Detecting Radiological Changes in Rheumatoid Arthritis That Are Considered Important by Clinical Experts: Influence of Reading With or Without Known Sequence

KARIN BRUYNESTEYN, DÉSIRÉE VAN DER HEIJDE, MAARTEN BOERS, ARIANE SAUDAN, PAUL PELOSO, HAROLD PAULUS, HARRY HOUBEN, BRIDGET GRIFFITHS, JOHN EDMONDS, BARRY BRESNIHAN, ANNELIES BOONEN, and SJEF VAN DER LINDEN

ABSTRACT. ***Objective.*** To evaluate whether knowledge of the chronological sequence influences the sensitivity and specificity of the Sharp/van der Heijde (SvH) and Larsen/Scott (LS) scoring method to detect clinically important progression of joint damage caused by rheumatoid arthritis (RA) in the individual patient and assess whether scoring in chronological order leads to better sensitivity at the cost of lower specificity.

Methods. For both scoring methods, progression scores obtained with (chronological) and without knowledge of the sequence of the films (paired) were compared with the judgment of an international expert panel. This panel assessed whether progression of joint damage seen on films with 1 year intervals was clinically relevant (defined as progression of joint damage that would make clinicians change therapy). The applied thresholds for clinical relevance were (1) the progression scores with the highest accuracy by receiver operating characteristics analyses for the expert opinion, and (2) the smallest progression score that can be detected apart from interobserver measurement error by the scoring method, i.e., the smallest detectable difference (SDD).

Results. Progression scores that detected clinically relevant progression most accurately (chronological: 3.0 SvH units and 2.0 LS units; paired: 2.5 SvH units and 1.5 LS units) were smaller than the SDD (chronological 5.0 SvH units and 5.8 LS units; paired 13.8 SvH units and 9.7 LS units). With the SDD as threshold, the sensitivity to detect clinically relevant progression increased significantly from 20 to 60% for the SvH method and from 23 to 33% for the LS method if the sequence of the films was known. The specificity remained good when scoring chronologically: 88% for the SvH and 100% for the LS.

Conclusion. Our results suggest that knowing the chronological sequence leads to an increase in detecting clinically relevant changes in the patient without serious overestimation of nonrelevant differences. Analyzing a clinical trial should be done preferably by reading films in chronological order. (J Rheumatol 2002;29:2306–12)

Key Indexing Terms:
RHEUMATOID ARTHRITIS
RADIOGRAPHIC SCORING METHODS
SMALLEST DETECTABLE DIFFERENCE
RADIOGRAPHS
READING ORDER
CLINICALLY IMPORTANT DIFFERENCE

From the Department of Internal Medicine, Division of Rheumatology, Maastricht University, Maastricht, the Department of Rheumatology, Atrium Medical Center, Heerlen, and the Department of Clinical Epidemiology and Biostatistics, VU Medical Center, Amsterdam, The Netherlands; Limburg University Center, Diepenbeek, Belgium; Centre Médical De L'Aeroport, Geneva, Switzerland; the Department of Rheumatology, University of Iowa Health Care, Iowa City, and the Department of Rheumatology, UCLA School of Medicine, Los Angeles, USA; the Muskuloskeletal Unit, Freeman Hospital, Newcastle Upon Tyne, United Kingdom; the Department of Rheumatology, St. George Hospital, Sydney, Australia; and the Department of Rheumatology, St. Vincent's Hospital, Dublin, Ireland.

Supported by the Dutch Arthritis Association.

