Annals of the Rheumatic Diseases
The EULAR Journal

REVIEW

Radiographic scoring methods as outcome measures in rheumatoid arthritis: properties and advantages

S Boini, F Guillemin

Abstract
***Background*—Use of scored radiographs as an outcome measure can help estimate the progression of rheumatoid arthritis (RA). Radiographs not only provide permanent records with which to evaluate RA serially, but can also be randomised and blinded, a major advantage in clinical trials.**
***Objectives and method*—Medline was searched for information about the principal methods of assessing joints affected by RA. Each technique was evaluated for its measurement properties, advantages, and limitations.**
***Main findings*—The most commonly used methods are those devised by Sharp, Larsen, and van der Heijde/Sharp, and their variants. Methods based on the Sharp technique provide separate scores for erosion and for joint space narrowing. Larsen and variants, together with the Simple Erosion Narrowing Score (SENS) method, provide an overall score. Each method's measurement properties (feasibility, time consumption, etc) depend on the degree of detail it considers. Authors consistently recommend taking a posteroanterior view of hand and foot radiographs, and the use of trained raters. Intra- and interrater reliability values are generally higher than 0.70 (less often assessed by the intraclass correlation coefficient than the correlation coefficient). Sensitivity to change is calculated by several techniques (standardised response mean (SRM), adjusted SRM, minimal detectable change, smallest detectable difference). Most methods assessed with SRM reach a value of 0.80 or more.**
***Conclusion*—Standardised procedures are available for performing and reading radiographs in RA. The choice of scoring method depends on the time and staff available, and the required degree of reliability and sensitivity to change.**
(*Ann Rheum Dis* 2001;**60**:817–827)

UPRES EA 1124-Ecole de Santé Publique, Faculté de Médecine, Nancy, France
S Boini
F Guillemin

Correspondence to:
Professor F Guillemin, Ecole de Santé Publique, Faculté de Médecine, 9 Avenue de la Forêt de Haye, BP 184, F 54500 Vandoeuvre les Nancy, France
francis.guillemin@sante-pub.u-nancy.fr

Accepted 12 April 2001

Radiographs can be used as an outcome measure to assess the severity and progression of rheumatoid arthritis (RA), and to establish the effects of treatment. They also provide a permanent record with which the disease can be serially evaluated. An additional advantage of radiographs is that they can be randomised and blinded for standardised scoring.[1–4]

Objectives and methods of the review

OBJECTIVES

Numerous ways of assessing RA radiologically have been developed and tested for degree of reliability and other characteristics. The objectives of this work are to review the principal joint damage scoring systems and compare their measurement properties, advantages, and limitations.

LITERATURE SEARCH

Medline was searched for articles reporting the use of scored radiographs in RA. Key words (Mesh, title and abstract) included "RA", "radiographs", "radiographic", "radiologic", "*x* ray", "scoring method", "score", "comparison", "progression", "reproducibility", "reliability", and "sensitivity to change". A selection of articles pertaining to the description of scoring methods and to the analysis of their measurement properties was obtained by the search and by examining relevant reference lists and other review articles. Particular care was taken to identify papers that compared two or more *x* ray reading and scoring methods.

Radiographic scoring methods

There are numerous radiographic scoring methods. Some give a global assessment for the entire patient (Steinbrocker[5] and Kellgren[6]), whereas others assess individual joints (Sharp and its variants, Larsen and its variants, and Simple Erosion Narrowing Score (SENS)).[3] Joint space narrowing (JSN) and erosions may or may not be scored separately. For example, Larsen and its variants give one overall score, and SENS provides scores for erosion and JSN that are summed thereafter to give a figure comparable to the Sharp total score. Recent techniques are more detailed than earlier ones.

In 1971, Sharp et al proposed a scoring method for the hands and wrists.[7] Twenty nine areas in each hand and wrist are considered forerosions, and 27 for JSN. Counts for erosion range from 0 to 5, to give an erosion score between 0 and 290. Counts for JSN range from 0 to 4, to give a score between 0 and 216. This original version is no longer used.

A modification proposed in 1985[8 9] is now considered the standard for the Sharp method. It considers 17 areas for erosion (five proximal interphalangeal (PIP), five metacarpophalangeal (MCP), 1st metacarpal base (MCB), multangular as one unit, navicular, lunate, triquetrum (and pisiform), radius, ulnar bone for each hand and wrist) and 18 areas for JSN (five PIP, five MCP, carpometacarpal (CMC) 3 to 5, multangular-navicular, lunate-triquetrum, capitate-navicular-lunate, radiocarpal, radioulnar joints for each hand and wrist). Each erosion scores one point, with a maximum of five points for each area (reflecting loss of more than 50% of either articular bone). Erosion scores range from 0 to 170. One point is scored for focal joint narrowing, two points for diffuse narrowing of less than 50% of the original space, and three points if the reduction is more than half of the original joint space. Ankylosis is scored as four. (Sub)luxation is not scored. The score for JSN ranges from 0 to 144.

In 1986 another modification was devised by Fries et al (with the participation of Sharp).[10] They tested several combinations with erosion and JSN using two strategies: "size and count" and "global". An erosion count is a simple sum of erosion in the assessed joints, whereas a weighted erosion count incorporates the size of each erosion (from 1 to 4, measured with a template). "Global" strategy uses an overall assessment of each joint. From the several comparisons made, the simple combination of scores of JSN in six selected sites (the worst PIP, the worst MCP, and the worst radiocarpal joints in each hand) plus either the weighted erosion count on 18 selected joints (four MCP, four PIP, and the ulnar styloid in each hand) or the global erosion score on the same 18 selected joints is recommended for clinical trials using radiographic progression as an end point. This method is more time consuming than the original, but probably adds little to the sensitivity and reliability.

In 1983, Genant et al developed a method of scoring hand and foot radiographs.[11] It considers erosive change at 16 sites in the hand and six in the foot, and JSN at 11 and six sites, respectively. The following joints are considered for erosions: interphalangeal (IP), five MCP, four PIP, mid-navicular, radial (styloid and ulnar), ulna (radial and styloid and outer aspect) in the hand, and IP, five metatarsophalangeal (MTP) in the foot. Joints considered for JSN are IP, five MCP, four PIP, radiocarpal compartment in the hand, and IP, five MTP in the foot. Erosions and JSN are separately graded from 0 to 4 (0 = normal; 1 = questionable; 2 = definite but mild; 3 = moderate; and 4 = severe). This method requires a standard reference set of radiographs for comparison. The range of erosion scores is from 0 to 128 in the hands, and from 0 to 48 in the feet. The JSN score ranges from 0 to 88 in the hands, and from 0 to 48 to in the feet.

