Variability of Precision in Scoring Radiographic Abnormalities in Rheumatoid Arthritis by Experienced Readers

JOHN T. SHARP, FREDERICK WOLFE, MARISSA LASSERE, MAARTEN BOERS, DÉSIRÉE van der HEIJDE, ARVI LARSEN, HAROLD PAULUS, ROLF RAU and VIBEKE STRAND

ABSTRACT. ***Objective.*** To determine the extent of precision and sources of variability among experts on scoring radiographic abnormalities in rheumatoid arthritis.

Methods. Radiographic scores from 6 datasets in which 2 or more readers had scored film sets were analyzed. Datasets included scores by 11 different readers, 6 of whom scored films by both the Larsen (global) and Sharp (composite) methods. Scores of each possible combination of 2 readers were compared in calculating the smallest detectable difference (SDD) on raw scores and on scores normalized for each individual reader (nSDD). Intraclass correlation (ICC), Pearson's r, and the correlation between differences in score and their mean scores were determined. Agreement on progression of radiographic damage scores was also examined.

Results. Variability among readers was greater than previous studies suggested. Agreement was better for intra- than interreader comparisons; average intrareader SDD was 24.4 for the composite method and 9.0 for the global. The larger SDD for the composite method reflect their greater range of possible scores. When normalized scores were used to adjust for the range difference, there was minimal difference in the SDD; nSDD was 10.1 for the composite method, 8.0 for the global. Interreader variability was larger: SDD of 53.7 for the composite method and 23.3 for the global; nSDD 12.9 and 14.4, respectively. ICC varied between 0.465 and 0.999, with all but one value below 0.925 occurring in composite scores with a range below 100. Differences in repeated scores were frequently associated with the mean of those scores and this was greater for inter- than for intrareader comparisons. Agreement between progression scores showed a similar pattern. The SDD was better for intrareader comparisons and smaller for global scores: compare 13.7 (composite, intrareader) and 5.4 (global, intrareader) to 18.1 (composite, interreader) and 8.7 (global, interreader). The ICC was lower for progression scores than for raw scores, averaging between 0.661 and 0.885.

Conclusion. The variability in scoring radiographic abnormalities is considerable among this group of 11 expert readers. This has important implications for power calculations in comparison studies such as therapeutic trials and for cross-trial comparisons. The correlation between the difference in repeated scores and their means indicates systematic error (bias), which, if corrected, may improve the detection of treatment effects when using a responder-type analysis. These and other design and analysis issues are discussed. (J Rheumatol 2004;31:1062–72)

Key Indexing Terms:
RHEUMATOID ARTHRITIS RADIOGRAPH SCORES PRECISION
SMALLEST DETECTABLE DIFFERENCE INTRACLASS CORRELATION

From the University of Washington School of Medicine, Seattle, Washington, USA; National Data Bank for Rheumatic Disease, Wichita, Kansas, USA; University Medical Center, Amsterdam, The Netherlands; University Hospital Maastricht, Maastricht, The Netherlands; Kongsvinger Hospital, Kongsvinger, Norway; St. George Hospital, University of New South Wales, Sydney, Australia; H. Paulus, University of California at Los Angeles, Los Angeles, California, USA; Evangelisches Fachkrankenhaus, Ratingen, Germany; and Stanford University, Palo Alto, California, USA.

J.T. Sharp, MD, University of Washington School of Medicine; F. Wolfe, MD, National Data Bank for Rheumatic Disease; M. Boers, MD, University Medical Center Amsterdam; D. van der Heijde, MD, University Hospital Maastricht; A. Larsen, MD, Kongsvinger Hospital; M. Lassere, MD, St. George Hospital, University of New South Wales; H. Paulus, MD, University of California at Los Angeles; R. Rau, MD, Evangelisches Fachkrankenhaus Ratingen; V. Strand, MD, Stanford University.

Address reprint requests to Dr. J.T. Sharp, 8387 NE Sumanee Place, Bainbridge Island, WA 98110, USA. E-mail: johntsharp@worldnett.att.net

Submitted February 19, 2003; revision accepted November 24, 2003.

Two methods of scoring radiographic abnormalities in rheumatoid arthritis (RA) have been used extensively in the last 30 years to describe the course of the disease in patients receiving the best treatment then available and to measure treatment effects in therapeutic trials. The Sharp method is a composite score that assigns separate scores for erosions and joint space narrowing (JSN) of multiple joints in the hands and sums these to give a total radiographic score[1]. The Larsen method is a single, global score for each joint, which, as originally described, represented erosions, juxtaarticular osteoporosis, and soft tissue swelling[2]. Both methods have been widely used and modified[3-6]. Most modifications have altered the selection and number of specific joints to be assessed. A particularly noteworthy addition to the composite method was introduced by van der Heijde, who emphasized the importance of including assess-

ment of toe joints and doubled the scale for erosion scores of the metatarsophalangeal joints while retaining the original scale for erosions in the hands and for JSN scores[4]. Modifications to the composite method have been proposed by Genant[7] and to the global method by Scott and Laasonen[6], Larsen[5], and Rau, *et al*[8]. Lassere recently summarized the key features of all published radiographic scoring methods[9].

The value of any measurement depends on its performance. The performance of a measurement is determined by its reliability, validity, responsiveness, and feasibility. Reliability, also known as precision, is a key but often overlooked element of performance[10]. Only with a thorough evaluation of interoccasion, intraobserver, and interobserver reliability under different conditions, and using different methods of statistical analysis, can we identify the sources of variability and quantify the amount of error inherent in a measurement method. Reducing measurement error (i.e., improving precision) is an important goal. Measurement error is a determinant of the number of subjects required to establish a difference in therapeutic trials as well as in studies to establish the predictive value of factors in longitudinal, observational studies of disease outcome. Further, a measurement's validity is dependent in part on its reliability.

The availability of new radiographic datasets in RA provided us the opportunity to reassess the precision of the composite (Sharp) and global (Larsen) scoring systems and to highlight important but often neglected design, analysis, and interpretation issues when reporting the precision of scoring radiographic abnormalities. Previous studies have reported the limits of agreement between duplicate readings of radiographic damage in RA for a small number of investigators. To determine how well these studies define the extent of variability between methods and a larger number of readers we invited investigators representing 10 databanks to participate in a pooled analysis; 6 agreed, 3 made up of composite scores and 3 including both composite and global scores.

