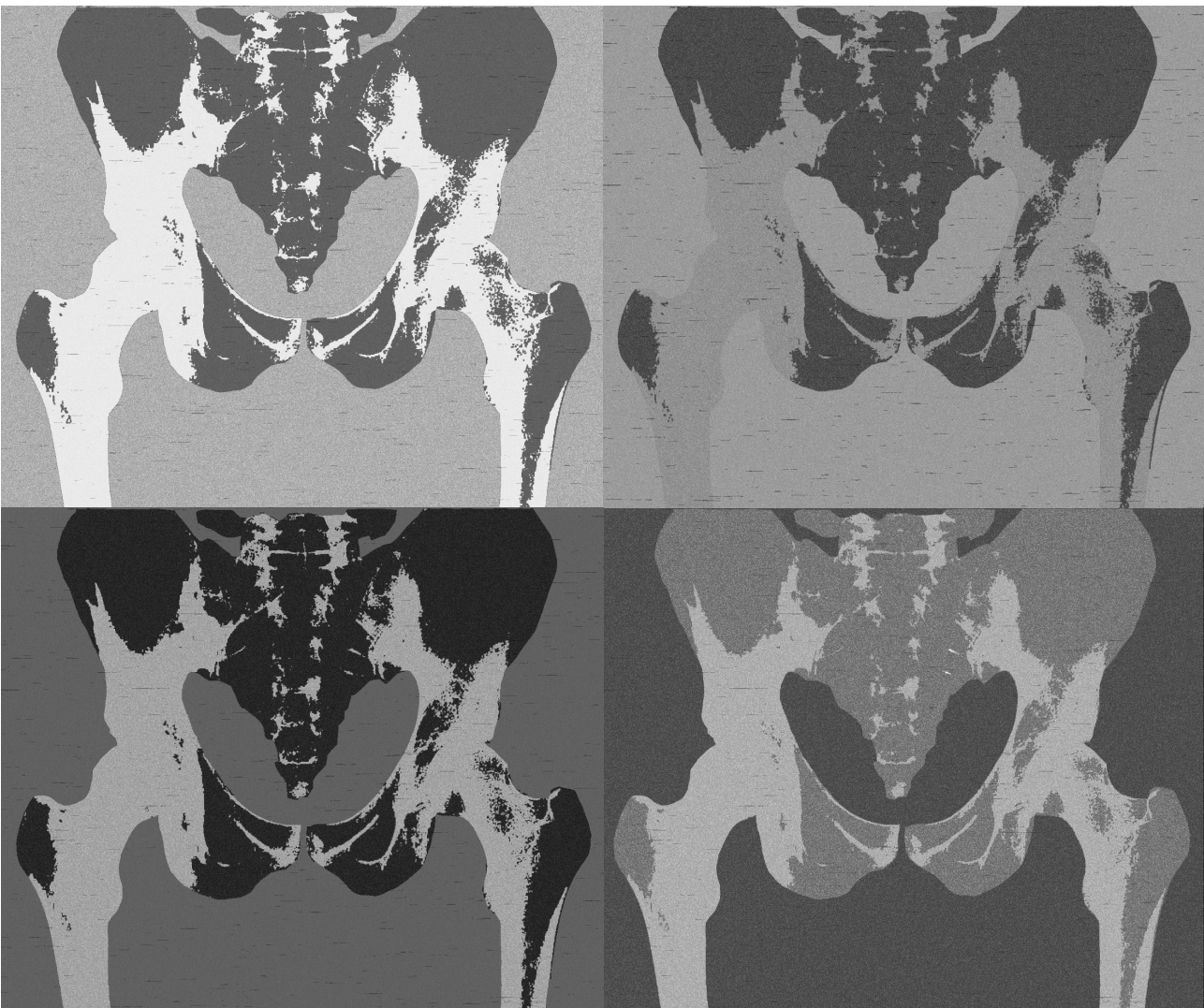
Reading of sacroiliac joints on plain pelvic radiographs: agreement between clinical practice and trained central reading. Results of the DESIR-cohort

Arthritis Rheumatol Published Online First: [6 June 2014]
doi: 10.1002/art.38738

R. van den Berg
G. Lenczner
A. Feydy
D. van der Heijde
M. Reijnierse
A. Saraux
A. Rahmouni
M. Dougados
P. Claudepierre



Abstract

Objective

Investigating agreement on presence/absence of radiographic sacroiliitis between local rheumatologists/radiologists and central trained readers (external standard).

Method

Inflammatory back pain patients (IBP; ≥3 months, <3 years) suggestive of axial spondyloarthritis (axSpA) were included in the DESIR-cohort. Baseline sacroiliac-joint-radiographs were read by two central readers (modified New York), adjudicated by a third reader in case of disagreement, yielding a positive or a negative result (central reading). The same radiographs were also read by local radiologists/rheumatologists rated 'normal', 'doubtful sacroiliitis', 'obvious sacroiliitis' or 'SI-joint fusion' (local reading); positivity defined as: 1) at least unilateral 'obvious sacroiliitis' 2) 'bilateral 'obvious sacroiliitis' or at least unilateral 'fusion''. Agreement and misclassifications between central readers and central reading versus local reading were calculated (kappas).

