Smallest Detectable Difference in Radiological Progression

MARISSA LASSERE, MAARTEN BOERS, DÉSIRÉE van der HEIJDE, ANNELIES BOONEN, JOHN EDMONDS, ARIANE SAUDAN, and ARCO C. VERHOEVEN

ABSTRACT. OMERACT IV started a discussion on the development of radiological response criteria in rheumatoid arthritis (RA). Such criteria depend on the definition of what constitutes the minimum clinically important progression of damage. Because such a definition is currently not available, as a first step we have used the concept of random measurement error to determine what is the smallest detectable difference (SDD) in radiological progression between 2 radiographs of a particular patient. Baseline and 12 month radiographs (hands, wrists, feet) of 52 patients representative of the spectrum of radiological progression were selected from a randomized controlled trial of early rheumatoid arthritis (COBRA study) and were read paired and chronologically by 2 observers using the van der Heijde modified Sharp method (0-448 scale) and another 2 observers using the Scott modified Larsen method (0-200). The measurement error of progression was determined using the metric 95% limits of agreement method of Bland and Altman. In the setting of early RA the SDD is 11 modified Sharp score units and 8 modified Larsen score units if there is an equal distribution of baseline damage and progression in the sample and the mean score of the same trained observers is always used. The SDD is 15.5 modified Sharp score units and 11 modified Larsen score units if there is an equal distribution of baseline damage and progression in the sample and the mean score of any 2 trained observers is used. Other SDD were determined depending on the context of measurement. Although this exercise needs repetition in other settings, the SDD is a useful starting point in the development of radiological response criteria. (J Rheumatol 1999;26:731–9)

Key Indexing Terms:
RELIABILITY RADIOGRAPHS MEASUREMENT ERROR
RHEUMATOID ARTHRITIS CLINICALLY IMPORTANT DIFFERENCE SENSITIVITY

There is nothing more difficult to carry out, nor more doubtful of success, nor more dangerous to handle, than to initiate a new order of things. — Niccolo Machiavelli

Radiographic indices are an important component of the assessment of rheumatoid arthritis (RA). Some consider them the premiere outcome measure in RA[1]. Others have pointed out the somewhat dangerous boundary between the gold standard and "fool's gold."[2] There are clear advantages to using radiological assessment as an outcome measure in RA. Radiographs reflect the history of joint damage. They can be used to assess damage at one point in time or can be assessed serially to evaluate the progression of damage over time. They are related to other measures of outcome such as physical function and clinical joint damage. They provide a permanent record. The scoring of radiographs can be performed blind, can be repeated on more than one occasion, and can be randomized to assess the effect of observer bias and random measurement error[3,4].

The disadvantages of radiological assessment are that many of the scoring methods were developed without access to our current knowledge of scale construction, and there is too much choice (which joints, views, scoring method, radiographic technique)[5]. Inappropriate study designs and statistical analyses have been used to ascertain the reliability of the radiological methods and the progression of radiological damage[6]. However, in the last few years, there has been a shift of opinion regarding the importance of reliability and its evaluation[7].

One of the objectives of the OMERACT IV RA Imaging Module was to start a discussion on the development of radiological response criteria. Such criteria depend on the definition of what constitutes the minimum clinically important progression of damage. Because such a definition is currently not available, our first step was to determine what would be the smallest detectable difference (SDD) in radiological progression. Given the recent interest in the reliability of radiographic measurement, this value was

From the Departments of Rheumatology, St. George Hospital, Sydney, Australia; University Hospital Maastricht, Maastricht, The Netherlands; Hôpital Beau-Séjour, Geneva, Switzerland; and the Department of Clinical Epidemiology, Amsterdam, The Netherlands.

M. Lassere, MBBS, FRACP, PhD, Research Fellow, Department of Rheumatology, St. George Hospital; M. Boers, MSc, MD, PhD, Professor of Clinical Epidemiology, VU University Hospital; D. van der Heijde, MD, PhD, Associate Professor of Rheumatology; A. Boonen, MD, Rheumatologist, Department of Rheumatology, University Hospital Maastricht; J. Edmonds, MBBS, FRACP, Professor of Rheumatology, Department of Rheumatology, St. George Hospital; A. Saudan, MD, Consultant Rheumatologist, Hôpital Beau-Séjour; A. Verhoeven, MD, Department of Rheumatology, University Hospital Maastricht.

Address reprint requests to Dr. M. Lassere, Department of Rheumatology, The St. George Hospital, Gray Street, Kogarah, NSW 2217, Australia. E-mail: p2112049@vmsuser.acsu.unsw.edu.au

based on the measurement error for scoring progression by 2 independent observers using the statistical methods proposed by Bland and Altman[8-10]. Data and radiographs from a recent multicenter, double blind randomized trial of COBRA (combination sulfasalazine, methotrexate and prednisolone with sulfasalazine alone[11]) were used to determine the measurement error for scoring progression.

