Reproducibility and Sensitivity to Change of 5 Methods for Scoring Hand Radiographic Damage in Patients with Rheumatoid Arthritis

FRANCIS GUILLEMIN, LAURENT BILLOT, STEPHANIE BOINI, NATHALIE GERARD, SIGRID ØDEGAARD, and TORE K. KVIEN

ABSTRACT. ***Objective***. To compare intrarater and interrater reproducibility and sensitivity to change of 5 scoring methods for radiographic damage on hand radiographs in patients with rheumatoid arthritis (RA).

Methods. Radiographs of 22 patients from Norway and France with average 2 years' disease duration at baseline and mean 30 months' followup were assessed by 2 readers according to Larsen, Larsen/Rau, Sharp, Sharp/van der Heijde, and Simple Erosion Narrowing Score (SENS) methods. Reproducibility at baseline and on progression was assessed using intraclass correlation coefficients (ICC) and Bland-Altman graphs. Sensitivity to change was compared across methods by computing the country-adjusted standardized response means (SRM) ratio.

Results. Intrarater reproducibility varied with the reader (ICC ranging from 0.90 to 0.97), with Larsen and Larsen/Rau ranking highest. Interrater reproducibility was highest with Sharp and Sharp/van der Heijde (ICC 0.76 to 0.93). Bland-Altman graphs showed a decrease of concordance in cases of more severe damage. Sensitivity to change was higher with Sharp and Sharp/van der Heijde modified for erosions (SRM ratio 1.44 and 1.70), than with Larsen/Rau and SENS. The differences between Sharp, Sharp/van der Heijde, and Larsen were less for joint space narrowing. There was a significant reader effect ($p < 0.05$) in all but the Sharp method. Expressed as percentage of the maximum score, the smallest detectable difference varied between 3.5% (Sharp/van der Heijde) and 14.2% (SENS erosion).

Conclusion. All methods have high intraobserver and interobserver reliability. The interrater reproducibility decreases with disease severity. Recent modified methods perform best to detect changes, but the advantages of SENS seemed to be lost when applied on hand radiographs alone. Training the readers appears to be essential. (J Rheumatol 2005;32:778–86)

Key Indexing Terms:
RHEUMATOID ARTHRITIS RADIOGRAPHS HAND SCORE

The assessment of radiographic joint damage is a major criterion for characterizing the severity and progression of the disease in rheumatoid arthritis (RA). It reflects impairment and is recommended as a main outcome variable in controlled clinical trials of disease modifying therapies[1] as well as in longitudinal observational studies[2].

From the School of Public Health, Nancy, France; and the Department of Rheumatology, Diakonhjemmet Hospital, Oslo, Norway.

Supported by a grant from Programme Hospitalier de Recherche Clinique of the French Ministry of Health, 1995, and from EC-COMAC.

*The EURIDISS study was supported in France by a grant from the Programme Hospitalier de Recherche Clinique, 1995, from the Ministry of Health. Internationally, the study was supported by the COMAC-Health Services Research, contract MR4*0344-NL, and DG contract BMH4-CT96-1580, from the European Community.*

F. Guillemin, MD, PhD, Professor of Epidemiology and Public Health; L. Billot, MSc, Statistician; S. Boini, MSc, Fellow; N. Gerard, MD, Rheumatologist, EA 3444, School of Public Health, Nancy; S. Ødegaard, MD, Rheumatologist; T.K. Kvien, MD, Professor of Rheumatology, Department of Rheumatology, Diakonhjemmet Hospital, Oslo.

Address reprint requests to Prof. F. Guillemin, Ecole de santé publique, Faculté de médecine BP 184, 54505 Vandoeuvre-les-Nancy, France. E-mail: francis.guillemin@sante-pub.u-nancy.fr

Accepted for publication November 22, 2004.

Standardization of joint damage assessment by reading radiographs as an outcome has inspired many efforts to improve quality of reading and scoring damage. It is crucial to develop a common tool for assessing this fundamental issue in RA so as to minimize the measurement error at its numerous sources. Several radiograph scoring methods have been proposed to standardize quantification of articular damage, in order to facilitate longitudinal observations and prognostic and therapeutic studies.

There is a recent trend toward favoring shortened methods at the expense of minimum loss of quality, but the measurement properties of a method depend on a number of factors, such as radiograph techniques, positioning, number of articular sites counted, grading systems, type of lesion considered, i.e., joint space narrowing (JSN), erosion or deformation, and skill of readers[3]. A recent review has tracked the various methods proposed and the degree of validity of each method[4]. Few methods have been directly compared for performance when used under similar conditions, although it is of interest to provide a comprehensive comparison of cross-sectional and longitudinal measurement properties of different methods in RA.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on October 20, 2021 from www.jrheum.org

We compared intrarater and interrater reproducibility and sensitivity to change of 5 scoring methods for radiographic damage focusing on hand radiographs in RA.

MATERIALS AND METHODS

Design. A trial examining the cross-sectional intrareader and interreader reproducibility, as well as the longitudinal interreader reproducibility and sensitivity to change, was designed in a successive-block repeated-measurement design.

Materials. A sample of hand radiographs from 22 patients — 10 in Norway and 12 in France — was selected from the EURIDISS database[5,6] by 2 senior rheumatologists (TKK, FG) to represent the spectrum of the disease damage in a range of 0 to 4 years' disease course of evolution on the basis of their disease severity, disability, and joint damage[7]. Hand radiographs were postero-anterior views of both hands and wrists on the same film. All patients had a clinical diagnosis of RA and satisfied the 1988 American College of Rheumatology (ACR) criteria for RA classification.

For each patient baseline and followup radiographs after 2–3 years were available for assessment.

Radiograph scoring. Five methods for scoring radiographic damage were compared on each set of radiographs. In this study, only joints in the hands and wrists were examined, even if the majority of the methods also included feet in their original description.

The Sharp method (1985) applies to 17 areas for erosions and 18 areas for JSN in each hand and wrist[6,8,9]. Each erosion scores 1 point, with a maximum of 5 points per area (reflecting loss of more than 50% of articular bone). Erosion scores range from 0 to 170. One point is scored for focal joint narrowing, 2 points for diffuse narrowing of less than 50% of the original space, and 3 points if the reduction is more than half of the original joint space. Ankylosis is scored 4. (Sub)luxation is not scored. The score for JSN ranges from 0 to 144.