MATERIALS AND METHODS

Data sources, scoring methods and readers

The 6 data sets (Table 1) made available were from the Denver Alpert Arthritis Center databank (from ARAMIS); the Wichita Arthritis Center; the COBRA trial; the Evangelische Fachkrankenhaus Ratingen; the Western Consortium of Practicing Rheumatologists dataset; and the Aventis data on controlled trials of leflunomide[11]. Three centers provided input to the Denver databank including the Wichita Arthritis Center, the Saskatoon arthritis program, and the Alpert Arthritis Center, and 3 readers scored films using the composite method; one reader scored a subset of films twice (intrareader) as well as films that were scored by each of the other 2 readers (interreader); the other 2 readers did not read any films in common. One set of films from the Wichita databank was utilized in the Denver databank study[12]. A second, separate set of Wichita films was read by both Larsen and Sharp using their respective methods (intermethod). The COBRA trial films were each scored by 2 van der Heijde-trained readers using the van der Heijde-modified composite method[13]. Six readers scored 210 films in the Ratingen study, 7 films from each of 30 patients with films at baseline and at 6 and 12 months and every one or 2 years thereafter. Ratingen patients were seen early after onset of disease; films of hands and feet were scored by both the composite and global methods by each of the 6 readers (interreader and cross-method). (Unpublished) Films for the Ratingen study were selected to represent a broad range of disease severity by choosing 20 patients at random from 128 patients with early RA (mean disease duration 11 months) participating in a trial with 5 years' followup, and 10 patients with established, severe disease and 10 years' followup. In addition, between 28 and 105 films, varying among the different readers, were scored twice by each method (intrareader). The Western Consortium films were collected from patients with disease duration less than one year and before disease modifying antirheumatic drug (DMARD) treatment was begun. Films were collected at entry and at 6 and 12 months and then yearly. Followup time in this cohort varied since enrollment extended over several years, resulting in the majority having low scores, since most films were taken early in the disease. The Western Consortium films were scored by 2 readers (interreader), with subsets scored a second and third time by each (intrareader). The leflunomide trial films were scored by Larsen and Sharp reading by their respective methods (intermethod). In addition, one reader scored one set twice using the composite method (intrareader). Datasets representing the radiographic scoring methods of Genant[7] and Rau, *et al*[8] were not available for study. The Ratingen dataset included here was read before development of the Rau method.

Table 1 outlines the analysis plan and shows the datasets on which 135 comparisons were made. Four datasets included patients with long followup, which ensured a wide spectrum of scores. The spectrum of scores was narrow and at the low end of the scoring range in one set of patients with early disease. The 6 available datasets include duplicate scores by 11 different readers, some of whom participated in multiple studies. These readers represent a significant proportion of investigators who have participated in numerous published reports of radiographic outcome in therapeutic trials and descriptive studies. The senior author scored films in all but one of the datasets, Larsen participated in 3 of the studies, and van der Heijde in one. Six readers participated in one study, reading films by each method, making possible 21 comparisons for each, 15 interreader and 6 intrareader, and 21 for cross-methods comparisons. All but 2 datasets included more than 50 film sets. In 3 studies patients were not selected for disease duration and included patients with a wide range of scores. In the other 3 studies enrollment was limited to patients with early disease.

Data analysis

Data were forwarded to the senior author as scores for individual joints and analyzed using Stata (Stata Corp., College Station, TX USA), SPSS (Chicago, IL, USA), and Excel (Microsoft, Redmond, WA, USA) software for all datasets except the leflunomide trials. Aventis Pharmaceuticals (Bridgewater, NJ, USA) provided total scores and erosion and JSN scores for the composite method, coded to maintain confidentiality of patient identity and without clinical information. In the Aventis set, Larsen scores were provided as average scores per joint scored. The Larsen scores were converted to total scores by multiplying by 40, the number of scores that would be included provided no joints were unscored. In the Western Consortium data, missing scores for individual joints were assigned a zero score if there were no more than 3 missing scores for all right or left erosion scores or for all right or left narrowing scores. A small number of out of range scores in the Ratingen dataset were assigned these scores as a code for unreadable joints based on the memory of several participants.

Our study primarily reports the results of individual readers' scores rather than the average of 2 or more readers' scores. The smallest detectable difference (SDD)[14] derived from Bland and Altman's Limits of Agreement[15] and the intraclass correlation coefficient (ICC)[16] were the primary summary statistics of reliability. The SDD quantifies the random error component of reliability using an absolute metric. The mean and standard deviation of the paired differences within-reader (intrareader scores), between-reader (interreader scores), or between-method (intermethod scores) were calculated. The SDD is defined as the 95% confidence interval

Table 1. Data sets analyzed: comparisons between raw scores reported in Tables 2, 3, and 4.

Aramis
3 readers scored 3 separate film sets by composite method
Ad1* compared to Ad2, As1 to Bs1, Aw1 to Cw1 — 3 comparisons

Aventis
2 readers, one by composite, one by global method
A1 to A2
A1 to B1** } 2 comparisons

Cobra
2 readers by composite method
A1 to B1 — 1 comparison

Ratingen
6 readers read by 2 methods and included replicate readings for each method. Thus there are composite to composite, global to global, composite to global comparisons.
Interreader comparisons:
A1 to A2, B1 to B2, C1 to C2, D1 to D2, E1 to E2, F1 to F2—18 comparisons, 6 intramethod by each method and 6 intermethod
Intrareader comparisons:
A to B, C, D, E, F
B to C, D, E, F
C to D, E, F
D to E, F,
E to F
} 45 comparisons, 15 intramethod by each method, 15 intermethod

Western Consortium
2 readers, both by composite, 3 readings each (each reading included smaller numbers)
A1 to A2, A1 to A3, A2 to A3
B1 to B2, B1 to B3, B2 to B3
A1 to B1, A1 to B2, A1 to B3
A2 to B1, A2 to B2, A2 to B3
A3 to B1, A3 to B2, A3 to B3
} 15 comparisons

Wichita
2 readers, one composite, one global
A to B— 1 comparison

Comparison between progression scores reported in Tables 5 and 6

Aventis
Intrareader comparisons, 1 by composite method

Cobra
Interreader comparison, 1 by composite method

Ratingen
Intrareader comparisons, 6 by composite and 6 by global method
Interreader comparisons, 15 by composite and 15 by global method

Western Consortium
Intrareader comparisons, 2 by composite method
Interreader comparisons, 4 by composite method

* A is reader A, d is the Denver dataset, 1 is the first reading of this set, s is Saskatchewan dataset, w is the Wichita dataset. Reader B did not read Wichita films and reader C did not read Saskatchewan films. ** Readers A and B are not necessarily the same individuals in different studies.

of the standard deviation of the paired differences, which is estimated by multiplying the standard deviation of the paired differences by 1.96. The SDD is biased toward smaller values if the data vary over a narrow range of values. SDD of 0 is perfect agreement, and there is no convention that anchors the upper limit[10]. The ICC described by Shrout and Fleiss[16] is a relative measure of reliability. With a constant amount of variation (error) ICC values are biased towards higher coefficients when scores have a wider range of values. An ICC of 1.0 is perfect reliability[10].

