



## Scoring of radiographic progression in randomised clinical trials in AS: a preference for paired reading order.

*Astrid JB Wanders, Robert BM Landewé, Anneke Spoorenberg, Kurt de Vlam, Herman Mielants, Maxime Dougados, Sjef van der Linden, and Désirée MFM van der Heijde*

### DISCLAIMER

The initial version of *ARD Online First* articles are papers in manuscript form that have been accepted and published in *ARD Online* but they have not been copy edited and not yet appeared in a printed issue of the journal. Copy editing may lead to differences between the *Online First* version and the final version including in the title; there may also be differences in the quality of the graphics. Edited, typeset versions of the articles may be published as they become available before final print publication.

Should you wish to comment on this article please do so via our eLetter facility on *ARD Online* (<http://ard.bmjournals.com/cgi/eletter-submit/ard.2004.022038v1>)

### DATE OF PUBLICATION

*ARD Online First* articles are citable and establish publication priority. The publication date of an *Online First* article appears at the top of this page followed by the article's unique Digital Object Identifier (DOI). These articles are considered published and metadata has been deposited with PubMed/Medline.

### HOW TO CITE THIS ARTICLE

Wanders A J B, Landewé R B M, Spoorenberg A, et al. Scoring of radiographic progression in randomised clinical trials in AS: a preference for paired reading order. *Ann Rheum Dis* Published Online First [date of publication]\*. doi: 10.1136/ard.2004.022038

\*Replace with date shown at the top of this page - remove brackets and asterisk

Online First articles are posted weekly at <http://ard.bmjournals.com/onlinefirst.shtml>

Scoring of radiographic progression in randomised clinical trials in Ankylosing Spondylitis: a preference for paired reading order.

Extended report

Astrid Wanders<sup>1</sup>  
Robert Landewé<sup>1</sup>  
Anneke Spoorenberg<sup>1</sup>  
Kurt de Vlam<sup>2</sup>  
Herman Mielants<sup>2</sup>  
Maxime Dougados<sup>3</sup>  
Sjef van der Linden<sup>1</sup>  
Désirée van der Heijde<sup>1</sup>

<sup>1</sup> Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, The Netherlands, <sup>2</sup> Department of Rheumatology, University Hospital Gent, Belgium, <sup>3</sup> Department of Rheumatology, Hôpital Cochin Paris, France.

Corresponding author:

Robert Landewé  
Department of Internal Medicine, Division of Rheumatology  
University Hospital Maastricht  
PO Box 5800  
6202 AZ Maastricht  
The Netherlands

Telephone: + 31 43 3875026  
Fax: + 31 43 3875006  
E-mail: [rlan@sint.azm.nl](mailto:rlan@sint.azm.nl)

Key words: radiographs, radiographic progression, reading order, ankylosing spondylitis, randomised controlled trial

## ABSTRACT

**Objectives:** To describe the influence of the reading order (chronological vs. paired) on radiographic scoring results in Ankylosing Spondylitis (AS). Since paired reading is requested for establishing drug efficacy in clinical trials, we investigated whether this method is sufficiently sensitive to change.

**Methods:** Films obtained from 166 patients (at baseline, 1 year and 2 years) were scored by one observer, using the modified Stoke Ankylosing Spondylitis Spinal Score. Films were first scored chronological, and were scored paired 6 months later.

**Results:** Reading chronological yields more radiographic progression; both at 1 year (mean progression: 1.3 (standard deviation (SD) 2.6) vs 0.5 (SD 2.4) units) and at 2 years (mean progression: 2.1 (SD 3.9) vs 1.0 (SD 2.9) units). Reading with chronological order yielded significantly more progression than paired reading (between- method difference:  $p < 0.001$  at 1 year, and  $p < 0.001$  at 2 years). After 1 year, progression ( $>0$  units) was found in 21% of patients after paired reading and in 33% after chronological reading. After 2 years, these numbers were 30% and 41%, respectively. Sample size calculations showed that 94 patients per treatment arm are required in a randomised clinical trial (RCT) to provide sufficient statistical power to detect a difference in 2-year progression if films are scored paired.