The Sharp modified method was developed by van der Heijde in 1989[10,11]. Erosion is assessed in 16 joints for each hand and wrist. One point is scored if erosions are discrete, rising to 2, 3, 4, or 5 depending on the surface involved (complete collapse of the bone is scored 5). The hand score for erosion ranges from 0 to 160. JSN is assessed in 15 joints for each hand and wrist. JSN is combined with a score for (sub)luxation and scored as follows: 0 = normal; 1 = focal or doubtful; 2 = generalized, less than 50% of the original joint space; 3 = generalized, more than 50% of the original joint space or subluxation; 4 = bony ankylosis or complete luxation. The hand score for JSN ranges from 0 to 120.

The Simple Erosion Narrowing Score (SENS) was recently developed by van der Heijde (1999) and is a simplified method by simply summing the number of eroded and narrowed joints on selected joints on hand and foot radiographs[12]. As specified, only hand radiographs were examined in this study. SENS assesses the same joints as the Sharp/van der Heijde method (1989). A joint is scored as "affected 1" if it displays any erosion, and as affected 1 for JSN if it scored 1 or more in the original method (at least focal JSN). The hand score per joint can therefore range from 0 to 2. Erosion is considered in 32 joints and JSN in 30 joints. The hand score ranges from 0 to 32 and from 0 to 30 for erosion and for JSN, respectively.

The Larsen original method has been modified several times by the author. The method recommended for longitudinal observation studies[13] was used here. The main differences from the original are deletion of scores for the thumbs and first metatarsophalangeal (MTP); subdivision of the wrist into 4; deletion of soft tissue swelling and osteoporosis; and distinction between erosions of different sizes. The grading scale ranges from 0 to 5: 0 = intact bony outlines and normal joint space; 1 = erosion < 1 mm in diameter or JSN; 2 = one or several small erosions (diameter > 1 mm); 3 = marked erosions; 4 = severe erosions (usually no joint space left and the original bony outlines are only partly preserved); and 5 = mutilating changes (the original bony outlines have been destroyed). The hand score ranges from 0 to 120.

The modified Larsen by Rau (1995) is restricted to definite erosions and the proportion of joint surface destruction[14]. Twenty-two joints are evaluated in the hands and wrists. The 6 stages are defined as follows: 0 = normal; 1 = soft tissue swelling and/or JSN/subchondral osteoporosis; 2 = erosions with destruction of the joint surface (DJS) < 25%; 3 = DJS 26–50%; 4 = DJS 51–75%; and 5 = DJS > 75%. The hand score ranges from 0 to 110. In this modification, the stages are described as a quantitative measure of the destroyed joint surface area and can therefore be applied more easily.

Reading strategy. Two readers (NG, SØ) performed all radiographic assessments. All 5 scoring methods were used to score all patients' radiographs at each baseline and followup time by both readers. Before the beginning of the study, one reader (NG) was familiar with the modified Sharp method, while the other (SØ) was experienced with the Larsen scoring. Both performed prestudy training for each of the 4 other methods on 10 other radiograph sets.

Interobserver reliability was assessed by having each of 20 radiographs (10 from Norway, 10 out of 12 from France) read at baseline by the 2 readers independently using each of the 5 scoring methods.

Intraobserver reliability was assessed by having each reader assess 10 baseline radiographs twice in random order by each of the 5 methods.

Longitudinal interobserver reliability and sensitivity to change were assessed by having each reader assess each 20-patient set at baseline and at followup. The longitudinal assessment was conducted in sequential order, according to OMERACT recommendations, with readers having baseline radiograph and score available when examining and scoring followup radiographs for each method, allowing for reduction in the score variance[15].

The whole set of radiographs was read 5 times using each scoring method in turn independently following a predefined order: Larsen, modified Larsen, Sharp, modified Sharp, and SENS.

Statistical analysis. Reproducibility was assessed by computing intraclass correlation coefficients (ICC) with their 95% confidence intervals. ICC were derived from a mixed 2-factor analysis of variance for interobserver and intraobserver reproducibility, cross-sectional and longitudinal (see Appendix for details). This information was completed by plotting difference in scores of each radiographic assessment against mean score according to the graphical method for assessing the degree of agreement by Bland and Altman[16]. The sensitivity to change of each method was assessed with the standardized response mean (SRM). A crude SRM was calculated as the ratio of the mean difference between baseline and followup score divided by the standard deviation of this difference. Since we were more likely to detect larger changes with longer followup, we also computed an adjusted SRM from a mixed model of repeated analysis of variance (ANOVA), including a fixed-time effect to control for heterogeneity in the followup duration between patients (see Appendix). Sensitivity to change was compared across methods by computing the ratio of standardized response means, using the Larsen score as reference. A ratio above (below) 1 indicated a sensitivity to change higher (lower) than the Larsen score.

To decide whether there was real progression or no progression at all, we calculated the smallest detectable difference (SDD) for each scoring method using the formula defined by one-side testing for a 95% confidence interval: $SDD = \sqrt{}*1.645*SEM$ applied to paired reading (the same patient on 2 occasions). The SEM is the standard error of measurement defined as

$$SEM = \sqrt{\qquad\qquad}$$

where score 1 corresponds to the score at first reading, and score 2 to the score at second reading, and n corresponds to the 20 patients used to calculate intrarater reliability. Progression scores smaller than the SDD cannot be distinguished reliably from measurement error. Then SDD was expressed as the percentage of the maximum score for each scoring method.

Feasibility. The feasibility of each scoring method was documented by subjective appreciation of the readers on the difficulty to separate out the low severity levels (stage 0 or 1 from stage 1 or 2), and by the time of reading.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on October 20, 2021 from www.jrheum.org

RESULTS

The 22 patients (77.3% female) were 59.7 years old with an average disease duration of 3 years at baseline. The average followup time was 2.5 years. The mean score observed on baseline radiographs was 12.6 (reader 1) and 17.9 (reader 2) by the Larsen method (on a 0–120 scale), 12.0 (reader 1) and 17.9 (reader 2) by the Larsen/Rau method (0–110 scale), 27.1 (reader 1) and 29.2 (reader 2) by the Sharp method (0–312 scale), 24.8 (reader 1) and 26.1 (reader 2) by the variant Sharp method (0–280 scale), and 13.5 (reader 1) and 16.0 (reader 2) by the SENS method (0–62 scale; Table 1).

