Which scoring method depicts spinal radiographic damage in early axial spondyloarthritis best? Five-year results from the DESIR cohort

S. Ramiro, P. Claudepierre, A. Sepriano, M. van Lunteren, A. Molto, A. Feydy, M.A. d'Agostino, D. Loeuille, M. Dougados, M. Reijnierse, D. van der Heijde

Sofia Ramiro, MD, PhD
Department of Rheumatology, Leiden University Medical Center, Leiden, the Netherlands.
Zuyderland Medical Center, Heerlen, the Netherlands
sofiaramiro@gmail.com

Pascal Claudepierre, MD
Department of Rheumatology, Henri Mondor Hospital, APHP, Créteil, France
Université Paris Est Créteil, EA 7379 – EpidermE, F-94010, Créteil, France.
pascal.claudepierre@aphp.fr

Alexandre Sepriano, MD
Department of Rheumatology, Leiden University Medical Center, Leiden, the Netherlands.
NOVA Medical School, Universidade Nova de Lisboa, Portugal.
alexsepriano@gmail.com

Miranda van Lunteren, Msc
Department of Rheumatology, Leiden University Medical Center, Leiden, the Netherlands.
m.van_Lunteren@lumc.nl

Anna Molto, MD, Msc, PhD
Department of Rheumatology, Paris Descartes University, Hôpital Cochin, Hôpitaux de Paris, Paris, France.
INSERM (U1153), Clinical Epidemiology and Biostatistics, PRES Sorbonne Paris-City, Paris, France.
anna.molto@aphp.fr

Antoine Feydy, MD, PhD
Department of Radiology, Paris Descartes University, Paris, France
antoine.feydy@aphp.fr

Maria Antonietta d'Agostino, MD, PhD
Department of Rheumatology, Ambroise Paré Hospital APHP, Boulogne-Billancourt, France
INSERM U1173, Laboratoire d'Excellence INFLAMEX, UFR Simone Veil, Université Versailles-Saint Quentin en Yvelines, 78180 Saint-Quentin en Yvelines, France.
maria-antonietta.dagostino@apr.aphp.fr

Damien Loeuille, MD, PhD
Department of Rheumatology, University of Nancy, Nancy, France
d.loeuille@chru-nancy.fr

Maxime Dougados, MD, Professor

Department of Rheumatology, Paris Descartes University, Hôpital Cochin, Hôpitaux de Paris, Paris, France
INSERM (U1153), Clinical Epidemiology and Biostatistics, PRES Sorbonne Paris-City, Paris, France
maxime.dougados@aphp.fr

Monique Reijnierse, MD, PhD
Department of Radiology, Leiden Univeristy Medical Center, Leiden, the Netherlands. m.reijnierse@lumc.nl

Désirée van der Heijde, MD, PhD, Professor
Department of Rheumatology, Leiden University Medical Center, Leiden, the Netherlands. mail@dvanderheijde.nl

Corresponding author:
Sofia Ramiro, MD, PhD
Department of Rheumatology, Leiden University Medical Center
P.O. Box 9600, 2300RC Leiden, the Netherlands
Telephone: +31 71 526 32 65
E-mail: sofiaramiro@gmail.com

Keywords
spondyloarthritis, radiology, epidemiology, outcome measures

Key messages

- The mSASSS and RASSS capture most structural change in early axSpA.
- There is no gain in scoring the thoracic spine in axSpA while noise is introduced.
- The mSASSS is the radiographic damage score depicting damage best in axSpA.

Abstract

Objectives: To compare the performance of different spinal radiographic damage scoring methods in patients with early axial spondyloarthritis (axSpA).

Methods: Five-year spinal radiographs from the DESIR cohort were scored by 3 readers (averaged) for the calculation of the SASSS, mSASSS, RASSS, BASRI-spine and BASRI-total, and following the OMERACT filter, scores were compared according to truth, discrimination (reliability and sensitivity to change) and feasibility. The proportion of patients with a net change>SDC and >1 was calculated. The proportion of total variance explained by the patient ('true variance') was calculated for the change scores as a measure of reliability, using ANOVA.

Results: In total 699 patients were included. Five-year net changes >SDC (>1) were: RASSS 17% (17%), mSASSS 12% (12%), BASRI-spine and BASRI-total 12% (9%), SASSS 11% (11%). The mSASSS and the RASSS performed the best in terms of capturing the signal (positive change) related to noise (negative change). The proportion of variance explained by the patient was highest for the mSASSS and RASSS (85% for both 5-year progression scores vs 50-55% for other methods). The proportion of patient variance in the thoracic segment of the RASSS was unsatisfactory (46% for progression).

Conclusions: The existing scoring methods to assess spinal radiographic damage performed well in early phases of axSpA. The mSASSS and RASSS captured most change. There was no clear gain in additionally scoring the thoracic spine for the RASSS. The mSASSS remains the most sensitive and valid scoring method in axSpA, including early phases of the disease.

Introduction

Structural damage is a core outcome in inflammatory rheumatic diseases and therefore included in all core sets of outcome domains and measures.[1-3] Structural damage is, across different inflammatory diseases and in the particular case of axial spondyloarthritis (axSpA), the best predictor of further damage and therefore a bad prognostic factor that needs to be objectively assessed when making treatment decisions.[4]

Several scoring methods capturing spinal radiographic damage have been developed in radiographic axial spondyloarthritis (r-axSpA). In chronological order these are: 1) the Stoke AS Spine Score (SASSS);[5] 2) the Bath AS Radiology Index (BASRI), with a score involving the spine only, the BASRI spine,[6] and 3) another one also including the hips, the BASRI total;[7] 4) modification of the SASSS to include the cervical column, the mSASSS;[8] 5) and a modification of the mSASSS, the Radiographic AS Spinal Score (RASSS)[9], including the lower part of the thoracic spine, under the hypothesis that most progression would occur in that segment. These scoring methods have been compared concerning their truth, discrimination, and feasibility, according to the Outcome Measures in Rheumatology (OMERACT) filter, and the mSASSS has been considered the most appropriate method, i.e. the most valid and sensitive to change, to assess radiographic damage in r-axSpA.[10-13]

So far, these scoring methods have not been assessed in early forms of the disease, namely in patients without radiographic sacroiliitis (i.e. non-radiographic axSpA (nr-axSpA)). To gain further insight into the development of structural damage and, particularly, into how to prevent or reduce progression, it is important to capture this in early stages of the disease. Moreover, axSpA is nowadays seen as a single disease, and it makes sense to analyse the performance of radiographic scoring methods in its whole spectrum, including both r-axSpA and nr-axSpA.[14, 15]

The importance of the assessment of structural damage is emphasized when investigating the efficacy of an intervention. Demonstrating a disease modifying effect, in principle, implies showing inhibition of structural damage. In axSpA this has proven to be a methodological challenge, with several studies throughout the last decade pointing towards it. Several factors can contribute to this, including the slow rate of radiographic progression in axSpA (even in r-axSpA) or the low sensitivity to change of the scoring methods used, particularly in a context of a low progression.[16] These aspects emphasize the importance of identifying the method that most efficiently captures radiographic progression.