Results

Interreader agreement between the central readers was moderate (κ=0.54); 108/688 radiographs (15.7%) were adjudicated. According to local reading (at least unilateral 'obvious sacroiliitis'), 183/688 patients (26.6%) had sacroiliitis; according to central reading, 145/688 patients (21.1%). Agreement between local reading and central reading was also moderate (κ=0.55); 76/183 patients (41.5%) with at least unilateral 'obvious sacroiliitis' (positive by local reading) and 32/109 patients (29.4%) with 'bilateral 'obvious sacroiliitis' or at least unilateral 'fusion'' (positive by local reading) were rated 'negative' by central reading; 38/505 patients (7.5%) and 68/579 (11.7%) respectively without sacroiliitis (negative by local reading) were read 'positive' by central reading.

Conclusion

In recent onset IBP-patients, both trained readers and local rheumatologists/radiologists agree only moderately in recognizing radiographic sacroiliitis. A significant proportion of locally recognized ankylosing spondylitis patients is not confirmed by central reading (false-positive), while a small minority is false-negative, indicating the necessity of re-evaluating the role of radiographic sacroiliitis as diagnostic criterion for axSpA.

Introduction

Sacroiliitis, detected on plain radiographs, is considered as the hallmark of ankylosing spondylitis (AS) and is mandatory for the classification of AS according to the modified New York criteria [1]. However, it is known that a major delay between symptom onset and the development of radiographic sacroiliitis exists [2]. Recently, the Assessment of SpondyloArthritis international Society (ASAS) published classification criteria to be able to classify patients with non-radiographic axial SpA (nr-axSpA) in addition to AS patients [3]. The classification of nr-axSpA, which was described for the first time already 29 years ago [4], is based on the absence of radiographic sacroiliitis but presence of SpA-features such as uveitis and dactylitis complementary to the presence of HLA-B27 and/or sacroiliitis visible on MRI. Although diagnostic criteria do not exist, radiographic sacroiliitis is also playing an important role in the diagnostic process of patients suspected of having axial SpA [5]. Axial SpA comprises the entire spectrum of patients with radiographic and non-radiographic disease, sacroiliitis on the radiograph being in fact the only discriminating feature. Though, the recognition of radiographic sacroiliitis is considered difficult because of the complex anatomy of the SI-joints, and the undulating articular surface makes the SI-joints hard to image on conventional radiographs, resulting in misinterpretations [6, 7]. A study including 100 rheumatologists and radiologists showed major discrepancies in grading of the SI-joints, especially in grades 1 and 2. Unfortunately, extensive training by workshops and self-education, did not enhance the performance of diagnosing sacroiliitis [7].

The distinction between AS and nr-axSpA based on the presence/absence of radiographic sacroiliitis becomes even more evident by the fact that in many countries TNF-inhibitors (TNFi) are currently approved for patients with established AS but not for nr-axSpA patients [8]. Moreover, the European Medical Agency has approved TNFi for nr-axSpA patients only if additional signs of objective inflammation such as a positive MRI and/or an elevated CRP are present, while this is not required for patients with radiographic axSpA. So there are major consequences depending on the judgement of a pelvis radiograph.

In daily practice, the diagnosis of AS is based on the judgement of the SI-joints on plain radiographs by the local radiologist and/or rheumatologist, frequently with knowledge of the clinical signs and symptoms. In cohorts and clinical trials on the other hand, the quantitative scoring of structural damage on radiographs of the SI-joints is usually done by one or more trained readers blinded for clinical information. In the DESIR (DEvenir des Spondylarthropathies Indifférenciées Récentes)-cohort, radiographs of the SI-joints at inclusion are scored by the local rheumatologist or radiologist and also by two trained central readers, including a third reader in case of discrepancy. Therefore, this cohort offers the unique opportunity to compare the evaluation of the local reading to the centralized reading, as external standard, in terms of agreement on abnormal versus normal SI-joints permitting to diagnose radiographic sacroiliitis.

Methods

Patients

For this analysis, baseline data from the DESIR-cohort were used. The DESIR-cohort is described extensively before 9. In short, consecutive patients aged 18-50 from 25 centers in France with inflammatory back pain (IBP) in the thoracic spine, lumbar spine and/or buttock area (≥3 months, but <3 years) based on either the Calin (4/5 items) or the Berlin (2/4 items) criteria [10, 11], suggestive of axSpA according to the rheumatologist with a score of ≥5 on a scale of 0 to 10 (where 0 was not suggestive of axSpA and 10 was very suggestive of axSpA), were included in this prospective longitudinal cohort to study the natural course and prognosis of axSpA starting at symptom onset. Between December 2007 and 29th of

April 2010, 708 patients were included.
The study fulfilled Good Clinical Practice Guidelines and was approved by the appropriate medical ethical committees. Participants gave written informed consent before they were included in the study. A detailed description of the study protocol is available at the website (http://www.lacohortedesir.fr/desir-in-english/). The research proposal for this particular analysis was approved by the scientific committee of the DESIR-cohort.

Data collection

A database was built by the use of standardized Case Record Form (CRF) on which the following, among others, needed to be filled out: physical examination, on-going treatment, co-morbidities, laboratory tests and questionnaires, according to the DESIR protocol [9]. The database used for this analysis was locked on October 30th 2012.