The cross-sectional interrater reliability was high, but did not show significant differences (Table 1). It was not uniform across methods. In more severe lesions, one rater provided higher erosion scores and lower JSN, with erosion predominant in Larsen and variant Rau method, as shown on Bland-Altman plots (Figures 1A to 1H).

The intrarater reliability was high by all methods, with ICC over 0.9. On average, the score at the second reading was lower than at the first reading (Table 2).

The longitudinal interrater reliability, i.e., the reproducibility of change assessment between raters, was high by ICC (Table 3). The graphical method showed that one rater gave higher erosion scores, particularly in the SENS erosion scale. For the Larsen method the direction (sign) of the difference between raters changed when severity increased (Figures 2A to 2H).

The sensitivity to change of methods compared by the adjusted SRM showed more ability of the Sharp and variant van der Heijde methods to detect changes, particularly in erosions, while change in JSN did not differ by method (Table 3). The SENS did not perform as well as the Larsen and other methods.

The SDD ranged from 3.9 in the SENS JSN score to 12.2 in the Sharp total score (Table 4). The Larsen and Larsen/Rau methods had similar SDD, about 5. All mean change scores (Table 3) were greater than SDD except for the SENS erosion score. Expressed as a percentage of the maximum score, SDD varied between 3.5% (Sharp/van der Heijde total score) and 14.2% (SENS erosion score).

The time for reading was appreciated concordantly by the 2 raters: the SENS scoring was faster, over the Sharp/van der Heijde method, while other methods took longer time. Readers had more difficulty clearly delineating between stages 0 and 1 by the Larsen method.

DISCUSSION

In this comparative survey of the properties of 5 scoring methods for hand radiographs in early RA, all methods showed high performance in both reliability and sensitivity to change comparatively.

A measure needs to have a high level of reliability to allow good sensitivity to change[17], as a prerequisite for optimizing signal to noise ratio. The high level of reproducibility of all methods allows consideration of all methods appropriate for longitudinal assessment of change.

Trained raters do perform better in the method they are trained for. Fries and others have long advocated for trained raters to improve the standard of quality of scoring[18]. Our results confirm such advice, and add the complementary information that training is not interchangeable, since training for one method does not confer universal skill for reading radiographs. Multiple raters have also been suggested as a means to improving reliability, but this does not preclude the influence of training, as untrained raters might fail and increase heterogeneity[19].

This study has some limitations. The number of radiographs was limited, so that the precision of estimates is moderate. However, the confidence intervals of the ICC obtained from a mixed model ANOVA were rather narrow. The site of radiographs was limited to hand radiographs and did not include foot radiographs. This is a strategy of the parent EURIDISS study, intended as a cost-saving in following cohorts of patients in longitudinal studies with repeated measurements. The complete SENS method has shown higher qualities when assessed on complete sets of

Table 1. Description of radiograph scores and interrater reliability by each scoring method (n = 22).

Method	Range	Reader 1		Reader 2		ICC	95% CI
		Mean (SD)	Median (Q1–Q3)*	Mean (SD)	Median (Q1–Q3)*		
Larsen	0–120	12.6 (15.0)	8 (3–17)	17.9 (14.4)	13.5 (8–24)	0.88	0.76–0.94
Larsen/Rau	0–110	12.0 (13.7)	9 (4–11)	17.9 (14.8)	13.5 (6–24)	0.88	0.76–0.94
Sharp	0–312	27.1 (30.8)	20 (5–39)	29.2 (30.1)	17.5 (12–34)	0.95	0.89–0.97
Erosion	0–170	12.6 (18.7)	5 (1–16)	17.3 (17.7)	10.5 (6–20)	0.93	0.86–0.97
JSN	0–142	14.5 (13.6)	12.5 (2–20)	11.9 (13.2)	7.5 (3–16)	0.92	0.84–0.96
Sharp/van der Heijde	0–280	24.8 (28.1)	16.5 (5–33)	26.1 (27.6)	13.5 (11–31)	0.93	0.86–0.97
Erosion	0–160	11.8 (17.1)	5 (2–24)	15.7 (17.3)	9 (5–19)	0.93	0.85–0.97
JSN	0–120	13.0 (12.3)	13 (2–19)	10.4 (11.4)	7.5 (3–13)	0.91	0.82–0.96
SENS	0–62	13.5 (14.1)	10 (3–17)	16.0 (13.1)	10.5 (7–16)	0.89	0.79–0.95
Erosion	0–32	4.2 (7.2)	1 (0–5)	10.3 (7.9)	8 (4–14)	0.80	0.63–0.90
JSN	0–30	9.3 (8.1)	7 (2–13)	5.8 (5.9)	4 (2–7)	0.77	0.57–0.88

* 1st quartile–3rd quartile. ICC: Intraclass correlation coefficient, JSN: joint space narrowing, SENS: Simple Erosion Narrowing Score.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on October 20, 2021 from www.jrheum.org

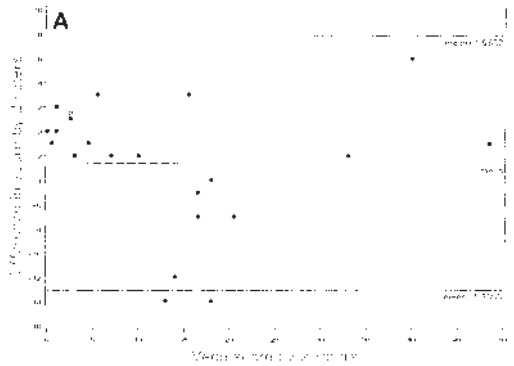

Figure 1a. Reliability at baseline (Sharp method: JSN). Difference against mean by 2 raters.