In 1998, Genant et al modified their method as follows[12]: erosion is scored according to an eight point scale with 0.5 increments, where 0 = normal; 0+ = questionable or subtle change; 1 = mild; 1+ = mild worse; 2 = moderate; 2+ = moderate worse; 3 = severe; and 3+ = severe worse. In each hand, IP of the thumb, PIP, MCP, 1st CMC, scaphoid, ulna, and radius are included. The score for erosion for both hands ranges from 0 to 98. JSN is scored according to a nine point scale with 0.5 increments, where 0 = normal; 0+ = questionable or subtle change; 1 = mild; 1+ = mild worse; 2 = moderate; 2+ = moderate worse; 3 = severe; 3+ = severe worse; and 4 = ankylosis or dislocation. In each hand, IP of the thumb, PIP, MCP, CMC 3 to 5, capitate-scaphoid-lunate, and the radiocarpal joint are included. The score for JSN for both hands ranges from 0 to 104. After separately summing the two scores for both hands, each score is normalised to a scale from 0 to 100. This variant method is in current use.[13]

Kaye et al combined the methods described by Genant[11] *and Sharp et al.*[8] Two scoring systems are used in this approach. The more detailed of the two[14] includes 21 joints (all PIP, all MCP, and 11 sites in the wrist) in the hand and wrist. Erosion is graded 0–4, with 0 for no evidence of articular erosion and 2, 3, and 4 for mild, moderate, and severe erosion, respectively. JSN is graded 0 to 5, with 0 for no evidence of JSN and 2, 3, 4, and 5 for mild, moderate, severe JSN, and bony ankylosis, respectively. Scores for erosion range from 0 to 168, and for JSN from 0 to 210. This system uses the standard reference set of radiographs developed by Genant.[11] The simplified system[15] assesses the same joints using a grading system ranging from 0 (normal joint) to 4 (marked erosion or JSN, including ankylosis, dislocation or marked (sub)luxation). Grade P is used for a postoperative joint, and grade X if a joint cannot be evaluated. When calculating the score, grade P is considered to be grade 4, and grade X is ignored. The absolute sum score ranges

from 0 to 168. The ultimate score is the absolute sum score divided by the number of evaluated joints. Neither of these two systems includes an equivocal grade (that is, grade 1).

In 1989, van der Heijde modified the method described by Sharp in 1985.[16–18] Erosion is assessed in 16 joints (five MCP, four PIP, IP of the thumbs, 1st MCB, radius and ulna bones, trapezium and trapezoid as one unit (multangular), navicular, lunate) for each hand and wrist, and six joints (five MTP, IP) for each foot. One point is scored if erosions are discrete, rising to 2, 3, 4, or 5 depending on the amount of surface area affected (complete collapse of the bone is scored as 5). The score for erosion ranges from 0 to 160 in the hands and from 0 to 120 in the feet (the maximum erosion score for a joint in the foot is 10). JSN is assessed in 15 joints (five MCP, four PIP, CMC 3 to 5, multangular navicular-lunate, radiocarpal) for each hand and wrist, and six joints (five MTP, IP) for each foot. JSN is combined with a score for (sub)luxation and scored as follows: 0 = normal; 1 = focal or doubtful; 2 = generalised, less than 50% of the original joint space; 3 = generalised, more than 50% of the original joint space or subluxation; 4 = bony ankylosis or complete luxation. The score for JSN ranges from 0 to 120 in the hands and from 0 to 48 in the feet. (The hand score has greater weight because more joints are scored.)

In 1999, van der Heijde described the SENS method.[19] It is a simplified method of scoring radiographs based on the Sharp/van der Heijde score: instead of grading, the number of joints with erosions and with JSN are simply summed. SENS assesses the same joints as the Sharp/van der Heijde method. 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 score for each joint can therefore range from 0 to 2. Erosion is considered in 32 joints in the hands and 12 in the feet, and JSN in 30 and 12 joints, respectively. The total SENS score ranges from 0 to 86.

In 1974, Larsen developed a method based on a set of standard films. It differentiates six stages from 0 (normal) to 5, reflecting gradual, progressive deterioration, and provides an overall measure of joint damage. This method was modified several times (1977, 1978, 1984, 1985, 1987, and 1995). In the 1977 version,[20 21] the six stages are as follows: grade 0 = normal; grade 1 = slight abnormalities (periarticular soft tissue swelling and periarticular osteoporosis and slight JSN); grade 2 = definite early abnormalities; grade 3 = medium destructive abnormalities; grade 4 = severe definite abnormalities; and grade 5 = mutilating abnormalities. The wrist is considered as one unit and the score is multiplied by five. Joints assessed include five distal interphalangeal (DIP), four PIP, five MCP, the wrist as one unit for each hand and wrist, and 10 MTP, two IP for the feet. The score ranges from 0 to 250.

In 1995, Larsen devised a method to evaluate radiographs in long term studies.[22] The main differences from the original are deletion of scores for the thumbs and 1st MTP; subdivision of the wrist into four quadrants (the joints considered are PIP 2 to 5 and MCP 2 to 5 in each hand, four quadrants in the wrist, and MTP 2 to 5 in each foot); deletion of soft tissue swelling and osteoporosis; distinction between erosions of different sizes. The grading scale ranges from 0 to 5: 0 = intact bony outlines and normal joint space; 1 = erosion less than 1 mm in diameter or JSN; 2 = one or several small erosions (diameter more than 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 score ranges from 0 to 160.

In 1995, Scott et al proposed a modification to the Larsen score descriptors.[23 24] Their new system for the hands and wrists consists in modified grading compared with the Larsen's method. Grade 1 is assigned for periarticular osteoporosis/joint swelling if these are major features or if suggested erosions/cysts at two sites in a joint are less than 1 mm in diameter; in grade 2, one or more erosions greater than 1 mm are present with a break in the cortical margin; in grade 3, erosions at both sides of the joint are of significant size, with preservation of some joint surface; in grade 4, subluxation is present. Grade 0 and grade 5 are left unchanged. The joints considered are as in the Larsen system.[20] The score ranges from 0 to 250. This modification permits a higher correlation of grade 1 between raters.