The aim of this study was to compare the performance of different spinal radiographic damage scoring methods in patients with early axSpA taking the three aspects of the OMERACT filter into account: feasibility, truth and discrimination.

Methods

Patients and radiographs

Patients from the previously described DESIR cohort were included.[17] In brief, DESIR is a cohort of 708 patients presenting with inflammatory back pain with ≥3 months but <3 years duration and with a high suspicion of axSpA. Following protocol, radiographs of the whole spine, pelvis (with hips) were performed and those from baseline, 2 and 5 years were read in the same reading campaign. Patients were included in this analysis provided they had at least one observation with at least one scoring method available. DESIR has been approved by the appropriate ethical committees and patients signed the informed consent upon participation.

Scoring methods

The existing 5 radiographic scoring methods were used, as well as 2 additional modifications of the BASRI scores to include the thoracic segment (Online Supplementary Table 1).

In the SASSS the anterior and posterior vertebral corners (VCs) of the lumbar spine (lower border of T12 to upper border of S1, total: 24 VCs) are scored, at a lateral view, for the presence of erosion and/or sclerosis and/or squaring (1 point), syndesmophyte (2 points) and bridging syndesmophyte (3 points).[5] The mSASSS is a modification of the SASSS including only anterior VCs of the cervical (lower border of C2 to upper border of T1) and lumbar (same as SASSS) segments (total: 24 VCs), with the same scoring rules and a total score from 0 to 72[8]. The RASSS, ranging from 0 to 84, is similarly scored to the mSASSS with 3 modifications: 1) inclusion of the lower thoracic spine (lower border of T10 to upper border of T12; total: 28 VCs); 2) erosions are not scored; 3) squaring is not scored in the cervical spine.[9] The BASRI-spine includes the sacroiliac joints (SIJ) (according to the New York criteria) and the lumbar and cervical segments.[6] Each spinal segment receives an overall score: 0=no change; 1=suspicious; 2=mild; 3=moderate and 4=severe. For the lumbar spine the view (lateral or anteroposterior) with the highest damage is used for the score, which ranges from 0 to 12. An adaptation of this score was used in the current study by adding an overall assessment of the thoracic spine, with the same scoring rules per segment, so that the final score (BASRI-spine-thoracic) varied between 0 and 16. The BASRI-total is similar to BASRI-spine, with an additional assessment of the hips (0=no change to 4=severe), resulting in a final score between 0 and 16.[7] Similar to the BASRI-spine, a modification was proposed to include the thoracic segment, the BASRI-total-thoracic (range 0-20).

The radiographs were independently scored according to all scoring methods in one reading campaign, by three trained experts (3 readers for spine and 3 readers for SIJ, one of them being the same for both modalities) who were blinded to chronological order, clinical characteristics and other imaging data.

Final scores per method were only calculated when at least three quarters of each segment had a score available.[11, 12] Individual missing VCs were imputed following a previously described method (details in Online Supplementary Text 1).[12] Averaged scores of the three readers VC/segment were calculated and the final sum scores computed.

Comparison of scoring methods following the OMERACT filter

Feasibility

The feasibility aspect of the OMERACT filter focuses on the question: 'Can the measure be applied easily, given constraints of time, money and interpretability?'[10] Information from a previous study on feasibility has been used.[11] Additionally, we assessed feasibility as indicated by the availability of each of the scoring methods analysed both in terms of status and progression scores. Availability of the score reflects a minimum number of VCs/sites available and readable. Progression scores were calculated at 2 and 5 years by subtracting the baseline score from the score at the corresponding time point.

Discrimination

The discrimination aspect, comprising reliability and sensitivity to change, addresses the question: 'Does the measure discriminate between situations of interest?'.[10] For reliability the variance components, namely patient, observer and residual variance, were analysed using a two-way analysis of variance (ANOVA) considering the 3 readers.[11, 18] The proportion of the total variance of the change scores (2- and 5-year) explained by the patient ('true variance') was used as a measure of reliability (the higher the better). Furthermore, reliability was investigated by means of Bland and Altman plots[19]. These are plotted for the different reader pairs from the total of the 3 readers. Additionally, the smallest detectable change (SDC) was calculated for each method. The SDC is the smallest change that can be detected beyond measurement error per individual patient and was calculated with the quantification of the measurement error of the change-score (SEM change score) derived from a two-way analysis of variance (ANOVA).[18]

To obtain insight into sensitivity to change of the methods, the means, medians and range of the status scores at all time points were calculated. Subsequently, mean 2- and 5-year progression was also analysed in the different spine segments and considering only the observations with all scoring methods available to enable a comparison. In the same subset of observations, the proportion of patients with a change above different cutoffs was calculated, namely above 0, 0.5, 1 and SDC. The proportion of change is presented as the change above the cut-off, change below the cut-off and net change. Net change corresponds to the number of patients with a positive change (e.g. ≥ 1) minus the number of patients with a negative change (e.g. ≥- 1) (numerator) divided by the total number of patients included in the analysis (denominator).[20] Cumulative probability plots of the 5-

year progression, ranking scores from the lowest to the highest and plotted as a cumulative proportion against the progression's actual value, provide further insight into the scoring methods by showing all individual data and enabling visualization of the internal coherence of the data.[21]