**Conclusion:** Reading with chronological time order is more sensitive to change than reading with paired time order, but paired reading is sensitive enough to pick up change with a follow-up of 2 years, resulting in an acceptable sample size for RCTs.

## INTRODUCTION

For evaluation of therapy in Ankylosing Spondylitis(AS), the ASessment in Ankylosing Spondylitis (ASAS) working group has developed core sets to be used in various settings(1), including the setting for disease controlling antirheumatic therapy (D-CART). One segment of the definition of D-CART reads: “prevent or significantly decrease the rate of progression of structural damage”(1). To assess progression of structural damage, radiographic outcome assessment is included in the D-CART core set. Radiology as outcome parameter in AS clinical trials is new, in contrast to clinical trials in rheumatoid arthritis (RA) in which radiographic outcome already has a prominent place. The methodology of radiographic scoring in AS is still developing. Recently, we performed a study comparing the existing radiographic scoring methods with respect to various aspects of validity (2). It was concluded that the modified Stoke Ankylosing Spondylitis Spinal Score (modified SASSS) is the most appropriate method for use in clinical trials.

It is known from studies concerning evaluation of radiographic damage in RA that the order in which films are presented to the observer influences results (3-6). Films can be grouped per patient and presented to the reader without knowing the chronological order of the films: paired scoring. Films can also be grouped per patient and presented in chronological order. The advantage of chronological reading is that it provides the reader with a maximum of information, thereby reducing ‘true’ measurement error. Reading films chronologically results in an increased ability to detect changes as compared to paired reading. In 1999 van der Heijde et al. showed that reading with chronological order was more sensitive to change than paired reading in RA (5). However, the possibility that chronological reading overestimated progression of joint damage because readers expected to see progression (expectation bias) could not be excluded. In a follow-up study Bruynesteyn et al (6) used progression considered as clinically relevant by rheumatologists as a proxy for true progression, and concluded that paired reading underestimates the true progression. The advantage of paired reading, however, is that expectation bias is almost ruled out: readers are not aware of the sequence of the films and therefore do not tend to score more progression in the follow-up film. The issue which of both reading orders should be used is therefore not unanimously answered by the above-mentioned studies. Despite this controversy, the reading of structural damage in RA clinical trials is predominantly performed by readers blinded for the sequence. This stems from the general epidemiological consensus that in order to prevent bias observers must be blinded as far as possible, and from the practical aspect that for registration purposes reading with blinded sequence is requested by the drug regulatory agencies. Therefore it seems obvious that radiographic progression in AS clinical trials should also be assessed by paired reading. However, there is some concern that paired scoring in AS is not sensitive enough, since progression occurs slowly, and only in a minority of patients (7). The aim of this study therefore was 1) to explore the differences with respect to sensitivity to change between paired and chronological scoring in AS, and 2) to investigate whether trials with radiographic progression as primary endpoint can be designed, that have sufficient statistical power with feasible patient numbers if films are read with paired order.

## METHODS

### *Patients and films*

Radiographs from an international longitudinal, observational study on outcome in AS, the OASIS cohort, were used (8). Originally 217 patients from four centres in the Netherlands, Belgium and France were included in this cohort. Radiographs were obtained at baseline, and

after one and two years of follow-up. After two years of follow-up, complete sets of radiographs of baseline, one and two year of 166 patients were available; only these patients were included in this study. The modified SASSS was assessed on lateral views of the lumbar and cervical spine.

#### *The scoring of films.*

The modified SASSS method scores every corner of the anterior site of the lumbar and cervical vertebrae on a scale from 0 – 3, in which 0 indicates no abnormalities, 1 is used for erosion, sclerosis or squaring, 2 indicates a syndesmophyte and 3 a bridging syndesmophyte. This yields a possible total score of 72 units. The lumbar spine is scored from the lower border of the 12<sup>th</sup> thoracic vertebra to the upper border of the first sacral vertebra, the cervical spine is scored from the lower border of the second cervical vertebra to the upper border of the first thoracic vertebra. In a previous study was shown that this method had a good inter- and intraobserver reliability (2). Intraclass correlation coefficients for inter- and intraobserver reliability for progression scores with a 2-year interval were 0.82 and 0.95, respectively. Films were available for three time points: baseline, 1 year and 2 years. First the films were scored in chronological order, and after 6 months the films were scored again by the same reader (AW), but now in a random time order (paired films per patient). The chronological scoring method allows negative progression scores.