K. Bruynesteyn, MD, Division of Rheumatology, Maastricht University; D. van der Heijde, MD, PhD, Professor of Rheumatology, Maastricht University and Limburg University Center; M. Boers, MSc, MD, PhD, Rheumatologist, Professor of Clinical Epidemiology, VU Medical Center; A. Saudan, MD, Rheumatologist, Centre Médical De L'Aeroport; P. Peloso, MSc, MD, FRCPC, Rheumatologist, University of Iowa Health Care; H. Paulus, MD, Professor of Medicine, UCLA School of Medicine; H. Houben, MD, PhD, Rheumatologist, Atrium Medical Center; B. Griffiths, MD, MRCP, Rheumatologist, Freeman Hospital; J. Edmonds, MBBS, FRACP, Professor of Rheumatology, St. George Hospital; B. Bresnihan, MD, FRCP, FRCPI, Professor of Rheumatology, St. Vincent's Hospital; A. Boonen, MD, Rheumatologist; S. van der Linden, MD, PhD, Professor of Rheumatology, Maastricht University.

Address reprint requests to Dr. K. Bruynesteyn, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands.
E-mail: kbru@sint.azm.nl

Submitted February 5, 2002; revision accepted April 18, 2002.

In rheumatoid arthritis (RA), several scoring methods have been developed to quantify radiological damage in the joints

of the hands and feet. In longitudinal studies, the order in which films are presented to the observer influences results[1–4]. Roughly, one can distinguish 3 ways of ordering films obtained at 2 or more points in time. First, films can be presented to the reader completely at random, i.e., a single film at a time. Second, films can be grouped per patient and presented to the reader without data on the chronological order of the films, which we call paired scoring. Finally, films can be grouped per patient and presented in chronological order.

Reading single films randomly has the major drawback that the reader cannot compare with contralateral joints or with identical joints taken at other moments. Hence, the reader will not be able to correct for variation in positioning of the hands and feet or for film quality, which may contribute to the introduction of measurement error. In 1986, Fries, *et al*[1] demonstrated that precision of paired scoring was greater than reading single films randomly. Two Italian studies confirmed this finding in 1997[2,3]. Chronological reading provides the reader with a maximum of information, thereby reducing measurement error. Theoretically, reading films chronologically results in an increased ability to detect changes than the paired reading order. Van der Heijde, *et al*[4] showed in 1999 that reading in chronological order was most sensitive to change. However, the possibility that chronological reading order overestimated progression of joint damage because the readers expected to see progression over time could not be excluded. In other words, it could not be excluded that the extra signal picked up by the chronological reading order was actually a false signal caused by expectation bias; especially because the Sharp/van der Heijde (SvH) progression scores used could not decrease by definition when applied chronologically[5].

An appropriate method to distinguish between a more precise signal, by reducing the measurement error, and false signal (bias), is to compare the progression scores of the chronological and paired reading order of the scoring methods with an external criterion for progression of joint damage. The amount of progression that would make clinicians change therapy, in other words the amount of progression that is considered clinically important, can be regarded as a relevant external criterion for this purpose. In routine clinical practice, films are judged with known sequence of the films. Hence, the judgments of a panel of rheumatologists aware of the sequence seems intuitively most appropriate as external criterion. However, overestimation by rheumatologists when judging films chronologically can of course not be excluded either, although decisions on therapy changes are presumably made with prudence, making overestimation of clinically relevant progression less likely. All things considered, it was decided to present the films to the panel with and without information on the sequence of the films. However, before analyzing the 2 readings, the validity and reliability of the chronological and paired panel must be examined.

Our aim was to evaluate the influence of the paired and chronological reading order on the ability of the SvH[5] and Larsen/Scott[6] (LS) scoring methods to detect clinically important changes on radiographs deemed important by an external panel of clinicians. We also wanted to assess whether scoring in chronological order leads to better sensitivity at the cost of lower specificity.

MATERIALS AND METHODS

The judgments of an international panel of rheumatologists on the clinical relevance of progression of joint damage seen on sets of films with 1 year intervals was used as external criterion. The majority opinion of 5 (3, 4 or 5 out of 5) clinicians was compared with the paired and chronological progression scores of the SvH and LS methods, each obtained from 2 different pairs of observers. The same expert members, films and SvH and LS readers have been used in another study. The subject of that study was the minimal clinically important difference (MCID) of the SvH and the LS scoring methods defined by clinical experts[7].