Images and scoring methods

Two central readers (RvdB and GL), both familiar with scoring SI-joints on plain radiographs (X-SI) in the anteroposterior view according to the modified New York (mNY) method, participated in a calibration session. A grade 0 is given for a normal SI-joint; 1 for suspicious changes; 2 for minimal abnormality - small localised areas with erosions or sclerosis without alteration in joint width; 3 for unequivocal abnormality - with one or more erosions, evidence of sclerosis, joint space narrowing or widening or partial ankylosis, and grade 4 for severe abnormality - a complete ankylosis of the SI-joint. Sacroiliitis is defined as grade ≥2 bilaterally or grade 3-4 unilaterally [1]. The calibration session was a systematic conducted exercise, executed by two senior radiologists (MR and AF) and two senior rheumatologists (PC and MD), who already did such calibration sessions before. The whole process was supervised by an expert in AS and imaging scoring (DvdH). During the first step of the calibration process, definitions of lesions, examples and pitfalls were discussed. The second step of the calibration session consisted of independently reading of training cases by the two readers, under the supervision of those radiologists and rheumatologists. The results of the readings were discussed plenary by the senior radiologists and rheumatologists, focussing on disagreement regarding specific lesions between the two readers in order to achieve agreement. In the third step of the calibration process, 30 X-SIs were read independently by the two central readers. The fourth step consisted of a consensus meeting in which the same four senior radiologists/rheumatologists participated as well. Again, during a plenary presentation the disagreements between the two readers were discussed by one of the senior radiologists/rheumatologist in order to achieve agreement. Next, in a fifth step a second set consisting of 20 X-SIs were read independently, again followed by the last step consisting of a consensus meeting with the same senior radiologists/rheumatologists executed in the same manner. At that time, interreader agreement largely improved even though we recognized that the kappas were still moderate (kappa=0.55). However, considering the results of the study by van Tubergen *et al*. showing that training did not improve performance [7], we didn't expect significant further improvements in agreement. Moreover the kappas were in the same range as what have been found in other studies as well (kappas ranging between 0.12 and 0.69) [6, 12], thereby justifying the decision that the two readers could start reading the DESIR-cohort.
Baseline X-SIs were acquired according to a standardized method, provided in the DESIR protocol. All available digital baseline X-SIs of the DESIR-cohort (n=688) were read independently by the two trained central readers according to the mNY criteria, blinded for all clinical and laboratory data, as well as for the results of the local reads for the X-SIs. Agreement on fulfilllment of the mNY criteria for radiographic sacroiliitis at the patient level between the two readers was calculated and in case the readers disagreed, a radiologist

experienced in the field of SpA (MR) served as adjudicator. A radiograph of the SI-joints was marked positive for sacroiliitis if 2/3 readers agreed on the fulfilllment of mNY criteria (ie: at least unilateral grade 3 sacroiliitis, or at least bilateral grade 2 sacroiliitis), hereafter called 'central reading'. Furthermore, the two central readers marked the type of lesions they recognized (erosions, sclerosis, joint space widening/narrowing and (partial) ankylosis).
Local radiologists or rheumatologists who might have access to all clinical information and lab test results at each study center, read all available baseline radiograph of the SI-joint in their own center, hereafter called 'local reading'. Since the local readers, who are working in regular clinical practice, were not trained experts it was considered more appropriate to use a scoring system that better resembles common clinical practice than the mNY criteria do: local readers were asked to rate each SI-joint either as 'normal' or as 'doubtful sacroiliitis' or as 'obvious sacroiliitis' or as 'SI-joint fusion' [13]. No specification of the type of lesions was provided by the local readers. In this scoring method, at least a unilateral rating of 'obvious sacroiliitis' was considered sufficient to fulfill the imaging criterion of sacroiliitis. This was our primary analysis with regard to the comparison with the fulfilllment of the mNY by central reading. In a second analysis, we have further compared a rating of 'bilateral 'obvious sacroiliitis' or unilateral 'fusion'' with the fulfilllment of the mNY criteria by central reading. Finally, we have also used a stricter definition of the mNY criteria of the central reading: at least bilateral grade 3 or unilateral grade 4.
To compare the grading of the *individual* SI-joints of the central readers to the scoring of the *individual* SI-joints by the local readers, central mNY grades 2 and 3 were combined and compared to local 'obvious sacroiliitis' and mNY grade 4 was compared to local 'SI-joint fusion' for each central reader separately. Moreover, central mNY grades 3 and 4 were combined and compared to the local read 'fusion' again for each central reader separately.

Statistical analysis

Agreement was calculated using cross-tabulation expressed in Cohen's Kappa (κ) or linear weighted kappa (κ_w) as appropriate, agreement on the positive cases (positive agreement) and agreement on the negative cases (negative agreement) for the following comparisons [14,15]: interreader agreement on the presence/absence of radiographic sacroiliitis between the two central readers, between local reading and central reading using the various definitions for sacroiliitis explained above, and interreader agreement on the type of lesion (erosions, sclerosis, joint space narrowing/widening and ankylosis) between the two central readers and on the grading of the SI-joints between the two central readers and between local reading and central reading as explained above. Sensitivity, specificity, positive predictive values (PPV), negative predictive values (NPV) as well as the number of false-positive and false-negative classifications by local reading versus central reading as external standard were calculated.
Among the patients with a positive X-SI according to local reading, it was investigated which type of lesion was most frequently scored by the central readers separately.
All kappas were interpreted according to the standards proposed by Landis and Koch; values <0 as indicating no agreement and 0-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as almost perfect agreement [16]. SPSS software version 20.0 was used for the statistical analysis.

