Minimal Clinically Important Difference in Radiological Progression of Joint Damage over 1 Year in Rheumatoid Arthritis: Preliminary Results of a Validation Study with Clinical Experts

KARIN BRUYNESTEYN, DÉSIRÉE van der HEIJDE, MAARTEN BOERS, MARISSA LASSERE, ANNELIES BOONEN, JOHN EDMONDS, HARRY HOUBEN, HAROLD PAULUS, PAUL PELOSO, ARIANE SAUDAN, and SJEF van der LINDEN

ABSTRACT. To determine the minimal clinically important difference (MCID) between hand and foot films with a 1 year interval assessed with the Sharp/van der Heijde or Larsen/Scott scoring method. Progression scores of the 2 methods were compared with the opinion of an international expert panel on clinical relevance of radiological joint damage in 4 predefined clinical settings. The expert panel consisted of 3 rheumatologists, who evaluated 46 pairs of hand and foot films, taken with 1 year intervals, of patients with early rheumatoid arthritis. Receiver operating characteristics curves analyzed the accuracy of different threshold values (progression scores) of the 2 scoring methods to detect the presence or absence of clinically important difference, as defined by the expert panel as external criterion. The threshold value with the highest accuracy was subsequently chosen as the score representing the MCID. Five Sharp/van der Heijde units and 2 Larsen/Scott units were the best cutoffs. The accompanying sensitivities ranged from 77% to 100% for the Sharp/van der Heijde method and from 73% to 84% for the Larsen/Scott method for the 4 clinical settings. The specificities were between 78% and 84% for the Sharp/van der Heijde method and between 74% and 94% for the Larsen/Scott method. The smallest progression score that can be detected apart from interobserver measurement error, the smallest detectable difference (SDD), was equal to or larger than the calculated MCID, 5 Sharp/van der Heijde units and 6 Larsen/Scott units in our study, if the mean progression scores of the same 2 observers were used. The SDD is a conservative estimate of the MCID; our panel rated progression at or below this level as clinically significant. (J Rheumatol 2001;28:904–10)

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

INTRODUCTION

Radiographs are important measurements to assess progression of joint damage caused by rheumatoid arthritis (RA). Moreover, they are included in the World Health Organization/International League of Associations for Rheumatology core set of measurements to evaluate clinical trials with duration of 1 year or more[1]. In clinical practice, a qualitative assessment of radiological damage is usually considered as sufficient, but for therapeutic trials or cohort studies quantitative methods are needed. Several scoring methods have been developed and evaluated through the years[2]. The methods used most widely are the ones developed by Larsen and Sharp and their modifications[3-10].

Besides results from mean group scores, the number of

From the Department of Rheumatology, University Hospital Maastricht, Maastricht, The Netherlands; Limburg University Center, Diepenbeek, Belgium; Department of Clinical Epidemiology, VU University Hospital, Amsterdam, The Netherlands; Department of Rheumatology, St. George Hospital, Sydney, Australia; Department of Rheumatology, Atrium Heerlen, Heerlen, The Netherlands; Department of Rheumatology, UCLA School of Medicine, Los Angeles, USA; Department of Rheumatology, Royal University Hospital, Saskatoon, Canada; and Centre médical de l'aeroport, Geneva, Switzerland.

Supported by a grant from the Dutch Arthritis Association.

K. Bruynesteyn, MD, Department of Rheumatology, University Hospital Maastricht; D. van der Heijde, MD, PhD, Department of Rheumatology, University Hospital Maastricht, and Limburg University Center; M. Boers, MSc, MD, PhD, Department of Clinical Epidemiology, VU University Hospital; M. Lassere, MB, BS, FRACP, PhD, FAFPHM, Department of Rheumatology, St. George Hospital; A. Boonen, MD, Department of Rheumatology, University Hospital Maastricht; J. Edmonds, MB, BS, FRACP, St. George Hospital; H. Houben, MD, Atrium Heerlen; H. Paulus, MD, Department of Rheumatology, UCLA School of Medicine; P. Peloso, MSc, MD, FRCPC, Department of Rheumatology, Royal University Hospital; A. Saudan, MD, Centre medical de l'aeroport; S. van der Linden, MD, PhD, Department of Rheumatology, University Hospital Maastricht.