To allow comparison with earlier studies of reliability of radiographic scoring methods we also report Pearson's correlation coefficient. We also calculated the correlation between readers' difference scores and mean scores.

Comparisons were made for intrareader, interreader, and for intermethod scores. For some calculations scores were normalized for individual readers using the range of scores employed in that dataset by that reader; this allowed calculation of the SDD limits of agreement between composite and global scores, which is not otherwise possible because they are scored on different scales. We calculated nSDD using the following: normalized score = [(raw score – minimal score)/(maximal score – minimal score)]*100, using the minimal and maximal scores employed by each individual reader.

RESULTS

There is considerable variation in the way different expert readers score the same set of films even when using the same method. Figure 1A shows the median and interquartile (IQR) scores of 6 different readers for 210 films, 7 time

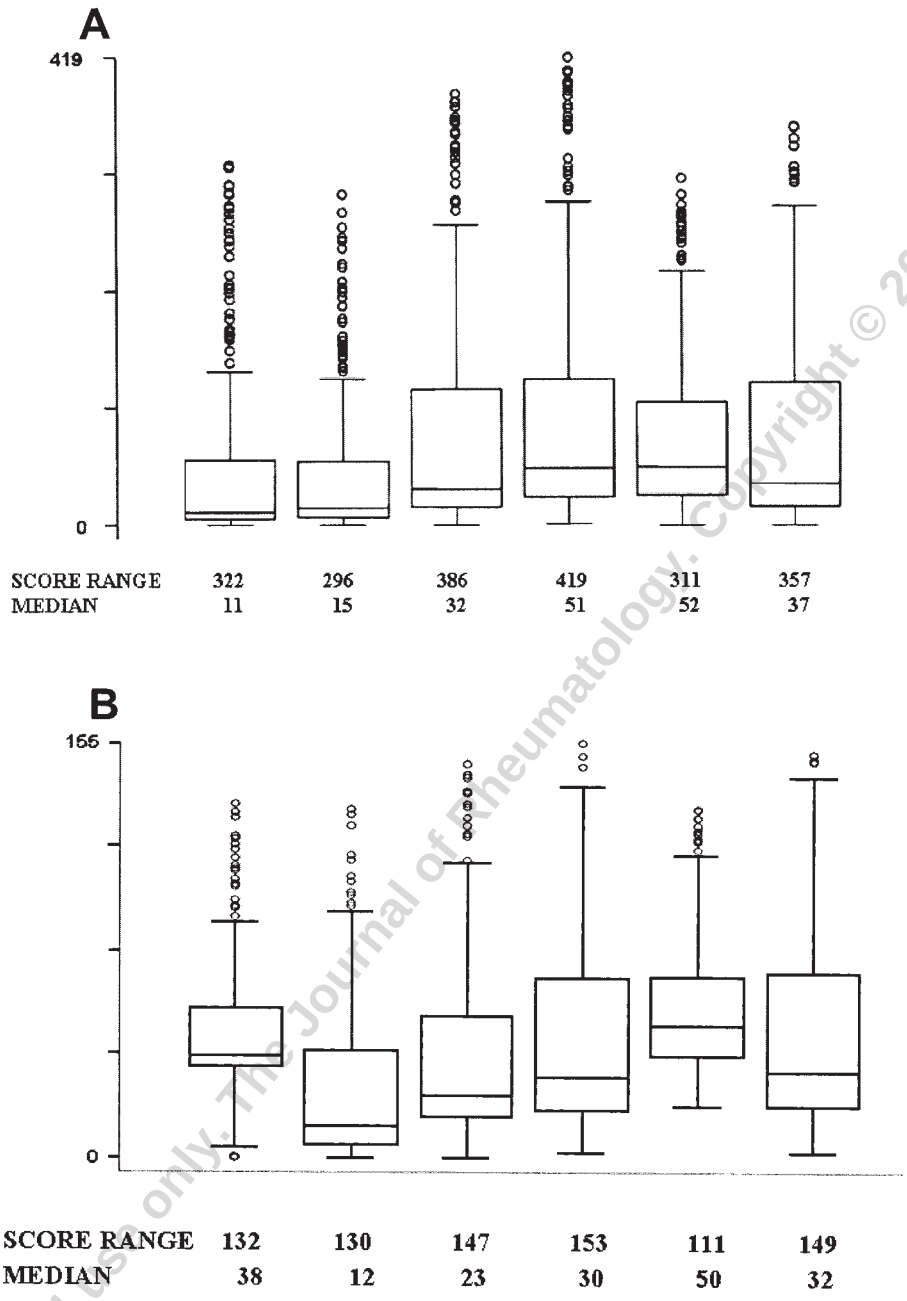

Figure 1. Box plots of 6 readers using the composite method (A) and the global method (B). All readers scored 210 film sets (30 patients with 7 time points each) once by each method. The box extends from 25th to 75th percentile (IQR). The line in the box is the median. Whiskers extend to upper and lower adjacent values, defined as the largest point equal to or less than X[75] + 1.5 × IQR for the upper one and the smallest point equal to or greater than X[25] – 1.5 × IQR for the lower.

points for each of 30 patients. The range of scores used to describe the abnormalities in this set of films extended from 296 to 419. Scores for the same film set recorded by the same readers using the global method, illustrated in Figure 1B, showed similar variation considering that the range of possible scores by this method is roughly half that for the composite method.

