Progression of Rheumatoid Arthritis on Plain Radiographs Judged Differently by Expert Radiologists and Rheumatologists

KARIN BRUYNESTEYN, SJEF van der LINDEN, ROBERT LANDEWÉ, FEIKJE GUBLER, RENÉ WEIJERS, and DÉSIRÉE van der HEIJDE

ABSTRACT. ***Objective.*** In a former study a panel of rheumatologists was used to assess which progression in radiological joint damage due to rheumatoid arthritis (RA) on hand and foot radiographs taken at one-year intervals was considered the minimally clinically important difference (MCID). We compare the judgments of the panel of rheumatologists with the judgments of 2 musculoskeletal radiologists.

Methods. Two experienced musculoskeletal radiologists evaluated independently the same hand and foot radiographs as assessed by the panel of rheumatologists. Progression was defined as important if the radiologist would state it as substantial progression in their report. Two readers, different from the radiologists and rheumatologists, independently obtained the Sharp/van der Heijde scores. Receiver operating characteristic curve analyses were performed to quantify the minimally important progression defined by the radiologists expressed in Sharp/van der Heijde change-scores. The change-score with the highest accuracy represented the minimally important progression and was compared with the MCID defined by the panel of rheumatologists for 4 different settings (early versus advanced RA and mild versus high disease activity).

Results. The minimally important progression defined by the radiologists was estimated at 6.5 Sharp/van der Heijde units. This was larger than the MCID defined by the panel of rheumatologists in 3 of the 4 clinical settings (3.0–4.5 units) and similar to the setting "advanced RA, mild disease activity." The panel of rheumatologists was inclined to change therapy in cases not reported as substantially progressive by the radiologists. The Sharp/van der Heijde progression scores of the radiographs on which the radiologists and rheumatologists disagreed related better with the rheumatologists' opinions.

Conclusion. Changes that were not regarded as substantial by the radiologists were judged clinically important by the rheumatologists in 3 of the 4 clinical settings. Thus, the radiologists appeared to be reserved in judging changes as important. (J Rheumatol 2004;31:1088–94)

Key Indexing Terms:
PANEL RADIOLOGISTS RHEUMATOLOGISTS
PLAIN RADIOGRAPHS MINIMAL CLINICALLY IMPORTANT DIFFERENCE

When analyzing clinical trials, the number of patients actually responding to the drug under investigation can provide important information, which adds to the information obtained from traditional statistical methods based on mean or median group changes. To assess whether a patient is a responder, a cutoff value needs to be chosen. In the ideal situation, the minimal clinically important difference (MCID) for the outcome measure in question is known, so that the outcome measure can be dichotomized. Several methods have been used to quantify which difference or change within an individual patient is clinically important. In a previous study[1], we used the opinion of a panel of rheumatologists to assess the MCID for rheumatoid arthritis (RA) related radiological joint damage. Clinically relevant progression was defined in this study as the amount of progression of joint damage that would make the rheumatologists change the second-line therapy prescribed. Because it was assumed that factors like disease duration and disease activity would influence this decision, the MCID was assessed for 4 different hypothetical settings: early RA with high disease activity, early RA with mild disease activity, advanced RA with high disease activity, and advanced RA with mild disease activity. It was shown that the rheumatologists were more inclined to change therapy in

From the Department of Internal Medicine, Division of Rheumatology and the Department of Radiology, University Hospital Maastricht, Maastricht; the Department of Radiology, Spaarne Hospital, Heemstede, The Netherlands; and Limburg University Center, Diepenbeek, Belgium.

Supported by the Dutch Arthritis Association.

K. Bruynesteyn, MD; S. van der Linden, MD, PhD, Professor of Rheumatology; R.B.M. Landewé, MD, PhD, Rheumatologist, Department of Internal Medicine, University Hospital Maastricht; F.M. Gubler, MD, PhD, Radiologist, Department of Radiology, Spaarne Hospital; R. Weijers, MD, Radiologist, Department of Radiology, University Hospital Maastricht; D.M.F.M. van der Heijde, MD, PhD, Professor of Rheumatology, Department of Internal Medicine, University Hospital Maastricht, Limburg University Center.