MATERIALS AND METHODS

The COBRA study. Summary of methods and results. Between May 1993 and May 1995, 155 patients with early RA (median disease duration was 4 mo, no patient had disease for greater than 2 yrs) were randomly assigned combined treatment (76 patients) or sulfasalazine alone (76 patients). The main outcome measures were a weighted change score of 5 disease activity measures (a pooled index) and the van der Heijde modified Sharp score of hands, wrists, and feet after 56 weeks. The study protocol and findings are described in detail[11]. The patients all met 1987 American College of Rheumatology criteria for RA, were aged between 16 and 70 years [mean age in the combined group: 49.5 (SD 11.9); in the sulfasalazine group: 49.4 (SD 12.3)], overall 41.3% were male, 58% had positive IgM rheumatoid factor, 54% were HLA-DR4 positive, and 73.5% had erosive disease. Patients were required to have ≥ 6 actively inflamed joints, located at ≥ 3 different sites.

Two trained observers (AV, AB) independently scored radiographs of hands and feet according to van der Heijde's modification of Sharp's method[12] unaware of the identity and treatment status of the patients. The radiographs were read in an ordered fashion (baseline, and at Weeks 28, 56, and 80), therefore scores could either be stable or increase but not decrease over time. The means of the 2 observer scores were used. At 56 weeks the radiographic damage score had increased by a median of 2 (0-43) modified Sharp units in the COBRA group and 6 (0-54) modified Sharp units in the sulfasalazine group ($p = 0.004$).

Further analysis of COBRA radiographs. In December 1997, the van der Heijde/Sharp disaggregated baseline and Week 56 radiographic scores (i.e., scores separated for joint regions) from the COBRA study (135 patients) were obtained for re-analysis to ascertain the measurement error of progression using the van der Heijde modified Sharp score. Rather than selecting radiographs at random from the original group (which had a disproportionate number of patients with minimal damage at baseline and little progression over 12 months), the 52 radiographs were chosen to reflect an equal distribution of baseline status and progression scores, because measurement error is influenced by the spectrum of scores as well as the underlying agreement between methods, observers, etc. Therefore, 13 patients had low baseline radiographic scores and minimal progression, 13 patients had low baseline radiographic scores and considerable progression, 13 patients had high baseline radiographic scores and minimal progression, and 13 patients had high baseline radiographic scores and considerable progression. The baseline and 56 week radiographs of these 52 patients were copied and sent to 2 independent observers (JE, AS), who were unaware of the identity and treatment status of the patients and were trained in scoring radiographs using the Scott modified Larsen method[13]. The 52 pairs of radiographs were read paired and chronologically. Observer JE also randomly selected a subset of the radiographs for repeat scoring, unaware of his previous scores. All initial analyses were performed without knowledge of the patients' treatment status.

Statistical methods. The measurement error of damage progression was determined using Bland and Altman's 95% limits of agreement method[8,9]. This method provides an absolute and metric estimate of random measurement error. If there are only 2 observations per subject, then the standard deviation (SD) of the differences estimates how well the observers are likely to agree for a subject, and about twice this value defines the 95% limits of agreement for the observers under comparison. If the mean of the 2 observers' progression scores is to be used as the final outcome measure, then SD of the differences are divided by the square root of 2[8,14] and about twice this value defines the 95% limits of agreement. These limits are then judged to be acceptable (or not), depending on the context of the measurement. The 95% confidence intervals (CI) of the upper and lower limits of agreement can also be estimated, if the results of the field study are to be generalized beyond the study sample and observers. The statistical methods are described in detail in Appendix 1.

Other methods that have been used to evaluate reliability and agreement were also reported. These were the fixed and random effects intraclass correlation coefficients (ICC), the Spearman correlation coefficient, and the paired t test. The fixed effects ICC (Type 1.3) and random effects ICC (Type 1.2) and their 95% CI were calculated using the formulae provided by Shrout and Fleiss[15] and the 95% CI for the random effects ICC was also estimated using bootstrapping[16]. Although the Spearman correlation coefficient and the paired t test are considered to be inappropriate methods of assessing agreement they were reported to allow comparison of the results with published data.

Finally, to directly compare the 2 scoring methods the van der Heijde modified Sharp and Scott modified Larsen scores were linearly transformed from their original scale to a scale from 0 to 100 (the former was multiplied by 0.2232, the latter was multiplied by 0.5).

RESULTS

To show the distribution of the radiographic scores by group and method, summary results for the average radiographic scores at baseline (denoted Time 0) and at 56 weeks (Time 1) are shown in Tables 1-3. As expected the baseline scores at Time 0 were lower than the scores at Time 1 (at Week 56) for all groups and methods. Furthermore, the modified Sharp score means and SD were smaller in the total group of 135 patients compared to the subset of 52 patients, indicating that the attempt to choose a subset of patients more representative of the range of damage was successful. However, the distribution of damage remained non-Gaussian, the median scores for all groups were smaller than the mean. Comparison of the relative amount of radiographic damage by joint region using the Sharp scores was difficult because each joint region had different potential maximal scores. However, in the Larsen method, each joint region contributed 50 units to the total score and the wrist appeared to have the most damage, followed by the feet, then the metacarpophalangeal joints and finally the proximal interphalangeal joints (data not shown).