Truth

The truth aspect deals with the question: 'Is the measure truthful, does it measure what is intended? Is the result unbiased and relevant?'[10] All scoring methods have previously been assessed with respect to construct validity.[11] The construct of radiographic damage is, expectedly, the same in early disease. In order to get insight into which parts of the skeleton were most affected in early disease and in which most change occurred, the proportion of patients with any baseline damage (>0) and any 5-year net change (>0) was analysed for each of the scoring methods and for the individual segments. Moreover, we were particularly interested in the potential additional value of some segments in one scoring method compared to another. For example, the additional value of the 4 thoracic VCs included in the RASSS or the posterior VCs in the SASSS. This was analysed by determining the relative contribution (in %) to the 5-year total score progression (RASSS or SASSS, respectively) of each spinal segment included – cervical, thoracic and lumbar for the RASSS and anterior and posterior lumbar for the SASSS. A balanced progression in every segment was assumed, i.e. balanced proportion to the contribution in terms of number of VCs to the score. The balanced expected contribution for the segments of the RASSS was 43% (12/28 VCs) for the cervical and lumbar segments and 14% (4/28 VCs) for the thoracic; for the SASSS: 50% for both segments. Observed and expected progression rates were compared with the chi-square test.

Stata SE version 12 was used for all above-mentioned analyses.

Results

In total, 699 patients were included, with a mean age of 34 (standard deviation (SD) 9) years, mean symptom duration 1.5 (0.9) years, 47% were males, and 59% HLA-B27 positive (Online Supplementary Table 2). For the analysis on sensitivity to change, only observations with progression scores from all scoring methods available were used, and the characteristics of included and excluded patients from these analyses were summarized: groups were very similar, except for the fact that older and female patients were slightly more likely to have all 2- (n=357) and 5-year (n=265) progression scores.

Feasibility

Out of all observations with at least one radiograph available (n=1617), the SASSS could be computed in 99.8% of them, followed by the mSASSS in 98%, RASSS, BASRI-spine and BASRI-total 97%, and BASRI-spine-thoracic and BASRI-total-thoracic in 82%. Availability of the 5-year progression scores was also above 94% for most of the methods, but 69% for the BASRI-spine-thoracic and BASRI-total-thoracic.

Discrimination

Of all radiographic scoring methods, the variance proportion explained by the patient was highest for the mSASSS and RASSS. For both status scores at 2 and 5 years, it was 86% and 89% (very good reliability), respectively, compared to 75-80% for the other methods (good reliability) (Table 1). For the progression scores, the difference was larger, with values of 70% and 69% for the mSASSS and RASSS 2-year progression (good reliability), respectively, and between 40% and 57% for the remaining methods (poor-moderate reliability). The proportion of the observer variance for all the BASRI scores, though with a low value (around 2% for status scores and 0.4-0.7% for progression scores), was substantially higher compared to the remaining scoring methods (0-0.1% for all scores). When comparing the proportion of variance explained by the patient across the different segments included in both the mSASSS and the RASSS, this was, as expected, similar for the cervical and lumbar segments. However, for the thoracic segment the proportion of patient variance was substantially lower and reflecting a poor reliability (e.g. 36% and 46% for the 2-year and 5-year progression score, respectively) (Table 1).

The same pattern of reliability was found in the 5-year cumulative probability plots, which show fewer zeros for the BASRI scores, i.e. showing more progression captured, but also at the cost of a higher proportion of negative scores (i.e. 'noise'/measurement error) (Figure 1). Bland and Altman plots of all progression scores (Online Supplementary Figure 1) across scoring methods are difficult to compare because of different scales. Nevertheless, plots from BASRI scores were more heteroscedastic, i.e. with a higher diversity of scores between the readers. The difference between readers was particularly large for higher scores and corresponding to an important part of the scale of the BASRI.

Table 2 shows the status scores of the different methods for all time points and for all patients included in the analysis. At baseline the mean mSASSS was 0.4 (SD 1.5) and its maximum value was 18.5, i.e. 26% of the maximum of the scale. The mean RASSS was 0.5 (1.6), and its maximum value (17.8) reflected 21% of the maximum of the scale. For the BASRI scores the maximum value at baseline corresponded to 57-67% of the scales.

A change in radiographic damage over time could be captured by all scoring methods (Table 3). After 5 years, the RASSS showed a change of 0.7 (2.5), the mSASSS 0.5 (2.0), the SASSS 0.4 (1.3) and the BASRI scores a change of around 0.3 (SD around 0.6).

Net change above 0 was highest for the BASRI scores (Table 4 and Online Supplementary Table 3). When defining net change above somewhat higher cutoffs, namely 0.5, 1 and SDC, the BASRI scores showed a lower proportion of patients captured and the mSASSS and the RASSS performed the best in terms of depicting the signal (i.e. positive change) in relation to the noise (i.e. negative change).

Truth

The presence of baseline damage and 5-year net progression in the different parts of the skeleton is presented in Table 5. Most radiographic damage and progression was found in the SIJ. Only a minority of the patients had hip involvement at baseline (11%) and progression in the hips occurred very rarely (2%). When looking at the spinal segments, progression was captured more frequently in the lumbar than in the cervical or thoracic segments. The comparison with the latter is true both for the few thoracic vertebral corners included in the RASSS and also for the whole thoracic spine included in a modification of the BASRI. Within the lumbar spine, progression took place mostly in the anterior site, with a very small progression in the posterior site.

Across the different scoring methods, the proportion of patients with net progression captured per segment did not differ. At an overall score level, more patients with progression were captured with the BASRI scores and this was mostly due to a higher progression in the SIJ, not included in the remaining methods. As a total score, the SASSS captured less patients with progression than the mSASSS or the RASSS since there was very little progression in the posterior site of the lumbar spine (only present in 3% of the patients, and only in 1 patient, 0.4%, was this progression in the posterior segment higher than in the anterior segment). Regarding the observed and expected progression across the segments of the SASSS, the former was substantially lower in the posterior segment (7% vs 50%) and higher in the anterior segment (93% vs 50%, p-value <0.0001).

Regarding observed and expected progression across the segments of the RASSS, the observed progression in the cervical segment was lower than expected (29% vs 43%, p=0.039), while it was numerically higher but without a statistically significant difference in the thoracic segment (24% vs 14%, p=0.071), and not different in the lumbar spine (46% vs 43%, p=0.669).