Expert panel. The expert panel consisted of 5 rheumatologists (BB, BG, HH, HP, and PP) who independently evaluated 46 pairs of hand and foot films, taken at 1 year intervals, of patients with early RA with varying followup duration (see also the section patients and films). The experts were chosen from several different countries based on their expertise in the treatment of RA. None had been trained in either of the scoring methods, but each was experienced in reading films in daily practice. The panel experts were first asked whether they noticed any progression of joint damage due to RA between the 2 sets of hand and foot films in 1 pair of films. Second, if they noticed progression, they had to state whether they considered that difference in joint damage clinically relevant in a typical patient with early RA: a 46-year old woman with a 2 year history of RA, with high disease activity, treated with methotrexate for 1 year. Clinically relevant progression was defined as that progression of joint damage that would make a clinician change second line therapy. In the original study, the panel also considered 3 other clinical scenarios[7]. As the results with these scenarios did not add relevant information for this study, they have been omitted here. Each panel expert viewed the radiographs 4 times: twice in chronological order and twice in paired order, at an interval of at least 4 weeks, to estimate the intrapanel reliability (variability between the first and second reading of the panel). Three panel members started with the chronological reading order, the other 2 with the paired reading order, to minimize possible bias caused by learning effects. The order in which patient sets of films were presented was different in the 4 viewing sessions. Unless stated otherwise, the opinion of the first viewing session of both reading orders was used in the analyses.

Radiographic scoring methods and observers. Two experienced readers scored the radiographs with and without knowledge of the sequence of the film, according to the SvH (DvH and AB) and LS method (JE and AS). Two different readers were used for each scoring method, because the readers can not be experienced in both scoring methods at the same time. The SvH method assesses erosions (hands 0-5; feet 0-10) and joint space narrowing (0-4) separately and has a range from 0 to 448[5]. The LS method has a range from 0 to 200 for hands and feet and applies one grade (0-5) to each joint[6]. The wrists are evaluated as single joints and are weighted by a factor of 5. The chronological SvH method was applied with the rule that scores cannot decrease by definition[8]. The scores of the chronological LS and the scores of the paired reading orders of both scoring methods could however decrease in time. The mean scores of each pair of observers were used for further analyses.

Patients and films. The film sets of a recent study on precision and sensitivity to change of the SvH method were used in this study[4]. The 46 pairs

of hand and foot films, made in posterioanterior view, were obtained from 22 patients. The films were selected in the previous study for high and low baseline scores and for high and low progression scores between the 2 sets of 1 pair of films. All patients fulfilled the 1987 ACR classification criterion for RA and had a disease duration of < 1 year at start. Ten patients had had a followup period of 1 year and supplied 1 pair of films each, 12 patients had had a followup period of 3 years and supplied 3 pairs each.

Statistical analyses. Receiver operating characteristics (ROC) curve analyses assessed the sensitivity and specificity of all available progression scores of the different reading orders of the 2 scoring methods to discriminate between clinically relevant and no (relevant) progression using judgment of the experts as standard[9]. An ROC curve plots the true positive rate (sensitivity) in function of the false positive rate (100 – specificity) at all possible cutoff levels. An overall index of goodness of the test, i.e. scoring method, is the area under the curve (AUC). A nondiscriminating test has an area of 0.5 and a perfect discriminating test has an area of 1.0. The threshold level that discriminates best is the progression score with the highest accuracy, i.e., that progression score with the best combination of sensitivity and specificity. In the ROC curve this is the cutoff point nearest the upper left corner. A second analysis assessed the sensitivity and specificity of the 2 reading orders if the smallest progression score that can be detected apart from measurement error, the smallest detectable difference (SDD), was applied as threshold. The SDD is a statistical measure based on the 95% limits of agreement as described by Bland and Altman[10-12]. This method helps to decide whether a difference between 2 scores of an individual patient is a real change, or one that cannot be separated reliably from differences caused by random variability (measurement error). Progression scores smaller than the SDD cannot be distinguished reliably from measurement error. Consequently, a threshold value should at least exceed the SDD, the measurement error in the study in question. In clinical trials it has been advised to report radiographic results as the mean score of 2 observers; in line with this, we applied the SDD based on this mean score of change. Besides being specific for the sets of films involved, the magnitude of this SDD also depends on the quality of the specific observers and on whether one wishes to generalize to other pairs of observers. The SDD applied in this study was restricted to the current pairs of observers. It has been shown that the interobserver SDD, restricted to the same observer pair, represents clinically relevant changes for the chronological reading order of both scoring methods: in the previous study with the same observers changes even smaller than these SDD were already considered clinically important by our panel[7].