Address reprint requests to Dr. D. van der Heijde, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands.
E-mail: dhe@sint.azm.nl

Submitted February 14, 2003; revision accepted December 3, 2003.

patients with early disease and high disease activity than in patients with advanced RA and mild disease activity, with the other 2 settings fitting in between.

Several studies have shown that expert panels can be useful methods to estimate health changes in patients[2-5]. One part of the evaluation of the usefulness of the panels is to assess the consistency of the panels' judgments. Another aspect that should be investigated is the validity of the panel. Without the availability of a "gold standard," comparing the results with other outcome measures that assess adjacent health attributes can assess the so-called criterion or concurrent validity of a method. For panel judgments, an alternative for this is using the judgments of a panel composed of panelists of another discipline or disciplines. For radiological joint damage, a valid choice seems to be musculoskeletal radiologists. In daily practice, some rheumatologists rely on the judgment of a musculoskeletal radiologist. Although radiologists do not make clinical judgments, they do interpret the changes observed in their reports.

We compared the results of the panel of rheumatologists used in the previous study with the judgments of a couple of radiologists. To determine how changes in radiological joint damage assessed as important by radiologists relate to the formerly defined MCID, we also estimated a minimally important difference defined by the radiologists and compared it with the 4 previously assessed MCID defined by the panel of rheumatologists. Finally, this study also investigated the influence of clinical information on the decision of the panel of rheumatologists whether observed progression was considered clinically relevant or not.

MATERIALS AND METHODS

Summary of methods of the previous study[6]. In this "MCID study" the expert panel consisted of 5 experienced rheumatologists of several nations. They independently evaluated 46 pairs of hand and foot radiographs, taken at one-year intervals. They were first asked whether they noted any progression of joint damage due to RA, and if they noted progression, they had to state whether they considered that difference clinically relevant in 4 hypothetical clinical settings: (1) advanced RA and mild disease activity, (2) early RA and mild disease activity, (3) advanced RA and high disease activity, and (4) early RA and high disease activity. Clinically relevant progression was defined as progression of joint damage that would make the rheumatologist change the second-line therapy (methotrexate) that has been started one year before. Radiographs were presented to the rheumatologist in chronological order. The majority opinion of the panel (score of 3, 4, or 5 out of 5) was the criterion applied in all analyses. The radiographs used in this study had been selected for high and low baseline joint damage and for high and low progression of joint damage. Radiographs were selected by an independent rheumatologist involved in the organization of the COBRA trial[6a]. A selection was made to represent a wide spectrum of baseline joint damage and progression. All 46 patients fulfilled the 1987 American College of Rheumatology classification criteria for RA. The interobserver reliability of the panel was assessed in the former study by average-measure intraclass correlation coefficients (ICC) and ranged between 0.60 (setting 4) and 0.74 (setting 3).

Methods in the present study. Two experienced musculoskeletal radiologists (FG and RW) evaluated the same radiographs as assessed by the rheumatologists. The radiologists were consultants with 5–10 years of experience in the field of musculoskeletal radiology. They judged the radiographs independently of each other. The radiologists were first asked whether they noted any progression of joint damage due to RA, and if they noted progression, they further had to state — during the same viewing session — whether they classified it as important progression, i.e., that they would record it as substantial progression in their report. As in clinical practice, the radiologists knew the chronological order of the hand and foot radiographs. No clinical information on disease duration and disease activity was given. The unanimous judgments of the 2 radiologists on the existence of (substantial) progression were used for the primary analyses. If one or both radiologists did not note (substantial) progression of the joint damage, the radiographs were defined as non- (substantial) progressive. To estimate intraobserver reliability, each radiologist viewed all radiographs twice, with an interval of at least 4 weeks. Sensitivity analyses were performed with the judgments of the radiologists defined as (substantial) progressive if one or both radiologists noted (substantial) progression.