In 1995, Rau and Herborn proposed another modification of the Larsen method.[25] Thirty two joints are evaluated: eight PIP, two IP of the thumbs, 10 MCP, two wrists, and 10 MTP. The six stages are defined as follows: 0 = normal; 1 = soft tissue swelling and/or joint space narrowing/subchondral osteoporosis; 2 = erosions with destruction of the joint surface (DJS) ≤25%; 3 = DJS 26–50%; 4 = DJS 51–75%; 5 = DJS >75%. The score ranges from 0 to 160. In this modification the stages are described as a quantitative measure of the destroyed joint surface area and can therefore be applied more easily.

In 1998, Rau et al developed a new method, the Ratingen score, derived from the Larsen score.[26] Scoring is performed in the following joints or areas: 10 PIP, 10 MCP, four sites in the wrist (navicular, lunate, radius, and ulna), eight MTP (2 to 5), and two IP on the great toe. This new method restricts scoring of an individual joint to definite changes of erosion and joint destruction. The extension of the erosion into the bone is not considered. The amount of joint surface destruction is defined by the length of the clearly visible interruption of the cortical plate in relation to the total joint surface. Grades are then assigned as follows: grade 1 = one or several definite erosions totalling destruction of ≤20% of the total surface;

Table 1 Features of rheumatoid arthritis included in the different radiographic scoring methods

Method	*Erosion*	*JSN**	*Osteoporosis*	*Malalignment*	*Soft tissue swelling*	*(Sub)luxation*	*Ankylosis*	*Cyst*
Sharp (1971)	√	√					√	√
Sharp (1985)	√	√				Excluded†	√	
Genant (1983)	√	√			√			
Genant (1998)	√	√				√	√	
Kaye/Sharp (1986)	√	√		√		√	√	
van der Heijde/Sharp (1989)	√	√				√	√	
SENS (1999)	√	√				√	√	
Larsen (1977)	√	√	√		√	Excluded†	Excluded†	
Larsen (1995)	√	√	Excluded†		Excluded†			
Scott/Larsen (1995)	√	√	√		√	√		√
Rau/Larsen (1995)	√	√	√		√			
Ratingen (1998)	√	√	Excluded†		Excluded†			
SES (2000)	√	√	Excluded†		Excluded†			

*JSN = joint space narrowing.
√Included in the scoring system; †not included in the scoring system.

grade 2 = joint surface destruction 21–40%; grade 3 = 41–60%; grade 4 = 61–80%; grade 5 >80%. Adding the scores from 38 areas gives a total score ranging from 0 to 190.

The carpo:metacarpal ratio (C:MC) is a quantitative measure of wrist involvement in RA proposed by Trentham and Masi.[27] It is calculated by dividing the carpal by the third metacarpal lengths. The longitudinal length of the carpus (from the dense volar-ulnar margin of the distal radius to the base of the 3rd metacarpal bone at its cortical midpoint) is measured (in mm) for each hand and wrist and then divided by the greatest length of the third metacarpal bone. The C:MC ratio is calculated for the right and the left hand and the average determined. It reflects joint space reduction in the wrist rather than erosive damage, and diminishes as carpal involvement progresses. Its value is 0.60 in a non-RA group. This approach is suitable for serial evaluation.

In 2000, Wolfe et al proposed a new approach to evaluating severity of RA, the Short Erosion Scale (SES), a modification of the Larsen method (1995).[28] Rasch analysis was used to determine the minimum number of joints necessary to produce a linear representation of radiographic severity with adequate fitting, scaling, and dimensionality. The SES considers 12 joints: three of four regions of the wrist as defined by Larsen (medial-proximal, medial-distal, and lateral-proximal) and MCP 2, 3, and 5. Each joint is graded as in the 1995 Larsen system.

Advantages

Radiographic scoring methods are used to measure and evaluate changes in RA; however, there is no universally accepted technique, and modifications are often proposed. The important issues to be taken into account when choosing between the different systems, particularly for use in therapeutic trials, are the features included, the joints counted, and the scale used. Although radiographs are considered the "gold standard", an awareness of the problems with which they are associated is necessary if they are not to become fool's gold.[2]

ABNORMALITIES IDENTIFIED (TABLE 1)

The numerous abnormalities that can be seen on radiographs of patients with RA (such as soft tissue swelling, osteoporosis, erosions, JSN, subluxation and malalignment, and ankylosis)[1] reflect disease severity and progression. No existing scoring method includes them all, but erosions and, to a lesser extent, JSN seem to be most widely accepted as important.[8] They are highly specific to RA, can be reliably assessed, and provide independent information.[4 12 19] The relative weight given to erosion versus JSN varies, and no consensus has yet been established.[10] The potential ratio of weighted erosion/JSN is always one or higher: 1.2 for Sharp, 1.7 for van der Heijde/Sharp, 1 for SENS. The Larsen methods combine erosion and loss of cartilage scores. Because osteoporosis is difficult to quantify and highly dependent on radiographic technique, it is generally excluded.[7 8 14 29] Substantial destruction related to cyst formation is not captured by Larsen's or Sharp's (1985) systems. Cyst formation can be counted as erosion. Subluxation and luxation are responsible for another difficulty in scoring films: very severe luxation makes it impossible to measure the degree of erosion or even to detect erosive changes. Healing phenomena are also a problem. The Sharp and Larsen methods do not consider healing as manifested by improvement in bone damage.[30] Most authors do not report whether they considered this problem in developing their scoring strategies. Those that did, did not allow for decreases in score.[31 32]

METRIC SCALING PROPERTY

By relating radiological abnormality scores to clinical features that reflect the severity of RA, it is possible to make a semiquantitative assessment of the value of radiographic change. The use of graphical statistical techniques allows the course of the disease to be described visually, and numerical assessment of the progression of radiological abnormalities has prognostic value.[7] However, semiquantitative scoring techniques in the evaluation of RA are limited by the association between the extent of a finding on the radiograph and the score assigned to it. An ideal scoring method supposes a linear relation between the quantity of change and the scale itself. This is not the case for the Larsen and the Sharp methods.[26 30] The SES method developed by Wolfe targets this goal.[28] Other methods use a count of abnormalities on a continuous scale.[33]

Table 2 Joints counted in the different radiographic scoring methods with separate scores