Descriptive analyses, kappa statistics, and intraclass correlation coefficients (ICC type 2,1) were performed by SPSS 10.0 for Windows. ROC curves were performed by MedCalc (Mariakerke, Belgium) statistical software. For both scoring methods, McNemar chi-square tests for paired proportions compared the accuracy (sensitivity/specificity) of paired and chronological reading methods at the 2 thresholds (optimum accuracy and SDD). The 95% confidence intervals of the differences in accuracy were calculated according to Gardner and Altman[13]. A 2 sided p value of ≤ 0.05 was considered significant.

RESULTS

Expert panel opinion. Table 1 shows the prevalence of film sets with (clinically important) progression of joint damage in the hands and/or feet according to the chronological and paired panels. The prevalence of the film sets judged by the expert panel as progressive dropped from 80 to 37% if the sequence of the films was not known by the experts. The number of sets with clinically relevant progression decreased or increased to the same extent. However, the reliability of the paired panel did not decrease the same way, demonstrated by the kappas, reflecting the variability between the first and second reading of the panel (Table 1). When the sequence of the films was not known by the panel members, the panel members refused *en masse* to give a judgment on the (relevance of the) progression unless the changes were huge. Consequently it was impossible to use the paired panel to evaluate the ability to detect clinically relevant change by the paired progressions scores, i.e., to evaluate the change scores as threshold for minimal clinically relevant change, which is the purpose of the study. These results made us decide that clinicians really are experts only if films are presented to them in chronological order. Consequently, only the opinions of the experts judging chronologically ordered sets of films were evaluated further and discussed.

Radiographic scores. Table 2 shows the distribution characteristics of the 2 reading orders for both scoring methods (note the different maximum obtainable scores). The chronological progression scores of the SvH method were evidently higher than the paired progression scores. The mean chronological SvH progression score was 7.6 (SD 10.0), the median 4.0 [inter quartile range (IQR): 2.4–8.6]; that is 1.7 and 0.9% of the maximally obtainable score, respectively. The mean paired SvH progression score was 4.5 (SD 10.2), the median 2.5 (IQR: –1.0–7.6); 1 and 0.6% of the maximally obtainable score, respectively. The difference between the paired and chronological progression scores of the SvH method was statistically significant (p = 0.001 Wilcoxon signed rank test). The influence of the reading orders on the LS progression scores was not that clear. The mean chronological progression score of the LS method was 4.0 (SD 8.0), the median 0.8 (IQR: 0.0–3.6),

Table 1. Prevalence of film sets with progression of joint damage in the hands and feet according to the opinions of the chronological and paired expert panels and the variability between the first and the second reading of the panels (observed agreement and chance-adjusted agreement).

	Chronological Panel			Paired Panel		
	Prevalence of Progression* % (n)	Observed Agreement (Proportion)	Kappa	Prevalence of Progression % (n)	Observed Agreement %	Kappa
Progression of joint damage	80 (37)	0.85	0.45	37 (17)	80	0.60
Clinically important progression	65 (30)	0.80	0.59	24 (11)	91	0.76

* Prevalence of film sets with progression seen in the first viewing session.

Table 2. Distribution characteristics and interobserver reliability of the different reading orders of the 2 scoring methods.