Discussion

The existing scoring methods to assess spinal radiographic damage perform well in capturing damage and its progression even in an early phase of axSpA. The mSASSS and RASSS capture most change. However, there is no clear gain in additionally scoring the thoracic spine for the RASSS while an increased ‘noise’ is introduced. Therefore, the mSASSS remains the most sensitive and valid scoring method in axSpA, including early phases of the disease. This conclusion, based on the aspects of the OMERACT filter, is the same as had been drawn for r-axSpA, so that we can consider the mSASSS the appropriate scoring method for the whole spectrum of axSpA.[11, 12]

With regard to our limited analysis on feasibility, no substantial differences were seen between the scoring methods. Only the BASRI modifications to include the thoracic segment showed less availability because of lack of a ‘complete’ thoracic segment. Previous information clearly favoured the mSASSS, particularly in what concerns this being the scoring method with the lowest exposure to radiation.[11] Altogether, the mSASSS stands out as the most feasible scoring method.

Concerning discrimination, there was a clear difference in the reliability of the scoring methods, with the mSASSS and RASSS outweighing the remaining methods. The reliability of the BASRI scores was particularly poor, as shown in the Bland and Altman plots, by the higher proportion of negative scores in the cumulative probability plots and by the SDCs that, despite having a low absolute value, represent a higher proportion of the smaller scale of the BASRI scores than the SDCs of other methods. Both mSASSS and RASSS showed a comparable reliability, but the individual reliability of the thoracic segment of the RASSS, the single major difference between both methods, was unacceptably low: the proportion of the true variance of its progression score, i.e. patient variance, was only 36% over 2 years and 46% over 5 years. This means that despite an acceptable reliability of the overall score, its addition compared to the mSASSS comes with an increase in measurement error and therefore potentially imprecise scores. Furthermore, the parallax associated with extending the view of the lumbar radiograph to include the thoracic VCs has been proposed as an explanation for the lower reliability in the thoracic segment.[12] However, in DESIR a separate radiograph of the thoracic spine was available and scored, and still a poor reliability was found, thus arguing against this previously proposed hypothesis.

In what concerns sensitivity to change, all scores demonstrated this property, even in this cohort of patients with early disease. Nevertheless, the magnitude of the change over time was small, which means that if we want to test the effect of interventions on structural damage or test associations with other outcomes, we will need a large group of patients, with repeated measurements in order to increase the statistical power, and likely over a long follow-up. Alternatively, new techniques with a higher sensitivity may gain a role. In the arena, we have the low dose computed tomography as a promising tool, but more research is still needed.[22] For the time being conventional radiographs are the cornerstone of the evaluation of structural progression and therefore a data-driven choice of the best scoring method is necessary.[1] The mSASSS and the RASSS captured the most ‘signal’ of structural progression. This was particularly true when looking at net changes, which take measurement error into account. The BASRI scores have shown to be sensitive to change, capturing more positive changes than the remaining methods, but it comes at a cost of higher ‘noise’, and therefore performed worse when comparing net changes particularly with progression defined >1 or >SDC. As we are ultimately interested in real changes (beyond measurement error), net changes are the correct method to consider.[20] Moreover, despite this being a cohort of patients with early disease and low incidence of structural damage, there were patients reaching 60-65% of the maximum of the BASRI scores already at baseline, with values of 80-85% at 5 years. This contrasts with the remaining methods with maximum baseline values of 20-25% of the scale and points towards a potential risk of a ceiling effect by the BASRI scores, already previously demonstrated.[11]

With regard to truth, most progression seemed to occur at the SIJ level in these patients with early axSpA. This contrasts to what is reported in r-axSpA, but is partly also explained by the fact that in this cohort with less SIJ baseline damage there is more room for change.[11] It is hypothesized that structural progression begins in the SIJ and continues to the spine. The SIJ segment was included in the BASRI scores, which showed a poor overall reliability. We also know that the assessment of the SIJ, particularly concerning the fulfilment of the modified New York Criteria, has a poor reliability.[23] So whether structural damage really occurs first in the SIJ and later in the spine or whether this conclusion is partly driven by differences in the sensitivity to change and reliability of the methods used for these outcomes needs to be confirmed. Additionally, in what concerns the truth aspect, no significant differences were observed between the expected and observed progression in the thoracic spine, though the latter was numerically higher. This, together with the poor reliability of the assessment of the thoracic segment confirms that there is no additional gain in scoring the thoracic spine.[12]

Some limitations of this study should be considered. First of all, not all patients could be included in all analyses because of loss to follow-up or lack of availability of all spinal segments to allow the calculation of all methods. Notwithstanding, there were no major differences between the patients that were included and excluded in the various analyses. Furthermore, this aspect limited the different scoring methods equally and comparisons were based on observations with all scoring methods available. Patients had in general low structural damage, which challenged the comparison across methods; nevertheless, this affected the scoring methods similarly and provides a good comparison of the methods in situations with minimal damage. Some clear strengths are the high number of patients, prospectively systematically followed, and having 3 readers scoring the radiographs, which approximates the average score to the truth.

In conclusion, according to the feasibility, discrimination and truth of the OMERACT filter, the mSASSS is the most valid, feasible and sensitive to change method to assess radiographic damage in all patients with axSpA, including those with early disease.