		Hand						Wrist										Forefoot	
Method		*DIP**	*PIP**	*MCP**	*CMC**	*IP**	*MCB**	*Scaphoid*	*Lunate*	*Triquetral*	*Pisiform*	*Trapezium*	*Trapezoid*	*Capitate*	*Hamate*	*Radio-ulnar*	*Radio-carpal*	*MTP**	*IP†**
Sharp (1971)	E*	√	√	√		√	√	√	√	√	√	√	√	√	√	√			
	JSN*	√	√	√	√	√		√	√	√		√		√	√	√	√		
Sharp (1985)	E		√	√			√ (1st)	√	√	√	√	√	√			√			
	JSN		√	√	√ (3rd–5th)			√	√	√		√	√	√		√	√		
Genant (1998)	E		√	√	√ (1st)	√		√								√			
	JSN		√	√	√ (3rd–5th)	√		√	√					√			√		
Kaye/Sharp (1986)	E		√	√	√ (3rd–5th)	√	√ (1st)	√	√	√		√	√	√	√	√	√		
	JSN		√	√	√ (3rd–5th)	√	√ (1st)	√	√	√		√	√	√	√	√	√		
van der Heijde/Sharp (1989)	E		√	√		√	√ (1st)	√	√			√	√			√		√	√
	JSN		√	√	√ (3rd–5th)			√	√			√	√	√			√	√	√
SENS (1999)	E		√	√		√	√ (1st)	√	√			√	√			√		√	√
	JSN		√	√	√ (3rd–5th)			√	√			√	√	√			√	√	√

*DIP = distal interphalangeal; PIP = proximal interphalangeal; MCP = metacarpophalangeal; CMC = carpometacarpal; IP = interphalangeal joint of the thumb; MCB = metacarpal base; MTP = metatarsophalangeal; IP† = interphalangeal; E = erosion; JSN = joint space narrowing.

Table 3 Joints counted in the different radiographic scoring methods with overall score

	Hand						Wrist										Forefoot	
Method	*DIP**	*PIP**	*MCP**	*CMC**	*IP**	*MCB**	*Scaphoid*	*Lunate*	*Triquetral*	*Pisiform*	*Trapezium*	*Trapezoid*	*Capitate*	*Hamate*	*Radio-ulnar*	*Radio-carpal*	*MTP**	*IP†**
Larsen (1977)	√	√	√		√		<—— As one unit ——									——>	√	√
Larsen (1995)		√ (2d–5th)	√ (2d–5th)						<—— 4 Quadrants ——							——>	√ (2d–5th)	
Scott/Larsen (1995)	√	√	√		√		<—— As one unit ——									——>	√	√
Rau/Larsen (1995)		√	√				<—— As one unit ——									——>	√	
Ratingen (1998)		√	√				<—— 4 Sites ——									——>	√ (2d–5th)	√
SES (2000)		√ (2d–5th)	√ (2d–5th)						<—— 3 Quadrants ——							——>		

*DIP = distal interphalangeal; PIP = proximal interphalangeal; MCP = metacarpophalangeal; CMC = carpometacarpal; IP = interphalangeal joint of the thumb; MCB = metacarpal base; MTP = metatarsophalangeal, IP† = interphalangeal.

HAND OR FOOT RADIOGRAPHS, OR BOTH?

With the exception of the Sharp method, all scoring techniques are based on the evaluation of hand and foot joints. In most studies comparing the Sharp method with other techniques, only hand radiographs were used, and the authors did not speculate on whether omission of the data for feet might have influenced their results.[29 34–37] Radiographic progression in RA has been shown to occur early, and the first erosions are more often found in the feet than in the hands. Van der Heijde underlined this fact in a study based on the Sharp/van der Heijde modification.[17] Plant *et al* justified extending the Sharp method to the joints of the foot[38] by pointing to the fact that erosions of MTP in the first year were almost as good predictors of outcome as the overall rate of radiological progression, whereas wrist and MCP erosions became better predictors beyond the first two years. This is consistent with MTP damage occurring at an earlier stage, and other joints being affected later. Paimela *et al* extended the Sharp method by including joints of the foot[39]; van der Heijde also successfully used this modification. These authors emphasised the importance of evaluating both hand and foot radiographs when assessing therapeutic intervention in early RA.

JOINT SITES (TABLES 2 AND 3)

Most scoring techniques assess the same joint areas. Early systems included more joints. It has been shown that eliminating areas that are technically difficult to read (hamate and capitate bones, radioulnar and lunate-triquetrum joints) and areas that are not commonly affected (DIP and MCB, for example) leaves an appropriate sample of the joints in the hands and wrists that still accurately represents the radiological abnormalities of patients with RA.[8 14] In the feet, MTP are the sites where erosions first develop and thus must be considered in any scoring method. Finally, the following joints are generally counted: PIP, MCP, MTP, IP, and joints of the wrist. Considerable information is gained from assessment of the PIP, MCP, and wrist joints, in particular.[1]

RADIOGRAPHIC TECHNIQUES

Any study based on radiographic assessment is highly dependent on the technical quality of the radiographs, particularly with regard to the reproducibility of the results. It is important to consider several technical variables. Firstly, accurate joint positioning is essential. Most investigators have opted for the posteroanterior view for hand and foot radiographs. However, other views such as the Norgaard view (a 45° supine view with straight finger) and the Brewerton view (a tangential view with the MCP joints flexed at 65° and with a 15° volar beam) have been used. When these three different approaches were compared in a prospective study,[40] the authors found no significant advantage in the Norgaard or the Brewerton view compared with the standard posteroanterior view used in most investigations.[12 17 38] Secondly, the exposure of the film is extremely important, as evidence of erosions is lost on both under- and overpenetrated films. Thirdly, high resolution films are essential in detecting early erosive disease. Single screen/film combinations give both an acceptable system speed and an acceptable exposure time.[3] Although excellent baseline film may be obtained, it is difficult to maintain high quality data because of technical variability. Even minuscule changes in the rotation of an imaged joint will cause modification of the film.

Thus accurate interpretation of radiographs depends on appropriate positioning, film exposure, screen film combination used, and reproducibility of radiographic production. Most technical problems are due to the fact that plain radiographs are two dimensional pictures of a three dimensional object.

READING STRATEGIES

Radiographs can be scored in a random order (that is, single order; radiographs are randomly ordered with regard to patient and sequence), paired without knowledge of the sequence (that is, paired order; two radiographs of the same patient without information about their chronological sequence), or ordered with known sequence (that is, chronological order; two radiographs of the same patient and of known chronological sequence). There are advantages and disadvantages for all these methods. Scoring in a chronological order probably gives the reader most information, but it introduces a bias in that he or she may expect progression of damage over time. Paired radiographs have the advantage that the positioning can be compared. Random order has the disadvantage that the quality of radiographs can vary greatly, but seems to give the best results and to be a good way of validating a scoring method. Most studies reviewed here used a known sequence—that is, chronological order, for reading radiographs.