Table 2 gives the averages for the SDD, ICC, Pearson r, and the correlation between score differences and score means for intrareader duplicate scores. Correlations are known to be sensitive to the range of values, and this was apparent in the analysis for the ICC. Figure 2 shows all ICC were above 0.92 when the range of score sets was > 120, and were < 0.9 when the range was < 100. Therefore data in

Table 2. Agreement between duplicate scores of same readers.

	n	ICC	SDD	nSDD	r (scores)	r (diffs & means)	Range
9 Readers Using Composite Method—Large-range Score Sets*							
Average	198	0.975	24.4	10.1	0.982	0.233	271
IQR	56–105	0.956–0.992	20.0–28.7	6.9–11.3	0.972–0.989	0.082–0.324	213–308
Minimum	28	0.942	16.4	5.4	0.962	0.028	136
Maximum	726	0.997	35.0	22.5	0.995	0.652	382
SD	245	0.021	6.5	5.4	0.012	0.196	81
Median	105	0.984	20.9	8.3	0.985	0.208	275
6 Readers Using Global Method**							
Average	79	0.986	9.0	8.0	0.982	0.178	110
IQR	28–105	0.984–0.993	7.7–9.7	6.5–8.7	0.974–0.992	0.036–0.189	103–131
Minimum	28	0.957	5.6	4.8	0.949	0.025	65
Maximum	105	0.999	13.6	12.9	0.999	0.712	132
SD	40	0.015	2.7	2.7	0.018	0.268	25
Median	105	0.992	8.7	7.7	0.990	0.054	114
6 Readers Using Composite Method—Small-range Score Sets***							
Average	121	0.648	13.2	NA†	0.690	0.184	46
IQR	36–222	0.601–0.697	12.2–14.2		0.636–0.771	0.134–0.255	27–56
Minimum	34	0.416	10.4		0.536	0.003	26
Maximum	295	0.842	15.8		0.793	0.287	94
SD	130	0.142	1.9		0.100	0.112	27
Median	41	0.676	13.3		0.699	0.228	34
6 Readers with 1 Score Set for Each Method							
Average	277	0.958	NA††	14.6	0.954	0.277	NA
IQR	224–315	0.956–0.965		13.5–16.1	0.940–0.964	0.147–0.428	
Minimum	224	0.934		12.9	0.937	0.064	
Maximum	315	0.966		16.8	0.967	0.557	
Median	290.5	0.952795		13.76508	0.9605	0.14065	

† The nSDD is artifactually elevated because the error term is large in comparison with the small range in scores. †† The SDD cannot be calculated for comparing data obtained using 2 different methods. Range of scores is always lower for the global method as determined by the highest possible scores in the 2 methods. * From Ratingen, Aventis, and ARAMIS datasets. ** From Ratingen dataset. *** From Western Consortium dataset. n: number of total radiographic scores included in the comparison; ICC: intraclass correlation; SDD smallest detectable difference, calculated as 1.96 × SD of the difference between 2 readers' scores for all members of a set; nSDD: smallest detectable difference on normalized scores calculated for each reader as (score–min score)/(max–min scores); r: Pearson correlation coefficient for the 2 scores; rho: Spearman rank correlation for the 2 scores; r for diffs & means: Pearson r calculated for difference between 2 scores and the mean of the 2 scores; Score range (maximum score–minimum score) is for the smaller set of scores. IQR: interquartile range; SD: standard deviation.

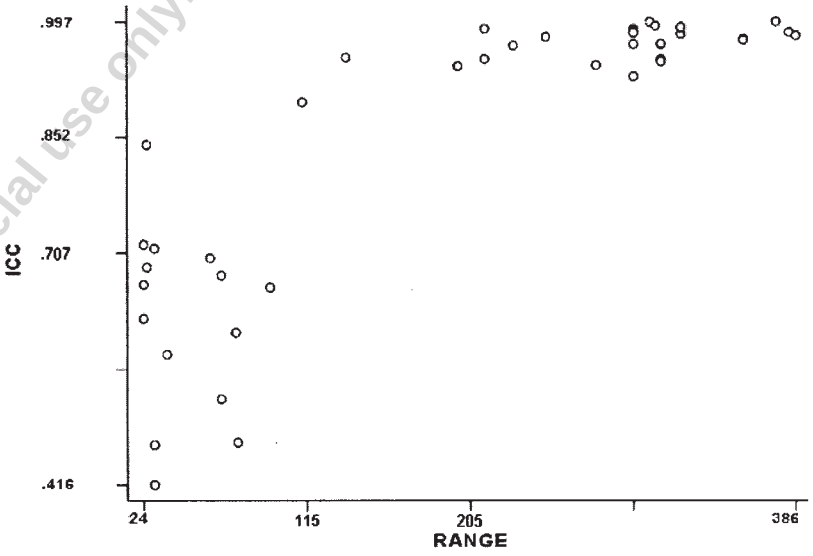

Figure 2. Intraclass correlation plotted against the range of the average of 2 duplicate scores.

Tables 2 and 3 are organized to show the analysis results by large-range scores (range > 100) and small-range scores (range < 100) using the smaller of the 2 ranges in the duplicate scores.

The intrareader comparisons shown in Table 2 had an average SDD for 9 sets of duplicate scores by the composite method in the large range of 24.4, with IQR extending from 20 to 28.7. The average score normalized for each individual reader (nSDD), which expresses the score as a percentage of the score range used, was 10.1. The average SDD for 6 duplicate score sets by the global method was 9, a value similar to that for the composite method considering the approximate 2-fold difference in score range. The ICC for the 2 methods also were similar, 0.975 for composite and 0.986 for global.

As expected, the small-range duplicate scores had an average SDD that was smaller, as well as a lower average ICC. Intrareader limits of agreement calculations for duplicate scores when one set of scores was by each method can only be performed using the normalized scores. The nSDD for 6 such comparisons was 14.6, reflecting less agreement than when a single method was used, whether composite or global.

Interreader agreement comparing the individual scores of 2 readers (see Figure 3) showed more variability in all analyses and all subsets, as illustrated by the average SDD of 53.7 for the large-range duplicate scores using the composite method (Table 3). The ICC of 0.971 is minimally different from 0.975, the ICC for intrareader scores. The SDD for interreader duplicate scores by the global method were also slightly more than twice that of the intrareader SDD.