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

patients actually responding to a particular drug, according to predefined criteria, is an important endpoint in trials. The American College of Rheumatology and the European League of Associations for Rheumatology criteria for clinical response in RA have found enthusiastic acceptance in trial reports[11-18]. The clinical relevance of a percentage of responders is usually easier to understand than a mean change on a group level. This is also, or perhaps even more, true for radiographic results. In order to develop the response criteria for radiographic data, a definition of the minimal clinically important difference (MCID) in joint damage between radiographs is needed. Patients with progression of joint damage exceeding this difference are the so-called progressors, or in response terminology the nonresponders. Lassere, *et al* recently proposed to use the smallest progression score that can be detected apart from interobserver measurement error, the smallest detectable difference (SDD), as minimum difference[19]. The SDD is based on statistical analyses only; it is not known whether it also represents a clinically relevant progression.

The aim of our study was to determine the MCID. We compared the progression scores of 2 scoring methods, the van der Heijde modified Sharp scoring method and the Scott modified Larsen method, with the opinion of an expert panel on the clinical relevance of the radiological joint damage in 4 predefined clinical settings. This expert panel is our external criterion or pseudo-gold standard. Receiver operating characteristic (ROC) curves analyzed the accuracy of different threshold values of the 2 scoring methods to detect the presence or absence of clinically important difference, as defined by the expert panel[20]. The threshold value with the highest accuracy was subsequently chosen as the score representing the MCID.

MATERIALS AND METHODS

Patients and radiographs. Forty-six pairs of hand (including the wrist) and foot films were used in this study, all taken with a 1 year interval. The films were obtained from 22 patients with early RA who fulfilled the 1987 ACR criteria for RA. Ten patients had had a followup period of 1 year and supplied 1 pair each, 12 patients had had a followup period of 3 years after diagnosis and supplied 3 pairs each. Radiographs were made in posterioanterior view and were blinded for identity. The pairs were offered in random order. However, the chronological time sequence within the sets of 1 patient was known by the readers as well as by the expert panel members, as is usual in clinical practice. The radiographs have also been used in a recent study on the influence of the reading order on precision and sensitivity to change of radiographic scoring methods[21].

Radiographic scoring methods. Radiographs were scored independently by experienced readers according to the Sharp/van der Heijde (DvdH and AB) and the Larsen/Scott method (JE and AS). The Sharp/van der Heijde method has a range from 0 to 448 and grades erosions and joint space narrowing separately. Thirty-two joints for erosions in the hands and 12 in the feet are included, with a maximum score of 5 per joint in the hands and 10 in the feet. Joint space narrowing is graded from 0 to 4 in 30 joints in the hands and 12 joints in the feet. Applying the Sharp/van der Heijde method with known sequence, scores cannot decrease by definition. The Larsen/Scott method has a range from 0 to 200 if the hands, wrists, and feet are scored. One grade is applied to each joint. Twenty joints in the hands and 10 joints in the feet are graded. The wrists are evaluated as single joints. Scores range from 0 to 5 and the score of the wrist is weighted by a factor of 5. Both scoring methods are widely used and validated[19,21-24]. The average scores of the 2 observers were used in all analyses for each scoring method.

Expert panel. The international expert panel consisted of 3 practising rheumatologists who routinely assessed their patients' hand and foot films but were not trained to apply the 2 radiologic scoring methods (HH, HP, and PP). The panel experts were first asked whether they noticed any progression of joint damage caused by RA between the hand and/or foot films in 1 pair of films. If they noticed progression, they had to state whether they considered that difference in joint damage clinically relevant in 4 specific clinical settings (Table 1), all based on the situation of a 46-year-old female patient with RA, treated with methotrexate for 1 year. Clinically relevant progression was defined as progression that should be taken into consideration in the decision on continuation or change of the second-line therapy. The majority opinion of the panel was the criterion applied in all analyses.

Viewing sessions were standardized: the experts had to pause for 30 minutes after every 2 hours and a maximum of 4 hours viewing was permitted per day. The same lightbox and the same room had to be used throughout the viewing session. Use of magnifying hand lenses was not allowed. All radiographs were viewed twice by each panel expert, with an interval of at least 4 weeks, to estimate the intraobserver reliability of the panel. The films were viewed in the first viewing session in a different random order than in the second session. If not stated otherwise, the opinion of the first viewing session was used in all analyses.