Results

The mean age of the 688 included patients was 31.6 (SD 8.6) years, the mean symptom duration was 17.7 (SD 10.5) months, 368 (53.5%) patients were men and 405 (58.9%) patients were HLA-B27 positive. In 648/688 patients imaging data was complete; in 40 additional patients data on MRI-SI was missing. In 582/648 patients with available imaging

data the ASAS axSpA criteria could be applied. The remaining 66 patients had an onset of IBP >45 years of age. Based on local scoring of imaging for sacroiliitis, 408/582 patients (70.1%) fulfilled the ASAS axSpA criteria (84.1% HLA-B27 positive). One hundred sixty-one fulfilled the ASAS axSpA criteria based on the presence of radiographic sacroiliitis (at least unilateral 'obvious sacroiliitis'; 75.8% HLA-B27 positive); 83/408 based on inflammatory sacroiliitis on MRI as judged by the local reading (67.5% HLA-B27 positive) and the remaining 164/408 patients fulfilled the ASAS axSpA criteria based on HLA-B27 positivity.

Table 1: Interreader agreement between central reader 1 and central reader 2 in SI-joints radiographs reading (n=688 patients).

		Central reader 2	
		modified New York +	**modified New York –**
Central reader 1	**modified New York +**	96	58
	modified New York –	50	484
	κ (95% CI): 0.54 (0.46-0.62)	Positive agreement: 64.0%	Negative agreement: 90.0%
		Erosions +	**Erosions –**
	Erosions +	132	131
	Erosions –	133	980
	κ (95% CI): 0.38 (0.32-0.44)	Positive agreement: 50.0%	Negative agreement: 88.1%
		Sclerosis +	**Sclerosis –**
	Sclerosis +	182	144
	Sclerosis –	133	917
	κ (95% CI): 0.44 (0.38-0.49)	Positive agreement: 56.8%	Negative agreement: 86.9%
		Joint space widening +	**Joint space widening –**
	Joint space widening +	13	137
	Joint space widening –	10	1216
	κ (95% CI): 0.13 (0.06-0.19)	Positive agreement: 15.0%	Negative agreement: 94.3%
		Joint space narrowing +	**Joint space narrowing –**
	Joint space narrowing +	14	127
	Joint space narrowing –	21	1214
	κ (95% CI): 0.12 (0.05-0.20)	Positive agreement: 15.9%	Negative agreement: 94.3%
		Ankylosis +	**Ankylosis –**
	Ankylosis +	27	63
	Ankylosis –	67	1219
	κ (95% CI): 0.24 (0.15-0.33)	Positive agreement: 29.3%	Negative agreement: 94.9%

Not possible to evaluate: n=27 for right SI-joint and n=29 for left SI-joint. Positive agreement is the agreement on positive cases. Negative agreement is the agreement on negative cases.

Agreement between the two central readers

Agreement between the two central readers regarding absence/presence of radiographic sacroiliitis (mNY) is moderate (κ=0.54; table 1). The adjudicator needed to read 108/688 X-rays (15.7%) because of disagreement between the 2 central readers.

Agreement regarding the grading of the SI-joints is also moderate (κ_W=0.56 for both left and right SI-joints; table 2). Most disagreement is seen in grade 0 versus grade 1, followed by grade 0 versus grade 2, and followed by grade 1 versus grade 2. Similar numbers of disagreement are seen for grade 2 versus grade 3 as for grade 1 versus grade 2. Depending on the grade of the other SI-joint, this could cause a different classification of a patient. The remaining numbers of disagreement are seen in grade 0 versus grade 3 and, and in grade 1 versus grade 3. With these types of disagreement, a patient is classified differently regardless of the grade of the other SI-joint of the patient.

Table 2: Interreader agreement of the grading of the SI-joints (mNY) between central reader 1 and central reader 2.

		Central reader 2				
Central reader 1	**Right SI-joint**	**Grade 0**	**Grade 1**	**Grade 2**	**Grade 3**	**Grade 4**
	Grade 0	345	51	35	4	0
	Grade 1	28	14	17	5	0
	Grade 2	26	16	40	15	0
	Grade 3	7	6	19	45	0
	Grade 4	0	0	0	5	2
	κ_W (95% CI): 0.56 (0.50-0.61)					
		Central reader 2				
Central reader 1	**Left SI-joint**	**Grade 0**	**Grade 1**	**Grade 2**	**Grade 3**	**Grade 4**
	Grade 0	369	48	22	6	0
	Grade 1	15	13	14	10	0
	Grade 2	24	26	33	20	0
	Grade 3	8	7	23	34	0
	Grade 4	0	0	0	1	5
	κ_W (95% CI): 0.56 (0.51-0.62)					

Not possible to evaluate: n=27 for right SI-joint and n=29 for left SI-joint.

Agreement between local reading and central reading

According to local reading (at least unilateral 'obvious sacroiliitis') 183/688 patients (26.6%) had radiographic sacroiliitis and according to central reading 145/688 patients (21.1%) had radiographic sacroiliitis. Agreement between local reading and central reading was very similar to the interreader agreement between the two central readers (κ=0.55). Comparing local reading to the scores of the individual central readers revealed very similar levels of agreement (table 3).