#### *Analysis and statistics*

Descriptive statistics (mean, standard deviation, median, 25<sup>th</sup> and 75<sup>th</sup> percentile) are given for the modified SASSS scores for both reading orders at the three time points, as well as for the progression scores. Also are descriptive statistics provided for those patients who had a radiographic progression greater than zero. To visualize the effects of scoring by the two reading orders, progression scores obtained by both methods were plotted by its cumulative frequency (expressed as percentage; cumulative probability) in probability plots (9). Wilcoxon's signed ranks test was used to test the null-hypothesis that 1- or 2-year progression is zero. Mann-Whitney test was used to investigate the null hypothesis that radiographic progression obtained by both reading orders was similar. Proportions of patients with progression (> 0 units) by reading order at 1- or 2 years were compared by chi-square test. Sample sizes for a putative RCT with one untreated control group and one active treatment group, and radiographic progression as primary endpoint, were calculated using the power calculator of the University of California, Los Angeles (<http://calculators.stat.ucla.edu/powercalc/>, significance level = 0.05, 2-sided, power = 0.80). This was done under the assumptions that an untreated control group will show progression as in the OASIS cohort, and that progression in the active treatment is zero, with a standard deviation equal to the standard deviation in the untreated control group. Van der Waerden-normalised progression scores were used to perform the sample size calculations.

## RESULTS

#### *Sensitivity to change*

Table 1 shows the patient characteristics at baseline.

**Table 1 Patient characteristics at baseline**

	mean	SD	median	p25	p75
Age (years)	43.9	12.5	43.1	33.6	52.9
Mean duration of complaints (years)	20.4	11.6	17.1	12.0	27.5
Mean duration of disease after diagnosis (years)	11.7	9.0	10.0	4.8	15.4
Male, %	71.5				

In table 2 the descriptive statistics of the modified SASSS scores according to chronological and paired reading are given.

**Table 2 Descriptive statistics of 1-year and 2-year follow-up of radiological damage scored according to the modified SASSS with paired and chronological reading order ( n = 166 patients).**

	mean	SD	median	p25	p75
<i>paired reading order</i>					
Baseline	13.1	18.0	4.9	0.0	18.1
1 year	13.6	18.4	4.9	0.0	21.2
2 years	14.1	18.6	6.0	0.0	23.3
progression after 1 year	0.5	2.4	0.0	0.0	0.0
progression after 2 years	1.0	2.9	0.0	0.0	1.6
<i>chronological reading order</i>					
Baseline	13.6	19.2	5.0	0.0	17.0
1 year	14.9	19.7	6.0	0.0	20.1
2 years	15.8	20.1	6.0	0.0	22.7
progression after 1 year	1.3	2.6	0.0	0.0	1.4
progression after 2 years	2.1	3.9	0.0	0.0	3.0

Modified SASSS = modified Stoke Ankylosing Spondylitis Spinal Score; SD = standard deviation; p25 = 25<sup>th</sup> percentile; p75 = 75<sup>th</sup> percentile.

Baseline scores are almost similar. Reading with chronological order yields more progression than paired reading, both at 1 year 1.3 (2.6) (mean (standard deviation)) units vs. 0.5 (2.4) units, and at 2 years 2.1(3.9) vs. 1.0(2.9) units.

Table 3 provides the descriptive statistics of the modified SASSS scores of those patients who showed a progression greater than zero.

**Table 3 Descriptive statistics of 1-year and 2-year follow-up of radiological damage scored according to the modified SASSS of the patients who showed progression greater than zero accordingly to the paired and chronological reading order.**

	N	mean	SD	median	p25	p75
<i>paired reading order</i>						
progression after 1 year	35	4.0	2.9	4.0	1.6	5.0
progression after 2 years	50	3.9	3.8	2.9	2.0	5.0
<i>chronological reading order</i>						
progression after 1 year	55	3.9	3.1	3.0	1.2	6.0
progression after 2 years	68	5.2	4.6	4.0	2.0	7.0

In the entire cohort of this study, both methods picked up progression from baseline significantly (chronological order:  $p < 0.001$  for 1 year, and  $p < 0.001$  for 2 years; paired order:  $p = 0.021$  for 1 year, and  $p < 0.001$  for 2 years). Reading with chronological order was significantly more sensitive than paired reading (between-method difference:  $p < 0.001$  at 1 year, and  $p < 0.001$  at 2 years). After 1 year of follow-up 21% of patients showed progression  $> 0$  units according to the paired reading results and 33 % of patients according to the chronological reading results. At 2-year follow-up these numbers were 30% and 41%, respectively.

