbar-cervical spine, SI New York, hip (2–16)	0.02 ± 0.59 P0 49.7% ≥ 1 grade change: P– 5.6%, P+ 2.8%, P(–) 28.2%, P(+) 1.5%	0.11 ± 0.86 P0 50.7% ≥ 1 grade change: P– 2.4%, P+ 0.6%, P(–) 30.1%, P(+) 19.3%	0.11 ± 0.73 P0 48.1% ≥ 1 grade change: P– 7.4%, P+ 1.8%, P(–) 32.7%, P(+) 19.8%
SASSS-total 1 wk post + ant (0–72)	SDD 8.2 P0 89.9% ≥ SDD change: P– 0.7%, P+ 0%, P(–) 8.6%, P(+) 0.7%	SDD 6.8 P0 89.8% ≥ SDD change: P– 0%, P+ 0.7%, P(–) 4.4%, P(+) 5.8%	SDD 98 P0 92.3% ≥ SDD change: P– 0%, P+ 0%, P(–) 6.2%, P(+) 1.5%
Modified SASSS 1 wk ant + cwk (0–72)	SDD 6.8 P0 89.9% ≥ SDD change: P– 1.3%, P+ 0%, P(–) 6.1%, P(+) 2.7%	SDD 6.2 P0 89.4% ≥ SDD change: P– 2.1%, P+ 0.7%, P(–) 4.2%, P(+) 3.5%	SDD 7.5 P0 92.1% ≥ SDD change: P– 0.8%, P+ 0%, P(–) 4.7%, P(+) 2.4%

[1] Mean and SD are calculated from the difference in scores of the 2 observers over 2 years. Example: mean [(score 2 years observer 1–score baseline observer 1)–(score 2 years observer 2–score baseline observer 2)]. [2] Level of reliability of at least 1 grade or SDD change. P0: % of patients who did not change according to both observers. P–: % of patients who deteriorated according to both observers. P+: % of patients who improved according to both observers. P(–): % of patients who deteriorated according to one observer. P(+): % of patients who improved according to one observer.

reported for the percentage of patients that changed according to only one or according to 2 observers. For the graded methods SI New York, SI SASSS, and BASRI-hip, 0.3–1.7% deteriorated 1 grade according to both observers (Table 4). Although there was some change in mean, median, and SD in SASSS spine over this 2 year period, only very few patients (0–1.1%) deteriorated more than the SDD in the combined SASSS scores (Table 4). Only BASRI-spine and BASRI-total were able to detect change over this 2 year period in a great number of patients, 7.5% and 7.4%, respectively. To avoid a possible ceiling effect we performed the same analysis excluding maximum scores. For the graded scales we excluded grade 4 and for the BASRI combined scores and SASSS scores we excluded all data above the 75th percentile from analysis. These analyses did not influence the results (data not shown).

DISCUSSION

Scoring radiographs in AS remains very difficult. For most radiological scoring methods developed for AS, moderate to excellent intra and interobserver reliability could be achieved by 2 well trained observers. However, only the combined BASRI scoring methods (BASRI-spine and BASRI-total) and especially the SASSS showed good to excellent reliability. Even with a scoring interval of 2 years the interobserver reliability remained very good. Because of this 2 year scoring interval, intraobserver agreement was less high in comparison with the interobserver agreement. The reliability of the relatively new scoring method for the hips, BASRI-hip, proved to be good. We found it more reliable than the Larsen scoring method for the hips used in our first study[4,16]. Hip involvement in AS often shows as bony formations, which cannot be scored properly using the Larsen method. The BASRI-hip seems to be more disease-specific, and the developers of the BASRI-hip also found good and even excellent intra and interobserver agreement using unweighted kappa[3]. In contrast to our first study and most other studies[3,4,9-11], we used linear weighted kappa statistics instead of unweighted kappa statistics in this present study. In comparison, the value of unweighted kappa is lower than weighted kappa because large and small differences in assessments between observers are judged equally in unweighted kappa statistics. Further, kappa indicates to what extent 2 observers are capable of perceiving differences between radiographs. So kappa often turns out to be relatively low in the case of a homogeneous group where every single radiograph receives more or less the same score. This could be an explanation for the relatively low intra and interobserver agreement of the SI scoring methods, because patients were included if they fulfilled the modified New York criteria. SI joints were at least scored grade 2 for both sites on a scale from 0 to 4. Measures that relate observed to expected agreement (such as kappa and ICC) are of only limited value in this situation because of high