When the average of 2 readers' scores was compared with averaged scores of a different 2 readers, the SDD was lower than when individual readers' scores were analyzed, and the ICC was higher (SDD 39.6 compared to 53.7, ICC 0.988 compared to 0.971 for the composite method; and SDD 16.6 compared to 23.3, ICC 0.978 compared to 0.951 for the global method).

The Larsen score as originally published did not assess JSN, limiting the scores of 2 through 5 to erosion damage and using the score of 1 to record the presence of soft tissue swelling and/or focal osteoporosis. Table 4 shows a close association between the global scores and both the erosion and JSN scores, with average r values comparing global scores to erosion scores of 0.956 for intrareader and 0.930

Table 3. Agreement between duplicate scores of different readers.

	n	ICC	SDD	nSDD	r (scores)	r (diffs & means)	Range
			17 Duplicate Sets from 8 Readers Using Composite Method Large-range Score Sets*				
Average	255	0.971	53.7	12.9	0.963	0.554	308
IQR	238–266	0.968–0.983	41.6–65.3	11.4–14.4	0.954–0.973	0.417–0.723	296–322
Minimum	134	0.928	36.2	9.1	0.932	0.050	198
Maximum	315	0.990	83.0	17.1	0.983	0.869	386
SD	45	0.018	14.8	2.3	0.015	0.225	42
Median	238	0.976	48.4	13.0	0.969	0.584	311
			15 Duplicate Sets from 6 Readers Using Global Method**				
Average	311	0.951	23.3	14.4	0.956	0.579	128
IQR	315–315	0.940–0.966	18.8–27.9	12.9–14.9	0.948–0.966	0.414–0.8	111–132
Minimum	252	0.925	15.9	10.4	0.927	0.093	111
Maximum	315	0.973	33.6	20.2	0.972	0.856	149
SD	16	0.015	5.6	2.8	0.013	0.234	14
Median	315	0.951	22.2	13.8	0.960	0.620	130
			19 Comparisons from 6 Readers 1 Score Set by Each Method***				
Average	360	0.927	NA†	16.8	0.920	0.317	NA
IQR	238–315	0.890–0.957		15.1–18.9	0.916–0.952	0.165–0.445	
Minimum	217	0.848		11.5	0.796	0.009	
Maximum	1029	0.970		21.8	0.970	0.674	
SD	243	0.043		2.5	0.053	0.200	
Median	266	0.950		16.8	0.942	0.313	
			10 Comparisons from 4 Readers—Small-range Score Sets††				
Average	227	0.636	14.1	NA†	0.696	–0.009	53
IQR	35–283	0.524–0.712	10.8–15.9		0.62–0.758	–0.253–0.281	24–75
Minimum	34	0.465	10.4		0.519	–0.460	24
Maximum	706	0.896	18.9		0.931	0.336	112
SD	241	0.131	2.9		0.113	0.290	31
Median	157	0.646	15.0		0.684	–0.055	49

See footnotes for Table 2 for explanation of column and row headings. † SDD cannot be calculated for score sets using different methods and scales. * Data from Ratingen and ARAMIS datasets. ** Data from Ratingen dataset. *** Data from Ratingen, Wichita, Aventis datasets. †† Data from Western Consortium and COBRA datasets.

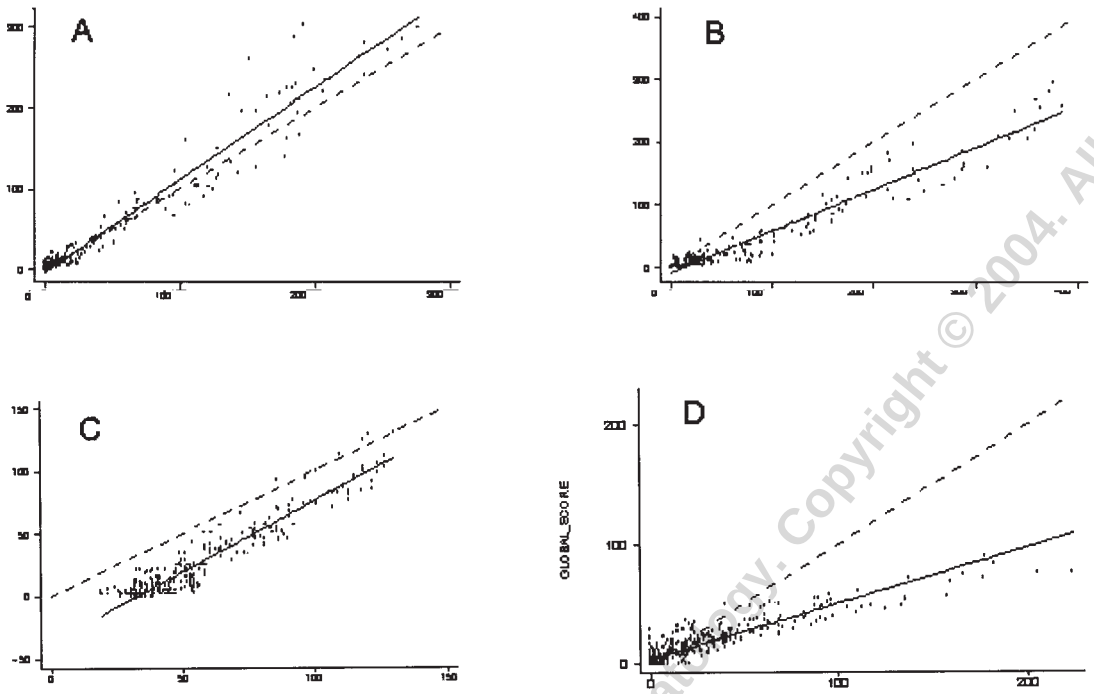

Figure 3. x-y plots of readers' scores. Panels A and B compare different pairs of readers' scores using the composite method. Panel C compares 2 readers' scores using the global method. Panel D shows scores of one reader using the composite method compared with another reader using the global method. Data must overlay the broken line for perfect concordance.

Table 4. Global scores compared to erosion and joint space narrowing (JSN) scores.