To be able to quantify the radiological progression (see also *statistical analyses*, below), the radiographs were also scored according to the Sharp/van der Heijde method[6] independently by 2 experienced readers, other than the radiologists or rheumatologists[1]. Both readers were researchers trained to score according to the Sharp/van der Heijde method by Dr. van der Heijde and have experience scoring radiographs in several trials. Radiographs were scored in chronological order and patient identity was blinded. The Sharp/van der Heijde method assesses erosions and joint space narrowing separately and has a range from 0 to 448. Thirty-two joints in the hands and 12 in the feet are scored for erosions, with a maximum score of 5 per joint in the hands and 10 per joint in the feet. Joint space narrowing is graded from 0 to 4 in 30 joints in the hands and in 12 joints in the feet. The principal score used in the analyses is the total score, which is the sum of the erosion score and the joint space narrowing score. Mean scores of the readers were used for the analyses.

Statistical analyses. The judgment on the presence of progression by the radiologists was compared with the judgment on progression by the rheumatologists with a 2-by-2 table; 2-by-2 tables were also made to describe the differences between the opinions on the importance of the progression by the radiologists with the opinion on the clinical importance of the progression by the rheumatologists. To quantify the change in damage of the radiograph sets on which the rheumatologists and radiologists agreed and disagreed, the medians and interquartile ranges (IQR) of the Sharp/van der Heijde change-scores were calculated for each cell of the 2-by-2 tables.

Receiver operating characteristic (ROC) curve analyses were performed to quantify the minimally important progression defined by the radiologists as Sharp/van der Heijde change-scores. The accuracy to discriminate between important progression and no (important) progression was assessed for every possible cutoff level of radiological joint damage expressed in Sharp/van der Heijde units. The ROC curve thus plotted the true positive rate (sensitivity) in function of the false positive rate (100 – specificity) at all possible cutoff levels of radiological joint damage. The change-score with the highest accuracy for detecting important progression as defined by the radiologists represented the minimally important progression. Note that the change-score that discriminated best between important progression and no progression represented the minimally important difference and not the lowest progression score judged as important by the radiologists. This is because the latter would be 100% sensitive, but not specific at all. Note further that this minimally important progression defined represents the minimally important progression of joint damage for an individual patient and not for a group of patients. The ROC analyses were performed using MedCalc statistical software. We compared the minimally important difference defined by the radiologists with the 4 MCID defined by the panel of rheumatologists.

Single-measure and average-measure random effects intraclass correlation coefficients (ICC) with 95% confidence intervals were calculated with SPSS 10.0 for Windows to evaluate the intraobserver and interobserver reliability of the radiologists. The interobserver reliability between judg-

ments of the panel of rheumatologists and those of both radiologists was also assessed by ICC. For dichotomous outcome, single-measures ICC are equal to kappa statistics. Average-measure ICC have the advantage that they can take into account whether the judgments of more than one observer or more than one viewing session are used in the analysis. Under the assumption that other radiologists would have had similar experience and training one can additionally simulate what the interobserver reliability (and thus generalizability) would have been if we used more than 2 radiologists (by dividing the variance components of the factor "radiologists" and its interactions by the simulated number of expert members). We simulated an expert panel that used the (majority) opinion of one, 3, and 5 radiologists.

RESULTS

The single-measure intraobserver ICC of the judgments of both radiologists were 0.82 (95% CI 0.70–0.90) for the detection of progression of joint damage and 0.87 (0.76–0.92) for the judgment of substantial progression. The average-measure interobserver ICC between the radiologists were 0.70 (0.42–0.84) for the detection of joint damage and 0.82 (0.67–0.90) for the judgment of substantial progression. These average-measures interobserver ICC based on random effect variance components estimate the generalizability of the results to other pairs of radiologists with similar experience and training. Using only one radiologist instead of 2 would have resulted in much lower interobserver ICC: simulations showed ICC of 0.54 for the detection of joint damage and 0.69 for the judgment on substantial progression. The simulated ICC for the panels of 3 or 5 radiologists were 0.78 and 0.85, respectively, for the detection of joint damage and 0.87 and 0.92 for the judgment of substantial progression.