In the comparison of local reading to central reading, 76 out of the 183 patients (41.5%) with a positive radiograph of the SI-joints according to local reading were read negative by central reading; 38 patients out of 505 (7.5%) negative radiographs of the SI-joints according to local reading were read positive by central reading (table 3). In daily practice, where local readers judge the X-SIs, this would mean that 41.5% of the AS patients are overclassified compared to central reading as external standard (45.9% compared to reader 1 and 42.1% compared to reader 2) and that AS (according to central reading) is not recognized in 7.5% of the patients compared to central reading as external standard (10.9% compared to reader 1 and 7.9% compared to reader 2).

Seventy-six of the 183 patients with 'obvious sacroiliitis' according to local reading had unilateral 'obvious sacroiliitis'; thirty of these were rated as 'negative' by central reading. If only patients with at least bilateral 'obvious sacroiliitis' or at least unilateral 'fusion' were considered 'positive' by local reading, still 32/109 patients (29.4%) were rated as 'negative' by central reading (according to the original mNY; κ=0.52). The proportion of negative ratings according to this definition of local reading that were rated 'positive' by central reading increased to 11.7% (68/579).

Thereafter we compared this more stringent definition of positivity by local reading with a more stringent definition of positivity by central reading (at least grade 3 bilateral or grade 4 unilateral involvement). This resulted in reduced agreement (κ=0.44 for reader 1 and κ=0.43 for reader 2), showing even a higher number of patients being overclassified as AS according to local reading compared to the central readers (62.4% and 65.1% reader 1 and 2 respectively). The interreader agreement between the two central readers for this stricter mNY definition of sacroiliitis was again only moderate (κ=0.56), but better than the agreement between local reading and the individual central readers. Even with the most stringent definition of sacroiliitis ('positivity' defined as bilateral fusion in the local reading and bilateral grade 4 in the central reading) still huge disagreement between local reading and the individual central readers was seen (table 3).

Table 4 shows the comparison of the individual ratings of local reading and the mNY gradings per SI-joint of central reading on a individual reader basis. For the determination of agreement, we have both combined mNY grades 2 and 3 (for 'positivity') as well as mNY grades 3 and 4, and compared these with 'obvious sacroiliitis' from the local read. The kappas were only moderate (range 0.36-0.51) with regard to both readers and both definitions.

Table 3: Agreement between local reading and central reading of presence/absence of sacroiliitis, for the various definitions of sacroiliitis.

		Central reader 1	
Local reading (at least unilateral 'obvious sacroiliitis')		**Sacroiliitis +**	**Sacroiliitis -**
	Sacroiliitis +	99	84
	Sacroiliitis -	55	450
	κ (95% CI): 0.46 (0.38-0.53)	Pos. agreem.: 58.8%	Neg. agreem.: 86.6%
		Central reader 2	
		Sacroiliitis +	**Sacroiliitis -**
	Sacroiliitis +	106	77
	Sacroiliitis -	40	465
	κ (95% CI): 0.53 (0.46-0.61)	Pos. agreem.: 64.4%	Neg. agreem.: 88.8%
		Central reading (2/3 readers)	
		Sacroiliitis +	**Sacroiliitis -**
	Sacroiliitis +	107	76
	Sacroiliitis -	38	467
	κ (95% CI): 0.55 (0.47-0.62)	Pos. agreem.: 65.2%	Neg. agreem.: 89.1%
Sensitivity: 73.8%	Specificity: 86.0%	NPV: 0.92	PPV: 0.58

Table 3: Continued

		Central reader 1	
Local reading (at least bilateral 'obvious sacroiliitis' or unilateral 'fusion')		**Sacroiliitis +**	**Sacroiliitis -**
	Sacroiliitis +	74	35
	Sacroiliitis -	80	499
	κ (95% CI): 0.46 (0.38-0.55)	Pos. agreem.: 56.3%	Neg. agreem.: 89.7%
		Central reader 2	
		Sacroiliitis +	**Sacroiliitis -**
	Sacroiliitis +	76	33
	Sacroiliitis -	70	509
	κ (95% CI): 0.51 (0.42-0.59)	Pos. agreem.: 59.6%	Neg. agreem.: 90.8%
		Central reading (2/3 readers)	
		Sacroiliitis +	**Sacroiliitis -**
	Sacroiliitis +	77	32
	Sacroiliitis -	68	511
	κ (95% CI): 0.52 (0.44-0.60)	Pos. agreem.: 60.6%	Neg. agreem.: 91.1%
		Reader 1, strict definition (≥3 bilaterally, or 4 unilaterally)	
Local reading (at least bilateral 'obvious sacroiliitis' or unilateral 'fusion')		**Sacroiliitis +**	**Sacroiliitis -**
	Sacroiliitis +	41	68
	Sacroiliitis -	15	564
	κ (95% CI): 0.44 (0.34-0.53)	Pos. agreem.: 49.7%	Neg. agreem.: 93.1%
		Reader 2, strict definition (≥3 bilaterally, or 4 unilaterally)	
		Sacroiliitis +	**Sacroiliitis -**
	Sacroiliitis +	38	71
	Sacroiliitis -	10	569
	κ (95% CI): 0.43 (0.33-0.53)	Pos. agreem.: 48.4%	Neg. agreem.: 93.4%
		Reader 1, strict definition (≥4 bilaterally)	
Local reading (bilateral 'fusion')		**Sacroiliitis +**	**Sacroiliitis -**
	Sacroiliitis +	2	3
	Sacroiliitis -	3	680
	κ (95% CI): 0.40 (0.01-0.78)	Pos. agreem.: 40.0%	Neg. agreem.: 99.6%
		Reader 2, strict definition (≥4 bilaterally)	
		Sacroiliitis +	**Sacroiliitis -**
	Sacroiliitis +	0	5
	Sacroiliitis -	1	682
	κ (95% CI): 0.00 (0.00-0.00)	Pos. agreem.: 0.00%	Neg. agreem.: 99.6%

Pos. agreem. is the agreement on positive cases. Neg. agreem. is the agreement on negative cases.