Salaffi *et al* compared these three different reading procedures using the Sharp method, with two hand radiographs for each patient (at baseline and after 18 months).[41] They concluded that paired order is preferable. In a randomised, controlled trial with multiple readers, Fries *et al* concluded that paired reading (reading the films pairs simultaneously) was preferable to separated reading (reading the films at separate times).[10] Using several statistical methods, Ferrara *et al* produced results suggesting that paired reading was the most suitable for evaluating the progression of joint damage.[42] Recently, van der Heijde *et al* devised two studies using the Sharp/van der Heijde method to score radiographs of hands and feet[32]: one compared random, chronological, and paired order; one evaluated random, chronological, and so called "single-paired" order (the films are grouped by anatomical region from a particular patient at a single point in time). It was concluded that chronological order was more sensitive to change than the other approaches, and that the difference was particularly pronounced with longer follow up.

Table 4 Reliability and sensitivity to change for different scoring methods

Method	Intrarater reliability		Intrarater reliability for progression		Interrater reliability		Interrater reliability for progression		Sensitivity to change	
Sharp (1971)	r*=0.97 (E*) r=0.94 (JSN*) r=0.95 (tot*)	[29] *			r=0.53–0.96 (E) r=0.58–0.95 (JSN) r=0.73–0.96 (tot)	[29]				
Sharp (1985)					r=0.95–0.95 ICC*=0.79–0.96 (E) ICC=0.72–0.88 (JSN) ICC=0.76–0.93 (tot)	[9] [41]	ICC=0.58–0.71 (E) ICC=0.46–0.71 (JSN) ICC=0.54–0.67 (tot)	[41]		
Kaye/Sharp (1986)	r=0.95–0.96	[14]			r=0.72–0.93 (E) r=0.84–0.96 (JSN) r=0.91–0.96 (tot)	[14]				
Genant (1983)	r=0.92–0.95	[15]			r=0.90–0.98 r=0.84–0.94 (E) r=0.78–0.90 (JSN)	[15] [11]	r=0.14–0.50 (E) r=0.12–0.30 (JSN)	[11]		
Genant (1998)	r=0.90–0.95 (E) r=0.90–0.95 (JSN) r=0.92–0.96 (tot)	[12]	–r=0.93–0.97 (E) r=0.87–0.95 (JSN) r=0.93–0.97 (tot)	[12]	r=0.82–0.96 (E) r=0.69–0.92 (JSN) r=0.80–0.95 (tot)	[12]	r=0.90–0.97 (E) r=0.80–0.92 (JSN) r=0.90–0.95 (tot)	[12]		
v d Heijde/Sharp (1989)	r=0.94–0.99 (E) r=0.94–0.99 (JSN) r=0.96–0.99 (tot)	[17]			r=0.92 (E) r=0.80 (JSN) r=0.90 (tot) G coeff=0.76–0.94	[17] [32]			G coeff=0.22–0.39	[32]
Scott/Larsen (1995)			r=0.92 and ICC=0.94	[31]						
Larsen (1977)			r=0.45	[34]	r=0.93	[34]				
Ratingen (1998)	ICC*=0.79–0.90	[26]	SD_{change}*/SD_{WR}*=2.2–3.2	[26]	ICC=0.69–0.80	[26]	SD_{change}/SD_{BR}*=1.70–2.95	[26]		
SENS (1999)									SRM*=1.15 (sequential) SRM=1.63 (paired) SRM=1.60 (random)	[47]

*r = correlation coefficients; ICC = intraclass correlation; E = erosions; JSN = joint space narrowing; tot = total; SD_{change} = standard deviation of radiographic change; SD_{wr} = SD within rater; SD_{br} = SD between raters; SRM = standardised response mean; [ ] = study reference.

NUMBER OF RATERS

The number of raters reading radiographs in reviewed studies ranged from one to more than 10. The optimum number was defined by the highest efficiency. Using the results of more than one reader reduced measurement error and gave more precise results (averaging scores increases the reliability of progression scores) but required more time for rater training and therefore cost more. The number of patients needed depends primarily on the amount of change within the subject group, and may be reduced as the number of raters increases. Finally, the number of raters varies according to the purpose of the study. A randomised, controlled trial conducted by Fries *et al* showed that two readers was the best compromise,[10] and that reader training was essential: of eight experienced readers, the four who were trained (that is, who had extensive experience in scoring radiographs in clinical trials) exhibited consistently higher agreement. Other comparisons between experienced and inexperienced readers for scoring methods showed no significant difference,[14 26 40] though agreement was slightly higher among experienced raters. The variability in the scores of experienced and inexperienced readers remained within a tolerable range. Most studies based on radiographic scoring systems have used trained readers.[12 19 36 39 43] The discipline of the reader (rheumatologist or radiologist) had no effect on the quality of the readings.[3 10]

TIME CONSUMPTION

Several authors have calculated the time needed to score radiographs with different methods. Wassenberg *et al* found that the time to score seven radiographs of hands and feet was 3.9 minutes for Larsen, 19 minutes for Sharp, 25 minutes for the Sharp/van der Heijde method, and 9 minutes for the Ratingen method.[43] Other studies gave similar results for the Ratingen score method and the Sharp/van der Heijde method.[19 26] The time needed to score seven radiographs of hands and feet was 7 minutes for SENS.[19] The time needed to score 12 radiographs of hands and feet with the Sharp/van der Heijde method ranged from 11.1 minutes to 20.5 minutes.[32] The time needed is one drawback of both the Sharp method and the Sharp/van der Heijde method; it is related to their higher degree of detail as compared with the Larsen and SENS methods.