Statistical analysis. Kappa coefficients and observed proportions of positive and negative agreement (P_{pos} and P_{neg}) were calculated to assess the intraobserver reliability of the expert panel[25-27]. The P_{pos} represents the probability that a patient regarded as a progressor the first viewing session will also be considered as a progressor the second session[28]. The P_{neg} represents the same for negative ratings, hence for films regarded as stable or improved. The accuracy of the scoring methods to discriminate between progression and stable damage was evaluated by ROC curve analyses. The ROC curve is a graph that plots the true positive rate (sensitivity) in function of the false positive rate (100 – specificity) at all possible decision thresholds. In other words, a ROC curve is the result of continuously varying the cutoff points (progression scores) for clinically important difference and calculating the accompanying sensitivity/specificity pairs. Figure 1 shows an example of a ROC curve. A scoring method that discriminates perfectly between progression and stability or healing of joint damage will have a curve that passes through the upper left corner, where the true-positive fraction is 1.0 (perfect sensitivity), and the false-positive fraction is zero (perfect specificity). The data points closest to the upper left corner of the ROC curves provide the threshold values that go together with the highest accuracies (minimal false negative and minimal false positive results) of the scoring methods — see asterisk Figure 1. These threshold values best represent the minimum values that are regarded as clinically relevant by the panel in the different clinical settings.

The SDD, as proposed by Bland and Altman[29,30] and applied to radiographs by Lassere[19], is based on the measurement error for scoring progression by 2 independent observers. This statistical method assesses

Table 1. Four clinical settings used in the radiographic assessment by the expert panel.

Clinical setting	
1	Disease duration 2 years, mild disease activity
2	Disease duration 2 years, high disease activity
3	Disease duration 8 years, mild disease activity
4	Disease duration 8 years, high disease activity

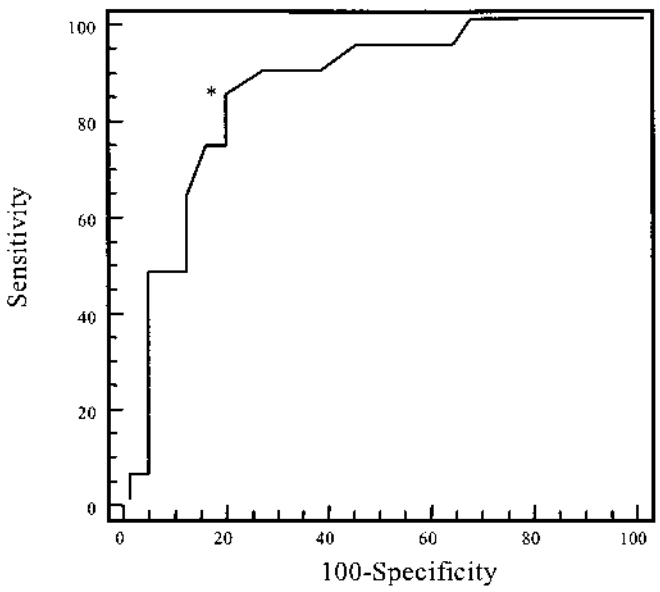

Figure 1. Example of a ROC curve (Sharp/van der Heijde in setting 1). *Cutoff point with the highest accuracy, [i.e., the highest sensitivity/specificity pair (in this example a cutoff point of 4.5 units with sensitivity 84 and specificity 82)].

whether an individual difference observed between 2 moments is a real change or that it is a change induced by the measurement error caused by the 2 observers. To calculate the SDD, the differences in the observers' scores of the same subjects are determined. The mean and standard deviation of these differences are then calculated. Ninety-five percent of the differences are expected to be less than 2 SD of the mean of the difference, the resulting limits are called the limits of agreement (mean $_{difference}$ ± 2SD $_{difference}$). Hence, a progression score that lies between the limits cannot be distinguished from interobserver measurement error. We calculated the sensitivities and specificities of the scoring methods if the SDD of the 2 scoring methods in this particular setting were used as threshold values. The magnitude of a SDD depends on the distribution of the radiographic progression in the population under study, the scoring method, the number of observers, and the choice whether to generalize the results to any other pair of observers. Both the SDD generalizable to any other pair of observers and those restricted to the current 2 observers were assessed. To calculate these "generalizable SDD" the limits of the 95% CI of the limits of agreement should be used (see Appendix for exact formulas).