This progression pattern is further illustrated by probability plots for the 1-year interval (figure 1) and for the 2-year interval (figure 2). In figure 1 it can be seen that for both scoring methods a majority of patients do not show progression. This was already represented by the median that was zero for both methods (this median value can be found on the x-axis at a proportion percentage of 0.50) The advantage of the probability plot is that it also easily represents the percentage of patients with progression: For instance, figure 1 for the chronological reading order shows that the curve deviates from zero at a value of 67%, indicating that 33 % of patients show progression. Although negative progression scores were allowed in the chronological scoring method, these are not seen in the two plots. For the paired scoring method negative scores are visible in both figures (15% at 1 year and 11% at 2 years).

A comparison of both plots shows that the curve for chronological reading lies most left, which indicates that with the chronological reading more patients are qualified as progressive. The difference between these two curves was statistically tested; both for the 1-year interval and for the 2-year interval the difference between both methods was statistically significant ( $p = 0.019$  and  $p = 0.051$  respectively).

#### *Sample size calculations for the paired reading order*

In table 1 and in the probability plots it is shown that the data of the paired scoring order have a skewed distribution. So before entering the data in sample size calculations a van der Waerden normalisation procedure was performed. The following assumptions were made in the sample size calculations: the mean progression in the intervention group is zero and the standard deviation is the same as in the control group (the OASIS cohort). With these assumptions the following sample sizes were obtained for a RCT in which radiographic progression is scored according to the modified SASSS by paired reading order; A RCT with a duration of 1 year requires 922 patients per arm, and a RCT with a duration of 2 years requires 94 patients per arm, in order to statistically underscore a true between-group difference of 0.5 units (1 year) resp. 1.0 units (2 years).

## DISCUSSION

The conclusion of this study is that the order by which films of AS patients are presented to the observer influences the reading results, which is in accordance with the findings in RA. Reading films in chronological order shows a higher mean progression and a greater proportion of patients with progression, in comparison with paired reading. However, we also showed that scoring with a paired reading order is sufficiently sensitive to pick up radiographic progression after 2 years of follow up, under the specific conditions set in this study. To illustrate the feasibility of paired scoring in trials with a radiographic endpoint, we demonstrated an acceptable sample size for a putative RCT, using real progression data from the OASIS cohort, provided that the duration of the trial is at least 2 years.

The theoretical assumption that chronological scoring in comparison to paired scoring would have a higher sensitivity to change, which is supported by data from research in RA, was confirmed in this study. It was also seen that the magnitude of the signal picked up by the chronological reading order is greater than by the paired reading order. However, which part of this signal is a 'true' effect and which part can be attributed to 'noise' is difficult to establish, especially for the chronological reading order. First of all, in the chronological

reading order expectation bias contributes to 'noise', whereas this bias is almost ruled out in paired scoring. It is impossible to determine in chronological reading which part of 'noise' is caused by expectation bias and which part by the remaining measurement error. The measurement error in paired reading can be visualised by means of probability plots. Since it is thought that the phenomenon of healing ("true negative scores") does not occur in AS (which is supported by the results of chronological reading data, in which no negative scores were found), the negative scores by paired scoring can be considered as measurement error. When this is applied to figure 1 with a 1-year time interval, then it is seen that 15 % of the patients have a negative score. The percentage of patients that have a positive score is 21%. Assuming that measurement error works equally in both directions, this would mean that only 6% of patients show 'real' progression. In figure 2, with a 2-year interval, it is seen that 11% of patients have a negative score and 30% of patients have a positive score, which means that 19% of patients show 'real' progression. This difference in signal-noise ratio is also reflected by the sample size calculations, after one year of follow-up a huge sample size is needed, 922 patients, versus 94 patients for a follow-up of 2 years.