	6 Comparisons by Same 6 Readers*				
	n	r	rho	r	rho
	Global vs Erosion Scores			Global vs JSN Scores	
Average	275	0.956	0.846	0.897	0.789
IQR	224–315	0.948–0.962	0.812–0.920	0.895–0.907	0.725–0.854
Minimum	217	0.948	0.686	0.870	0.575
Maximum	315	0.964	0.948	0.908	0.890
Median	291	0.958	0.855	0.902	0.846
	17 Comparisons by 6 Different Readers**				
Average	312	0.930	0.741	0.881	0.746
IQR	238–315	0.934–0.957	0.668–0.829	0.851–0.893	0.676–0.795
Minimum	217	0.793	0.298	0.837	0.571
Maximum	669	0.964	0.944	0.918	0.866
Median	238	0.946	0.768	0.883	0.740

* Data from Ratingen dataset. ** Data from Ratingen, Wichita, and Aventis datasets. r: Pearson correlation coefficient; rho: Spearman rank correlation.

for interreader scores. Association with narrowing scores was less close, averaging 0.881 and 0.897.

Radiographic progression in randomized control trials (RCT) is measured by the change in scores from baseline to followup. Tables 5 and 6 show the SDD was smaller for both intra- and interreader progression scores than for the raw scores by both methods, reflecting the smaller range in the progression scores. The ICC for progression scores also were lower than for raw scores.

It is general practice in recent RCT for 2 readers to score all films and for the average of the 2 scores to be used in the analysis. Progression scores for the average of 2 scores using all combinations of readers compared to different reader combinations in the data set with 6 readers found the SDD was moderately better for the averaged scores than for the single scores (15.8 compared with 21.1).

The results from the Ratingen dataset show that the ICC and SDD for individual scores (SDDi) for interreader (Table 2) and intermethod scoring (Table 3) were only minimally different.

Table 5. Agreement on progression for duplicate scores by same readers.

	n	ICC	SDD	r (scores)	r (diffs & means)	Progression	Range†
			9 Comparisons for 6 Readers Using Composite Method*				
Average	118	0.833	15.0	0.758	0.200	13.2	80
IQR	48–93	0.808–0.969	11.7–18	0.679–0.935	0.038–0.316	7.4–15.1	56–112
Minimum	24	0.440	9.6	0.229	0.006	1.9	33
Maximum	510	0.993	22.4	0.971	0.603	26.6	120
SD	150	0.221	4.6	0.297	0.209	9.0	33
Median	90	0.946	13.4	0.917	0.113	14.3	67
			6 Comparisons for 6 Readers Using Global Method**				
Average	68	0.885	5.3	0.909	0.169	4.5	30
IQR	41–90	0.847–0.913	4.3–5.9	0.891–914	0.073–0.262	3.6–4.9	24–40
Minimum	24	0.816	3.8	0.882	0.045	3.2	12
Maximum	90	0.984	7.4	0.959	0.328	6.6	43
SD	34	0.060	1.3	0.028	0.119	1.3	12
Median	90	0.873	5.3	0.907	0.144	4.3	30

See footnotes for Table 2 for explanation of column and row headings. † Range is the smaller range of the 2 progression scores. * Data from Ratingen, Western Consortium, and Aventis datasets. ** Data from Ratingen dataset.

Table 6. Agreement on progression for duplicate scores by different readers.

	n	ICC	SDD	r (scores)	r (diffs & means)	Progression	Range†
			20 Comparisons for 9 Readers Using Composite Method*				
Average	203	0.736	21.1	0.734	0.259	10.8	87
IQR	191–228	0.695–0.827	19.2–24	0.783–0.841	0.067–0.456	7.9–14.5	70–116
Minimum	90	0.215	9.1	0.039	0.003	2.1	45
Maximum	398	0.898	27.3	0.898	0.640	16.2	148
SD	73	0.178	4.2	0.223	0.218	4.8	28
Median	204	0.812	21.0	0.816	0.201	12.3	80
			15 Comparisons for 6 Readers Using Global Method**				
Average	227	0.658	8.8	0.769	0.311	4.5	38
IQR	204–270	0.594–0.706	8.3–9.2	0.740–0.787	0.132–0.447	4.1–4.9	36–39
Minimum	192	0.485	7.3	0.670	0.034	3.5	36
Maximum	270	0.785	11.0	0.870	0.515	5.5	45
SD	34	0.082	1.0	0.053	0.176	0.6	3
Median	204	0.676	8.8	0.771	0.395	4.3	37

See footnotes for Table 2 for explanation of column and row headings. † Range in this table is the range of progression scores. * Data from Ratingen, Western Consortium, and COBRA datasets. ** Data from Ratingen dataset.

The relationship of difference in scores to mean scores varied. With the exception of the small-range score sets, average correlations between differences and means shown in Tables 2 and 3 were lower for intrareader than for inter-reader comparisons, ranging from 0.178 to 0.277 compared with 0.317 to 0.579.

DISCUSSION

The availability of new radiographic datasets in RA provided us the opportunity to reassess the precision of the composite (Sharp and its modifications) and global (Larsen and its modifications) scoring systems. It also provided an opportunity to evaluate design, analysis, and interpretation issues. Using the limits of agreement analysis, i.e., the SDD, as the primary summary statistic of precision, we found greater variability between readers than reported in previous studies[17]. However, there was lesser variability when the SDD was expressed as a percentage of the range of scores used by the individual reader. These findings have important implications in the interpretation of RCT, both within trials (for power calculations) and particularly for cross-trial comparisons. We confirmed that variability between readers was greater than that within readers, and found that the reader's expertise in scoring radiographs, irrespective of method, may be even more important than the method of scoring applied. We also confirmed that the SDD is subject to clinical heterogeneity. The SDD increases with increasing magnitude of the score in the population under study. This increase may contain systematic error (or bias, as demonstrated by the substantial correlation between difference scores and their mean in many comparisons) as well as random measurement error. Correcting for bias may improve the detection of treatment effects when using a responder-type analysis. In addition, the combination of bias

and random error further challenges the interpretation of cross-trial comparisons. We provide additional evidence that the SDD and therefore the precision in scoring radiographic abnormalities is influenced by multiple factors and therefore is context-specific. We clearly demonstrate that the statistical methods used to report the precision of a measurement can give very different results, and thereby lead to different conclusions. Importantly, we show that the ICC, as a summary measure of precision, can hide significant variation between readers, a variation that only becomes apparent using the SDD as the statistical method of analysis. Finally, we offer the nSDD as another method of overcoming a SDD weakness. This facilitates the comparison of SDD across measurement methods and clinically heterogeneous populations. We recommend that both ICC and limits of agreement summary statistics be reported in all studies evaluating reliability of radiographic scores.