Acknowledgements

The DESIR cohort was sponsored by the Département de la Recherche Clinique et du Développement de l'Assistance Publique–Hôpitaux de Paris. This study is conducted under the umbrella of the French Society of Rheumatology and INSERM (Institut National de la Santé et de la Recherche Médicale). The database management is performed within the department of epidemiology and biostatistics (Professor Paul Landais, D.I.M., Nîmes, France). An unrestricted grant from Pfizer was allocated for the 10 years of the follow-up of the recruited patients. The authors would like to thank the different regional participating centres: Pr Maxime Dougados (Paris – Cochin B), Pr André Kahan (Paris - Cochin A), Pr Olivier Meyer (Paris - Bichat), Pr Pierre Bourgeois (Paris – La Pitié Salpetrière), Pr Francis Berenbaum (Paris - Saint Antoine), Pr Pascal Claudepierre (Créteil), Pr Maxime Breban (Boulogne Billancourt), Dr Bernadette Saint-Marcoux (Aulnay-sous-Bois), Pr Philippe Goupille (Tours), Pr Jean-Francis Maillefert (Dijon), Dr Xavier Puéchal, Dr Emmanuel Dernis (Le Mans), Pr Daniel Wendling (Besançon), Pr Bernard Combe (Montpellier), Pr Liana Euller-Ziegler (Nice), Pr Philippe Orcel, Dr Pascal Richette (Paris - Lariboisière), Pr Pierre Lafforgue (Marseille), Dr Patrick Boumier (Amiens), Pr Jean-Michel Ristori, Pr Martin Soubrier (Clermont-Ferrand), Dr Nadia Mehsen (Bordeaux), Pr Damien Loeuille (Nancy), Pr René-Marc Flipo (Lille), Pr Alain Saraux (Brest), Pr Corinne Miceli (Le Kremlin Bicêtre), Pr Alain Cantagrel (Toulouse), Pr Olivier Vittecoq (Rouen). The authors would also like to thank URC-CIC Paris Centre for the coordination and monitoring of the study.

Funding

An unrestricted grant from Pfizer was allocated for the 10 years of the follow-up of the recruited patients.

Contributors

All authors contributed and finally approved the current manuscript.

Competing interests

The authors have no competing interests to declare.

Table 1 - Inter-observer reliability of the different radiographic methods or their segments*

	2 years						5 years					
	Status scores (n = 440-534)			**Progression scores** (n = 334-405)			**Status scores** (n = 358-499)			**Progression scores** (n = 266-389)		
	Residual variance	Observer variance	Patient variance	Residual variance	Observer variance	Patient variance	Residual variance	Observer variance	Patient variance	Residual variance	Observer variance	Patient variance
SASSS	19.5	0.4	80.2	43.0	0.0	57.0	23.1	0.7	76.2	38.6	0.3	61.1
mSASSS	14.0	0.1	85.9	30.2	0.1	69.7	11.0	0.1	89.0	15.1	0.0	84.8
Cervical segment	15.7	0.0	84.3	30.3	0.1	69.6	12.5	0.0	87.4	20.5	0.1	79.5
Lumbar segment	18.5	0.4	81.2	44.8	0.0	55.2	15.2	0.5	84.3	22.7	0.2	77.2
RASSS	14.7	0.1	85.2	31.4	0.0	68.5	11.3	0.0	88.5	13.8	0.0	86.2
Cervical segment	15.3	0.1	84.7	27.3	0.1	72.6	11.7	0.1	88.3	16.5	0.1	83.4
Lumbar segment	19.1	0.3	80.6	46.3	0.1	53.7	15.0	0.5	84.5	21.3	0.2	78.5
Thoracic segment	33.0	0.3	66.8	63.6	0.0	36.4	36.3	0.5	63.3	53.8	0.2	46.0
BASRI-spine	18.5	2.0	79.5	56.6	0.7	42.7	17.9	1.6	80.5	48.6	0.7	50.7
BASRI-spine-thoracic	21.8	1.9	76.4	59.3	0.7	40.0	18.2	1.7	80.0	45.9	0.4	53.4
BASRI-total	20.2	2.7	77.1	56.4	0.7	42.9	19.9	2.2	77.9	48.6	0.7	50.7
BASRI-total-thoracic	22.8	2.2	75.0	58.3	0.7	41.0	19.4	1.9	78.7	46.2	0.4	53.4

* Results reflect the proportion of the different components of the variance obtained through ANOVA (summing up to 100%, or otherwise close to 100% due to rounding), in which the score is the outcome and the between-patient effects are analysed, as well as the within-patient effects, i.e. within measurement reflecting the observer/reader effect and the residual error.

SASSS: Stoke Ankylosing Spondylitis Spine Score; mSASSS: modified Stoke Ankylosing Spondylitis Spine Score; RASSS: Radiographic Ankylosing Spondylitis Spinal Score; BASRI: Bath Ankylosing Spondylitis Radiology Index

Table 2 – Structural damage throughout follow-up according to all scoring methods

	N	Mean (SD)	Median	25th perc.	75th perc.	95th perc.	Min	Max	% of the scale maximum*
SASSS (0-72)									
Baseline	675	0.17 (0.67)	0	0	0	1.3	0	7.8	11
2 years	534	0.30 (1.25)	0	0	0	1.7	0	17.7	25
5 years	405	0.56 (1.99)	0	0	0	3.0	0	23.0	32
mSASSS (0-72)									
Baseline	655	0.43 (1.46)	0	0	0	2.4	0	18.5	26
2 years	522	0.73 (2.75)	0	0	0.3	3.3	0	30.1	42
5 years	399	1.05 (3.55)	0	0	0.7	4.7	0	38.6	54
RASSS (0-84)									
Baseline	651	0.46 (1.62)	0	0	0	2.0	0	17.8	21
2 years	521	0.80 (2.97)	0	0	0.3	3.3	0	31.4	37
5 years	399	1.23 (4.11)	0	0	0.7	6.0	0	42.6	51
BASRI-spine (0-12)									
Baseline	652	0.97 (1.19)	0.7	0	1.5	3.2	0	8.0	67
2 years	519	1.13 (1.41)	0.7	0.2	1.5	4.0	0	9.3	78
5 years	399	1.33 (1.55)	0.8	0.2	2.0	4.2	0	10.3	86
BASRI-spine-thoracic (0-16)									
Baseline	551	1.12 (1.36)	0.7	0.2	1.7	3.7	0	9.5	59
2 years	442	1.25 (1.56)	0.8	0.2	1.8	4.0	0	10.7	67
5 years	335	1.54 (1.88)	1.0	0.3	2.2	5.0	0	12.7	79
BASRI-total (0-16)									
Baseline	650	1.03 (1.27)	0.7	0.2	1.5	3.5	0	9.7	60
2 years	517	1.20 (1.50)	0.7	0.2	1.7	4.2	0	10.5	66
5 years	398	1.41 (1.62)	1.0	0.3	2.0	4.3	0	10.3	65
BASRI-total-thoracic (0-20)									
Baseline	551	1.18 (1.43)	0.7	0.2	1.7	3.8	0	11.3	57
2 years	440	1.33 (1.65)	0.8	0.3	1.8	4.3	0	12.5	63

5 years	334	1.61 (1.95)	1.0	0.3	2.3	5.3	0	12.7	63

* Maximum value found in the dataset divided by the theoretical maximum of the score, to give an idea of a ceiling effect. E.g. for mSASSS at baseline: maximum value found in the dataset was 18.53 and the theoretical maximum value is 72, so 26%.