To evaluate the interobserver reliabilities of the 2 scoring methods, 2 types of intraclass correlation coefficients (ICC) were calculated[31]. They were calculated using a 2 way analysis of variance (ANOVA) model with the observers defined as either fixed (type 3,1) or random (type 2,1).

Descriptive analyses, crosstabs, kappa statistics, and the 2 way ANOVA were performed by SPSS 9.0 for Windows. ROC curves were analyzed by MedCalc statistical software.

RESULTS

Distribution of radiographic scores. The Sharp/van der Heijde total scores had an average baseline score of 24.6 (standard deviation, SD, 16.5), with a median of 19.5 (interquartile range, IQR, 11.9–5.4). The Larsen/Scott total scores had an average baseline score of 14.5 (SD 10.4), with a median of 15.3 (IQR 5.8–19.8). The progression of the Sharp/van der Heijde total scores was on average 7.6 (SD 9.9), with a median of 4.0 (IQR 2.4–8.6). The corresponding figures for the Larsen/Scott total progression scores were 4.0 (SD 8.0), with a median of 0.8 (IQR 0.0–3.6).

Prevalence of progression of joint damage seen by the expert panel (Table 2). The expert panel reported progression of joint damage in 36 pairs of radiographs (78%). The decision whether this progression was clinically important depended on the setting. In setting 2 (early RA, high disease activity), 26 of these 36 film pairs (72%) were judged to show clinically relevant progression. In settings 4 (late RA, high disease activity) and 1 (early RA, mild disease activity), roughly half the films were judged clinically relevant, dropping to only a quarter in setting 3. The panel was very consistent in this order of clinical settings, deviating from it only in one of the 46 pairs of films. Hence, progression in a patient under methotrexate treatment with early RA and high disease activity was most frequently viewed as clinically relevant. The judgment of clinically important progression was very consistent, slightly higher in the second viewing session.

Reproducibility of the expert panel's opinion. The panel's consistency is also reflected by the kappas and observed proportions of agreement (see Table 3). The panel was more consistent in judging the relevance of progression in late RA ($\kappa_{s\,3} = 0.74$ and $\kappa_{s\,4} = 0.78$) than in early RA ($\kappa_{s\,1} = 0.52$ and $\kappa_{s\,2} = 0.42$), or when presented as the observed proportion of

Table 2. Prevalence of radiological progression of joint damage (difference) in the hands and feet by the opinion* of the expert panel.

	Opinion		Concordant Opinion†
	Viewing Session 1	Viewing Session 2	
	N (%)	N (%)	N (%)
Progression of joint damage	36 (78)	39 (85)	34 (74)
CID in setting 1 (early RA, mild disease activity)	19 (41)	21 (46)	14 (30)
CID in setting 2 (early RA, high disease activity)	26 (57)	27 (59)	20 (44)
CID in setting 3 (late RA, mild disease activity)	10 (22)	10 (22)	8 (17)
CID in setting 4 (late RA, high disease activity)	21 (46)	19 (41)	17 (37)

*The majority opinion of the panel was used for all analyses. †Concordant opinion: progression seen in the first and second viewing session. CID: clinically important difference between 2 pairs of hand and foot films.

Table 3. Intraobserver variability of the expert panel for clinically relevant progression of joint damage in the hands and feet.

	Kappa	P_{pos}	P_{neg}
CID in setting 1 (early RA, mild disease activity)	0.47	0..70	0.77
CID in setting 2 (early RA, high disease activity)	0.42	0.75	0.67
CID in setting 3 (late RA, mild disease activity)	0.74	0.80	0.94
CID in setting 4 (late RA, high disease activity)	0.74	0.85	0.88

CID: clinically important difference between 2 pairs of hand and foot films. P_{pos}: observed proportion of positive agreement, i.e., the presence of CID. P_{neg}: observer proportion of negative agreement, i.e., the absence of CID.

positive agreement, 0.80 and 0.88 in late RA versus 0.73 and 0.75 in early RA.

Interobserver reliability of the scoring methods. The interobserver reliabilities for the total scores of both methods were good. If we are only interested in the 2 current observers, the interobserver reliabilities were 0.91 and 0.85 for the Sharp/van der Heijde and Larsen/Scott methods, respectively. If the results are to be generalized to other observers, the ICC were 0.89 for the Sharp/van der Heijde and 0.82 for the Larsen/Scott method. Because one is usually more interested in progression scores, we also calculated the ICC for progression scores. The interobserver reliabilities for the Sharp/van der Heijde progression scores were both 0.94, with the observers defined as fixed and as random, respectively. The corresponding numbers for the Larsen/Scott progression scores were 0.88 and 0.85.