Assessing precision of scoring radiographic abnormalities in RA has been approached by different methods reflecting uncertainty about which analysis is optimal[15,18]. Many studies have compared 2 measurements using correlation methods, although Bland and Altman have discussed the limited value of this approach, which is sensitive to outliers, is influenced by the range of values, and measures association[15,19,20]. They have emphasized the value of limits of agreement analysis, which is based on the standard deviation of the differences between 2 or more measurements of the same objects and expresses variability in units of the measurement. The SDD[14], which is 1.96 times the standard deviation of the differences, is the smallest number of units in individual serial measurements that represent a true difference with 95% confidence. The SDD is specific to the observer(s) and the set of objects, in this case the individuals scoring a set of films and the film set being scored[21].

Most previous reports have not included limits of agreement, but some have published sufficient data to calculate this statistic, as reported by Lassere, *et al*[14,17]. In 13 interobserver comparisons of raw data using the composite method, the mean SDDi was 21.8 with an IQR of 15–24 (Table 7)[22-25]. Intraobserver comparisons for the composite method were fewer in number, with 9. Their average SDDi was 20.7 with an IQR of 16–26. There were fewer comparisons using the global method, with 3 each for inter- and intraobserver comparisons, in which the SDDi varied from 7 to 25 for intraobserver and 10.7 to 19.6 for interobserver scores. There are even fewer published limits of agreement data on progression scores. Lassere, *et al* found the SDDi to be 11 for global scores, 12.3 and 15.3 for composite scores[14]. In a later publication, Lassere, *et al* compared the limits of agreement of radiological progression using the composite method across 5 different datasets[21]. The SDD was study-specific and varied from 4.7 to 15.5, the variation depending on the patient population, reader calibration, and the method of analysis.

The data presented here have revealed greater variability between readers, as illustrated in the box plots (Figures 1 and 2) and the SDDi shown in Tables 2 and 3, than in previous studies. In this study there was a greater difference between the composite and the global methods. Differences between the 2 methods seen in this study are largely accounted for by the difference in range of possible scores, which is roughly 2-fold. There are no established standards for an acceptable level of variability. Even though this study reveals a greater variability than previously shown, this extent of variability does not preclude effective use of scoring in RCT. Results from 2 trials included in the analysis showed highly significant differences between treatment groups, suggesting that other factors, primarily the effectiveness of the drug and sample size, are of greater importance in designing trials.

One proposed use of the SDD is in dichotomizing data. For example, a change in radiographic score greater than the SDD can be used to segregate patients in a RCT into "progressors" and "non-progressors," and the possibility that a negative progression score exceeding the SDD can represent healing or repair is under investigation[14,17]. Another use, less frequently commented on, is to determine the number of subjects required in a study to observe a specific difference between groups at a given confidence level. The SDD for individual scores in further discussion here will be referred to as the SDDi and for groups, the SDDg. The SDDi represents possible error in individual scores, which in an RCT with adequate randomization should be approximately evenly distributed within and between groups. The power calculation for an RCT with a given number of subjects per treatment group is usually calculated from the ratio of the standard deviation of progression scores in the control (or comparison) group to the progression scores expected in the treatment group. It is also possible to calculate the number of patients required in a study using the SDDi. Dividing the SDDi by the square root of the number of subjects proposed to be included will determine the SDDg, which is the smallest difference expected between groups with 95% confidence. This is a valid post hoc analysis. Using it to predict outcome in a prospective trial assumes that the control group in the proposed study will have a similar rate and distribution of progression scores, which is also an assumption in the more usually employed power calculation. From Tables 5 and 6 it can be appreciated that the choice of readers and probably the choice of pairs of readers can affect the number of patients required to demonstrate a specific difference in radiographic outcome by 2- to 3-fold.

We calculated a nSDD to allow comparisons across the different radiographic scoring methods. Three different methods for standardizing the SDD have been reported in the literature, and we offer a fourth. If the raw, individual scores of different measures are available for different patients, they can be changed to a common metric, and the

Table 7. Agreement on radiographic scoring reported in the literature.

	n*	Raw Scores Mean SDD	IQR	Minimum	Maximum
Intraobserver, composite method	9	20.7	16–26	11.4	34
Intraobserver, global method	3	13.1		7.0	25.0
Interobserver, composite method	13	21.8	15–24	10.4	42.3
Interobserver, global method	3	15.7		10.7	19.6
		Progression			
	n	SDD			
Composite method	1	10.6			
Global method	1	12.3			

* Number of separate studies reviewed and summarized. Data are from Guth[24] (quoted by Lassere), Lassere[14], O'Sullivan[22], Ruckman[26], Sharp[27].

SDD then can be calculated and compared[9]. If the raw individual scores of different measures are available for the same patients, a regression equation converts the measures to a common metric, and the SDD can be calculated and compared[9]. If the raw scores are not available, the reported SDD can be compared by calculating the SDD as a percentage of the actual score range, provided the minimum and maximum values are also reported[17,21]. In this discussion we offer a nSDD, which is similar to the percentage SDD, but is calculated using the raw individual scores. Although the 4 methods have not been directly compared, the conclusions drawn from their use here and in previous publications are comparable.

Many more of the SDDi than previously appreciated were significantly related to magnitude of the score, as shown by the correlation between the differences in scores and their means (Tables 2, 3, 5, and 6), and was closer for interreader (IQR 0.414 to 0.8) than intrareader comparisons (IQR 0.08 to 0.32). This relationship was less marked in the progression scores, which is a reflection of their lower range. Bland and Altman state that when there is a relationship between the differences and the mean scores, it can sometimes be ignored, but it may be better to try and remove this relationship by log-transformation of the data or more general measures such as curve fitting, but they do not express an opinion about what level of relationship should be considered to require adjustment[15]. In those instances in which the relationship is close, adjusting for the magnitude of the score would result in a smaller SDDi for low raw and progression scores and a higher SDDi for high scores than is shown in the tables for unadjusted SDDi. If a clinical trial were undertaken in which the control group had relatively little progression, provided the progression was fairly uniformly distributed among subjects, correcting for the relationship between error and magnitude of progression scores would assure a greater sensitivity in detecting a difference between treatment groups.