	Sharp/van der Heijde Method (range 0–448)		Larsen/Scott Method (range 0–200)	
	Chronological	Paired	Chronological	Paired
Baseline scores				
Mean (SD)	24.6 (16.5)	25 (16.0)	14.5 (10.4)	16.8 (11.2)
Median	19.5	21.5	15.3	15.8
IQR	11.9–35.4	14.4–31.8	5.8–19.8	9.4–23.1
Range	2.0–62.5	1.5–59.0	0.0–37.5	0.0–52.0
Progression total scores				
Mean (SD)	7.6 (10.0)	4.5 (10.2)	4.0 (8.0)	3.7 (10.3)
Median	4.0	2.5	0.8	1.0
IQR	2.4–8.6	–1.0–7.6	0–3.6	–0.6–5.9
Range	0–51.0	–13.0–52.0	–3.5–43.5	–12.0–55.0
Interobserver reliability				
SDD progression score	5.0	13.8	5.8	9.7
ICC progression socre	0.94	0.63	0.88	0.80

IQR: interquartile range; SDD: smallest detectable difference.

i.e., 2 and 0.04% of the maximally obtainable score. The corresponding figures for the paired LS progression scores were 3.7 (SD 10.3) and 1.0 (–0.6–5.9), i.e., 1.8 and 0.05% of the maximally obtainable score. The agreement between the 2 readers of each scoring method diminished if the order in time was not known by the readers, shown by the increase in SDD and decrease in intraclass correlation coefficients (Table 2).

Influence of the reading order on the sensitivity and specificity to detect clinically important differences in the individual. Figure 1 shows the ROC curves of the chronological and paired progression scores of the SvH and LS scoring method. The SvH AUC for the chronological reading order was 0.83 and for the paired order 0.82. The LS AUC were 0.84 and 0.79, respectively. So the AUC of the scoring methods were virtually identical and quite acceptable for both reading orders.

At the most accurate threshold for clinically important progression, little accuracy was lost when moving from chronological to paired reading in both scoring methods (Table 3). The thresholds were slightly lower for the paired reading order (2.5 SvH units and 1.5 LS units) than for those of the chronological order (3.0 and 2.0, respectively), like the majority of other progression scores.

At the SDD threshold for clinically important progression, the results were different (Table 3). Because the SDD of the paired reading order was much higher than that of the chronological order, applying this threshold resulted in a decrease in sensitivity of both the SvH method (40%, from 60 to 20%) and the LS method (10%, from 33 to 23%) and an increase of specificity of the SvH (12%, from 88 to 100%). The difference in sensitivity between the chronological and paired SvH was highly statistically significant ($p < 0.0001$), in contrast to the other differences.

Sensitivity analysis. One might argue that the paired reading order of the SvH was too insensitive because our SvH readers were not experienced enough in scoring films in paired order and as a result disagreed too much with each other, in contrast with LS readers. Therefore, we also analyzed the accuracy of the paired scoring method as if the readers had agreed more often and as a consequence the SDD would have been smaller. An imaginary SDD of 8.5 units was chosen, based on the ratio of 1.7 (9.7:5.8) between the paired (9.7) and chronological SDD (5.8) of the LS method. With this SDD as threshold the sensitivity of the paired SvH increased to 33%.

All analyses were repeated with the concordant opinion of the panel as standard (i.e., progression seen in the first as well as in the second session by the majority of the panel), which led to results very similar to those presented here for the first reading only (data not shown).

DISCUSSION

Underlying this study was the understanding that chronological reading order seemed to be more sensitive to change, but that this increased sensitivity might be due to expectation bias rather than diminishing of measurement error. In order to assess if the extra signal picked up by reading in chronological order was principally due to bias or indeed represented a more precise measure, we determined the influence of the paired and chronological reading order on the ability of the SvH and LS scoring methods to detect clinically relevant progression (defined by an expert panel) of radiological damage of hands and feet in the individual patient. To make a totally fair comparison between the reading orders and the opinion of the panel, we decided first to present the films to the expert members with and without information on the sequence of the films. *A priori* we