Reliability and sensitivity to change (Table 4)

RELIABILITY (TABLE 5)

The value of any scoring method used to measure a clinical variable depends on its reliability as shown by intra- and interreader reproducibility.[1] This can be assessed at a given point in time using absolute scores, or between two time points using values related to progression. Spearman and Pearson correlation coefficients, intraclass correlation coefficients (ICC), and κ statistics are used to assess reliability. However, the most appropriate methods are ICC and κ statistics because

Table 5 Reliability: comparison between scoring methods

Method	Intrarater reliability					Intrarater reliability for progression	Interrater reliability				Interrater reliability for progression	
	[19]*	[36]	[38]	[39]	[49]	[38]	[8]	[35]	[38]	[39]	[31]	[38]
Sharp (1971) Sharp (1985)		ICC*=0.96 (E*) ICC=0.94 (JSN*) ICC=0.97 (tot*)	r=0.96	ICC=0.97–0.98		r=0.87	r*=0.73–0.96 r=0.75–0.97	r=0.94	r=0.90–0.97	ICC=0.94		r=0.80–0.96
v d Heijde/Sharp (1989)	G coeff=0.97–0.98				ICC=0.96						r=0.84–0.89 ICC=0.83–0.86	
Larsen (1977) Larsen (1995) Scott/Larsen (1995)		ICC=0.88	r=0.92	ICC=0.94–0.98		r=0.75		r=0.93	r=0.91–0.99	ICC=0.88 ICC=0.90	r=0.82 ICC=0.84–0.85	r=0.39–0.69
SENS (1999) C: MC Kellgren	G coeff=0.89–0.98	ICC=0.97	r=0.98		ICC=0.92				r=0.91–0.95			

*r = correlation coefficients; ICC = intraclass correlation; E = erosions; JSN = joint space narrowing; tot = total; [ ] = study reference.

Spearman and Pearson correlation coefficients measure only association and not agreement.[44]

The most detailed methods, such as Sharp, are most reliable. The level of intrarater reliability is higher than those of interrater reliability and reliability for progression. Most studies in this area compared the Sharp and the Larsen methods. Plant *et al* compared the Sharp, Larsen, and C:MC methods.[38] Expressed as percentages of the maximum score, Sharp and Larsen scores showed substantial changes in the first two years (negligible for C:MC). The Spearman correlation coefficients for the three methods were similar: intrarater and interrater reliability ranged from 0.90 to 0.99; reproducibility of changes in scores between successive films was better with Sharp than with Larsen. The κ values were higher for JSN than erosions or Larsen and better for intrarater agreement. In conclusion, the Sharp and Larsen methods performed similarly in early RA, but the Sharp system reproducibility was better for change in score. Similar results were found in an other study.[45]

This was confirmed when Guth *et al* compared the same three methods using ICC and Bland and Altman graphics (see below).[36] The results for intrarater reliability were similar to those in the earlier study. Paimela *et al* compared the Sharp, Larsen, and Larsen's modified (especially for long term studies) methods.[39] Reliability assessed by ICC exceeded 0.88 and 0.94 for interrater and intrarater agreement, respectively. When van der Heijde scored radiographs of 20 patients,[19] she found ICC (as defined by Streiner and Norman) of 0.99 for Sharp/van der Heijde and 0.98 for SENS. This indicated high intrarater reliability for both methods during the first five years. Expressed as percentages of maximum scores, the means of SENS were higher than the means of Sharp/van der Heijde. Wassenberg *et al* compared the Larsen, Sharp, Sharp/van der Heijde, and the Ratingen score methods.[43] He determined intrarater reliability using the quotient of intrapatient SD (standard deviation) to intrarater SD. Higher values indicated better results. As expected, the most detailed methods—that is, Sharp and Sharp/van der Heijde, were more reliable, followed by the Ratingen score and, finally, the Larsen method.

Rau *et al* studied the reliability of their method[26] and found it to range from 0.70 to 0.90, with higher values for intrarater reliability. The ratio of radiographic change SD to intra- or interrater SD defined the reliability over time. If this is equal to 1, detected radiographic change is due to measurement error; if it is over 1, the method can detect true change from one point to the next, with higher values corresponding to higher sensitivity. The ratio ranged from 1.7 to 3.2, indicating that the method could detect real radiographic change over time.

A graphical method described by Bland and Altman is used to assess reliability between two methods[46] by plotting the difference in progression between the two observers for each method. The y axis represents the difference in progression as assessed by the two observers,

Table 6 Sensitivity to change: comparison between scoring methods

	Sensitivity to change						
Method	*[19]**	*[26]*	*[39]*	*[43]*	*[48]*	*[49]*	*[28]*
Sharp (1971)							
Sharp (1985)			SRM*=0.72	MDC*=2.3%	SRM ratio=1.44		
v d Heijde/Sharp (1989)	SDD*=7 to 24			MDC=2.3%	SRM ratio=1.70	SRM=0.81–1.06	
Larsen (1977)		MDC=6.6%–9.6%	SRM=0.80	MDC=3.2%	SRM ratio=1 (as reference)		
Larsen (1995)			SRM=0.88				ES*=0.34
Rau/Larsen (1995)					SRM ratio=1.25		
Ratingen (1998)		MDC=3.3%–4.6%		MDC=3.3%			
SENS (1999)	SDD=4 to 6				SRM ratio=0.66		
Kellgren						SRM=0.82–1.12	
SES (2000)							ES=0.38

*SRM = standardised response mean; MDC = minimal detectable change; SDD = smallest detectable difference; ES = effect size; [ ] = study reference.

and the x axis the mean of progression as assessed by the two observers. The ideal situation would be for all points to be situated on or close to y=0. A confidence interval is calculated using a normal approximation. This method determines whether reliability is sufficient and whether one reader scores higher than the other, thus providing an absolute and metric estimate of random measurement error. This additional method was used in some studies.[31 32 36]

SENSITIVITY TO CHANGE (TABLE 6)

Sensitivity to change is measured to determine whether a radiographic scoring method can detect a real change over time. Several authors have compared the sensitivity to change of various methods. Some used the standardised response mean (SRM), whereas others used the minimal detectable change (MDC) (also called the smallest detectable difference (SDD)), or the G coefficient. SRM is defined as a unitless expression of change and is calculated by dividing the mean of the difference between scores at two times by the SD of change score. A value above 0.80 is considered to reflect high potential to detect changes. The G coefficient is also defined as a unitless expression of change, but corresponds to the ratio of the signal SD to the noise SD derived from a mixed effect analysis of variance model.[32] It can range from 0 to 1, and values above 0.80 are considered to be good.