Summary statistics of the difference between observers' absolute scores at Time 0 (Baseline) and at Time 1 (Week 56) by scoring method, observer, and joint region are also shown in Tables 1-3. Over all there was little systematic bias between observers (mean of the difference scores). The SD of the difference scores determined random measurement error and twice this value approximates the 95% limits of agreement. Therefore, the 95% limit of agreement on the absolute scores for the modified Sharp was 12.6 Sharp units and for the modified Larsen score was 17.1 units at Time 0, and 19.0 Sharp units and 20.1 Larsen units at Time 1.

However, our main interest concerned the progression scores. The distribution of the radiographic progression scores (Time 1 scores minus Time 0 scores) by method and group is shown in Table 4. Summary statistics of the differ-

Table 1. Summary statistics of the average scores of the 2 observers and summary statistics of the difference scores between the 2 observers' scores at Time 0 (Baseline) and at Time 1 (Week 56) for modified Sharp method (52 selected patients).

	Average of the 2 Observers[a]		Difference Between the 2 Observers[b]	
	Mean (SD)	Median (min, max)	Mean (SD)[c]	Median (min, max)
Time 0: Baseline				
Sharp erosion score (ES)	8.6 (10.7)	3.3 (0,48)	–2.0 (5.0)	0 (–18,7)
Sharp joint space narrowing score (JSNS)	5.4 (7.4)	1.5 (0,25)	–0.1 (3.6)	0 (–16, 10)
Total Sharp score (ES + JSNS) for hands, wrists, and feet (0–448)	13.9 (16.0)	5.8 (0,58)	–2.1 (6.3)	0 (–21,6)
Time 1: Week 56				
Sharp erosion score (ES)	17.5 (16.2)	15.0 (0,91)	–1.7 (7.1)	–1.5 (–18,14)
Sharp joint space narrowing score (JSNS)	9.7 (9.7)	6.0 (0,30)	1.2 (6.2)	1 (–21,26)
Total Sharp score (ES + JSNS) for hands, wrists, and feet (0–448)	27.2 (22.7)	21.5 (0,97)	–0.5 (9.5)	0.5 (–28,21)

[a]Average score of the 2 observers = (observer 1 score + observer 2 score) divided by 2;
[b]Difference score between the 2 observers = observer 1 score minus observer 2 score;
[c]Standard deviation of the difference scores (i.e., the $SD_{difference}$), which is an estimate of random measurement error used to calculate the 95% limits of agreement.

Table 2. Summary statistics of the average scores of the 2 observers and summary statistics of the difference scores between the 2 observers' scores at Time 0 (Baseline) and at Time 1 (Week 56) for modified Larsen methods (52 selected patients).

	Average of 2 Observers[a]		Difference Between the 2 Observers[b]	
	Mean (SD)	Median (min, max)	Mean (SD)[c]	Median (min, max)
Time 0: Baseline				
Total Larsen score for hands, wrists, and feet (0–200)	15.1 (15.4)	9.5 (0,56)	–0.3 (8.6)	0 (–35,16)
Time 1: Week 56				
Total Larsen score for hands, wrists, and feet (0–200)	24.3 (16.8)	20.5 (1,65)	–1.1 (10.0)	1 (–38,15)

[a]Average score of the 2 observers = (observer 1 score + observer 2 score) divided by 2;
[b]Difference score between the 2 observers = observer 1 score minus observer 2 score;
[c]Standard deviation of the difference scores (i.e., the $SD_{difference}$), which is an estimate of random measurement error used to calculate the 95% limits of agreement.

ence between observers' scores for radiological progression by scoring method and group are also shown in Table 4. Systematic bias between observers was present but small, but random measurement error (the SD of the difference scores) was greater than the mean progression score for both scoring methods.

To directly compare the status and progression scores of the Sharp and Larsen methods, the scores were transformed to a 0-100 scale and the data were analyzed by scoring method and observer (data not shown). After this transformation the SD of the interobserver Sharp difference progression score (1.8) was much smaller than the interobserver and intraobserver Larsen difference progression scores (2.8 and 1.9, respectively). However, the better results for the Sharp scores were confounded: the Sharp method used a smaller part of the scale than the Larsen method.

The normality of the difference scores for radiological progression was assessed statistically using a test of skewness[16]. Skewness was still significant in the subset of 52 patients using the Sharp method (–0.65, $p = 0.003$), but there was no significant skewness using the Larsen method (0.28, $p = 0.37$).

The statistical analysis of reliability using the various statistical methods is summarized in Table 5. The scatterplots

Table 3. Summary statistics of the average scores of the 2 observers and summary statistics of the difference scores between the 2 observers' scores at Time 0 (Baseline) and at Time 1 (Week 56) for modified Sharp method (all patients, n = 135).

	Average of the 2 Observers[a]		Difference Between the 2 Observers[b]	
	Mean (SD)	Median (min, max)	Mean (SD)[c]	Median (min, max)
Time 0: Baseline				
Sharp erosion score (ES)	5.6 (8.3)	2.5 (0,48)	–0.9 (3.8)	0 (–18,11)
Sharp joint space narrowing score (JSNS)	2.8 (5.1)	1.0 (0,26)	0.4 (3.2)	0 (–16,14)
Total Sharp score (ES + JSNS) for hands, wrists, and feet (0–448)	8.4 (11.9)	3.5 (0,59)	–0.4 (5.3)	0 (–21,15)
Time 1: Week 56				
Sharp erosion score (ES)	11.9 (14.2)	7.0 (0,91)	–0.2 (5.7)	0 (–26,14)
Sharp joint space narrowing score (JSNS)	5.9 (8.0)	3.0 (0,38)	1.6 (5.3)	0 (–21,26)
Total Sharp score (ES + JSNS) for hands, wrists, and feet (0–448)	17.8 (20.2)	11 (0,97)	1.3 (8.4)	1 (–36,21)

[a]Average score of the 2 observers = (observer 1 score + observer 2 score) divided by 2;
[b]Difference score between the 2 observers = observer 1 score minus observer 2 score;
[c]Standard deviation of the difference scores (i.e., the $SD_{difference}$), which is an estimate of random measurement error used to calculate the 95% limits of agreement.