The lack of expectation bias and the possibility of assessing measurement error are advantages of paired reading. Apart from these advantages, it is also a fact that this scoring method is requested by the agencies for registration purposes. Therefore the feasibility of this method is relevant with respect to the number of patients needed to demonstrate a significant difference in radiographic progression. The problem with sample size calculations is that they are dependent on the assumptions, which are arbitrary. Determining the assumptions underlying a RCT with radiographic progression in AS as outcome parameter is particularly difficult, because not much is known about the effect of interventions on radiographic progression. Data from a study in RA (10) showed that anti-TNF treatment inhibited radiographic progression, which might be supportive for our assumption of a progression of zero. However, despite all the assumptions and uncertainties associated with sample size calculations, there is a precedent that shows that a sample size of 94 patients may provide sufficient statistical power. Recently a RCT in AS was performed in which radiographic progression of 2- years was used as primary outcome parameter (11). In this RCT continuous versus on demand intake of non-steroidal-anti-inflammatory drugs was compared with regard to radiographic progression. Radiographic progression was assessed by the modified SASSS with a paired scoring order. The two treatment groups consisted of 74 and 76 patients, and a between-group difference of 1.1 was found to be statistically significant in this study.

Therefore, based on theoretical arguments and on the results of this study we recommend RCTs in AS with radiographic progression as an endpoint to be designed with 2 years duration and to be scored by paired reading order.

Figure 1 Probability plot of 1-year progression in modified SASSS scores for paired and chronological reading order

Figure 2 Probability plot of 2-year progression in modified SASSS scores for paired and chronological reading order

## REFERENCES

1. van der Heijde D, van der Linden S, Bellamy N, Calin A, Dougados M, Khan MA. Which domains should be included in a core set for endpoints in ankylosing spondylitis? Introduction to the ankylosing spondylitis module of OMERACT IV. *J Rheumatol* 1999;26(4):945-7.
2. Wanders AJB, Landewé RBM, Spoorenberg A, Dougados M, van der Linden S, et al. What is the most appropriate radiological scoring method in Ankylosing spondylitis clinical trials. A comparison based on the OMERACT filter. *EULAR* 2003;abstract SP0173.
3. Ferrara R, Priolo F, Cammisa M, Bacarini L, Cerase A, Pasero G, et al. Clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from the GRISAR Study. Gruppo Reumatologi Italiani Studio Artrite Reumatoide. *Ann Rheum Dis* 1997;56(10):608-12.
4. 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(10):2055-6.
5. van der Heijde D, Boonen A, Boers M, Kostense P, van der Linden S. Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. *Rheumatology (Oxford)* 1999;38(12):1213-20.
6. Bruynesteyn K, van der Heijde D, Boers M, Saudan A, Peloso P, Paulus H, et al. Detecting radiological changes in rheumatoid arthritis that are considered important by clinical experts: influence of reading with or without known sequence. *J Rheumatol* 2002;29(11):2306-12.
7. Spoorenberg A, de Vlam K, van der Linden S, Dougados M, Mielants H, van de Tempel H, et al. Radiological scoring methods in Ankylosing Spondylitis. Reliability and sensitivity to change over one and two years. *J Rheumatol* 2003;In press.
8. Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, et al. Relative value of erythrocyte sedimentation rate and C-reactive protein in assessment of disease activity in ankylosing spondylitis. *J Rheumatol* 1999;26(4):980-4.
9. Landewé R, van der Heijde DF. Radiographic progression visualised by probability plots. *Arthritis Rheum* 2004;in press.
10. Lipsky PE, van der Heijde DM, St Clair EW, Furst DE, Breedveld FC, Kalden JR, Smolen JS, Weisman M, Emery P, Feldmann M, Harriman GR, Maini RN. INfliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. *N Engl J Med* 2000; 343 (22) : 1594-602.
11. Wanders A, van der Heijde D, Landewé R, Béhier J-M, Calin A, Olivieri I, et al. Inhibition of radiographic progression in ankylosing spondylitis by continuous use of NSAIDs. *ACR* 2003;abstract 518.