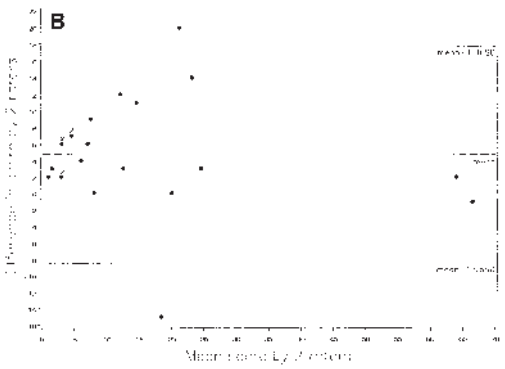

Figure 1b. Reliability at baseline (Sharp method: erosion). Difference against mean by 2 raters.

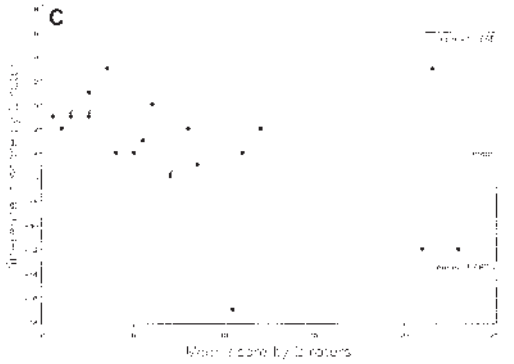

Figure 1c. Reliability at baseline (SENS method: JSN). Difference against mean by 2 raters.

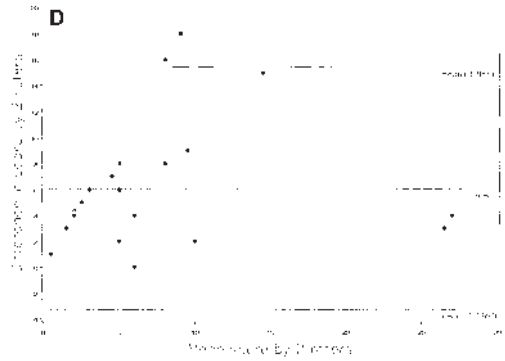

Figure 1d. Reliability at baseline (SENS method: erosion). Difference against mean by 2 raters.

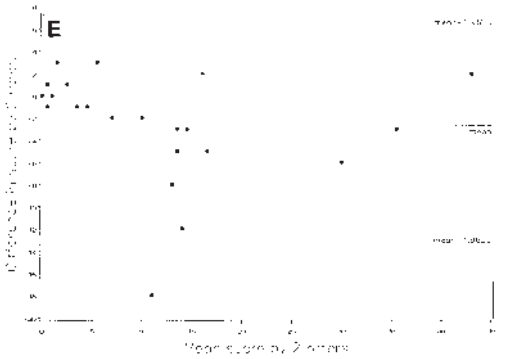

Figure 1e. Reliability at baseline (Sharp/van der Heijde method: JSN). Difference against mean by 2 raters.

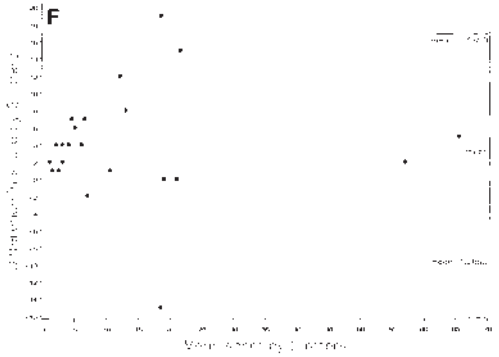

Figure 1f. Reliability at baseline (Sharp/van der Heijde method: erosion). Difference against mean by 2 raters.

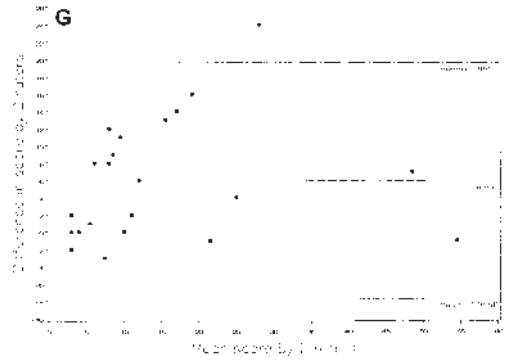

Figure 1g. Reliability at baseline (Larsen/Rau method). Difference against mean by 2 raters.

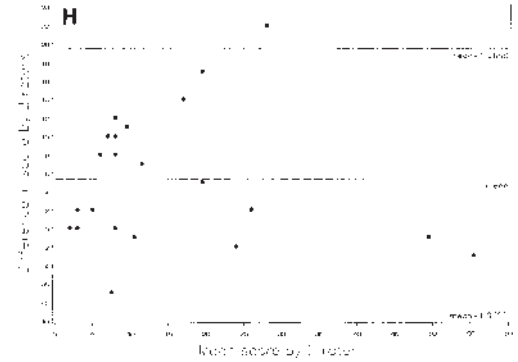

Figure 1h. Reliability at baseline (Larsen method). Difference against mean by 2 raters.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on October 20, 2021 from www.jrheum.org

Table 2. Interrater reliability of scoring method (n = 20).