Table 4. Summary statistics of the observers' mean scores of radiological progression and the difference between observer scores of radiological progression for modified Sharp method (52 selected patients), modified Larsen method (52 selected patients), and modified Sharp method (all patients).

	Average of the 2 Observers[a]		Difference Between the 2 Observers[b]	
	Mean (SD)	Median (min, max)	Mean (SD)[c]	Median (min, max)
Selected patients (n = 52)				
Sharp erosion score (ES)	8.9 (9.1)	5.5 (0,43)	0.3 (5.4)	0 (–18,11)
Sharp joint space narrowing score (JSNS)	4.4 (6.2)	2.3 (0,30)	1.3 (5.0)	0 (–10,26)
Total Sharp score (ES + JSNS) for hands, wrists, and feet (0–448)	13.3 (13.1)	9.8 (0,51)	1.5 (7.8)	2 (–25,24)
Total Larsen score for hands, wrists, and feet (0–200)	9.2 (9.7)	5.0 (0,40)	–0.9 (5.6)	–1 (–13,15)
All patients (n = 135)				
Sharp erosion score (ES)	6.2 (7.8)	3.0 (0,43)	0.6 (4.1)	0 (–18,22)
Sharp joint space narrowing score (JSNS)	3.0 (5.1)	1.0 (0,30)	1.1 (4.2)	0 (–16,26)
Total Sharp score (ES + JSNS) for hands, wrists, and feet (0–448)	9.2 (11.5)	4.0 (0,56)	1.7 (6.3)	1 (–25,24)

[a]Average progression score of the 2 observers = (observer 1 score + observer 2 score) divided by 2;
[b]Difference score between the 2 observers = observer 1 score minus observer 2 score;
[c]Standard deviation of the difference scores (i.e., the $SD_{difference}$), which is an estimate of random measurement error used to calculate the 95% limits of agreement.

of the observer difference by mean progression scores and the 95% limits of agreement (for 2 observers and the mean of 2 observers) are shown in Figure 1.

The reliability as judged by the indirect methods using the various intraclass correlation coefficients was reasonable: both fixed and random effects intraclass correlation coefficients were > 0.84. Furthermore, the 95% CI for the respective ICC were also tolerably narrow. The direct methods of evaluation showed no or negligible systematic bias as judged by the mean difference values for each of the 4 analyses (the 95% CI nearly always included zero difference). The SD of the difference scores ($SD_{difference}$), an indicator of random measurement error, depended on the scoring method, the distribution of progression scores in the

Table 5. Summary statistics of reliability: observer studies of radiological damage progression.

Study	Analysis 1, Modified Sharp Method	Analysis 2, Modified Larsen Method	Analysis 3, Modified Sharp Method	Analysis 4, Modified Larsen Method
Study design	Interobserver, paired chronological	Interobserver, paired chronological	Interobserver, paired chronological	Intraobserver, paired chronological
Sample size	52	52	135	26
Spearman correlation (95% CI)	0.84 (0.75,0.91)	0.82 (0.70,0.88)	0.89 (0.84,0.91)	0.92 (0.83,0.99)
Fixed effects ICC (95% CI)	0.84 (0.73,0.90)	0.85 (0.75,0.91)	0.86 (0.79,0.90)	0.94 (0.87,0.97)
Random effects ICC (95% CI)	0.83 (0.75,0.90)	0.84 (0.77,0.91)	0.85 (0.81,0.90)	0.94 (0.88,0.99)
Bland and Altman's methods				
Scatterplot of difference vs mean scores	See Figure 1b	See Figure 1c	See Figure 1a	See Figure 1d
$Mean_{difference}$ (95% CI) (systematic bias)	1.5 (–0.6,3.7)	–0.8 (–2.4,0.8)	1.7 (0.7,2.8)	–0.04 (–1.6,1.5)
$SD_{difference}$ (random measurement error)	7.8	5.6	6.3	3.8
SDD ~ 95% limits of agreement (assuming no systematic bias)	± 15.6	± 11.2	± 12.6	± 7.6
95% CI of 95% limits of agreement	–20.4, 23.4	–16.5, 14.9	–13.7, 17.2	–12.2, 12.1
Mean score of 2 observers				
$SD_{difference}/2$ (random measurement error)	5.5	4.0	4.4	2.7
SDD ~ 95% limits of agreement (assuming no systematic bias)	± 11.0	± 8.0	± 8.8	± 5.4
95% CI of 95% limits of agreement, based on $SD_{difference}/2$	–13.9, 17.0	–11.9, 10.3	–9.2, 12.7	–8.6, 8.5