The radiologists labeled 20 of the 46 (43%) sets as progressive. On 16 of these 20, substantial progression was seen, which is on 35% of all film pairs (16/46). Table 1 shows a 2-way table in which the judgment on the presence of progression by the radiologists is compared with the judgment by the rheumatologists. All 20 sets judged as progressive by the radiologists were also labeled progressive by the rheumatologists. Seventeen sets judged as nonprogressive by the radiologists were judged as progressive by the rheumatologists, which is 65% (17/26) of all sets judged as nonprogressive by the radiologists. To estimate the amount of change in damage of the 17 sets on which the rheumatologists and radiologists disagreed, the median Sharp/van der Heijde change-scores were determined (Table 2). A median change-score of 3.5 units was found for the 17 sets judged as nonprogressive by the radiologists but progressive by the rheumatologists. A median change-score of 2.0 units was found for the radiograph pairs judged nonprogressive by both the radiologists and rheumatologists.

Table 3 shows 2-way tables comparing the opinion of the radiologists on the importance of the progression seen with the opinion of the panel of rheumatologists for the 4 clinical settings. When considering a patient with advanced RA with mild disease activity (Table 3A), the panel wanted to change the treatment strategy in only 9 of the 16 cases (56%) labeled as importantly progressive by the radiologists. The panel wanted to change treatment strategy in one of the 30 patients labeled as having no important progression by the radiologists. In the other 3 settings, the rheumatologists were inclined to change treatment in many patients that were not classified as important progressive by the radiologists. In patients with early RA with high disease activity (Table 3D), the rheumatologists judged the amount of joint damage such that they even wanted to change therapy in 14 of the 30 (47%) cases defined as not importantly progressive by the radiologist.

To quantify the agreement in judgments between the panel of rheumatologists and both radiologists, interobserver ICC were assessed. The single-measure interobserver ICC between the panel of rheumatologists and both radiologists was 0.42 (95% CI 0.15–0.63) for the detection of progression and ranged between 0.53 (0.19–0.71) and 0.64 (0.43–0.78) when comparing substantial progression defined by the radiologists with the clinically relevant progression defined by the rheumatologists.

To quantify the change in damage of the film pairs on which the rheumatologists and radiologists agreed or disagreed, the median Sharp/van der Heijde change-scores were also determined for each cell of Table 3, as shown in Table 4. Table 4 shows that the median change-score ranged between 3.3 and 8.5 for the film sets labeled non-importantly progressive by the radiologist and as clinically importantly progressive by the rheumatologists. The median change-scores of the radiograph set that made the rheumatologists change therapy, but which were not stated as

Table 1. Judgment of the radiologists on the presence of progression of joint damage compared with the judgment of the rheumatologists.

	Radiologists: Progression		
	Yes	No	Total
Panel of rheumatologists: progression			
Yes	20	17	37
No	0	9	9
Total	20	26	46

Table 2. Median (interquartile range) Sharp/van der Heijde change-scores of the radiograph sets of each cell in Table 1.

	Radiologists: Progression		
	Yes	No	Total
Panel of rheumatologists: progression			
Yes	7.5 (4.1–16.3)	3.5 (2.3–6.0)	5.5 (3.0–13.0)
No	—	2.0 (0.5–2.5)	2.0 (0.5–2.5)
Total	7.5 (4.1–16.3)	2.5 (1.4–4.4)	4.0 (2.4–8.6)

Table 3. Judgment of the radiologists on the importance of the progression* compared with the judgement of the rheumatologists on the clinical importance† in 4 clinical settings.

Panel of Rheumatologists Setting	Radiologists: Sets with Important Progression Yes	 No	 Total
A: Patients with advanced RA with mild disease activity			
Sets with clinical important progression			
Yes	9	1	10
No	7	29	36
Total	16	30	46
B: Patients with early RA with mild disease activity			
Sets with clinical important progression			
Yes	14	7	21
No	2	23	25
Total	16	30	46
C: Patients with advanced RA with high disease activity			
Sets with clinical important progression			
Yes	15	8	23
No	1	22	23
Total	16	30	46
D: Patients with early RA with high disease activity			
Sets with clinical important progression			
Yes	16	14	30
No	0	16	16
Total	16	30	46

* Important progression: amount of progression of joint damage noted as substantial in the radiologists' report.
† Clinically important progression: amount of progression of joint damage that would make the rheumatologists change the second-line therapy prescribed.