Rau *et al* determined that MDC = 3.3% of the maximum score (intrarater) and MDC = 4.6% of the maximum score (interrater) for their method. Seven points were scored in 20 patients: 14 patients selected randomly from 128 subjects in a prospective trial were followed up over five years, and six patients with severe progressive disease were followed up over 10 years in an outpatient department. MDC was defined by $1.96 \times 2 \times$ intra- (or inter-)rater SD with 95% confidence if the variances were from a normal distribution. Below the value of the MDC, the difference between two readings of the same set will lie with a probability of 97.5%. A difference above this value confirms true change in the radiographs. They found that the MDC was substantially better with the new method than published data gathered using the Larsen method.[26]

The sensitivity to change of the Sharp/van der Heijde method using the G coefficient was shown by van der Heijde to range from 0.22 to 0.39[32] (10 patients, one year follow up). In another study the same author determined the SRM of SENS (1.15, 1.63, and 1.60 for sequential, paired, and random order, respectively).[47] Paimela *et al* calculated SRM by dividing the mean change between the baseline and year 1 scores by the SD of change score in 83 patients.[39] Their result was 0.80 for the Larsen method, 0.88 for the variant Larsen method (1995), and 0.72 for the Sharp method. The variant Larsen method was the most responsive system, but the differences between the three methods were small. These findings indicated that all three methods were sensitive to change during the first year of RA.

When Wassenberg used the same formula as Rau to calculate MDC[43] the results were as follows: 2.3% for the Sharp and Sharp/van der Heijde methods, 3.2% for Larsen, and 3.3% for Ratingen (20 patients). All results were expressed as a percentage of maximum scores. Smaller values indicate better precision. In a comparison between SENS and Sharp/van der Heijde, van der Heijde used the Norman's quasi-classical ICC formula and SDD to calculate the sensitivity to change.[19] She found G coefficient = 0.88 for SENS and 0.84 for Sharp/van der Heijde modification. From derived results of SDD, it was concluded that SENS was approximately as good as Sharp/van der Heijde in detecting progression.

More recently, Guillemin *et al* compared scoring methods (Sharp, Sharp/van der Heijde, Larsen, Rau/Larsen, and SENS methods) using hands radiographs of 20 patients at two different times by applying adjusted SRM.[48] As elsewhere, adjusted SRM was computed as the mean of the difference between scores at two times over its standard deviation. The difference between this value and "classical" SRM lies in the adjustment, based on a mixed model, for fixed rater and country effects, and the variance was adjusted according to interpatient variability. When the Larsen method was used as reference, the ratios between SRMs showed that all other methods (excluding SENS, but restricted to the hands) were more sensitive to change than the reference (van der Heijde/Sharp over Sharp over Rau/Larsen).

Drossaers-Bakker *et al* compared three scoring systems for the long term assessment of

RA.[49] Sensitivity to change was calculated using SRM for the van der Heijde/Sharp, Kellgren and the "Sharp max" scores. To look for a possible ceiling effect, the Sharp max score consists of the extension of the erosive score for each hand joint to 10 (rather than 5), as for the erosive score for the feet. The SRM invariably exceeded 0.80. SRM values over 0.20, 0.50, and 0.80 are considered to reflect low, moderate, and high potential to detect changes, respectively. The Kellgren and Sharp scores appeared to be equally sensitive to change over time in early RA, whereas the Kellgren score proved to be more sensitive late in the disease.

Wolfe *et al* compared the ability of the SES score and the Larsen score to distinguish change over time.[28] Two radiographs were randomly selected from each of 857 patients who had at least two radiographs over an average three years' follow up. They calculated for the two methods an effect size (that is, standardised change score) defined by the ratio of the difference between the two films to the pooled standard deviation. They found 0.38 and 0.34 for SES and Larsen, respectively. The SES is as sensitive at detecting change as the Larsen scale.

Conclusions

The aim of the present review, which may be considered exhaustive, was to elucidate the major issues to be considered in the radiographic evaluation of RA development. The findings indicate that:

- Erosion and JSN are suitable measures of RA severity and progression, and can be used to provide separate or combined scores
- Changes at the PIP, MCP, MTP, IP, and joints of the wrist provide a good indication of the evolution of RA (in particular, more information is gained when MCP and PIP are considered)
- Posteroanterior radiographs of joints should be performed
- Hand and foot *x* rays capture the early process of RA development (erosions of MTP and then other joints)
- Having films read by two raters is the best compromise; readers should invariably be trained
- The scoring method should be chosen according to the objective of the study:
- The time taken for a method, and the differences it detects, depend on its degree of detail
- The degree of reliability should preferably be assessed by ICC. Bland and Altman graphics provide excellent visual information
- The sensitivity to change of a method should be compared with those of other techniques, preferably using SRM.

Other issues are less clear. In particular, radiographic production and reproduction are the most random factors; however, new digital technologies may help to standardise imaging.[50 51] All the available reading strategies have advantages and disadvantages. The choice is for investigators to make, but the technique selected must be specified in order to facilitate the interpretation of results.

We thank L Billot for his advice and help with the editing of various passages, including "statistical methods", and we thank William Francis for his pertinent queries while editing the English language of a previous version of this manuscript.

Source of support: Programme Hospitalier de Recherche Clinique 1995 from the French Ministry of Health.