SASSS: Stoke Ankylosing Spondylitis Spine Score; mSASSS: modified Stoke Ankylosing Spondylitis Spine Score; RASSS: Radiographic Ankylosing Spondylitis Spinal Score; BASRI: Bath Ankylosing Spondylitis Radiology Index

Table 3 – Mean baseline and progression scores (2- and 5-years) per scoring method and per segment*

BASELINE STATUS SCORE							
	SASSS **n = 550**	**mSASSS** **n = 550**	**RASSS** **n = 550**	**BASRI spine** **n = 550**	**BASRI spine with thoracic** **n = 550**	**BASRI total** **n = 550**	**BASRI total with thoracic spine** **n = 550**
Total score	0.18 (0.70)	0.43 (1.51)	0.46 (1.69)	0.98 (1.20)	1.12 (1.36)	1.03 (1.27)	1.18 (1.43)
Cervical segment	--	0.22 (1.14)	0.20 (1.08)	0.15 (0.44)	0.15 (0.44)	0.15 (0.44)	0.15 (0.44)
Lumbar segment	0.18 (0.70)	0.21 (0.75)	0.16 (0.64)	0.17 (0.41)	0.17 (0.41)	0.17 (0.41)	0.17 (0.41)
Lumbar segment with thoracic segment included	--	--	0.27 (1.09)	--	--	--	--
Thoracic segment	--	--	0.10 (0.58)	--	0.14 (0.40)	--	0.14 (0.40)
Lumbar anterior	0.17 (0.67)	--	--	--	--	--	--
Lumbar posterior	0.01 (0.11)	--	--	--	--	--	--
SI joints	--	--	--	0.66 (0.84)	0.66 (0.84)	0.66 (0.84)	0.66 (0.84)
Hips	--	--	--	--	--	0.06 (0.22)	0.06 (0.22)
2-YEAR PROGRESSION SCORES							
	SASSS **n = 357**	**mSASSS** **n = 357**	**RASSS** **n = 357**	**BASRI spine** **n = 357**	**BASRI spine with thoracic** **n = 357**	**BASRI total** **n = 357**	**BASRI total with thoracic spine** **n = 357**
Total score	0.09 (0.51)	0.16 (0.96)	0.20 (1.03)	0.10 (0.35)	0.11 (0.42)	0.10 (0.36)	0.11 (0.43)
Cervical segment	--	0.09 (0.67)	0.09 (0.71)	0.03 (0.19)	0.03 (0.19)	0.03 (0.19)	0.03 (0.19)
Lumbar segment	0.09 (0.51)	0.07 (0.51)	0.08 (0.45)	0.03 (0.22)	0.03 (0.22)	0.03 (0.22)	0.03 (0.22)
Lumbar segment with thoracic segment included	--	--	0.10 (0.51)	--	--	--	--
Thoracic segment	--	--	0.02 (0.17)	--	0.01 (0.18)	--	0.01 (0.18)
Lumbar anterior	0.08 (0.46)	--	--	--	--	--	--
Lumbar posterior	0.01 (0.11)	--	--	--	--	--	--
SI joints		--	--	0.04 (0.18)	0.04 (0.18)	0.04 (0.18)	0.04 (0.18)
Hips	--	--	--	--	--	0.00 (0.06)	0.00 (0.06)
5-YEAR PROGRESSION SCORES							

	SASSS **n = 265**	**mSASSS** **n = 265**	**RASSS** **n = 265**	**BASRI spine** **n = 265**	**BASRI spine with thoracic** **n = 265**	**BASRI total** **n = 265**	**BASRI total with thoracic spine** **n = 265**
Total score	0.38 (1.31)	0.51 (2.04)	0.68 (2.47)	0.29 (0.59)	0.34 (0.69)	0.29 (0.60)	0.35 (0.70)
Cervical segment	--	0.24 (1.32)	0.23 (1.36)	0.08 (0.32)	0.08 (0.32)	0.08 (0.32)	0.08 (0.32)
Lumbar segment	0.38 (1.31)	0.27 (1.09)	0.34 (1.15)	0.11 (0.32)	0.11 (0.32)	0.11 (0.32)	0.11 (0.32)
Lumbar segment with thoracic segment included	--	--	0.45 (1.48)	--	--	--	--
Thoracic segment	--	--	0.11 (0.52)	--	0.05 (0.25)	--	0.05 (0.25)
Lumbar anterior	0.33 (1.13)	--	--	--	--	--	--
Lumbar posterior	0.05 (0.51)	--	--	--	--	--	--
SI joints	--	--	--	0.10 (0.28)	0.10 (0.28)	0.10 (0.28)	0.10 (0.28)
Hips	--	--	--	--	--	0.01 (0.06)	0.01 (0.06)

* in observations with all scoring methods available

-- means that the given spinal segment is not included in the scoring method

SASSS: Stoke Ankylosing Spondylitis Spine Score; mSASSS: modified Stoke Ankylosing Spondylitis Spine Score; RASSS: Radiographic Ankylosing Spondylitis Spinal Score; BASRI: Bath Ankylosing Spondylitis Radiology Index; SI: sacroiliac

Table 4 - Percentage of patients with 2-year (5-year) change from baseline >SDC and >1*