Sensitivity, specificity, NPV (negative predictive value) and PPV (positive predictive value) applies to the comparison of the local reading (at least unilateral 'obvious sacroiliitis') to the central reading (≥2 bilaterally, or ≥3 unilaterally, 2/3 readers).

Table 4: Agreement of the grading of the SI-joints between local reading and the individual central readers

		Central reader 1				
Local read	**Right SI-joint**	**Grade 0**	**Grade 1**	**Grade 2***	**Grade 3*^**	**Grade 4^**
	Normal	321	36	37	15	0
	Doubtful sacroiliitis	80	16	28	10	1
	Obvious sacroiliitis	34	10	31	49	3
	Fusion	0	2	1	3	3
	* Grade 2 and 3 of central reader 1 combined: κ_w (95% CI): 0.40 (0.34-0.46) ^ Grade 3 and 4 of central reader 1 combined: κ_w (95% CI): 0.36 (0.30-0.41)					
		Central reader 1				
Local read	**Left SI-joint**	**Grade 0**	**Grade 1**	**Grade 2***	**Grade 3*^**	**Grade 4^**
	Normal	336	21	36	15	0
	Doubtful sacroiliitis	77	12	24	9	1
	Obvious sacroiliitis	31	18	42	44	1
	Fusion	0	1	2	4	4
	* Grade 2 and 3 of central reader 1 combined: κ_w (95% CI): 0.45 (0.38-0.51) ^ Grade 3 and 4 of central reader 1 combined: κ_w (95% CI): 0.41 (0.35-0.46)					
		Central reader 2				
Local read	**Right SI-joint**	**Grade 0**	**Grade 1**	**Grade 2***	**Grade 3*^**	**Grade 4^**
	Normal	324	46	32	10	0
	Doubtful sacroiliitis	64	27	33	10	1
	Obvious sacroiliitis	21	14	45	50	0
	Fusion	1	1	1	5	1
	* Grade 2 and 3 of central reader 2 combined: κ_w (95% CI): 0.48 (0.42-0.54) ^ Grade 3 and 4 of central reader 2 combined: κ_w (95% CI): 0.47 (0.42-0.53)					
		Central reader 2				
Local read	**Left SI-joint**	**Grade 0**	**Grade 1**	**Grade 2***	**Grade 3*^**	**Grade 4^**
	Normal	330	50	23	7	0
	Doubtful sacroiliitis	60	29	25	12	1
	Obvious sacroiliitis	28	17	45	48	1
	Fusion	0	0	0	8	3
	* Grade 2 and 3 of central reader 2 combined: κ_w (95% CI): 0.51 (0.45-0.57) ^ Grade 3 and 4 of central reader 2 combined: κ_w (95% CI): 0.44 (0.39-0.49)					

Not possible to evaluate: n=27 for right SI-joint and n=29 for left SI-joint. Positive agreement is the agreement on positive cases. Negative agreement is the agreement on negative cases.

Types of lesions

Regarding the type of lesion, agreement between the two central readers varied from κ=0.12 for joint space narrowing to κ=0.44 for sclerosis (table 5). The prevalence of joint space alterations and ankylosis is low in this cohort of recent onset IBP patients; among the SI-joints graded as at least grade 3 the prevalence of joint space widening was 8.7% (reader 2) and 24.6% (reader 1) and the prevalence of ankylosis was 13.7% (reader 1) and 27.3% (reader 2). The frequency of erosions and sclerosis in SI-joints graded as 2 was very similar to the frequency of erosions and sclerosis in SI-joints graded as 3, for both readers.
Among the patients with a positive radiograph of the SI-joints according to local reading (at least unilateral 'obvious sacroiliitis'), sclerosis was the most frequently reported lesion by the central readers (56.8% according to reader 1 and 72.7% according to reader 2), followed by erosions (50.3% according to reader 1 and 61.2% according to reader 2) (table 5).

Table 5: Frequency of lesions read by the central reader 1 and central reader 2 among patients with a positive radiograph of the SI-joints according to local reading (at least unilateral 'obvious sacroiliitis') (n=183).