Table 4. Median (interquartile range) Sharp/van der Heijde change-scores of the radiograph sets of each cell in Table 3, A to D.

Panel of Rheumatologists Setting	Median (IQR) Sharp/van der Heijde Change-Scores Radiologists: Sets with Important Progression* Yes	 No	 Total
A: Patients with advanced RA with mild disease activity			
Sets with clinical important progression†			
Yes	14 (9.8–23.5)	8.5 (8.5)††	14 (8.3–21.8)
No	6.0 (4.0–17.0)	2.5 (1.3–4.0)	3.0 (1.6–5.4)
Total	13 (6.3–19.3)	2.5 (1.4–4.3)	4.0 (2.4–8.6)
B: Patients with early RA with mild disease activity			
Sets with clinical important progression			
Yes	14 (7.4–21.8)	5.0 (3.5–8.5)	8.5 (5.5–16.3)
No	3.5 (3.0–4.0)††	2.5 (1.0–3.5)	2.5 (1.0–3.5)
Total	13 (6.3–19.3)	2.5 (1.4–4.3)	4.0 (2.4–8.6)
C: Patients with advanced RA with high disease activity			
Sets with clinical important progression			
Yes	14 (7.0–20.0)	4.5 (3.1–7.9)	7.5 (4.5–14.5)
No	4.0 (4.0)††	2.5 (0.9–3.5)	2.5 (1.0–3.5)
Total	13 (6.3–19.3)	2.5 (1.4–4.3)	4.0 (2.4–8.6)
D: Patients with early RA with high disease activity			
Sets with clinical important progression			
Yes	13 (6.3–19.3)	3.3 (1.9–6.6)	6.5 (3.1–14.1)
No	—	2.5 (0.6–3.5)	2.5 (0.6–3.5)
Total	13 (6.3–19.3)	2.5 (1.4–4.3)	4.0 (2.4–8.6)

* Important progression: amount of progression of joint damage stated as substantial in the radiologists' report.
† Clinically important progression: amount of progression of joint damage that would make the rheumatologists change the second-line therapy prescribed. †† Median (IQR) based on one or 2 radiograph sets.

importantly progressive by the radiologists, were higher than the median change-scores of the radiograph sets judged as important by the radiologists, but did not result in change of treatment by the rheumatologists (difference of 0.5 to 3.3 in medians for the different settings). The median change-score for the radiograph sets judged as non-importantly progressive by both the radiologists and rheumatologists was 2.5 in all settings.

The ROC analyses showed that a cutoff level of 6.5 Sharp/van der Heijde units discriminated best between important progression and no important progression as assessed by the radiologists. Thus, the minimal individual change in radiological joint damage deemed important by the radiologists was estimated at 6.5 Sharp/van der Heijde units. In the previous study the following MCID had been found: 3.0 units for early RA patients with high disease activity, 4.5 units for early RA patients with mild disease activity and advanced RA patients with high disease activity, and 6.5 units for advanced RA patients with mild disease activity. The minimally important progression defined by the radiologists was larger than the MCID defined by the panel of rheumatologists in 3 of the 4 clinical settings and similar to the 4th setting, "advanced RA with mild disease activity".

Sensitivity analyses. Sensitivity analyses were performed with (substantial) progression defined as positive if one or both radiologists noted (substantial) progression. As expected, more films were now judged as (substantial) progressive: in 67% (31/46) one or both radiologists noted progression, and 50% (23/46) were labeled as substantial progressive. However, the majority of the sets judged nonprogressive by the radiologist were still judged as progressive by the panel of rheumatologists (53%, 8/15).

Regarding the importance of the progression noted, the sensitivity analyses showed similar results when comparing the judgments of the radiologists with the judgments of the panel for the extreme settings (advanced RA with mild disease activity, early RA with high disease activity). In a considerable number (13/23, 56%) of the sets labeled "substantial" progressive by the radiologists the rheumatologists were not inclined to change treatment if the patients had advanced RA and mild disease activity. But in patients with early RA with high disease activity, the rheumatologists were again inclined to change treatment in a substantial number of patients that were not classified as important progressive by the radiologists (8/23, 35%), despite the higher percentage of cases labeled as importantly progressive by the radiologists.