Receiver operating characteristics of the scoring methods. The threshold values corresponding with the highest accuracy of the scoring methods and their accompanying sensitivities, specificities, and positive likelihood ratios are listed in Table 4 for each setting separately. Using Sharp/van der Heijde progression scores of 4–6 (average 4.7) units as decision thresholds for clinically relevant progression in the different settings led to sensitivities ranging from 77% to 100% and specificities between 78% and 84%. The Larsen/Scott method was most accurate in detecting clinically relevant progression if threshold values of 1.0–4.5 (average 1.8) were used (sensitivities 73–84%, specificities 74–94%). The 4.7 Sharp/van der Heijde and the 1.8 Larsen/Scott units are both 1% of their maximum scores. Films that showed progression equal to or greater than the highest accuracy thresholds of the Sharp/van der Heijde method were 3.9 to 5.1 times as likely to be regarded as progressive than as stable by the panel. The corresponding figures for the Larsen/Scott method ranged from 3.4 to 14.4.

If decision thresholds are varied over the spectrum of possible scores, the sensitivities and specificities move in opposite directions. This can be seen in Tables 5 and 6, which sum the operating characteristics if the SDD of the scoring methods on this particular study were chosen as cutoff points. In Table 5 we used the SDD not generalizable to other observers as cutoff points. The SDD were 5.0 Sharp/van der Heijde score units and 5.8 Larsen/Scott score units. This resulted in higher specificities (Sharp/van der Heijde 72–85%; Larsen/Scott 94–96%) and the expected lower sensitivities (Sharp/van der Heijde 65–100%, Larsen/Scott 35–80%). The sensitivities were especially low for the Larsen/Scott method, but they were low to satisfactory for the Sharp/van der Heijde method. The accompanying ratios between the "true-progressive" and "false-progressive" fractions were between 3.6 and 4.8 for the Sharp/van der Heijde method. The positive likelihood ratios for the Larsen/Scott method ranged between 6.9 and 14.4. The SDD used in Table 6 are generalizable to any other couple observers and are consequently larger than the SDD used in Table 5. This resulted again in a decrease of sensitivities.

The highest accuracy thresholds for the total scores of the hands and feet separately and for the Sharp/van der Heijde erosion scores were also calculated. The Sharp/van der Heijde method was most accurate in detecting clinically relevant progression in the hands if threshold values of 3.0–4.5 units were used. The highest accuracy thresholds of the feet were 1.5 units for settings 1, 2, and 4 and 4.5 units

Table 4. Highest accuracy characteristics of the progression scores of the scoring methods.

	Sharp/van der Heijde				Larsen/Scott			
	Threshold Value	Sensitivity (95% CI)	Specificity (95% CI)	Likelihood Ratio†	Threshold Value	Sensitivity (95% CI)	Specificity (95% CI)	Likelihood Ratio†
CID in setting 1 (early RA, mild disease activity)	4.5	84 (60–96)	82 (62–94)	4.6	1.0	84 (60–96)	74 (54–89)	3.3
CID in setting 2 (early RA, high disease activity)	4.0	77 (56–91)	80 (56–94)	3.9	1.0	73 (52–88)	80 (56–94)	3.7
CID in setting 3 (late RA, mild disease activity)	6.0	100 (100)	78 (61–90)	4.5	4.5	80 (44–97)	94 (81–99)	14.4
CID in setting 4 (late RA, high disease activity)	4.5	81 (58–94)	84 (64–95)	5.1	1.0	81 (58–94)	76 (55–91)	3.4

CID: clinically important difference between 2 pairs of hand and foot films. †Positive likelihood ratio.

Table 5. Operating characteristics if the smallest detectable difference* (same 2 observers) is used as threshold value.

