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)

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 and Kellgren), whereas others assess individual joints (Sharp and its variants, Larsen and its variants, and Simple Erosion Narrowing Score (SENS)).

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. Twenty-nine areas in each hand and wrist are considered for erosions, 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 1983 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, radio-ulnar 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 88.

In 1986 another modification was devised by Fries et al (with the participation of Sharp). 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 1998, Genant et al modified their method as follows: erosion is scored according to an eight point scale with 0.5 increments, where 0 = normal; 0.5 = questionable or subtle change; 1 = mild; 1.5 = mild worse; 2 = moderate; 2.5 = moderate worse; 3 = severe; and 3.5 = severe worse. In each hand, IP of the thumb, PIP, MCP, CMC 3 to 5, carpal-striquetrum, lunate-triquetrum, capitate-scaphoid-lunate, and the radiocarpal joint 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; 0.5 = questionable or subtle change; 1 = mild; 1.5 = mild worse; 2 = moderate; 2.5 = moderate worse; 3 = severe; 3.5 = severe worse; 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.

Kaye et al combined the methods described by Genant and Sharp et al. Two scoring systems are used in this approach. The more detailed of the two 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. The simplified system 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 foot (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 foot, 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 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 foot. The score ranges from 0 to 250.

In 1995, Larsen devised a method to evaluate radiographs in long term studies.21 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.22–24 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 there 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 totaling destruction of ≤20% of the total surface;
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) 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. They are highly specific to RA, can be reliably assessed, and provide independent information. The relative weight given to erosion versus JSN varies, and no consensus has yet been established. 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. 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 formation can be counted as erosion. Subluxation related to cyst formation is not captured by Sharp or Larsen methods. The SES method 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.

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. 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. The SES method developed by Wolfe targets this goal. Other methods use a count of abnormalities on a continuous scale.
### Table 2  Joints counted in the different radiographic scoring methods with separate scores

<table>
<thead>
<tr>
<th>Method</th>
<th>Hand</th>
<th>Wrist</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIP*</td>
<td>PIP*</td>
</tr>
<tr>
<td>Sharp (1971)</td>
<td>E*</td>
<td>v</td>
</tr>
<tr>
<td>JSN*</td>
<td></td>
<td>v</td>
</tr>
<tr>
<td>Genant (1998)</td>
<td>E</td>
<td>v</td>
</tr>
<tr>
<td>Kaye/Sharp (1986)</td>
<td>E</td>
<td>v</td>
</tr>
<tr>
<td>van der Heijde/Sharp (1989)</td>
<td>E</td>
<td>v</td>
</tr>
<tr>
<td>SENS (1999)</td>
<td>E</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>JSN*</td>
<td>v</td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th>Method</th>
<th>Hand</th>
<th>Wrist</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIP*</td>
<td>PIP*</td>
</tr>
<tr>
<td>Larsen (1977)</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>Larsen (1985)</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>Scott/Larsen (1995)</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>Rau/Larsen (1995)</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>Ratingen (1998)</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>SBS (2000)</td>
<td>v</td>
<td>v</td>
</tr>
</tbody>
</table>

*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. 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. Plant et al justified extending the Sharp method to the joints of the foot 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; 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, radiolunar and lunatricketrum 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. 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.

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, the authors found no significant advantage in the Norgaard or the Brewerton view compared with the standard posteroanterior view used in most investigations. 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. 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). 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). Using several statistical methods, Ferrara et al produced results suggesting that paired reading was the most suitable for evaluating the progression of joint damage. Recently, van der Heijde et al devised two studies using the Sharp/van der Heijde method to score radiographs of hands and feet: 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Intrarater reliability</th>
<th>Intrarater reliability for progression</th>
<th>Interrater reliability</th>
<th>Interrater reliability for progression</th>
<th>Sensitivity to change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharp (1971)</td>
<td>$r^* = 0.97$ (E*)</td>
<td>$r = 0.53–0.96$ (E)</td>
<td>$r^* = 0.95$ (JSN*)</td>
<td>$r = 0.58–0.95$ (JSN)</td>
<td></td>
</tr>
<tr>
<td>Sharp (1985)</td>
<td></td>
<td>$r^* = 0.95$ (tot*)</td>
<td></td>
<td>$r = 0.75–0.96$ (JSN)</td>
<td></td>
</tr>
<tr>
<td>Kaye/Sharp (1986)</td>
<td></td>
<td>$r^* = 0.95–0.96$ (E)</td>
<td></td>
<td>$r = 0.53–0.96$ (E)</td>
<td></td>
</tr>
<tr>
<td>Genant (1983)</td>
<td></td>
<td>$r^* = 0.92–0.95$ (E)</td>
<td></td>
<td>$r = 0.91–0.96$ (E)</td>
<td></td>
</tr>
<tr>
<td>Genant (1998)</td>
<td>$r^* = 0.90–0.95$ (E)</td>
<td>$r = 0.90–0.98$</td>
<td>$r^* = 0.72–0.93$ (E)</td>
<td>$r = 0.94–0.96$</td>
<td>$r = 0.80–0.96$ (JSN)</td>
</tr>
<tr>
<td>Scott/Larsen (1995)</td>
<td>$r = 0.90–0.94$ (E)</td>
<td>$r = 0.90–0.98$</td>
<td>$r^* = 0.93–0.97$ (E)</td>
<td>$r = 0.94–0.96$</td>
<td>$r = 0.92–0.95$ (E)</td>
</tr>
<tr>
<td>Larsen (1977)</td>
<td></td>
<td>$r = 0.92$ and ICC = 0.94</td>
<td></td>
<td>$r = 0.94–0.95$</td>
<td></td>
</tr>
<tr>
<td>Ratingen (1998)</td>
<td></td>
<td>$r = 0.90–0.96$ (E)</td>
<td></td>
<td>$r = 0.86–0.96$</td>
<td></td>
</tr>
<tr>
<td>SENS (1999)</td>
<td>ICC $^* = 0.79–0.90$</td>
<td>$r = 0.92$</td>
<td>ICC $^* = 0.69–0.80$</td>
<td>$r = 0.78–0.90$</td>
<td>$r = 0.90–0.96$ (JSN)</td>
</tr>
</tbody>
</table>

* $r^*$ = 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.
Spearman and Pearson correlation coefficients measure only association and not agreement.\textsuperscript{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.\textsuperscript{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 $\kappa$ 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.\textsuperscript{45}

This was confirmed when Guth et al compared the same three methods using ICC and Bland and Altman graphics (see below).\textsuperscript{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.\textsuperscript{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,\textsuperscript{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 percentage 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.\textsuperscript{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\textsuperscript{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\textsuperscript{28} 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity to change</th>
<th>[19]*</th>
<th>[26]</th>
<th>[39]</th>
<th>[43]</th>
<th>[48]</th>
<th>[49]</th>
<th>[28]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharp (1971)</td>
<td>SRM=0.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sharp (1985)</td>
<td>SRM=0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v d Heijde/Sharp (1989)</td>
<td>SRM=0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larsen (1977)</td>
<td>SRM=0.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratingen (1998)</td>
<td>SRM=0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rau/Larsen (1995)</td>
<td>SRM=0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SENS (1999)</td>
<td>SRM=0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SES (2000)</td>
<td>SRM=0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*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.

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. 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 (inter-rater) 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 x 2 x 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.

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 (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 SENS was approximately as good as Sharp/van der Heijde in detecting progression. When Wassenberg used the same formula as Rau to calculate MDC they 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. 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, Guillenin 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. 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. 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. 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 andPIP 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.

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.


Methods for scoring radiographs in RA