ICC: Intraclass correlation coefficient;
$Mean_{difference}$: mean of the difference between observers' scores for progression;
$SD_{difference}$: SD of the difference between observers' scores for progression.

sample, and whether the study design was interobserver or intraobserver. The modified Sharp score $SD_{difference}$ was smaller when the entire dataset of 135 patients was used (Analysis 3) compared to the $SD_{difference}$ calculated from the smaller subset of 52 patients (Analysis 1), because a greater proportion of patients overall had little damage and minimal progression, and agreement was generally better in such circumstances. Table 5 also shows the 95% limits of agreement and their CI based on the assumption that the mean of 2 observers' scores is used. In this situation the SD of the difference scores is divided by 2 (i.e., by about 1.414) and measurement error is therefore smaller.

If random measurement error of progression is calculated using the 95% limits of agreement and if the smallest detectable difference in radiological progression is determined from random measurement error, then the smallest detectable difference in radiological progression is the 95% limits of agreement. Table 6 shows the various smallest detectable differences for the Sharp and Larsen methods conditional on the distribution of radiographic progression in the population of interest (because the reliability of a measure in a field trial depends on the distribution of that measure in the trial) and whether the results are to be generalized to other observers.

These SDD only apply if the radiographs are scored paired and chronologically, the observers are trained, and patients have early RA. Furthermore, these SDD are a robust summary measure of measurement error if the relationship between the difference scores and the mean scores is constant. However, because measurement error was smaller where there was little damage and minimal progression (as shown in Figure 1a), the smallest detectable difference is marginally overestimated at lower progression scores and underestimated at higher progression scores.

Finally, the performance of these SDD was used to evaluate the effect of treatment on radiological damage (Sharp method) using the COBRA study results of combination treatment versus sulfasalazine group. Using a SDD > 8.5 (mean of 2 progression scores, same 2 observers, little baseline damage, and minimal radiological progression), 75% of the COBRA treatment group had no radiographic progression compared to 55% of the sulfasalazine alone group [chi-squared = 5.3 (1 df), p = 0.021]. Using a SDD > 11.0 (mean of 2 progression scores, same 2 observers, equal distribution of baseline damage and radiological progression or mean of 2 progression scores, any observers, little baseline damage and minimal radiological progression), 77% of the COBRA treatment group had no radiographic progression compared