	Sharp/van der Heijde			Larsen/Scott		
	Sensitivity (95% CI)	Specificity (95% CI)	Likelihood Ratio‡	Sensitivity (95% CI)	Specificity (95% CI)	Likelihood Ratio‡
CID in setting 1 (early RA, mild disease activity)	79 (54–94)	82 (62–94)	4.3	47 (25–71)	96 (81–99)	12.8
CID in setting 2 (early RA, high disease activity)	65 (44–83)	85 (62–97)	4.4	35 (17–56)	95 (75–99)	6.9
CID in setting 3 (late RA, mild disease activity)	100 (100)	72 (55–86)	3.6	80 (44–97)	94 (81–99)	14.4
CID in setting 4 (late RA, high disease activity)	76 (53–92)	84 (64–95)	4.8	43 (21–66)	96 (80–99)	10.7

*The SSD if the mean scores of the same 2 (fixed) observers are used: 5 Sharp/van der Heijde units and 6 Larsen/Scott units in this study. CID: clinically important difference between 2 pairs of hand and foot films. ‡Positive likelihood ratio.

Table 6. Operating characteristics if the smallest detectable difference* (any 2 observers) is used as threshold value.

	Sharp/van der Heijde			Larsen/Scott		
	Sensitivity (95% CI)	Specificity (95% CI)	Likelihood Ratio‡	Sensitivity (95% CI)	Specificity (95% CI)	Likelihood Ratio‡
CID in setting 1* (early RA, mild disease activity)	58 (43–80)	89 (71–98)	5.2	42 (20–67)	96 (81–99)	11.4
CID in setting 2 (early RA, high disease activity)	50 (30–70)	95 (75–99)	10.0	31 (14–52)	95 (75–99)	6.2
CID in setting 3 (late RA, mild disease activity)	80 (44–97)	83 (67–94)	4.8	70 (35–93)	94 (81–99)	12.6
CID in setting 4 (late RA, high disease activity)	53 (34–78)	92 (74–99)	7.1	38 (18–62)	96 (80–99)	9.5

*The SSD if the mean scores of a random pair of observers are used: 7 Sharp/van der Heijde units and 8 Larsen/Scott units in this study. CID: clinically important difference between 2 pairs of hand and foot films. ‡Positive likelihood ratio.

for setting 3. The Larsen/Scott method was most accurate in detecting clinically relevant progression in the hands if threshold values of 2.5–3.0 units were used. The highest accuracy thresholds for the Larsen/Scott method for the feet were 1.0 for settings 2 and 4 and 3.5 units for settings 1 and 3. If erosion scores of the Sharp/van der Heijde method were used as tests for clinically relevant progression of radiological joint damage, the threshold values leading to highest accuracy of the test were 2.5–3.0 Sharp/van der Heijde units. In other words, they are about half of those of the total scores (erosions + joint space narrowing).

In addition, we assessed the highest accuracy thresholds for the 2 scoring methods, if the concordant opinion of the panel was used as gold standard. These results were very similar to those presented here for the first reading (data not shown).

Sensitivity of the expert panel. The sensitivity of the expert panel for progression of joint damage, defined as a mean progression score equal to or greater than 0.5, was 80% (Sharp/van der Heijde) and 92% (Larsen/Scott). In other words, the panel picked up a change in mean total score (hands + feet) of 0.5 or more in 83% of the cases when scored according to the Sharp/van der Heijde method. A change in mean Larsen/Scott score of 0.5 or more was picked up in 92% of the cases.

DISCUSSION

One of the problems of scoring radiographs is the difficulty in interpreting the progression observed. What is the meaning of an increase in radiological joint damage of 10 Sharp/van der Heijde or 5 Larsen/Scott units? A first attempt to define the 1 year MCID was made by Lassere, *et al*, who determined the SDD on statistical grounds[19]. Our study attempted to constitute the lower boundaries of the *clinically* important difference. The expert panel's opinion on clinical relevance of the progression observed served as an external criterion (gold standard) and made it possible to construct ROC curves. The cutoff points that led to the highest accuracy of the scoring methods gave us an indication of the MCID. In this study the minimum progression of radiological joint damage considered clinically relevant by the expert panel turned out to be equal to or smaller than the calculated interobserver measurement error of both scoring methods. A possible explanation is that the experts based their judgement on progression of joint damage also on other features of joint damage than erosion and joint space narrowing, which are the dominant features included in the scoring systems. Although differences smaller than the SDD are already considered clinically relevant, these cannot be picked up reliably by the scoring methods due to measurement error. The sensitivity of the SDD as a threshold to detect meaningful differences is therefore lower than the MCID derived from the expert panel, but its specificity is higher. This makes one wonder whether the 95% interval currently used to calculate the SDD is a correct choice. Perhaps 90% or even 80% intervals will be more appropriate when assessing the MCID by the SDD. This issue should be addressed in future discussions and research. For now, the percentage of progressors based on the (95%) SDD will be too insensitive to use as primary outcome measure and should be used as secondary outcome measure.