Type of lesion (%)	Reader 1	Reader 2
Erosions	50.3	61.2
Sclerosis	56.8	72.7
Joint space widening	24.6	8.7
Joint space narrowing	24.0	11.5
Ankylosis	13.7	27.3

Discussion

In the DESIR-cohort, the interreader agreement between two trained central readers and between local reading and central reading of X-SIs is moderate at best. The two central trained readers showed only moderate agreement with regard to presence/absence of radiographic sacroiliitis, grading of the SI-joints and about type of lesion, yet comparable to levels of agreement reported in previous studies where central readers were also trained (κ=0.12-0.69 for absence/presence of sacroiliitis; κ=0.22 for erosions; κ=0.26 for sclerosis and κ=0.19 for joint space alterations) [6, 12, 17]. Furthermore, interreader agreement was at a similar level as the agreement between local reading and central reading about presence/absence of radiographic sacroiliitis and grading of the SI-joints.
As it was not specified which X-SIs were judged by local radiologists and which by local rheumatologists, it was not possible to compare the readings of radiologists and rheumatologists separately to the central reading. However, we did not expect a difference in number of misclassified patients between radiologists and rheumatologists based on the findings of van Tubergen *et al.*, and based on our own findings regarding presence/absence of sacroiliitis on MRI in the DESIR-cohort [7, 18]. Moreover, the results of an unpublished ASAS survey pointed out that more than 55% of the rheumatologists rely on both the judgement of the radiologist as well as their own judgement in assessing sacroiliitis on radiographs (M. Rudwaleit, personal communication, January 18 2014, unpublished data from an ASAS survey).
Misclassifications could have major implications for a patient, as the presence/absence of sacroiliitis is the only difference in the classification of AS (mNY) versus nr-axSpA or even no SpA. This is indeed what the results show; 41.5-45.9% of the patients classified as AS by the local readers are falsely classified with the central read as external standard and 7.5-10.9% of the AS patients according to the central read is not recognized in daily practice.

The percentages of misclassified patients are somewhat higher than reported in another study where 11.4% of AS patients according to local readers were reclassified as nr-axSpA by central trained readers, and 15.5% of nr-axSpA patients according to local readers were reclassified as AS by central readers [19]. However, these lower percentages can probably be explained by the fact that this study also included patients with longstanding AS (>10 years) showing more severe lesions which are easier to recognize, the fact that the presence/absence of syndesmophytes in the spine was taken into account as well, and by the fact that both readers had to agree on the absence/presence of sacroiliitis thereby not looking at a possible reclassification of patients in whom the two central readers disagreed regarding the presence/absence of sacroiliitis [19].

As our aim was to compare the diagnostic performance of readers in daily clinical practice to the performance of trained expert readers using the mNY grading system for the classification of patients in studies, we have applied a "daily practice definition" of sacroiliitis: in daily clinical practice, a rheumatologist will consider a diagnosis of axial SpA when there is 'obvious sacroiliitis' at least unilaterally, but obviously the rheumatologist is more convinced of a diagnosis of axial SpA in case of bilateral involvement. Here we have examined both definitions of 'positivity': one more lenient and one more stringent definition. If 'at least unilateral 'obvious sacroiliitis'' in the local read was required for positivity, 41.5% of the patients with radiographic sacroiliitis could not be confirmed by central reading (using the original mNY definition). However, even if 'bilateral 'obvious sacroiliitis' or at least unilateral 'fusion'' by the local reading was required for positivity, still 29.4% of the patients with a local diagnosis of AS could not be confirmed by central reading. This percentage of misclassification increased to 62.4% (reader 1) and 65.1% (reader 2) if we compared the local rating of either 'bilateral 'obvious sacroiliitis' or unilateral 'fusion'' with a similarly strict definition of sacroiliitis by central reading (at least grade 3 bilaterally or unilateral grade 4).

As the local and central reading are not identical, the true percentage of misclassified patients must be between 29.4% and 65.1%, and likely around 40-45% as the use of 'at least unilateral 'obvious sacroiliitis'' by the local reader is what is required in clinical practice to classify a patient as having AS. While this study has been performed in the DESIR centers in France, and generalizability is formally restricted, there is no valid reason to assume that clinical rheumatologists in other countries apply different diagnostic reasoning.

As the local readers did not mark a specification of the type of lesions, the reads of the central readers were used to gain insight in which type of lesion was best recognized by the local readers. However, because of the low prevalence of joint space alterations and ankylosis it is difficult to investigate the agreement on recognizing this type of lesion. Yet, if joint space alterations are present, the two central readers recognized ankylosis more easily than widening or narrowing of the joint space. The prevalence of sclerosis and erosions, on the other hand, is higher and both types of lesions are more easily recognized than joint space alterations as shown by the higher kappas.

Although the question whether training improves recognition of radiographic sacroiliitis was not addressed directly in this study, the fact that the agreement between local reading and central reading is so similar to the interreader agreement between the two trained central readers seems to confirm the findings of van Tubergen *et al.* [7] that training does not improve recognition of radiographic sacroiliitis. This arises the question whether it is necessary for DESIR and similar cohorts to have a central reading of the radiographs of the SI-joints instead of a local reading. It could be argued based on the discovered levels of agreement that there is no preference for central readers over local readers since both trained central readers and local readers can only poorly recognize radiographic sacroiliitis. Nonetheless, central reading consisting of a judgement of 2 out 3 agreeing readers suggests being more robust than a local reading based on the opinion of a single reader. Depending on the research question, a choice could be made which reads to use: either the reads that would have been