	2-year change (n = 357)							5-year change (n = 265)						
		Progression >SDC			Progression >1				Progression >SDC			Progression >1		
	SDC	Positive change N (%)	Negative change N (%)	Net change N (%)	Positive change N (%)	Negative change N (%)	Net change N (%)	SDC	Positive change N (%)	Negative change N (%)	Net change N (%)	Positive change N (%)	Negative change N (%)	Net change N (%)
SASSS	0.75	11 (3)	0 (0)	**11 (3)**	11 (3)	0 (0)	**11 (3)**	1.17	30 (11)	0 (0)	**30 (11)**	30 (11)	0 (0)	**30 (11)**
mSASSS	0.88	22 (6)	2 (0.6)	**20 (6)**	18 (5)	1 (0.3)	**17 (5)**	1.10	34 (13)	1 (0.4)	**33 (12)**	34 (13)	1 (0.4)	**33 (12)**
RASSS	1.00	17 (5)	0 (0)	**17 (5)**	17 (5)	0 (0)	**17 (5)**	1.19	44 (17)	0 (0)	**44 (17)**	46 (17)	0 (0)	**46 (17)**
BASRI-spine	0.59	30 (8)	14 (4)	**16 (4)**	9 (3)	0 (0)	**9 (3)**	0.74	32 (12)	1 (0.4)	**31 (12)**	22 (8)	0 (0)	**22 (8)**
BASRI-spine-thoracic	0.59	35 (10)	16 (4)	**19 (5)**	12 (3)	2 (0.6)	**10 (3)**	0.89	31 (12)	2 (1)	**29 (11)**	25 (9)	1 (0.4)	**24 (9)**
BASRI-total	0.61	31 (9)	14 (4)	**17 (5)**	9 (3)	0 (0)	**9 (3)**	0.75	33 (12)	1 (0.4)	**32 (12)**	23 (9)	0 (0)	**23 (9)**
BASRI-total-thoracic	0.72	19 (5)	4 (1)	**15 (4)**	13 (4)	2 (0.6)	**11 (3)**	0.91	32 (12)	2 (1)	**30 (11)**	26 (10)	1 (0.4)	**25 (9)**

* in observations with all 2-year (or 5-year) progression scores available from all scoring methods

Net change results are highlighted as these are the ones best reflecting the real change, taking measurement error into account

SASSS: Stoke Ankylosing Spondylitis Spine Score; mSASSS: modified Stoke Ankylosing Spondylitis Spine Score; RASSS: Radiographic Ankylosing Spondylitis Spinal Score; BASRI: Bath Ankylosing Spondylitis Radiology Index; SDC: smallest detectable change

Table 5 - Percentage of patients with baseline damage (>0) and a 5-year net change (>0) per radiographic score and per segment*

	SASSS		mSASSS		RASSS		BASRI spine		BASRI spine with thoracic		BASRI total		BASRI total with thoracic	
	% pts with baseline damage	% pts with any change from baseline	% pts with baseline damage§	% pts with any change from baseline±	% pts with baseline damage	% pts with any change from baseline	% pts with baseline damage	% pts with any change from baseline	% pts with baseline damage	% pts with any change from baseline	% pts with baseline damage	% pts with any change from baseline	% pts with baseline damage	% pts with any change from baseline
Total score	16	18	26	21	25	25	77	37	81	39	78	37	82	39
Cervical segment	--	--	12	9	9	8	15	9	15	9	15	9	15	9
Lumbar segment	16	18	18	15	16	17	25	15	25	15	25	15	25	15
Lumbar segment with thoracic segment included	--	--	--	--	20	21	--	--	--	--	--	--	--	--
Thoracic segment	--	--	--	--	6	7	--	--	18	7			18	7
Lumbar anterior	16	18	--	--	--	--	--	--	--	--	--	--	--	--
Lumbar posterior	1	3	--	--	--	--	--	--	--	--	--	--	--	--
SI joints	--	--	--	--	--	--	67	20	67	20	67	20	67	20
Hips	--	--	--	--	--	--	--	--	--	--	11	2	11	2

* analysed in patients with all 5-year progression available from all scoring methods (n = 265)

§ Presence of damage is defined as a status score at baseline >0

±Change>0 from baseline to year 5 (scores at 2 years are disregarded here)

SASSS: Stoke Ankylosing Spondylitis Spine Score; mSASSS: modified Stoke Ankylosing Spondylitis Spine Score; RASSS: Radiographic Ankylosing Spondylitis Spinal Score; BASRI: Bath Ankylosing Spondylitis Radiology Index; SI: sacroiliac