Method		Reader 1		Reader 2		ICC	95% CI
		1st Reading	2nd Reading	1st Reading	2nd Reading		
Larsen	Mean (SD)	10.8 (16.6)	10.4 (15.3)	13.0 (8.8)	11.5 (8.2)	0.97	0.93–0.99
	Median (Q1–Q3)*	4 (3–11)	5 (3–11)	12.5 (4–22)	9.5 (4–16)		
Larsen/Rau	Mean (SD)	10.6 (16.0)	10.4 (15.0)	12.6 (8.7)	11.1 (8.0)	0.97	0.93–0.98
	Median (Q1–Q3)*	4.5 (3–10)	5 (3–11)	12.5 (4–21)	9.5 (4–16)		
Sharp	Mean (SD)	23.8 (33.2)	24.6 (33.0)	21.2 (17.6)	16.6 (10.4)	0.96	0.91–0.98
	Median (Q1–Q3)*	14 (5–24)	17.5 (7–22)	15.5 (9–25)	16 (7–23)		
Erosion	Mean (SD)	10.7 (19.3)	11.1 (19.8)	11.4 (9.4)	9.6 (6.8)	0.96	0.92–0.98
	Median (Q1–Q3)*	3.5 (1–9)	3.5 (1–8)	8 (4–14)	10 (3–14)		
JSN	Mean (SD)	13.1 (15.0)	13.5 (14.2)	9.8 (9.3)	7.0 (4.6)	0.91	0.81–0.96
	Median (Q1–Q3)*	8 (2–20)	9.5 (5–19)	6.5 (3–13)	7 (2–11)		
Sharp/van der Heijde	Mean (SD)	21.5 (28.7)	22.0 (27.3)	18.0 (15.1)	14.9 (9.6)	0.96	0.92–0.98
	Median (Q1–Q3)*	12 (6–22)	16 (7–20)	12.5 (9–20)	12.5 (7–20)		
Erosion	Mean (SD)	9.9 (17.2)	10.0 (17.0)	9.9 (8.7)	8.9 (6.5)	0.96	0.92–0.98
	Median (Q1–Q3)*	2.5 (1–9)	2.5 (2–8)	7 (4–11)	8.5 (3–12)		
JSN	Mean (SD)	11.6 (12.8)	12.0 (11.5)	8.1 (8.0)	6.0 (4.2)	0.90	0.79–0.95
	Median (Q1–Q3)*	8 (2–16)	9.5 (5–17)	6.5 (2–11)	7 (2–9)		
SENS	Mean (SD)	11.4 (12.7)	10.2 (12.1)	12.8 (9.0)	10.1 (6.0)	0.94	0.87–0.97
	Median (Q1–Q3)*	8.5 (4–14)	7.5 (5–10)	10 (7–13)	9.5 (5–14)		
Erosion	Mean (SD)	4.6 (7.4)	4.8 (7.3)	7.6 (5.5)	6.2 (4.5)	0.90	0.80–0.95
	Median (Q1–Q3)*	3 (1–4)	2.5 (1–4)	5 (4–11)	5 (3–9)		
JSN	Mean (SD)	6.8 (5.6)	5.4 (5.2)	5.2 (4.3)	3.9 (2.6)	0.90	0.80–0.95
	Median (Q1–Q3)*	6 (3–10)	4.5 (3–6)	4 (2–7)	3 (2–6)		

* 1st quartile–3rd quartile. ICC: Intraclass correlation coefficient, JSN: joint space narrowing, SENS: Simple Erosion Narrowing Score.

Table 3. Interrater longitudinal reliability and sensitivity to change.

Scoring Method	Mean Change	SD of Change	SRM	SRM Ratio*	Adjusted SRM Ratio**	ICC	95% CI
Larsen	9.30	7.72	1.21			0.67	0.40–0.83
Larsen/Rau	7.30	6.26	1.17	0-.97	1.25	0.75	0.53–0.88
Sharp	22.68	17.91	1.27	1.05	1.65	0.85	0.70–0.93
Erosion	11.95	9.76	1.22	1.02	1.44	0.77	0.56–0.89
JSN	10.85	10.16	1.07	0.89	1.08	0.81	0.63–0.91
Sharp/van der Heijde	19.90	16.13	1.23	1.02	1.78	0.86	0.73–0.93
Erosion	10.95	9.35	1.17	0.97	1.70	0.83	0.67–0.92
JSN	9.1	8.58	1.06	0.88	1.07	0.80	0.61–0.90
SENS	6.90	6.18	1.12	0.93	0.94	0.62	0.32–0.81
Erosion	2.93	3.38	0.87	0.72	0.63	0.26	–0.12–0.57
JSN	4.15	4.04	1.03	0.85	0.69	0.53	0.19–0.75

* Ratio of each method SRM to Larsen SRM. ** Adjusted on rater and duration of followup. SRM: standardized response mean, ICC: intraclass correlation coefficient, JSN: joint space narrowing, SENS: Simple Erosion Narrowing Score.

hand and foot radiographs, and should not be ruled out on the basis of our findings.

The duration of followup was not equal in all patients. This variability in time intervals between baseline and second assessment may have influenced the degree of change to be detected differentially between methods. For that reason, this heterogeneity was controlled for by adjusting the SRM on the duration of followup in a mixed model of repeated ANOVA. Adjustment of SRM ratio has been used in previous analysis of variance-covariance (ANCOVA)[19] and allows control for additional sources of heterogeneity, enlarging the generalizability of the results. In these patients from the early 1990s, we assumed no decrease in score over time.

We assumed that cross-sectional intraobserver reliability would not differ at other time points and assessed it only at baseline. The design of our study did not include repetition of readings at followup, thus we could not assess intraobserver reliability for progression.

The order of reading per patient set was sequential, as recently advised, and proved logical[10,20]. On the other hand, the order of methods was fixed by family of methods and corresponding variants. This fixed order could possibly influence the results if one reading could further alter the

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on October 20, 2021 from www.jrheum.org

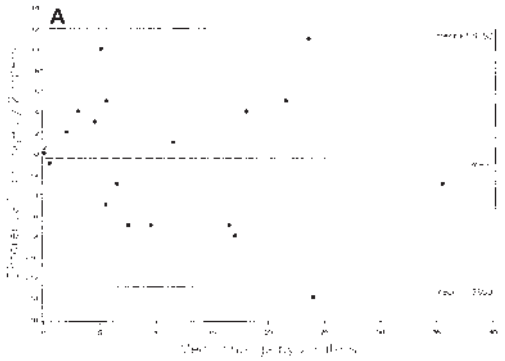

Figure 2a. Reliability for progression (Sharp method: JSN). Difference in change against mean change by 2 raters.

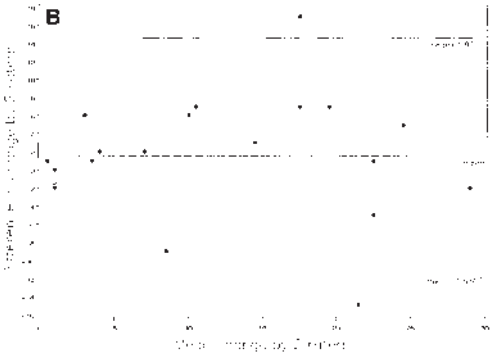

Figure 2b. Reliability at baseline (Sharp method: erosion). Difference in change against mean change by 2 raters.