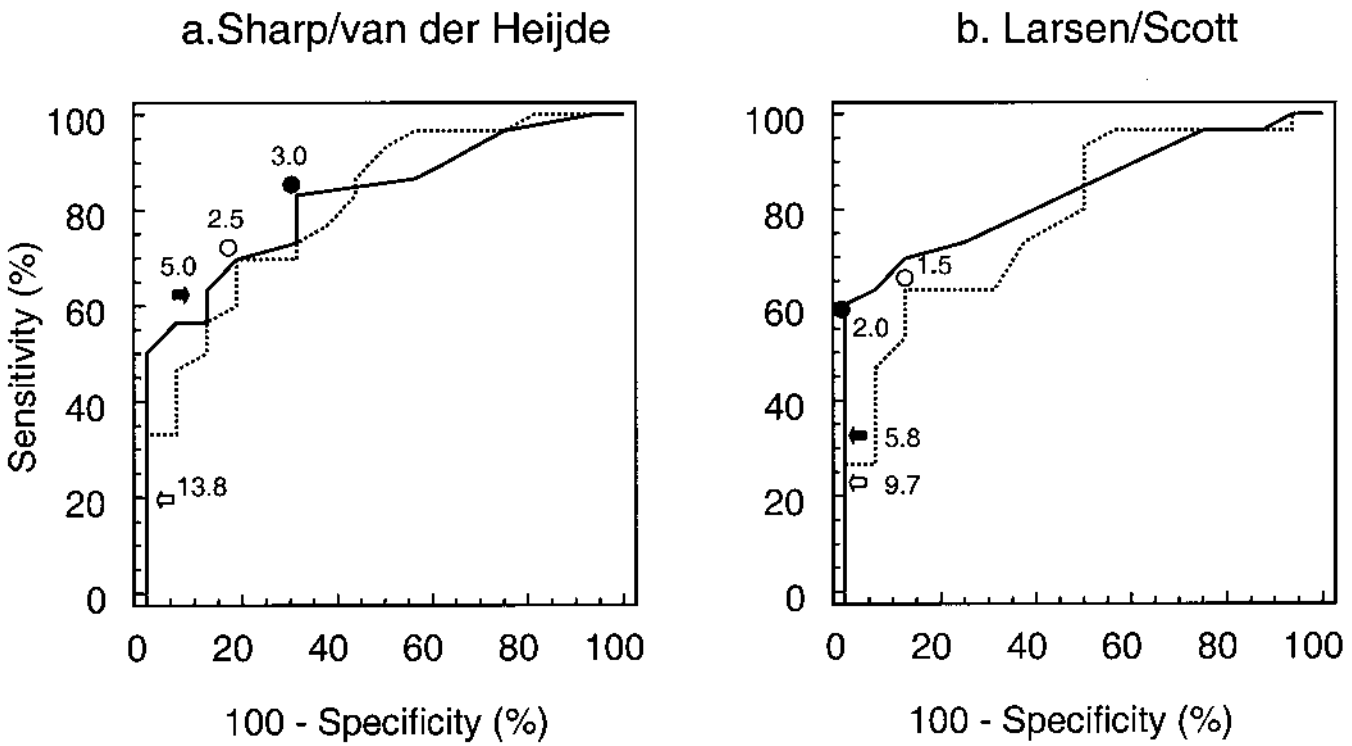

Figure 1. Receiver operator characteristic (ROC) curves of the chronological and paired progression scores of the Sharp/van der Heijde (a) and the Larsen/Scott (b) scoring methods, using the opinion of the expert panel as external criterion. Black line in curve: chronological reading order; dotted line in curve: paired reading order; n: chronological progression score with highest accuracy; n: paired progression score with highest accuracy; white arrows: SDD chronological progression scores; black arrows: SDD paired progression scores; Sensitivity: percentage of patients with clinically important progression of damage correctly labeled by the scoring method (true positive rate); Specificity: percentage of patients without clinically important progression of damage correctly labeled by the scoring method (true negative rate); accuracy: combination of sensitivity and specificity; SDD: smallest detectable difference.

considered the judgment of the panel that was not aware of the sequence not appropriate as external criterion. The results in Table 1 supported this view and allowed us to abandon the paired panel as external criterion for the paired reading order. Clinicians are not experts in judging clinical relevance of progression if the chronological sequence is not known. If they don't know the chronological sequence they are only willing to change therapy if they are absolutely sure that the damage worsened, i.e., a large difference that uncovers the sequence. The somewhat higher percentage of observed agreement of the paired panel is also a reflection of this: it is much easier to agree on large changes than on small ones.