Although the judgment on clinically relevant change varied across the different clinical settings, the threshold scores leading to the highest accuracy of the scoring methods were very similar. Thus, the MCID does not seem to be influenced much by disease activity or disease duration.

MRI seems to be a very promising and powerful tool and will certainly play a major role in the valuation of joint damage in the future. Nevertheless, we have chosen to use an international expert panel as external criterion and not MRI because the process of validating the scoring methods for MRI is yet in its early stage. Furthermore, the clinical relevance of damage seen on MRI is still totally unclear.

The choice of the panel is of course a subjective one. We tried to represent the average expert by choosing an international panel and presenting the opinion of the majority. On first sight, the intraobserver agreement of the panel may seem somewhat disappointing. This may reflect uncertainty in the clinical decision making process. However, the consistency of the panel is satisfactory if we look not only at the kappas but also at the observed agreements and prevalence of the concordant opinions compared to the prevalence on the separate viewing sessions.

A definition of the MCID in radiological progression of joint damage based on the judgement of a clinical expert panel is the product of its time and shall depend on the efficacy of the drug at that particular time. With the recent development of more powerful DMARDs and the expectations of promising future drugs, the MCID for radiological progression of joint damage is expected to become smaller. Nevertheless, the MCID of radiological progression already turned out to be smaller than the measurement error in this study. If in the future the fortunate circumstance will be achieved that joint damage will heal and won't progress if patients are treated with the appropriate drug, the SDD will then be usable to determine the cut off level for the minimum detectable change in healing.

The finding that the highest accuracy thresholds for the Sharp/van der Heijde erosion scores were about half of the total scores suggests that the expert panel's opinion was influenced both by the degree of erosive damage and by joint space narrowing. This confirms the opinion that scoring joint space narrowing provides additional independent information.

The relation between the SDD of the Sharp/van der Heijde and the SDD of the Larsen/Scott is remarkably different in comparison with the relation between the SDD found in the work by Lassere, *et al*. They reported SDD (same 2 observers) of 11 Sharp/van der Heijde units and 8 Larsen/Scott units, whereas our SDD were 5 and 6, respectively. The magnitude of the SDD depends on several aspects of the study design, such as the distribution of radiographic progression in the population under study, the scoring method, and the number of observers, as previously noted. The differences in distribution of progression between our study and Lassere, *et al* very likely affect the relation between the SDD of the 2 scoring methods. The smaller SDD of the Sharp/van der Heijde can be explained by the fact that another pair of observers scored by the Sharp/van der Heijde method. This pair showed higher interobserver reliability than the previous pair.

The MCID and the SDD are individual response measures like the ACR20 and the EULAR, the 2 response criteria most widely used. These indices were developed rather differently, but in the development processes of both criteria the opinions of experts were used[11,12]. An expert's opinion approach like ours is one step in the process of defining the appropriate MCID and, comparable to the development of the disease activity response criteria, future steps in defining the MCID of radiological progression (or even healing) should be data driven.

In conclusion, this preliminary report suggests that the smallest detectable difference is a conservative estimate of the minimal clinically important difference in radiographs as judged by an expert panel. A subsequent report will describe the data based on a panel with 5 experts and will evaluate the influence of compounding and size of the panel. Also, the influence of distribution of radiographic progression will be explored further. Finally, further research in other and larger data sets is needed to confirm these findings.

REFERENCES

1. Boers M, Tugwell P, Felson DT, et al. World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol 1994;21 Suppl 41:86-9.
2. 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.
3. Larsen A, Dale K, Eek M. Radiographic evaluation of rheumatoid arthritis and related conditions by standard reference films. Acta Radiol 1977;18:481-91.
4. Sharp JT, Young DY, Bluhm GB, et al. How many joints in the hands and wrists should be included in a score of radiologic abnormalities used to assess rheumatoid arthritis? Arthritis Rheum 1985;28:1326-35.
5. van der Heijde DM, van Riel PL, Nuver Zwart IH, Gribnau FW, van de Putte LB. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. Lancet 1989;1:1036-8
6. Larsen A. How to apply Larsen score in evaluating radiographs of rheumatoid arthritis in longterm studies. J Rheumatol 1995;22:1974-5.
7. Rau R, Herborn G. A modified version of Larsen's scoring method to assess radiologic changes in rheumatoid arthritis. J Rheumatol 1995;22:1976-82.
8. Scott DL, Houssien DA, Laasonen L. Proposed modification to Larsen's scoring methods for hand and wrist radiographs. Br J Rheumatol 1995;34:56.
9. Edmonds J, Saudan A, Lassere M, Scott D. Introduction to reading radiographs by the Scott modification of the Larsen method. J Rheumatol 1999;26:740-2.
10. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 1999;26:743-5.

11. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727-35.
12. van Gestel AM, Prevoo ML, van 't Hof MA, van Rijswijk MH, van de Putte LB, van Riel PL. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. Arthritis Rheum 1996;39:34-40.
13. Strand V, Cohen S, Schiff M, et al. Treatment of active rheumatoid arthritis with leflunomide compared with placebo and methotrexate. Leflunomide Rheumatoid Arthritis Investigators Group. Arch Intern Med 1999;159:2542-50.
14. Bankhurst AD. Etanercept and methotrexate combination therapy. Clin Exp Rheumatol 1999;17:S69-72.
15. Taylor WJ, Rajapakse CN, Harris KA, Harrison AA, Corkill MM. Inpatient treatment of rheumatoid arthritis with synacthen depot: a double blind placebo controlled trial with 6 month followup. J Rheumatol 1999;26:2544-50.
16. Anderson JJ, Wells G, Verhoeven AC, Felson DT. Factors predicting response to treatment in rheumatoid arthritis: the importance of disease duration. Arthritis Rheum 2000;43:22-9.
17. Rojkovich B, Hodinka L, Balint G, et al. Cyclosporin and sulfasalazine combination in the treatment of early rheumatoid arthritis. Scand J Rheumatol 1999;28:216-21.
18. Felson DT, LaValley MP, Baldassare AR, et al. The prosorba column for treatment of refractory rheumatoid arthritis: a randomized, double-blind, sham-controlled trial. Arthritis Rheum 1999;42:2153-9.
19. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731-9.
20. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561-77.
21. 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 1999;38:1213-20.
22. van der Heijde D, Dankert T, Nieman F, Rau R, Boers M. Reliability and sensitivity to change of a simplification of the Sharp/van der Heijde radiological assessment in rheumatoid arthritis. Rheumatology 1999;38:941-7.
23. Kuper IH, van Leeuwen MA, van Riel PL, et al. Influence of a ceiling effect on the assessment of radiographic progression in rheumatoid arthritis during the first 6 years of disease. J Rheumatol 1999;26:268-76.
24. Molenaar ET, Edmonds J, Boers M, van der Heijde DM, Lassere M. A practical exercise in reading RA radiographs by the Larsen and Sharp methods. J Rheumatol 1999;26:746-8.
25. Cohen J. A coefficient agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
26. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991.
27. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551-8.
28. Samsa GP. Sampling distributions of p(pos) and p(neg). J Clin Epidemiol 1996;49:917-9.
29. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
30. Bland JM, Altman DG. Measurement error. BMJ 1996;313:744.
31. Shrout P, Fleis J. Intraclass correlations: use in assessing rater reliability. Psychol Bull 1979;86:420-8.

Appendix

Formulas used to calculate the SDDs:

1. SDD restricted to the current observers:

$$t_{0.05,n-1(df)} * \frac{SD_{difference}}{\sqrt{r}}$$

2. SDD generalizable to any other reader (pair): – [lower limit] + upper limit of the 95% CI of the limits of agreement divided by 2

$$([(mean_{diff} - t_{0.05,n-1(df)} * \frac{SD_{diff}}{\sqrt{r}}) - t_{0.05,n-1(df)} * \sqrt{\frac{(3 * (\frac{SD_{diff}}{\sqrt{r}}))^2}{n}}] +$$

$$((mean_{diff} + t_{0.05,n-1(df)} * \frac{SD_{diff}}{\sqrt{r}}) + t_{0.05,n-1(df)} * \sqrt{\frac{(3 * (\frac{SD_{diff}}{\sqrt{r}}))^2}{n}}) / 2$$

Where r is the number of observers if the mean progression score of the readers is used and n is the sample size