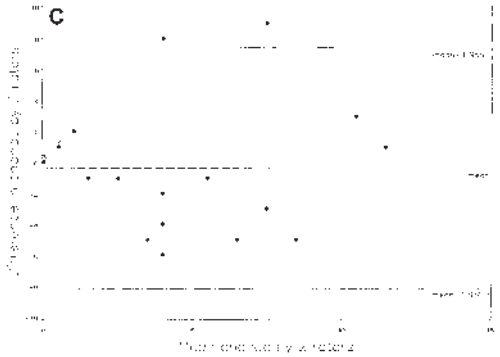

Figure 2c. Reliability for progression (SENS method: JSN). Difference in change against mean change by 2 raters.

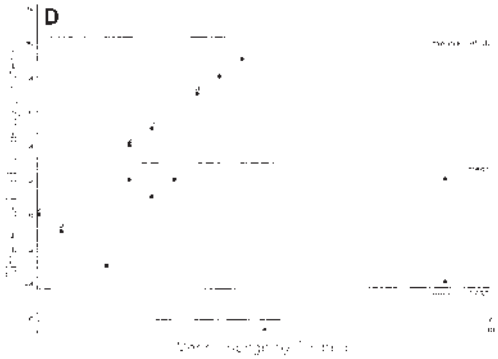

Figure 2d. Reliability for progression (SENS method: erosion). Difference in change against mean change by 2 raters.

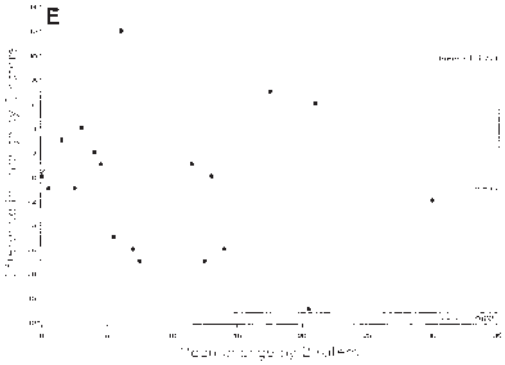

Figure 2e. Reliability for progression (Sharp/van der Heijde method: JSN). Difference in change against mean change by 2 raters.

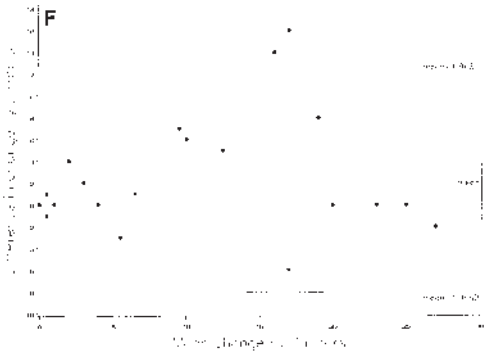

Figure 2f. Reliability for progression (Sharp/van der Heijde method: erosion). Difference in change against mean change by 2 raters.

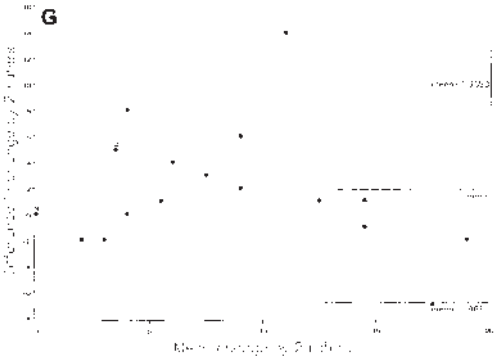

Figure 2g. Reliability for progression (Larsen/Rau method). Difference in change against mean change by 2 raters.

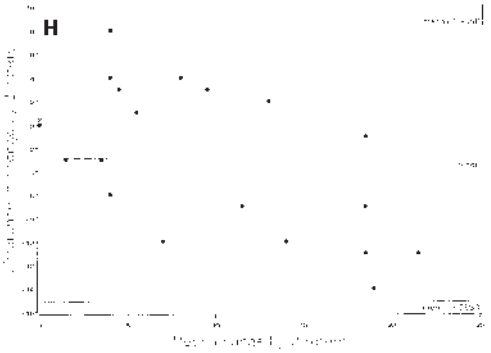

Figure 2h. Reliability for progression (Larsen method). Difference in change against mean change by 2 raters.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on October 20, 2021 from www.jrheum.org

Table 4. Smallest detectable difference (SDD) and percentage of total score, according to the scoring methods.

Scoring Method	SEM	SDD	%*
Larsen	2.28	5.29	4.41
Larsen/Rau	2.21	5.14	4.7
Sharp	5.25	12.21	3.9
Erosion	2.77	6.43	3.8
JSN	3.53	8.21	5.8
Sharp/van der Heijde	4.17	9.69	3.5
Erosion	2.49	5.80	3.6
JSN	3.09	7.18	6.0
SENS	2.78	6.51	10.5
Erosion	1.95	4.53	14.27
JSN	1.68	3.91	13.0

* Percentage of maximum score. SEM: standard error of measurement, JSN: joint space narrowing, SENS: Simple Erosion Narrowing Score.

rating of this radiograph by another method. Randomizing the order of methods could have prevented this. However, it was likely counterbalanced by the different skill of each trained rater, and is not likely to have much biased the comparative performance of methods.

Patients were at early stage of disease, and sensitivity to change may be different in late disease, in which results may differ. Documenting sensitivity to change at the early stage appears useful in light of current recommendations for early referral and treatment[21].

We used 2 different versions of the Larsen method for scoring hand radiographs: the Larsen modification (Larsen/Rau) that still included soft tissue swelling as grade 1 (as in the original Larsen method), and a Larsen modification (by Larsen himself) in which soft tissue swelling had been eliminated.

Intrarater reliability was estimated for the Sharp and the Larsen methods in a cohort of RA patients with 6.7 years' mean disease duration[22]. ICC were 0.96 for Sharp erosion score, 0.94 for Sharp JSN score, and 0.97 for Sharp total score and 0.88 for Larsen score. In another study, intrarater reliability for the Sharp/van der Heijde method was 0.96[23].