The much wider variation in the SDDi than observed in the ICC confirms that information provided by correlation methods and limits of agreement analysis is qualitatively different. The correlation methods reveal that experienced readers appreciate differences in severity. The uniformly high ICC seen for film sets having a broad range of scores that also demonstrate considerable variability in the SDDi indicate that correlations are not sensitive to disagreement between scores of individual films.

The inclusion of the Ratingen dataset raises important questions about study design. Six readers, all of whom had previous experience scoring films by one method, were asked to score by both methods, one they were familiar with and another with which they had no or very limited experience. Three readers had regularly used the composite method and 2 the global method. One declared that he had limited experience with both methods. The 3 with the most experience, 2 with the composite method and one with the global, had never read films by the other method. These readers stated they scored films by their usual method and transformed scores based on their concept of the relationship of their accustomed method to the opposite one. For example, a reader accustomed to using the composite scale determined what the erosion score would be by that usual method, and then assigned a global score by transforming this score from a 1 to 5 scale to a 2 to 5 scale. Examination of the data reveals a slightly lower ICC and higher SDDi when the same readers scored by 2 different methods (Table 3) in contrast to the same readers scoring by the same methods (Table 2). However, the differences are minimal, which is interpreted as indicating the most important factor is the experience and consistency of the individual reader.

The observed variability between readers is clear proof

that absolute scores between different readers and between different studies cannot be compared. Comparison between studies would be possible if all future RCT and longterm followup studies incorporated a standard set of films interleaved with study films in a blinded fashion and read along with the study cohort. This would permit calculation of a standardized score unit of joint damage that would provide a reasonable basis for comparisons, probably the best we can achieve until truly quantitative measurements are widely used. Limits of agreement between readers should be reported in all studies using radiographic damage scores.

In summary, the precision and sources of variability of scoring radiographic abnormalities in patients with RA were evaluated using intra- and interreader duplicate scores of individual readers and found to be extensive but similar for the composite (Sharp, Sharp/van der Heijde) and global (Larsen) methods. Limits of agreement and correlation statistics provide different information when comparing duplicate scores by the same or by different readers. The SDDi is frequently related to the magnitude of the score, and may require correction when large, if using a SDD-responder-based analysis of treatment effects.

ACKNOWLEDGMENT

Individual readers were Annalies Boonen, Richard Gold, Desiree van der Heijde, Gertraud Herborn, Arvi Larsen, Don Mitchell, John T. Sharp, Arco C. Verhoeven, Seigfreid Wassenberg, Maths Wijnands, and Fred Wolfe. The authors gratefully acknowledge those who shared data for this analysis: the COBRA Study Group, The Netherlands; Frederick Wolfe, MD, the Wichita Arthritis Center, Wichita, Kansas; Harold Paulus, MD, Professor of Medicine, Director, The Western Consortium of Practicing Rheumatologists, The University of California at Los Angeles, California; the Department of Rheumatology, Evangelisches Fachkrankenhaus Ratingen, Ratingen, Germany; and Karen Simpson, The Aventis Corporation.

REFERENCES

1. Sharp J, Lidsky MD, Collins LC, Moreland J. Methods of scoring the progression of radiologic changes in rheumatoid arthritis. Correlation of radiologic, clinical and laboratory abnormalities. Arthritis Rheum 1971;14:706-20.
2. Larsen A. A radiological method for grading the severity of rheumatoid arthritis [thesis]. Helsinki: University of Helsinki; 1974:133.
3. Sharp J, Young DY, Bluhm GB, et al. How many joints in the hand and wrist need to be included in a score of radiologic abnormalities? Arthritis Rheum 1985;28:1326-35.
4. van der Heijde D, 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.
5. Larsen A. How to apply Larsen score in evaluating radiographs of rheumatoid arthritis in longterm studies. J Rheumatol 1995;22:1974-5.
6. Scott DL, Houssien DA, Laasonen L. Proposed modification to Larsen's scoring methods for hand and wrist radiographs. Br J Rheumatol 1995;34:56.
7. Genant H. Methods of assessing radiographic change in rheumatoid arthritis. Am J Med 1983;75:35-47.
8. Rau R, Wassenberg S, Herborn G, Stucki G, Gebler A. A new method of scoring radiographic change in rheumatoid arthritis. J Rheumatol 1998;25:2094-107.
9. Lassere M. Pooled meta-analysis of radiographic progression: comparison of Sharp and Larsen methods. J Rheumatol 2000;27:269-75.
10. Lassere M, Edmonds J. The evaluation of damage. The gold standard — the X-ray. Rheumatoid arthritis and osteoarthritis scoring reliability. APLAR J Rheumatol 1998;2:123-8.
11. Sharp J, Strand V, Leung H, Hurley F, Loew-Friedrich I. Treatment with leflunomide slows radiographic progression of rheumatoid arthritis. Arthritis Rheum 2000;43:495-505.
12. Sharp J, 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.
13. Boers M, Verhoeven AC, Markusse HM, et al. Randomized comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet 1997;350:309-18. [Erratum appears in Lancet 1998;351:220].
14. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731-9.
15. Bland J, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
16. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
17. Lassere M, van der Heijde D, Johnson K, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: Implications for smallest detectable difference, minimum clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol 2001;28:892-903.
18. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring quantitative variable. Comput Biol Med 1989;19:61-70.
19. Bland J, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990;20:337-40.
20. Bland J, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-60.
21. Lassere M, van der Heijde D, Johnson K, et al. Robustness and generalisability of smallest detectable difference in radiological progression. J Rheumatol 2001;28:911-3.
22. O'Sullivan M, 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.
23. Sharp J, Wolfe F, Corbett M, et al. Radiologic progression in rheumatoid arthritis: How many patients are required in a therapeutic trial to test disease modification? Ann Rheum Dis 1993;52:332-7.
24. 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.
25. Nance EJ, Kaye JJ, Callahan LF, et al. Observer variation in quantitative assessment of rheumatoid arthritis: Part I. Scoring erosions and joint space narrowing. Invest Radiol 1986;21:922-7.
26. Ruckmann A, Ehle B, Trampisch H-J. How to evaluate measuring methods in the case of non-defined external validity. J Rheumatol 1995;22;1998-2000.
27. Sharp J, Bluhm GB, Brook A, et al. Reproducibility of scoring radiologic abnormalities in the hands and wrists of patients with rheumatoid arthritis by multiple observers. Arthritis Rheum 1985;28:16-24.