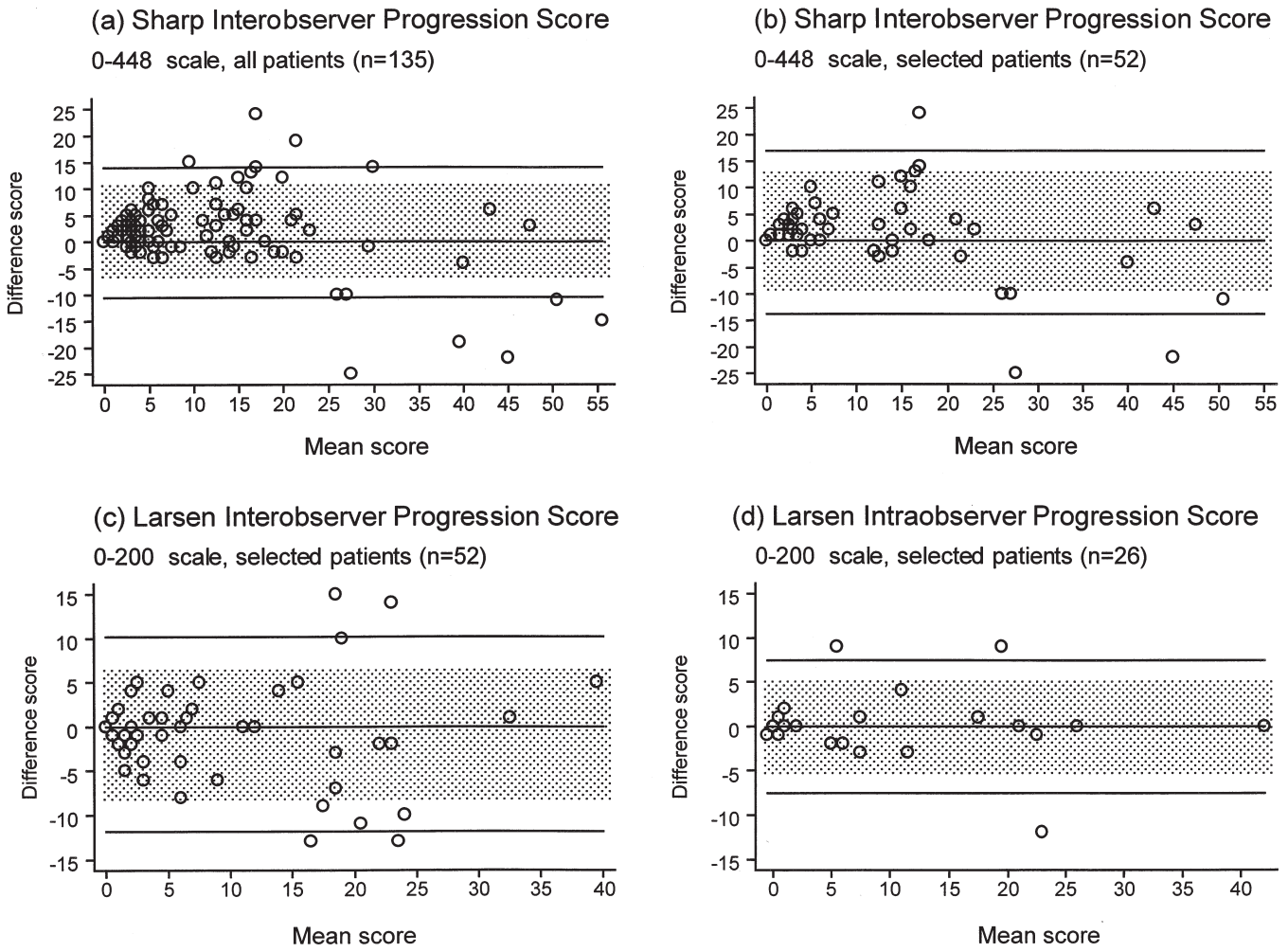

Figure 1. Graph of the difference scores against the mean score of radiographic progression. (a) Interobserver modified Sharp score result for all 135 patients. (b) The interobserver modified Sharp score result for the 52 selected patients. (c) Interobserver modified Larsen score result for the 52 selected patients. (d) The intraobserver modified Larsen score result for 26 selected patients. The shaded, narrower 95% limits of agreement interval represents the result for the mean score of 2 fixed observers; the outermost lines represent the 95% limits of agreement for 2 observers.

to 57% of the sulfasalazine alone group [chi-squared = 6.3 (1 df), p = 0.012]. Using a SDD > 15.5 (mean of 2 progression scores, any observers, equal distribution of baseline damage, and radiological progression), 81% of the COBRA treatment group had no radiographic progression compared to 72% of the sulfasalazine alone group [chi-squared = 1.6 (1 df), p = 0.21].

DISCUSSION

We have used the concept of random measurement error to determine the smallest detectable radiological progression in a one year period in early RA. We arbitrarily defined the smallest detectable difference in radiological progression as that number greater than the interobserver measurement error of progression as determined by Bland and Altman's 95% limits of agreement method. More than one SDD can be calculated depending on the scoring method, the distribution of damage and progression, the number of scorers, the way their scores are combined, and whether the same observers are always used (vs any random sample of observers chosen from all possible observers). In our suggested setting where the mean score of 2 fixed observers is reported using paired chronological reading, the smallest detectable difference in radiological progression is 11 van der Heijde modified Sharp units and 8 Scott modified Larsen units, in a sample where baseline damage and progression is relatively evenly distributed. Other conditional-specific SDD were also determined.

We are not the first to interpret radiological progression by considering measurement error. In 1990 O'Sullivan, *et al* evaluated one year progression scores, and defined progression as a change greater than the interobserver status score measurement error[17]. Furthermore, in 1995, Menninger, *et al*[18] and Ruckmann, *et al*[19] determined the responsiveness of

Table 6. Smallest detectable difference in radiological progression by method and group, same 2 observers or any 2 observers. Values rounded to nearest 0.5 number.

Distribution of Baseline Damage and Progression in Population	Scoring Method				
	Van der Heijde Modified Sharp (0–448 scale)		Scott Modified Larsen (0–200 scale)		
	Score of 1 Observer	Mean Score of 2 Different Observers	Score of 1 Observer	Mean Score of 2 Different Observers	Mean Score of 1 Observer Scoring on 2 Occasions
Same 2 observers (results are not generalizable to other observers)					
Little baseline damage and minimal progression in population	12.5	8.5	NA	NA	NA
Baseline damage and progression more evenly distributed in population	16	11	11.5	8	5.5
Any 2 observers (results are generalizable to other observers)					
Little baseline damage and minimal progression in population	15.5	11	NA	NA	NA
Baseline damage and progression more evenly distributed in population	22	15.5	15.5	11	8.5

NA: not available.

the Larsen score by counting the number of patients that showed radiological progression greater than the intraobserver status score measurement error. The following year, Dougados, *et al*[20] in an elegant study of hip osteoarthritis, determined a cutoff value for change in radiological progression based on intraobserver progression score measurement error. In our study, we defined radiological progression by the interobserver progression score measurement error.

How do the reliability results of our study compare with other studies of RA? Although most studies included an evaluation of the reliability of the scoring method, few used an appropriate method of analysis. Fewer still used the limits of agreement method, and no study used this method to evaluate the reliability of progression. O'Sullivan, *et al*[17] reported an intraobserver 95% limits of agreement for the Larsen status score (hands, wrists, feet, 0-210) as ± 8 and an interobserver agreement of ± 11. The observers had undergone considerable training. Ruckmann, *et al*[19] reported an even smaller intraobserver agreement of ± 7 (Larsen score range not provided), whereas Guth, *et al*'s[21] intraobserver agreement was ± 25 (Larsen score 0–150). The discrepancy between Ruckmann, *et al*'s and Guth, *et al*'s results is explained by examination of the graphical analysis (difference score versus mean score). Ruckmann, *et al* evaluated 24 patients and 21 of these patients had a Larsen score of less than 10. Whereas Guth, *et al* assessed 71 patients, and 66 patients had a Larsen score between 6 and 75. These results demonstrate how measurement error is usually small if there is negligible abnormality. By comparison our modified Larsen status score interobserver agreement was good (± 10), particularly as the Larsen scores were evenly distributed between 0 and 60.