Interrater reliability for the Sharp/van der Heijde method ranged from 0.76 to 0.94, according to study design[20]. Concerning the Sharp method, Salaffi, *et al* found 0.79–0.96 and 0.58–0.71 for erosion score, 0.72–0.88 and 0.46–0.71 for JSN score, and 0.76–0.93 and 0.54–0.67 for total score for interrater reliability and interrater reliability for progression, respectively[24]. Interobserver reproducibility assessed by the Larsen method in 10 representative RA hand radiographs ranged from 0.78 to 0.92[25]. Intra and inter-rater reproducibility of the Larsen/Rau method assessed by experienced readers in patients with 2.7 years of disease duration were up to 0.80[26].

SRM for the Sharp/van der Heijde method varied from 0.81 to 0.85 according to disease duration for all patients, and from 1.03 to 1.06 for patients with erosions[23]. In another study, SRM for the SENS method ranged from 1.15 to 1.63[27]. When radiographs of hands and feet of 30 patients with early RA were assessed by Larsen/Rau and Larsen methods, the corresponding SRM were 0.83 and 0.88, respectively[28]. Another study showed that both Sharp and Larsen methods were sensitive to change in the first year of RA [median scores (at baseline vs after one year) of 15.5 vs 7.5 for Sharp method and median of 30.5 vs 22.5 for Larsen method; $p < 0.001$ for all comparisons][29].

We found results similar to those of van der Heijde, *et al*: SDD were around 10 for the Sharp/van der Heijde method and varied from 4 to 6 for the SENS method[12]. Moreover, Lassere, *et al* calculated interrater reliability for progression and SDD for the Sharp/van der Heijde and the Larsen (modified by Scott) methods[30]. They found ICC equal to 0.86 and 0.85 for Sharp/van der Heijde and Larsen methods, respectively. SDD were ± 12.6 and ± 11.2, respectively, when estimated by the 95% limits of agreement of the Bland-Altman method and ± 8.8 and ± 8.0, respectively, estimated by the 95% limits of agreement of mean score of 2 observers.

According to reading strategies, for the Sharp/van der Heijde method, Bruynesteyn, *et al*[31] found SDD = 5 with mean score at baseline of 24.6 (16.5) and mean progression of 7.6 (10.0) in chronological reading and SDD = 13.8 with mean score at baseline of 25 (16) and mean progression of 4.5 (10.2) in paired reading. The Larsen method (as modified by Scott) was also considered in this study: SDD was 5.8 [mean score at baseline 14.5 (10.4), mean progression 4.0 (8.0)] in chronological reading and 9.7 [mean score at baseline 16.8 (11.2), mean progression 3.7 (10.3)] in paired reading[31].

In a study assessing the minimal clinically important difference (MCID) in RA-related radiologic joint damage measured by the Sharp/van der Heijde method, the authors found similar MCID (4.6) and SDD (5.0), suggesting that the SDD can be used as the threshold for individual clinically relevant change in trials[32]. The minimal detectable radiographic change of the Larsen/Rau method was around 6 for intrareader and 7.7 for interreader[26].

SDD is based on measurement error. It corresponds to the minimal amount of progression that can reliably be distinguished from random measurement error. Mean scores of change in our study are greater than SDD (except SENS erosion score), suggesting that there is real progression. SDD is study-specific but should be reported for all endpoints as a quality control[33,34].

In summary, all methods have high intraobserver reliability. Although our results appear to be consistently in favor of the Sharp and Sharp/van der Heijde methods, the magnitude of difference does not strongly discriminate between methods. Interobserver reliability decreases when damage severity increases. Sensitivity to change was similar, with the exception of the SENS method on hands, which had a lower SRM. Overall, any choice of a method should have

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on October 20, 2021 from www.jrheum.org

quality ensured by training the readers, over and above trusting a particular method, since the method's performance relies more on the reader's skill than the scoring itself.