A possible overestimation of progression by the chronological reading order would result theoretically in higher sensitivity at the cost of lower specificity. At their most accurate threshold, the accuracy of paired and chronological reading orders was roughly similar for both scoring methods. However, these thresholds comprised progression scores that were smaller than the SDD and can thus not be

Table 3. Highest accuracy characteristics and accuracy characteristics of the reading orders of the 2 scoring methods applying the smallest detectable difference (SDD) as threshold.

	Sharp/van der Heijde (Range 0–448)			Larsen/Scott (Range 0–200)		
	Chronological, % (SDD 5.0)	Paired, % (SDD 13.8)	Difference, % (95% CI)	Chronological, % (SDD 5.8)	Paired, % (SDD 9.7)	Difference, % (95% CI)
Highest accuracy characteristics						
Sensitivity	83	70	13 (–6–20)*	60	63	–3 (–22–17)*
Specificity	69	81	–12 (–35–21)*	100	88	12 (–9–13)*
Accuracy characteristics of the SDD as threshold						
Sensitivity	60	20	40 (19–40)†	33	23	10 (–4–10)*
Specificity	88	100	–12 (–13–9)*	100	100	0 (n/a)*

* $p > 0.20$; † $p < 0.0001$ (McNemar Chi-square tests).

distinguished reliably from measurement error in the individual case. Without knowledge of the sequence, scoring became less reproducible, increasing measurement error and thus the SDD. If the SDD was subsequently applied as threshold for clinically relevant progression, sensitivity of paired reading proved remarkably lower than that of chronological reading, especially for the SvH method. As the specificity of chronological reading by this method was already good (88%), the increase to 100% specificity by paired reading still represents only a modest gain. Hence, the advantage of higher sensitivity to pick up change caused by less measurement error surpassed the adverse consequence (loss in specificity) by the possible introduction of bias. At 33%, the sensitivity of the chronological LS method was already much less than that of the SvH method, and this decreased further to 23% in paired reading. Specificity remained perfect regardless of reading order. With the prospect of more powerful disease modifying antirheumatic drugs, future trials will document very low rates of progression or arrest of joint damage, thus requiring increasingly sensitive methods. This would be an argument for the chronological SvH scoring method.

The van der Heijde modification of the Sharp score included the rule that the scores could not decrease by definition if scoring in chronological order. This was based on the frequent experience that at one moment an erosion would be clearly visible, at the next it would be gone, only to reappear at the next assessment. Theoretically the erosion could of course have healed at moment 2 and reappeared at moment 3, but this was thought unlikely from a pathophysiological point of view. Rather, variation in film quality or positioning of the hands/feet was deemed more likely. So this rule was instituted to further reduce measurement error. The drawback is of course that this rule enhances the possibility of overestimation of progression (bias) and that healing of erosions could not occur. As the rule was not applied for the chronological LS method, improvement of scores could occur. However, in our set of films this was a rare phenomenon: in only 1 set of films improvement was observed by both Larsen readers. One may therefore conclude that the influence of the rule on our results was probably negligible. Finally, the fact that the chronological SvH method was applied in this study with the rule that the scores cannot decrease, doesn't imply that we don't believe that healing can occur. However, we believe that regarding the assessment of healing, scores should not simply be subtracted.

Our panel was made up of clinicians from different continents to ensure generalizability of the results. The intrapanel agreement on clinically relevant changes was satisfactory, as shown by a kappa value of 0.59. The intrapanel agreement on progression alone, irrespective of clinical relevance, on the other hand, was low (0.45) in a setting of a high level of observed agreement (0.85). The paradox of "high observed agreement and low kappa" is a well known feature of kappa statistics and is caused by the fact that kappas are affected by the prevalence of the disease or condition concerned[14,15]. Very low or high disease prevalence can result in misleadingly low kappas despite good agreement.