levels of expected agreement. This is confirmed by the relatively low median scores for the SASSS-spine methods (Table 3). The low prevalence of radiological damage in SASSS inflates the ICC statistics with a tendency to overestimate the ICC.

We also decided to show perfect concordance rates as a measure of (complete) agreement between the 2 observers, not depending on statistical techniques used such as kappa and ICC.

For all scoring methods the perfect concordance rates for the 2 observers were rather low. The developers of the BASRI method found good to excellent perfect concordance rates for the hips, between 78% and 95%[3,11]. They found good concordance rates (73–81%) for BASRI applied on lumbar and cervical spine, and they reached comparable concordance rates for the SI New York method (78–86%)[10,11]. Concordance rates for the SASSS method were not reported by the developers. Visual presentation of a Bland and Altman method aids understanding the data, especially because it visualizes the distribution of the data and outliers over the entire range of observed data. Visual presentation of agreement using the Bland and Altman method can be applied reliably only in scores with large ranges such as the SASSS, and not for the various BASRI scores.

In this study only BASRI-spine and BASRI-total were able to detect change in a significant number of patients over a 2 year period. This change could not be identified by the other graded and detailed scoring methods. For the BASRI-spine and BASRI-total, observers agreed in up to 52% that no change occurred. Unfortunately, we may still conclude that relevant change occurred rarely, because observers agreed in only 7.5% of cases that real change of at least 1 grade occurred. A reason for this could be that observer variation or error cannot be distinguished from radiological progression. An important reason could be that we followed an unselected group of patients, without a particular request for disease activity. In a group of AS patients selected for high disease activity the situation might be different, with a better signal-to-noise ratio. The developers of the BASRI-hip found significant change after 1 year using the Wilcoxon signed-rank test for nonparametric data (n = 60)[3]. For BASRI-spine they found significant change after 2 years (n = 31), and after 1 year 30% of 20 cases showed change of at least 1 grade, but this was not significant[10]. In 1999 they reported the magnitude of change for the BASRI-spine was from 7.0 to 7.9 in 2 years and 42% of 31 patients showed change in BASRI-spine score[11]. In these studies, change over time was not specified for BASRI-total. These results are based on a small number of patients because of our selection of only severe cases. The developers of the SASSS methods found significant change over a group of 28 patients in 1 year using the Mann-Whitney U test, with a mean change of 4.1 points (range 0–72) in SASSS-total and a mean change of 1.02 grade in SASSS for the SI joints[7]. In Dawes' study the order in which the radiographs were scored was known, in contrast with our study. This can influence the results markedly, as has been shown for rheumatoid arthritis (RA)[17-19]. All these sensitivity to change studies report results for a relatively small number of patients.

Comparing all radiological AS studies available at the moment, we recommend use of the New York method for the SI joints because it is most widely used and the reliability is similar to the SASSS score for SI joints. The BASRI-hip should be used because it is the only AS disease-specific method for the hips available, and it has good intra and interobserver reliability. To score the spine the choice is not unequivocal. The BASRI-spine and BASRI-total are preferred above the SASSS methods by their feasibility. According to face validity, the BASRI and the modified SASSS score the highest because both include the cervical spine in addition to the lumbar spine.