Comparing the radiologists' judgments with the judgments of the panel for the other 2 settings (early RA patients with mild disease activity or advanced RA patients with high disease activity), the number of cases that were not regarded as "substantial" progressive by the radiologists but in which the rheumatologists wanted a treatment change decreased in comparison with the primary analysis to 4 sets in both settings. However, not all extra sets judged as important progressive by the radiologists were also judged as progressive by the rheumatologist. Further, the median change-scores of the sets in which the rheumatologists did not want to change treatment but which were judged as substantial progressive by the radiologists were lower than those of the sets in which the rheumatologists wanted to change therapy but were not judged as important progressive by the radiologists (3.0 and 2.5 vs 5.0 Sharp/van der Heijde units in both settings, respectively). The minimal individual change in radiological joint damage deemed important by the radiologists became 4.5 Sharp/van der Heijde units in the sensitivity analysis, which equals the MCID defined by the panel of rheumatologists for the intermediate settings, but is larger than the MCID for patients with early RA with high disease activity and smaller than the MCID for patients with advanced RA with mild disease activity.

DISCUSSION

That the panel of rheumatologists was inclined to change therapy in cases not reported as substantially progressive by the radiologists raises the question whether the rheumatologists based their decision to change therapy on the clinical information rather than on the extent of the radiological change. Clinical information was given to the panel of rheumatologists to evaluate its influence on the MCID for RA-related radiological joint damage. Implicitly, we thus evaluated the influence of clinical information on the therapy strategies based on radiological joint damage. Apart from influencing therapy strategies, clinical information may also influence actual recognition of the features on the radiographs. Previous studies investigating the influence of clinical information on the interpretation of roentgenographic examinations have shown mixed results: an increase in the true-positive rate[7-9], an increase in false positives[10,11], or just no effect. In our study, we first asked the panelists whether they observed progression of joint damage or not. If they noted progression, they were asked in the same session to judge whether they considered that level of progression clinically relevant for 4 clinical settings. Thus clinical information was only given after they viewed the radiographs. This makes variability in accuracy to observe progression due to clinical information most unlikely. Moreover, the rheumatologists were not only inclined to change therapy in cases not reported as substantially progressive by the radiologists, they also judged more radiographs as progressive than did radiologists. These radiographs had a higher median Sharp/van der Heijde change-score than the radiographs judged as nonprogressive by both the radiologists and the rheumatologists. Apparently the radiologists in this study were more reserved in labeling radiograph sets as progressive than were the rheumatologists. This was also

reflected in the fact that in the primary analyses the minimal difference on hand and foot radiographs taken with one-year intervals that was judged as substantial progression of joint damage by the radiologists appeared to be larger than the MCID for 3 of the 4 settings defined by the panel of rheumatologists, and similar to the setting of "advanced RA, mild disease activity."

A better explanation for the higher percentage of cases defined as clinically important progression in the settings with high disease activity or recently diagnosed RA may be the following. In patients with advanced RA with mild disease activity the rheumatologists did not consider minor changes as clinically important, but they did consider them clinically important in patients with early RA or patients with high disease activity. These "minor changes" were not considered substantial by the radiologists. Another explanation may be that in patients with early RA or patients with high disease activity the rheumatologists were inclined to change therapy in case of "ambiguous changes" instead of "minor changes." So the clinical information might have introduced bias, namely so-called expectation bias[12]: they expect progression of joint damage in patients with active disease and therefore are inclined to judge "ambiguous" changes as "minor, but clinically important" changes if the patients have high disease activity. Evaluating the Sharp/van der Heijde progression scores of the pairs on which the radiologists and rheumatologists disagreed gave more insight in this matter. After all, the Sharp/van der Heijde readers were blinded for patient's disease activity and disease duration. These analyses showed that change-scores of the sets judged as "positive" by the panel of rheumatologists and "negative" by the radiologists were higher than in the sets judged "positive" by the radiologists and "negative" by the rheumatologists. Further, the median change-scores of the radiograph sets judged "positive" by the rheumatologists and "negative" by the radiologists were for all 4 cases higher than the sets judged as "negative" by both radiologists and rheumatologists. The judgments by the rheumatologists were thus supported by the independently obtained Sharp/van der Heijde scores, which have been documented to be related to outcome like physical functioning[13].