The film sets we used were selected for high and low baseline scores and for high and low progression scores to reflect the spectrum of damage found in early RA trials. Our data selection was, however, restricted in that it did not include many film sets with major progression of joint damage. The sample, although rather small, was large enough to detect differences between the chronological and paired progression scores of the SvH method. However, the difference in sensitivity of the LS reading orders, also in favor of the chronological reading order, did not reach statistical significance in our rather small sample.

We used the opinion of an expert panel on the clinical relevance of progression in a patient with recent onset RA and high disease activity as external criterion. However, the influence of progression of radiological joint damage on the rheumatologists' treatment strategies could have differed if the disease duration had been longer or if the patient had only mild disease activity. In our previous study[7], which involved the same panel members, the clinical importance of radiological joint progression was also assessed for 4 different clinical settings (early vs late RA and mild vs high disease activity). In that study, the level of progression scores with the highest accuracy indeed varied somewhat per setting: the panel was more likely to judge progression relevant for patients with early disease and high disease activity than for patients with late RA and mild disease activity, with the other 2 settings fitting in between. However, these highest accuracy progression scores also remained smaller than the SDD throughout, leading to similar conclusions for the different settings.

In this study, clinically important progression was defined as that progression in radiological joint damage that makes the rheumatologist change therapy. This decision of course also depends on other factors such as patient's history of toxic reactions or availability of alternative treatment (such as tumor necrosis factor-α inhibitors). However, the panel was explicitly instructed not to include these factors. We asked them to state only their intention to change treatment, without taking into account other factors except the radiological progression of joint damage and the specification of the setting given. From the results of the previous study it was also clear that their judgments were indeed based on what they saw on the films and not merely on the clinical description.

In conclusion, our study confirms greater sensitivity in detecting differences by scoring films in chronological order. These differences were defined as clinically relevant by an international expert panel. Therefore, knowing the sequence of films did not lead to overestimation of nonrele-

vant differences, but enabled better detection of clinically relevant changes. Scoring films without knowing their chronological sequence substantially decreased sensitivity in the detection of clinically relevant changes as defined by an expert panel in comparison with scoring films in chronological order. In clinical trials of early RA, our results strongly suggest that radiographs should be read in chronological order.

REFERENCES

1. Fries JF, Bloch DA, Sharp JT, et al. Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial. Arthritis Rheum 1986;29:1-9.
2. Ferrara R, Priolo F, Cammisa M, et al. Clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from the Grisar study. Ann Rheum Dis 1997;56:608-12.
3. Salaffi F, Carotti M. Interobserver variation in quantitative analysis of hand radiographs in rheumatoid arthritis: comparison of 3 different reading procedures. J Rheumatol 1997;24:2055-6.
4. 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;38:1213-20.
5. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 2000;27:261-3.
6. Scott DL, Houssien DA, Laasonen L. Proposed modification to Larsen's scoring methods for hand and wrist radiographs. Br J Rheumatol 1995;34:56.
7. Bruynesteyn K, van der Heijde D, Boers M, et al. Determination of the minimal clinically important difference in rheumatoid arthritis joint damage of the Sharp/van der Heijde and Larsen Scott scoring methods by clinical experts and comparison with the smallest detectable difference. Arthritis Rheum 2002;46:913-20.
8. van der Heijde DM. Plain X-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. Baillieres Clin Rheumatol 1996;10:435-53.
9. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283-98.
10. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
11. Bland JM, Altman DG. Measurement error. BMJ 1996;313:744.
12. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731-9.
13. Gardner M, Altman D. Statistics with confidence: confidence intervals and statistical guidelines. London: British Medical Journal; 1989.
14. Lantz CA, Nebenzahl E. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. J Clin Epidemiol 1996;49:431-44.
15. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543-9.