There has been only one report on the reliability of Sharp's method of scoring using the limits of agreement method[21], although Sharp, *et al*'s 1985 publication[22] on the reliability of his method provided sufficient data on the first author's (Sharp) own intraobserver agreement, which was acceptable at ± 15 (scaled from 0 to 314). The random effects ICC was also calculated using the available data and it appeared to verify these findings (intraobserver ICC: 0.95). Guth, *et al*'s[21] intraobserver agreement was ± 30, but about one-third of his scores were > 100, whereas all but one of Sharp's scores were < 100 and the majority were < 60. However, Guth, *et al*'s ICC of the Sharp method was 0.97, which was better than Sharp's, emphasizing that the more heterogenous the group (i.e., the greater the spectrum), the better the ICC, whereas in the limits of agreement method, the more heterogenous the group, the wider the limits of agreement, indicating poorer agreement. One further point of relevance to the analysis of reliability is the regression to the mean effect, illustrated by Sharp's own results. His second reading was lower than the first for most films and for 5 films the difference was at least 20 Sharp units, although the mean difference was only 1.5 points.

How do the SDD of our study compare with published studies of radiographic progression? Dawes, *et al* in a longitudinal observational study stated that 10 Larsen score units in one year was a significant increase[23]. O'Sullivan used the interobserver error of the status score (11 Larsen score units) to evaluate progression in another longitudinal observational study, and found that 11% of their patients had progressive radiographic damage in one year[17]. Hassell reported the annual median change in Larsen score was 1.42 units per year in a low disease activity group (as judged by

the Stoke index) and 6.5 units per year in a high disease activity group[24]. Menninger, *et al* used the intraobserver measurement error to define progression, and found it to be greater than 10 Larsen score points in 60% of patients[18]. In an early disease RA cohort, followed over 5 years, Fex, *et al* found that the rate of progression was about 5 Larsen score units per year in the first 2 years, then fell to 2-3 Larsen score units per year[25]. Sharp reported a mean rate of progression of 4 Sharp units per year over a 25 year period. The rate was more rapid in the early years, about 9 units per year[26]. Plant, *et al* in another early RA disease cohort found that median radiological progression in the first year was 6.5 Larsen score units and 8 (Plant modified) Sharp score units[27]. These results indicate that on average at least one year followup is required to reliably detect progression in an individual patient. The reported numbers more or less agree with our findings. However, the SDD needs to be measured in several other settings (e.g., longstanding RA, more heterogenous groups, other scoring methods) as it can be expected that in those settings the SDD will be greater.

Although this study did not attempt to compare the Sharp and Larsen methods of scoring, it is the first study that has compared and reported the interobserver reliability of radiographic progression for the Sharp and Larsen methods using the limits of agreement method, in a representative group of patients with early RA. We found that agreement for the van der Heijde modified Sharp method appeared to be better than for the Scott modified Larsen method. This is clearly seen when the scores were linearly transformed to the same 0-100 scale. However, even this direct comparison of the 2 methods fails to take into account the differing distributions of the status and progression scores of the Larsen and Sharp methods, and other factors need to be considered before deciding which of the 2 methods had better agreement. Moreover, apart from discrimination, the choice of method also depends on the other elements of the OMERACT filter: truth and feasibility[28].

In conclusion, we have determined the minimum one-year radiological progression that can be reliably detected in an individual patient with early RA. Although further study in other settings is required, these results serve as a useful starting point for defining radiological response criteria and for stimulating further debate on what constitutes the lower boundary of a clinically important progression or difference in progression in an individual patient.

APPENDIX

Limits of Agreement Method

The steps for the limits of agreement method were as follows:

1. The mean and the difference for each paired observation were calculated.

2. The difference of the paired observations (ordinate) was graphed against the mean of the paired observations (abscissa) to show the scatter and the shape of the agreement (or disagreement).

3. The mean of the difference scores, $mean_{difference}$, and the standard deviation of the difference scores, $SD_{difference}$, were calculated.

4. The $mean_{difference}$ was an estimate of the mean systematic bias of observer 1 relative to observer 2. If the $mean_{difference}$ was zero, or considered to be some value that was negligible, then the observers agreed very well on average.

5. The $SD_{difference}$ estimated how well the observers agreed for an individual. An interval of the limits of agreement for the observers under comparison was constructed. If the differences were normally distributed then 95% of the differences would lie between $mean_{difference} \pm 1.96\ (SD_{difference})$, and this interval is the 95% limits of agreement of the observers under comparison. Note that the SD of the difference scores is equivalent to $(2*MS_{error})$, where MS is the mean square error from the single factor repeated measures ANOVA.