In the sensitivity analyses the percentage of sets judged as (substantial) progressive was logically higher. Comparison of the radiologists' judgment with the rheumatologists' judgments, however, showed similar results for the extreme settings as compared to the primary analysis. For the intermediate settings, the percentage of patients in which the rheumatologists were inclined to change treatment but were not labeled "substantial progressive" by the radiologists decreased compared to the primary analyses. However, for the intermediate settings as well, the Sharp/van der Heijde progression scores continued to relate better to the rheumatologists' opinion than to the radiologists'.

To ensure the reliability and generalizability of an outcome measurement it is customary to use standardized scoring methods and well trained experts. When determining (clinically) important differences, however, standardization is of course not possible, and the "training" occurred in the form of years of medical education and daily practice. In our previous study it was therefore decided to use a panel of 5 rheumatologists instead of just 2 observers, as is common in the field of scoring radiological joint damage due to RA. In this study, however, we only used 2 radiologists, because the opinion about important change was expected to differ less between radiologists than the opinions on changing a therapy strategy due to radiological joint damage progression by rheumatologists. In addition, we anticipated that the high number of hand and foot radiographs seen daily by musculoskeletal radiologists would ensure consistency of their opinion. The intraobserver and interobserver ICC of the radiologists were indeed moderate to good, and were higher than those of the panel of rheumatologists. However, because we realize that the number of radiologists is a limitation of this study we also simulated how the interobserver reliability would have been with one, 3, and 5 radiologists, under the assumption that these radiologists would have had comparable experience and training. These simulations showed that the greatest gain in generalizability was found by increasing the number of radiologists from one to 2 (0.54 to 0.70 and 0.69 to 0.81 for scoring progression and substantial progression, respectively). By adding more radiologists the generalizability would have increased further, to 0.85 and 0.92, when using the majority opinion of 5 radiologists, in comparison with 0.70 and 0.82 obtained in our study, based on 2 radiologists.

When constructed properly, panels can give reliable estimates of health outcomes. Because nonstandardized judgments of experts tend to vary widely, panels by definition contain more than one expert. For panels, formal consensus methods are often used to assess the health outcome in question[14]. These consensus methods derive quantitative estimates through qualitative approaches. Part of the approach is to give the panelist feedback on the decisions made by other panelists. In the case of radiological joint damage due to RA in the hands and feet, this is difficult to do without specifying the joints judged by the other panelists. Such a specification resembles official scoring of radiographs instead of nonstandardized judgments. The structure of an informal consensus meeting was also not thought to be appropriate because of the risk of being dominated by the more powerful member(s). Therefore, the radiographs were judged independently, and in the analyses the majority opinions of both panels were used. That the judgments of single panelists tend to vary and consequently consensus or majority opinion methods are used to express the health outcome in question, means that the results of such panels do not lend themselves to inferences for clinical practice[5,14,15].

Both the expert panel of rheumatologists and the radiologists were consistent over time. However, to determine whether the judgment of a panel of rheumatologists, besides producing a consistent outcome, is also a valid method of assessing the MCID, the panel results should ideally be compared with a gold standard. As such a standard is not available for radiological joint damage due to RA, an alternative method should be sought. We have illustrated the comparison with the judgments of a panel composed of panelists of another discipline, namely radiologists. The radiologists could not be asked to derive inferences on the clinical importance of the progression, so they were asked to state whether they noted progression that they would judge as substantial in their reports in daily practice. That there was an (unavoidable) difference in the definitions of relevant progression in the assessments of the radiologists and rheumatologists, however, is not likely the cause of the results found in this study. Moreover, the fact that the radiologists judged fewer radiograph sets as progressive, regardless of the clinical relevance, revealed that the difference in definitions did not cause the differences found between the judgments of the 2 professions. However, which profession gave the most valid judgment remains debatable and cannot be definitively determined by this study. However, the study did reveal that the type of profession can strongly influence panels' judgments and — although not feasible for research like ours in which the MCID is defined as that progression of the outcome measure that would make the professional want to change treatment of that patient — from a generalizability point of view it seems important to include more than one profession in an expert panel.