1 Sharp JT. Radiologic assessment as an outcome measure in rheumatoid arthritis. Arthritis Rheum 1989;32:221–9.
2 Brower AC. Use of the radiograph to measure the course of rheumatoid arthritis. The gold standard versus fool's gold. Arthritis Rheum 1990;33:316–24.
3 Van der Heijde D. Plain X-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. Baillieres Clin Rheum 1996;10:435–53.
4 Van der Heijde D, Boers M, Lassere M. Methodological issues in radiographic scoring methods in rheumatoid arthritis. J Rheumatol 1999;26:726–30.
5 Steinbrocker O, Traeger C, Batterman R. Therapeutic criteria in rheumatoid arthritis. JAMA 1949;140:659–62.
6 Kellgren J, Bier F. Radiological signs of rheumatoid arthritis: a study of observer differences in the reading of hand films. Ann Rheum Dis 1956;15:55–60.
7 Sharp JT, Lidsky MD, Collins LC, Moreland J. Method of scoring the progression of radiologic changes in rheumatoid arthritis. Arthritis Rheum 1971;14:706–20.
8 Sharp JT, Young DY, Bluhm GB, Brook A, Brower AC, Corbett M, *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 Fries JF, Bloch DA, Sharp JT, McShane DJ, Spitz P, Bluhm GB, *et al*. Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial. Arthritis Rheum 1986;29:1–9.
11 Genant HK. Methods of assessing radiographic change in rheumatoid arthritis. Am J Med 1983;75:35–47.
12 Genant HK, Jiang Y, Peterfy C, Lu Y, Redei J, Countryman PJ. Assessment of rheumatoid arthritis using a modified scoring method on digitized and original radiographs. Arthritis Rheum 1998;41:1583–90.
13 Jiang Y, Genant H, Watt I, Cobby M, Bresnihan B, Aitchison R, *et al.* A multicenter, double-blind, dose-ranging, randomised, placebo-controlled study of recombinant human interleukin-1 receptor antagonist in patients with rheumatoid arthritis. Arthritis Rheum 2000;43: 1001–9.
14 Nance EPJ, Kaye JJ, Callahan LF, Carroll FE, Winfield AC, Earthman WJ, *et al*. Observer variation in quantitative assessment of rheumatoid arthritis. Part I. Scoring erosions and joint space narrowing. Invest Radiol 1986;21:922–7.
15 Kaye JJ, Nance EPJ, Callahan LF, Carroll FE, Winfield AC, Earthman WJ, *et al*. Observer variation in quantitative assessment of rheumatoid arthritis. Part II. A simplified scoring system. Invest Radiol 1987;22:41–6.
16 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;i:1036–8.
17 Van der Heijde DMFM, Van Leeuven MA, Van Riel PL, Koster AM, Van't Hof MA, Van Rijswijk MH, *et al*. Biannual radiographic assessments of hands and feet in a three year prospective followup of patients with early rheumatoid arthritis. Arthritis Rheum 1992;35:26–34.
18 Van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 1999;26: 743–5.
19 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.
20 Larsen A, Dale K, Eek M. Radiographic evaluation of rheumatoid arthritis and related conditions by reference films. Acta Radiol Diagn 1977;18:481–91.
21 Larsen A, Edgren J, Harju E, Laasonen L, Reitamo T. Interobserver variation in the evaluation of radiologic changes of rheumatoid arthritis. Scand J Rheumatol 1979;8:109–12.
22 Larsen A. How to apply Larsen score in evaluating radiographs of rheumatoid arthritis in long-term studies. J Rheumatol 1995;22:1974–5.
23 Scott DL, Houssien DA, Laasonen L. Proposed modification to Larsen's scoring methods for hand and wrist radiographs. Br J Rheumatol 1995;34:56
24 Edmonds J, Saudan A, Lassere M, Scott DL. Introduction to reading radiographs by the Scott modification of the Larsen method. J Rheumatol 1999;26:740–2.
25 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.
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 Trentham DE, Masi AT. Carpo:metacarpal ratio. A new quantitative measure of radiologic progression of wrist involvement in rheumatoid arthritis. Arthritis Rheum 1976;19:939–44.
28 Wolfe F, van der Heijde DMFM, Larsen A. Assessing radiographic status of rheumatoid arthritis: introduction of a short erosion scale. J Rheumatol 2000;27:2090–9.
29 Sharp JT, Bluhm GB, Brook A, Brower AC, Corbett M, Decker JL, *et al*. Reproducibility of multiple-observer scoring of radiologic abnormalities in the hands and wrists of patients with rheumatoid arthritis. Arthritis Rheum 1985; 28:16–24.
30 Wassenberg S, Rau R. Problems in evaluating radiographic findings in rheumatoid arthritis using different methods of radiographic scoring: examples of difficult cases and study design to develop an improved scoring method. J Rheumatol 1995;22:1990–7.
31 Lassere M, Boers M, Van der Heijde D, Boonen A, Edmonds J, Saudan A, *et al*. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731–9.
32 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.
33 Sharp JT. An overview of radiographic analysis of joint damage in rheumatoid arthritis and its use in metaanalysis. J Rheumatol 2000;27:254–60.
34 Scott DL, Coulton BL, Bacon PA, Popert AJ. Methods of X-rays assessment in rheumatoid arthritis: a re-evaluation. Br J Rheumatol 1985;24:31–9.
35 Cuchacovich M, Couret M, Peray M, Gatica H, Sany J. Precision of the Larsen and the Sharp methods of assessing radiologic change in patients with rheumatoid arthritis. Arthritis Rheum 1992;35:736–9.
36 Guth A, Coste J, Chagnon S, Lacombe P, Paolaggi J. Reliability of three methods of radiologic assessment in patients with rheumatoid arthritis. Invest Radiol 1995;30: 181–5.
37 Pincus T, Larsen A, Brooks RH, Kaye J, Nance P, Callahan LF. Comparison of 3 quantitative measures of hand radiographs in patients with rheumatoid arthritis: Steinbrocker stage, Kaye modified sharp score, and Larsen score. J Rheumatol 1997;24:2106–12.
38 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.
39 Paimela L, Laasonen L, Helve T, Leirisalo-Repo M. Comparison of the original and the modified Larsen methods and the Sharp method in scoring radiographic progression in early rheumatoid arthritis. J Rheumatol 1998; 25:1063–6.
40 Mewa AAM, Pui M, Cockshott WM, Buchanan WW. Observer differences in detecting erosions in radiographs of rheumatoid arthritis. A comparison of posterioanterior, Norgaard and Brewerton views. J Rheumatol 1983;10: 216–21.
41 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.
42 Ferrara R, Priolo F, Cammisa M, Bacarini L, Cerase A, Pasero G, *et al*. Clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from GRISAR study. Ann Rheum Dis 1997;56:608–12.
43 Wassenberg S, Herborn G, Larsen A, Sharp JT, van der Heijde DMFM, Wijnands M, *et al*. Reliability, precision and time expense of four different radiographic scoring methods [abstract]. Arthritis Rheum 1998;41(suppl): S104.
44 Nunnally JC. *Psychometric theory*. New York: McGraw-Hill; 1978:701pp.
45 de Carvalho A. Discriminative power of Larsen's grading system for assessing the course of rheumatoid arthritis. Acta Radiol Diagn 1981;22:77–80.
46 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
47 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;40:S287.
48 Guillemin F, Oedegaard S, Gérard N, Billot L, Boini S, Kvien TK. Reproducibility and sensitivity to change of 5 scoring methods for hand X-ray damage in rheumatoid arthritis [abstract]. Ann Rheum Dis 2000;59(suppl I):214.
49 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.
50 Sharp JT, Gardner JC, Bennett EM. Computer-based methods for measuring joint space and estimating erosion volume in the finger and wrist joints of patients with rheumatoid arthritis. Arthritis Rheum 2000;43:1378–86.
51 Solymossy C, Dixey J, Utley M, Gallivan S, Young A, Cox N, *et al*. Larsen scoring of digitized X-ray images. Rheumatology (Oxford) 1999;38:1127–9.