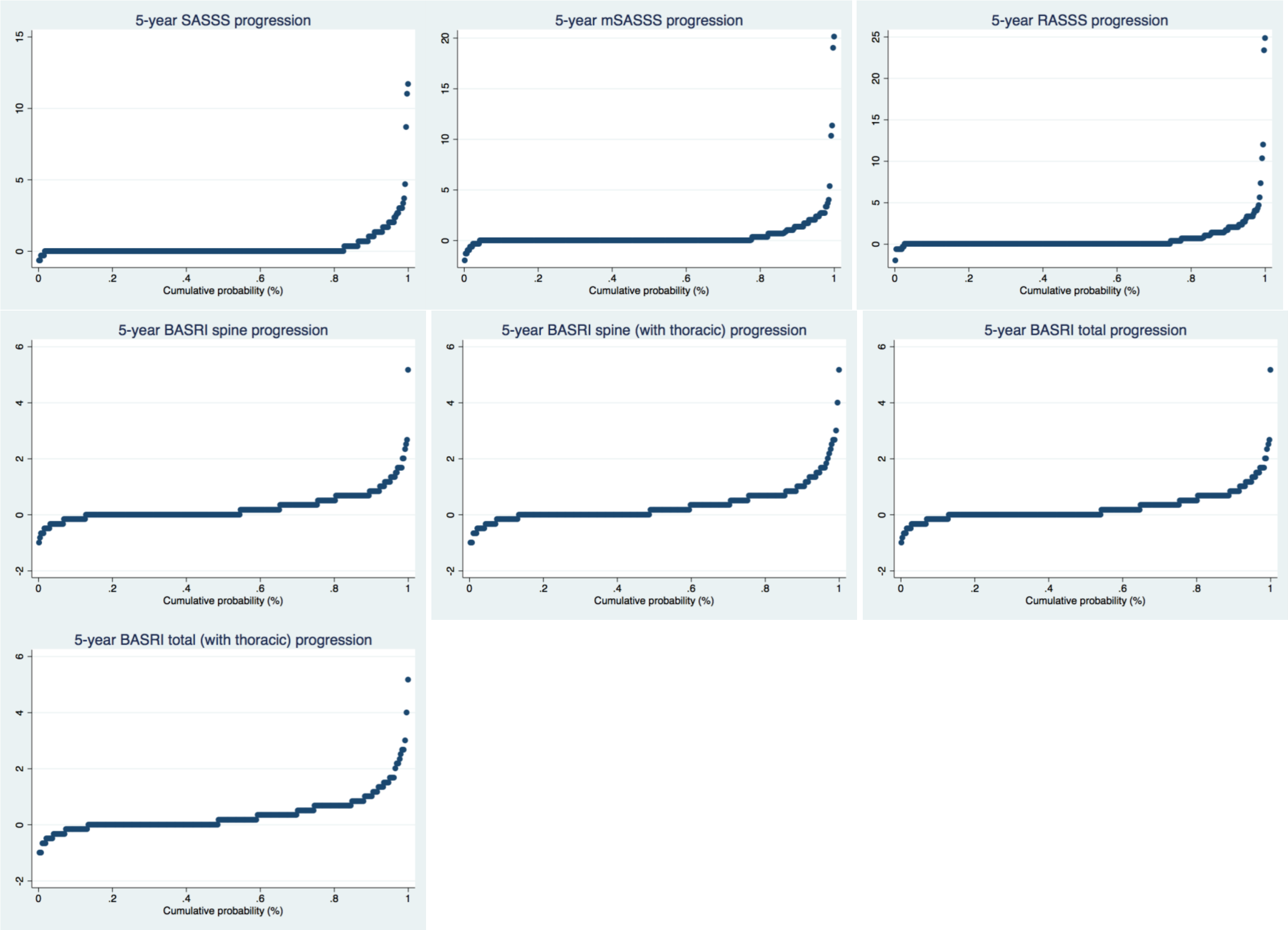

Figure 1 - Cumulative probability plots for the 5-year progression scores of the different scoring methods

SASSS: Stoke Ankylosing Spondylitis Spine Score; mSASSS: modified Stoke Ankylosing Spondylitis Spine Score; RASSS: Radiographic Ankylosing Spondylitis Spinal Score; BASRI: Bath Ankylosing Spondylitis Radiology Index

References

1 van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden S, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy, and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. Assessments in Ankylosing Spondylitis. J Rheumatol 1999;26(4):951-4.

2 Gladman DD. Clinical aspects of the spondyloarthropathies. Am J Med Sci 1998;316(4):234-8.

3 Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36(6):729-40.

4 Ramiro S, Stolwijk C, van Tubergen A, et al. Evolution of radiographic damage in ankylosing spondylitis: a 12 year prospective follow-up of the OASIS study. Ann Rheum Dis 2015;74(1):52-9.

5 Averns HL, Oxtoby J, Taylor HG, Jones PW, Dziedzic K, Dawes PT. Radiological outcome in ankylosing spondylitis: use of the Stoke Ankylosing Spondylitis Spine Score (SASSS). Br J Rheumatol 1996;35(4):373-6.

6 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(12):2263-70.

7 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(12):2866-72.

8 Creemers MC, Franssen MJ, van't Hof MA, Gribnau FW, van de Putte LB, van Riel PL. Assessment of outcome in ankylosing spondylitis: an extended radiographic scoring system. Ann Rheum Dis 2005;64(1):127-9.

9 Baraliakos X, Listing J, Rudwaleit M, Sieper J, Braun J. Development of a radiographic scoring tool for ankylosing spondylitis only based on bone formation: addition of the thoracic spine improves sensitivity to change. Arthritis Rheum 2009;61(6):764-71.

10 Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. J Rheumatol 1998;25(2):198-9.

11 Wanders AJ, Landewe RB, Spoorenberg A, et al. What is the most appropriate radiologic scoring method for ankylosing spondylitis? A comparison of the available methods based on the Outcome Measures in Rheumatology Clinical Trials filter. Arthritis Rheum 2004;50(8):2622-32.

12 Ramiro S, van Tubergen A, Stolwijk C, et al. Scoring radiographic progression in ankylosing spondylitis: should we use the modified Stoke Ankylosing Spondylitis Spine Score (mSASSS) or the Radiographic Ankylosing Spondylitis Spinal Score (RASSS)? Arthritis research & therapy 2013;15(1):R14.

13 van der Heijde D, Landewe R. Selection of a method for scoring radiographs for ankylosing spondylitis clinical trials, by the Assessment in Ankylosing Spondylitis Working Group and OMERACT. J Rheumatol 2005;32(10):2048-9.

14 Rudwaleit M, van der Heijde D, Landewe 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(6):777-83.

15 Sieper J, Poddubnyy D. Axial spondyloarthritis. Lancet 2017;390(10089):73-84.

16 van der Heijde D, Landewe R, van der Linden S. How should treatment effect on spinal radiographic progression in patients with ankylosing spondylitis be measured? Arthritis Rheum 2005;52(7):1979-85.

17 Dougados M, Etcheto A, Molto A, et al. Clinical presentation of patients suffering from recent onset chronic inflammatory back pain suggestive of spondyloarthritis: The DESIR cohort. Joint, bone, spine : revue du rhumatisme 2015;82(5):345-51.

18 Bruynesteyn K, Boers M, Kostense P, van der Linden S, van der Heijde D. Deciding on progression of joint damage in paired films of individual patients: smallest detectable difference or change. Ann Rheum Dis 2005;64(2):179-82.

19 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1(8476):307-10.

20 Dougados M, Sepriano A, Molto A, et al. Sacroiliac radiographic progression in recent onset axial spondyloarthritis: the 5-year data of the DESIR cohort. Ann Rheum Dis 2017;76(11):1823-8.

21 Landewe R, van der Heijde D. Radiographic progression depicted by probability plots: presenting data with optimal use of individual values. Arthritis Rheum 2004;50(3):699-706.

22 de Koning A, de Bruin F, van den Berg R, et al. Low-dose CT detects more progression of bone formation in comparison to conventional radiography in patients with ankylosing spondylitis: results from the SIAS cohort. Ann Rheum Dis 2018;77(2):293-9.

23 van den Berg R, Lenczner G, Feydy A, et al. Agreement between clinical practice and trained central reading in reading of sacroiliac joints on plain pelvic radiographs. Results from the DESIR cohort. Arthritis & rheumatology 2014;66(9):2403-11.