The concurrent validity of expert panels to assess MCID can also be assessed by comparing their judgments with the judgments of the actual patients. In many cases, the judgments of the patients can even be considered the gold standard (i.e., in the case of pain and quality of life), but it is clear that patients cannot decide whether progression of radiological damage is clinically important or not. Although we realize that radiological joint damage is an intermediate outcome measure it is still believed to be important to assess in trials in addition to the patient-reported outcomes. From a research point of view, it is consequently important to estimate the clinical relevance of a certain progression score of radiological joint damage. Deriving the MCID from a clinician's global assessment based on experience and knowledge, however, is not the final step. A data-driven approach will have to lead to more scientific evidence for clinical relevance of a certain progression score of radiological joint damage. The question remains, however, which data-driven approach will provide unambiguous answers, particularly because outcome measures like disability also largely depend on factors external to joint damage caused by RA.

In this study, the radiologists were consistent over time and with each other, and were able to differentiate patients with more progression from those with less progression, but were reserved in judging important changes compared to the panel of rheumatologists. It seems that minor changes that were not evaluated as substantial by the radiologists were judged to be clinically important by rheumatologists in patients with early RA, whatever the disease activity, and in patients with advanced RA with high disease activity.

REFERENCES

1. 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.
2. Bernstein SJ, Hofer TP, Meijler AP, Rigter H. Setting standards for effectiveness: a comparison of expert panels and decision analysis. Int J Qual Health Care 1997;9:255-63.
3. Maillefert JF, Gueguen A, Nguyen M, et al. Relevant change in radiological progression in patients with hip osteoarthritis. I. Determination using predictive validity for total hip arthroplasty. Rheumatology Oxford 2002;41:142-7.
4. Maillefert JF, Nguyen M, Gueguen A, et al. Relevant change in radiological progression in patients with hip osteoarthritis. II. Determination using an expert opinion approach. Rheumatology Oxford 2002;41:148-52.
5. Hotvedt R, Lossius HM, Kristiansen IS, Steen PA, Soreide E, Forde OH. Are expert panel judgments of medical benefits reliable? An evaluation of emergency medical service programs. Int J Technol Assess Health Care 2003;19:158-67.
6. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 2000;27:261-3.

6a. Boers M, Verhoeven AC, Markusse HM, et al. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet 1997;350:309-18. Erratum in: Lancet 1998;351:220.

7. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. AJR Am J Roentgenol 1981;137:1055-8.
8. Berbaum KS, Franken EA Jr, Dorfman DD, et al. Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs. Invest Radiol 1986;21:532-9.
9. Berbaum KS, Franken EA Jr, Dorfman DD, Lueben KR. Influence of clinical history on perception of abnormalities in pediatric radiographs. Acad Radiol 1994;1:217-23.
10. Eldevik OP, Dugstad G, Orrison WW, Haughton VM. The effect of clinical bias on the interpretation of myelography and spinal computed tomography. Radiology 1982;145:85-9.
11. Babcook CJ, Norman GR, Coblentz CL. Effect of clinical history on the interpretation of chest radiographs in childhood bronchiolitis. Invest Radiol 1993;28:214-7.
12. Brealey S, Scally AJ. Bias in plain film reading performance studies. Br J Radiol 2001;74:307-16.
13. van der Heijde D. Radiographic progression in rheumatoid arthritis: does it reflect outcome? Does it reflect treatment? Ann Rheum Dis 2001;60 Suppl 3:47-50.
14. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ 1995;311:376-80.
15. Vella K, Goldfrad C, Rowan K, Bion J, Black N. Use of consensus development to establish national research priorities in critical care. BMJ 2000;320:976-80.