REFERENCES

1. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729-40.
2. Wolfe F, Lassere M, van der Heijde D, et al. Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. J Rheumatol 1999;26:484-9.
3. Van der Heijde D. Plain X-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. Baillieres Clin Rheumatol 1996;10:435-53.
4. Boini S, Guillemin F. Radiographic scoring methods as outcome measures in rheumatoid arthritis: properties and advantages. Ann Rheum Dis 2001;60:817-27.
5. Guillemin F, Suurmeijer T, Krol B. Functional disability in early rheumatoid arthritis: description and risk factors. J Rheumatol 1994;21:1051-5.
6. Smedstad LM, Moum T, Guillemin F, et al. Correlates of functional disability in early rheumatoid arthritis: a cross-sectional study of 706 patients in four European countries. Br J Rheumatol 1996;35:746-51.
7. Guillemin F, Gerard N, van Leeuwen M, Smedstad LM, Kvien TK, van den Heuvel W. Prognostic factors for joint destruction in rheumatoid arthritis: a prospective longitudinal study of 318 patients. J Rheumatol 2003;30:2585-9.
8. Sharp JT, Young DY, Bluhm GB, et al. How many joints in the hands and wrists should be included in a score of radiologic abnormalities used to assess rheumatoid arthritis? Arthritis Rheum 1985;28:1326-35.
9. 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.
10. Van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 1999;26:743-5.
11. Van der Heijde D, van Riel PL, Nuver-Zwart IH, Gribnau FW, van de Putte L. Effects of hydroxychloroquine and sulfasalazine on progression of joint damage in rheumatoid arthritis. Lancet 1989;1:1036-8.
12. Van der Heijde D, Dankert T, Nieman F, Rau R, Boers M. Reliability and sensitivity to change of a simplification of the Sharp/van der Heijde radiological assessment in rheumatoid arthritis. Rheumatology Oxford 1999;38:941-7.
13. Larsen A. How to apply Larsen score in evaluating radiographs of rheumatoid arthritis in long-term studies. J Rheumatol 1995;22:1974-5.
14. Rau R, Herborn G. A modified version of Larsen's scoring method to assess radiologic changes in rheumatoid arthritis. J Rheumatol 1995;22:1976-82.
15. Van der Heijde D, Boers M, Lassere M. Methodological issues in radiographic scoring methods in rheumatoid arthritis. J Rheumatol 1999;26:726-30.
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
17. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992;45:1341-5.
18. Fries JF, Bloch DA, Sharp JT, et al. Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial. Arthritis Rheum 1986;29:1-9.
19. Bombardier C, Raboud J, The Auranofin Cooperating Group. A comparison of health-related quality of life measures for rheumatoid arthritis research. Control Clin Trials 1991;12:243S-56S.
20. Van der Heijde D, Boonen A, Boers M, Kostense P, van der Linden S. Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. Rheumatology Oxford 1999;38:1213-20.
21. Emery P, Breedveld FC, Dougados M, Kalden JR, Schiff MH, Smolen JS. Early referral recommendation for newly diagnosed rheumatoid arthritis: evidence based development of a clinical guide. Ann Rheum Dis 2002;61:290-7.
22. 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.
23. Drossaers-Bakker KW, Amesz E, Zwinderman AH, Breedveld FC, Hazes JMW. A comparison of three radiologic scoring systems for long-term assessment of rheumatoid arthritis. Arthritis Rheum 2000;43:1465-72.
24. Salaffi F, Carotti M. Interobserver variation in quantitative analysis of hand radiographs in rheumatoid arthritis: Comparison of 3 different reading procedures. J Rheumatol 1997;24:2055-6.
25. Matsuno H, Yudoh K, Hanyu T, et al. Quantitative assessment of hand radiographs of rheumatoid arthritis: interobserver variation in a multicenter radiographic study. J Orthop Sci 2003;8:467-73.
26. Rau R, Wassenberg S, Herborn G, Stucki G, Gebler A. A new method of scoring radiographic change in rheumatoid arthritis. J Rheumatol 1998;25:2094-106.
27. Van der Heijde D, Boonen A, van der Linden S, Boers M. Reading radiographs in sequence, in pairs or random in rheumatoid arthritis: Influence of sensitivity to change [abstract]. Arthritis Rheum 1997;Suppl 40:S287.
28. Tanaka E, Yamanaka H, Matsuda Y, et al. Comparison of the Rau method and the Larsen method in the evaluation of radiographic progression in early rheumatoid arthritis. J Rheumatol 2002;29:682-7.
29. 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.
30. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731-9.
31. Bruynesteyn K, van der Heijde D, Boers M, et al. Detecting radiological changes in rheumatoid arthritis that are considered important by clinical experts: influence of reading with or without known sequence. J Rheumatol 2002;29:2306-12.
32. 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.
33. Van der Heijde D, Lassere M, Edmonds J, Kirwan J, Strand V, Boers M. Minimal clinically important difference in plain films in RA: group discussions, conclusions, and recommendations. OMERACT Imaging Task Force. J Rheumatol 2001;28:914-7.
34. Lassere MN, van der Heijde D, Johnson K, et al. Robustness and generalizability of smallest detectable difference in radiological progression. J Rheumatol 2001;28:911-3.
35. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;36:420-8.
36. Ravaud Ph, Giraudeau B, Audeley GR, Edouard-Noël R, Dougados M, Chastang CL. Assessing smallest detectable change over time in continuous structural outcome measures: application to radiological change in knee osteoarthritis. J Clin Epidemiol 1999;52:1225-30.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on October 20, 2021 from www.jrheum.org

APPENDIX

Intraclass correlation coefficients (ICC) and standardized response means (SRM) were used to assess respectively reproducibility and sensitivity to change. All calculations were derived from the same two-way mixed model:

$$Y_{ij} = \mu + p_i + r + \varepsilon_{ij} \qquad \text{(A)}$$

where Y_{ij} represents the j[th] observation on subject i, with $i = 1, \ldots, 20$ and $j = 1,2$

μ is the overall effect common to all observations,

$p_i \sim N(0, \sigma_p^2)$ is the random effect of the i[th] patient,

r is the fixed effect of rater or reading (see below),

$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ is the residual error associated with observation (i,j).

1. Reproducibility

The ICC was obtained as the ratio of the between-subject variability on the overall variability, it was estimated by $\frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_\varepsilon^2}$. A confidence interval was computed with the formula given by Shrout and Fleiss[35]. This summary information was completed by plotting the difference between two measures against their mean according to the method of Bland and Altman [16].

1.1 INTER-RATER REPRODUCIBILITY

Both raters read the 20 patients' baseline x-rays once. Y_{ij} was the score affected to patient i by rater j. The parameter r represented the fixed rater effect.

1.2 INTRA-RATER REPRODUCIBILITY

Each rater read the baseline x-rays of half of the patients in two occasions. Here Y_{ij} denoted the score of patient i at the j[th] reading and r represented the fixed reading effect.

1.3 LONGITUDINAL INTER-RATER REPRODUCIBILITY

Each patient x-rays were read by the two raters at baseline (time 1) and after a follow-up period (time 2). We modelized $Y_{ij} = Y_{ij2} - Y_{ij1}$ the difference between the score at time 2 and score at time 1 according to model (A), r being the fixed rater effect.

2. Sensitivity to change

The SRM was computed as the mean of the difference between score at time 2 and score at time 1 over its standard deviation. That is, the SRM was estimated by : $\frac{\bar{Y}_2 - \bar{Y}_1}{\hat{\sigma}_{Y_2 - Y_1}}$ ref [36]

One can calculate the SRM by modeling $Y_{ij} = Y_{ij2} - Y_{ij1} = \mu + \varepsilon_{ij}$,

with μ representing an overall effect and ε_{ij} the measurement error. The SRM would be estimated by $\frac{\hat{\mu}}{\hat{\sigma}_\varepsilon}$.

This formulation of the SRM gives us 'crude' SRM in the sense that it is not adjusted on other effects such as the rater effect. In our study, the duration of follow-up was not the same in all subjects ranging between two and three years. Based on the hypothesis that damage severity can not decrease with time, we were more likely to detect a larger difference in patients having a longer follow-up period. Furthermore, we know that a non-negligible part of the variability is due to the between-patient variance, which is not taken into account by the above SRM.

We proposed an adjusted SRM based on a mixed model derived from model (A) in which we added a parameter c corresponding to the fixed effect of the time:

$$Y_{ij} = \mu + p_i + r + c + \varepsilon_{ij} \qquad \text{(B)}$$

This model allowed us to adjust the estimated mean on fixed rater and follow-up time effects and also to take into account the between-patient variability. The adjusted SRM was then estimated from model (B) as:

$$SRM_{adj} = \frac{\hat{\mu}_{adj}}{\hat{\sigma}_{adj_\varepsilon}}$$

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on October 20, 2021 from www.jrheum.org