6. The standard error of the $mean_{difference}$, the $SE_{diffmean}$, was $(SD_{difference})^2 / n$, where n is the sample size. The 95% confidence interval for the $mean_{difference}$ was $mean_{difference} \pm t_{0.05,\ n-1\ (df)}\ (SE_{diffmean})$. The standard error of the limits of agreement, SE_{limits}, was approximately $[(3SD_{difference})^2 / n]$. The 95% confidence interval for the lower limit of agreement was: $[mean_{difference} - 1.96(SD_{difference})] \pm t_{0.05,\ n-1\ (df)}\ (SE_{limits})$. The 95% confidence interval for the upper limit of agreement was: $[mean_{difference} + 1.96(SD_{difference})] \pm t_{0.05,\ n-1\ (df)}\ (SE_{limits})$.

7. If the mean of the 2 observers' progression scores is to be used as the final outcome measure (such as in the COBRA study), then the variance of the mean score is multiplied by $(1/2)^2$ (i.e., 1/4)[8,14]. The variance of the mean score is therefore equal to $\sigma^2/2$, and the standard deviation of the mean score is now $\sigma/2$. The limits of agreement (and their 95% confidence intervals) were calculated and $\sigma/2$ was substituted in the above calculations.

REFERENCES

1. Sharp, JT. Radiological assessment of joint damage: the premier outcome measure in rheumatoid arthritis. Current status and future potential. In: Wolfe F, Pincus T, editors. Rheumatoid arthritis: pathogenesis, assessment, outcome and treatment. New York: Marcel Dekker; 1994:167-89.
2. Brower AC. Use of the radiograph to measure the course of rheumatoid arthritis: gold standard versus fool's gold. Arthritis Rheum 1990;33:316-24.
3. Sharp JT. Scoring radiographic abnormalities in rheumatoid arthritis. J Rheumatol 1989;16:568-9.
4. Dawes PT. Radiological assessment of outcome in rheumatoid arthritis. Br J Rheumatol 1988;27 Suppl 1:21-36.
5. van der Heijde D, Boers M, Lassere M. Methodological issues in radiographic scoring methods in rheumatoid arthritis. J Rheumatol 1999;26:726-30.
6. Lassere M, Edmonds JP. The measurement of reliability [abstract]. Arthritis Rheum 1997;40 Suppl:S109.
7. van der Heijde D. Plain x-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. Bailliere's Clin Rheumatol 1996;10:435-53.
8. Altman D, Bland M. Measurement in medicine: the analysis of method comparison studies. Statistician 1983;32:307-17.
9. Bland M, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
10. Bland J, Altman D. Measurement error. BMJ 1996;312:1654.
11. Boers M, Verhoeven A, Markusse A, et al. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet 1997;350:309-18.
12. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 1999;26:743-5.
13. 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.
14. Chinn S. The assessment of methods of measurement. Stat Med 1990;9:351-62.

15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
16. StataCorp. Stata Reference Manual: Release 5.0, Vol 3. College Station, TX: Stata Press; 1997.
17. O'Sullivan MM, Lewis PA, Newcombe RG, et al. Precision of Larsen grading of radiographs in assessing progression of rheumatoid arthritis in individual patients. Ann Rheum Dis 1990;49:286-9.
18. Menninger H, Meixner C, Sondgen W. Progression and repair in radiographs of hands and forefeet in early rheumatoid arthritis. J Rheumatol 1995;22:1048-54.
19. Ruckmann A, Ehle B, Trampisch HJ. How to evaluate measuring methods in the case of non-defined external validity. J Rheumatol 1995;22:1998-2000.
20. Dougados M, Gueguen A, Nguyen M, et al. Radiological progression of hip osteoarthritis: definition, risk factors and correlations with clinical status. Ann Rheum Dis 1996;55:356-62.
21. Guth A, Coste J, Chagnon S, Lacombe P, Paolaggi JB. Reliability of three methods of radiologic assessment in patients with rheumatoid arthritis. Invest Radiol 1995;30:181-5.
22. Sharp J, Bluhm G, Brook A, et al. Reproducibility of multiple observer scoring of radiologic abnormalities in hands and wrists of patients with RA. Arthritis Rheum 1985;28:16.
23. Dawes PT, Fowler PD, Clarke S, Fisher J, Lawton A, Shadworth MF. Rheumatoid arthritis: treatment which controls the C-reactive protein and erythrocyte sedimentation rate reduces radiological progression. Br J Rheumatol 1982;25:44-9.
24. Hassell AB, Davis MJ, Fowler PD, et al. The relationship between serial measures of disease activity and outcome in rheumatoid arthritis. Q J Med 1993;86:601-7.
25. Fex E, Johnson U, Johnson K, Eberhardt K. Development of radiographic damage during the first 5-6 years of rheumatoid arthritis. A prospective follow up study of a Swedish cohort. Br J Rheumatol 1996;35:1106-15.
26. Sharp JT, Wolfe F, Mitchell DM, Bloch DA. The progression of erosion and joint space narrowing scores in rheumatoid arthritis during the first twenty-five years of disease. Arthritis Rheum 1991;34:660-8.
27. Plant MJ, Saklatvala J, Borg AA, Jones PW, Dawes PT. Measurement and prediction of radiological progression in early rheumatoid arthritis. J Rheumatol 1994;21:1808-13.
28. Boers M, Brooks P, Strand V, Tugwell P. The OMERACT filter for outcome measures in rheumatology. J Rheumatol 1998;25:198-9.