used in clinical practice as well, or the reads of the central reading based on a majority read of 2 out of 3, which are closer to the truth.
For both local and trained central readers, the recognition of radiographic sacroiliitis remains challenging. Nevertheless, the only difference between AS and nr-axSpA is the presence of radiographic sacroiliitis. A patient can be classified as AS if only IBP is present in addition to radiographic sacroiliitis while in the absence of radiographic sacroiliiitis and in the absence of a positive MRI (ASAS definition) a minimum of 3 other SpA-features must be present in order to classify the patient as axSpA [3]. The fact that a patient is classified differently, based on a different read of the same radiograph of the SI-joints - which is shown to happen frequently - arises the question how 'gold' this distinction between AS and nr-axSpA is. It is worrisome that such a small factor can have major consequences for a patient, not only in terms of diagnosis but also in terms of treatment as based on the presence/absence of radiographic sacroiliitis TNFi can be administered or not. Moreover, inclusion of patients in clinical trials is based on the presence/absence of radiographic sacroiliitis as judged by either local or central readers. Rereading the radiographs by different readers or even blinded rereading of the same radiographs by the same readers could lead to significant change in classification of the patients [6].
A limitation of this study is the lack of a gold standard as CT, to confirm the presence/absence of sacroiliitis [20]. Furthermore, this study focussed on sacroiliitis on conventional radiographs only. The role of MRI in the diagnosis and classification of axSpA should be investigated in more detail as well as its correlation to conventional radiographs.
In conclusion, in patients with recent onset IBP, individually trained central readers disagree as much as clinical practice local rheumatologists/radiologists in recognizing radiographic sacroiliitis. While the two central readers disagree with each other in a balanced manner (disagreement in both directions, reflecting measurement error), the local readers primarily overrate sacroiliitis in comparison with central readers, which results in an unacceptably high percentage of false-positive diagnoses of AS. A small minority of patients with a classification of AS according to central reading is not recognized in daily clinical practice. Independently of the precise definition of sacroiliitis, the disagreement regarding the presence/absence of sacroiliitis is so significant that the role of radiographic sacroiliitis as a diagnostic criterion for axSpA should be re-evaluated.

References

1. van der Linden S, Valkenburg H, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.
2. Feldtkeller E, Khan M, van der Heijde D, *et al.* Age at disease onset and diagnosis delay in HLA-B27 negative vs. positive patients with ankylosing spondylitis. Rheumatol Int 2003;23:61-6.
3. Rudwaleit M, van der Heijde D, Landewé R, *et al.* The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part II): validation and final selection. Ann Rheum Dis 2009;68:777-83.
4. Khan M, van der Linden S, Kushner I, *et al.* Spondylitic disease without radiologic evidence of sacroiliitis in relatives of HLA-B27 positive ankylosing spondylitis patients. Arthritis Rheum 1985;28:40-3.
5. van den Berg R, de Hooge M, Rudwaleit M, *et al.* ASAS modification of the Berlin algorithm for diagnosing axial spondyloarthritis: results from the SPondyloArthritis Caught Early (SPACE)-cohort and from the Assessment of SpondyloArthritis international Society (ASAS)-cohort. Ann Rheum Dis 2013;72:1646-53.
6. Spoorenberg A, de Vlam K, van der Linden S, *et al.* Radiological scoring methods in ankylosing spondylitis. Reliability and change over 1 and 2 years. J Rheumatol 2004;31:125-32.
7. van Tubergen A, Heuft-Dorenbosch L, Schulpen G, *et al.* Radiographic assessment of sacroiliitis by radiologists and rheumatologists: does training improve quality? Ann Rheum Dis 2003;62:519-25.
8. van den Berg R, Stanislawska-Biernat E, van der Heijde DM. Comparison of recommendations for the use of anti-tumour necrosis factor therapy in ankylosing spondylitis in 23 countries worldwide. Rheumatology (Oxford) 2011;50:2270-7.
9. Dougados M, D'Agostino M, Benessiano J, *et al.* The DESIR-cohort: a 10-year follow-up of early inflammatory back pain in France: study design and baseline characteristics of the 708 recruited patients. Joint Bone Spine 2011;78:598-603.
10. Calin A, Porta J, Fries J, *et al.* Clinical history as a screening test for ankylosing spondylitis. JAMA 1977;237:2613-4.
11. Rudwaleit M, Metter A, Listing J, *et al.* Inflammatory back pain in ankylosing spondylitis: a reassessment of the clinical history for application as classification and diagnostic criteria. Arthritis Rheum 2006;54:569-78.
12. Poddubnyy D, Gaydukova I, Hermann K, *et al.* Magnetic resonance imaging compared to conventional radiographs for detection of chronic structural changes in sacroiliac joints in axial spondyloarthritis. J Rheumatol 2013;40:1557-65.
13. Blachier M, Coutanceau B, Dougados M, *et al.* Does the site of magnetic resonance imaging abnormalities match the site of recent-onset inflammatory back pain? The DESIR-cohort. Ann Rheum Dis 2013;72:979-85.
14. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70(4):213-20.
15. Dice L. Measures of the amount of ecologic association between species. Ecology 1945;26:297-302.
16. Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
17. http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/ArthritisAdvisoryCommittee/UCM366519.pdf. (24 December 2013, date last accessed).
18. van den Berg R, Lenczner G, Thévenin F, *et al.* Concordance between 'a positive MRI of the sacroiliac joints' based on the local reading versus a centralised reading: experience from the DESIR-cohort. Arthritis Rheum 2013;65[10]:S475.
19. Rudwaleit M, Haibel H, Baraliakos X, *et al.* The early disease stage in axial spondylarthritis: results from the German Spondyloarthritis Inception Cohort. Arthritis Rheum 2009;60:717-27.
20. Geijer M, Gothlin G, Gothlin J. The clinical utility of computed tomography compared to conventional radiography in diagnosing sacroiliitis. A retrospective study on 910 patients and literature review. J Rheumatol 2007;34:1561-5.