In our study the BASRI was the only method that showed change in a considerable number of patients over a 2 year period. However, this might be misleading information, as we set a change of 1 grade arbitrarily as a cutoff. Looking at the concordance rates within 1 grade difference, the observers agreed in only about 70% of the cases. For the SASSS the comparable data for concordance within 6 points is somewhat higher (in 78% of the cases). However, the calculated SDD for the SASSS is higher (9.8 for SASSS and 7.5 for modified SASSS). So the cutoff used for SASSS is very strict and that for BASRI is much looser. This might be an important reason why we were unable to detect changes if we applied the SASSS. Further study is needed, with sets of radiographs in which progression of damage is likely, e.g., sets with a 5 year interval or in a population with AS with a short disease duration, because these patients tend to show more radiological change or are selected for high disease activity. Additional studies where AS radiographs are scored in both random and chronological order are warranted to assess the difference in methodology, as done for RA[17-19].

Given the conditions used in our study (paired reading without information on sequence, average score of 2 observers, cutoff based on SDD on interobserver data, unselected patient population), the scoring methods are unable to detect change over 2 year interval reliably in most patients.

REFERENCES

1. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S, for the Assessment in Ankylosing Spondylitis Working Group. Preliminary core sets for endpoints in ankylosing spondylitis. J Rheumatol 1997;24:2225-9.
2. Calin A. Ankylosing spondylitis. Seronegative spondylarthropathies. Clin Rheum Dis 1985;11:41-61.
3. Calin A, Makay K, Santos H, Brophy S. A new dimension to outcome. Application of the Bath Ankylosing Spondylitis Radiology Index. J Rheumatol 1999;26:988-92.

4. Spoorenberg A, de Vlam K, van der Heijde D, et al. Radiological scoring methods in ankylosing spondylitis: reliability and sensitivity to change over one year. J Rheumatol 1999;26:997-1002.
5. Taylor HG, Wardle T, Beswick EJ, Dawes P. The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. Br J Rheumatol 1991;30:330-5.
6. Dale K. Radiographic gradings of sacroiliitis in Bechterews syndrome and allied disorders. Scand J Rheumatol 1979;32 Suppl 32:92-7.
7. Dawes PT. Stoke Ankylosing Spondylitis Spine Score. J Rheumatol 1999;26:993-6.
8. Creemers MCW, Franssen MJAM, van 't Hof MA, Gribnau FWJ, van de Putte LBA, van Riel PLCM. A radiographic scoring system and identification of variables measuring structural damage in ankylosing spondylitis [thesis]. Nijmegen, The Netherlands: University of Nijmegen; 1994.
9. Kennedy LG, Jenkinson TR, Mallorie PA, Whitelock HC, Garrett SL, Calin A. Ankylosing spondylitis: The correlation between a new metrology score and radiology. Br J Rheumatol 1995;34:767-70.
10. MacKay K, Mack C, Brophy S, Calin A. The Bath Ankylosing Spondylitis Radiology Index (BASRI). A new validated approach to disease assessment. Arthritis Rheum 1998;41:2263-70.
11. MacKay K, Brophy S, Mack C, Doran M, Calin A. The development and validation of a radiographic grading system for the hip in ankylosing spondylitis: the Bath Ankylosing Spondylitis Radiology Hip Index. J Rheumatol 2000;27:2866-72.
12. van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis: A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.
13. Shrout P, Fleiss J. Intraclass correlation: use in assessing rater reliability. Psychol Bull 1979;86:420-8.
14. Bland JM, Altman DG. Comparing methods of measurement: why plotting differences against standard methods is misleading. Lancet 1995;346:1085-87.
15. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731-9.
16. Larsen A, Dale K, Morten E. Radiographic evaluation of rheumatoid arthritis and related conditions by standard reference film. Acta Radiologica Diagnostica 1977;18:481-91.
17. van der Heijde D, Boonen A, van der Linden S, Boers M. Reading radiographs in sequence, in pairs or random in rheumatoid arthritis: influence on sensitivity to change [abstract]. Arthritis Rheum 1997;40 Suppl:S287.
18. Ferrara S, Priolo F, Cammisa M, et al. Clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from the Grisar study. Ann Rheum Dis 1997;56:608-12.
19. Salaffi F, Carotti M. Interobserver variation in quantitative analysis of hand radiographs in rheumatoid arthritis: comparison of 3 different reading procedures. J Rheumatol 1997;24:2055-6.