Reliability and sensitivity to change of a simplification of the Sharp/van der Heijde radiological assessment in rheumatoid arthritis

D. van der Heijde, T. Dankert, F. Nieman, R. Rau\(^1\) and M. Boers\(^2\)

University Hospital Maastricht, Maastricht, The Netherlands, \(^1\)Evangelisches Fachkrankenhaus, Ratingen, Germany and \(^2\)Free University, Amsterdam, The Netherlands

Abstract

Objective. To determine the reliability and sensitivity to change of a simplified radiological scoring method [simple erosion narrowing score (SENS)] for rheumatoid arthritis (RA). SENS was compared to the Sharp/van der Heijde score (SHS) as a gold standard.

Methods. Sets of seven radiographs of hands and feet were taken of 20 RA patients with a wide spectrum of radiological damage. For 14 patients, these seven radiographs were taken during a follow-up period of 5 yr, and for six patients during a follow-up of 10 yr. Each set of radiographs was scored twice by the same observer (DvdH). Erosions and joint space narrowing were scored with SHS (range 0–448) in 32 and 30 joints in the hands, respectively, and both in 12 joints in the feet. SENS gives a score of 1 if there is any erosion in a joint and also 1 if there is any narrowing in the joint (range 0–86). In each case, SENS was derived from SHS. To analyse data, generalizability theory and repeated measurements ANOVA were used.

Results. The overall reliability coefficient was 0.81 for SHS and 0.80 for SENS. Intra-observer reliability [intraclass correlation coefficient (ICC)] was 0.99 and 0.98 for SHS and SENS, respectively. The ICC for the sensitivity to change was 0.84 for SHS and 0.88 for SENS. The smallest detectable difference (SDD) could be determined for both methods. The presence of progression based on this SDD was very comparable between the two methods.

Conclusion. The measurement properties of SENS are good and comparable to SHS. This makes SENS suitable for use in clinical practice and in large (epidemiological) studies, especially in the first years of disease.

Key Words: Rheumatoid arthritis, Radiological assessment, Sharp/van der Heijde score, Simplification, Reliability, Sensitivity to change.

The chronic inflammation of rheumatoid arthritis (RA) leads to swelling of soft tissue, and to lesions in articular cartilage and subchondral bone. Radiographs are important to assess the progression of joint damage over time. Hands and feet are most frequently imaged for several reasons: joints in these sites are affected in most patients with RA, changes in these small joints correlate with the abnormalities in large joints, and changes can be scored reliably \([1, 2]\). Erosions and joint space narrowing (JSN) are the abnormalities which should be scored because both give independent information, are caused by RA and indicate damage \([3, 4]\).

Several methods are available to score progressive joint damage. Some methods give a global assessment for the whole patient. Other methods give a global joint score and again others score specific joint abnormalities \([1, 4–8]\). The methods used most widely are the ones developed by Larsen and Sharp \([1, 4, 6, 7]\). Modified versions of both methods have been developed to overcome some disadvantages of the original methods \([9, 10]\). In particular, one of the authors (DvdH) has modified Sharp’s method to include the joints of the feet \([9]\). Several studies have shown higher intra- and interobserver reliability, and a higher sensitivity to change in Sharp’s method than in Larsen’s method \([11–13]\). Larsen’s method has the advantage of being less time consuming than Sharp’s method \([13]\). The disadvantage of both methods is that they require trained observers. For clinical trials, this disadvantage generally is no problem. However, in clinical practice, a less time-consuming and simplified method, with adequate reliability and sensitivity to change, would be desirable.

We propose a simplified method of scoring radiographs for clinical practice based on the Sharp/van der Heijde score (SHS): instead of grading, the number of
joints with erosions and the number of joints with JSN are simply summed [simple erosion narrowing score (SENS)]. Similar suggestions have been made in the literature, but have until now never been fully validated [1, 15]. We have tested the simplified method against the reference standard (SHS) in several ways. First, we compared the reproducibility (intra-observer consistency) and the sensitivity to change. Next, we investigated how often progression was seen with SENS and not with SHS (false positive) and vice versa (false negative). We also determined how much the patient’s score with SENS had to increase to measure progression reliably. Finally, we checked for a ceiling effect in progression of SENS scores.

Patients and methods

Patients

Seven hand and foot radiographs were available in 20 patients fulfilling the ACR criteria (1987). These patients showed a large spectrum of radiographic damage. Fourteen patients were recruited through probability sampling from a pool of 128 patients who participated in a clinical trial comparing methotrexate and parenteral gold. The probability of a patient being selected was higher if they showed more damage after 5 yr in an earlier evaluation. These 14 patients were followed for 5 yr: three radiographs at half-yearly intervals and four radiographs yearly thereafter. Another six patients were selected because they showed severe progressive disease and were followed for 10 yr: three radiographs at yearly intervals and four radiographs each second year thereafter.

Radiographic analysis

Radiographs were made in posteroanterior view and scored twice by the same trained observer (DvdH) according to a randomization list including 40 sets of radiographs (each set twice). Per patient, the radiographs were scored in chronological order. The observer was unaware of each patient’s identity. The number of erosions and JSN were scored according to SHS [9]. In SHS, erosions are counted in the 10 metacarpophalangeal (MCP) joints, the eight proximal interphalangeal (PIP) joints, the two interphalangeal joints of the thumbs, the right and left first metacarpal bone, the right and left radius and ulnar bones, the right and left trapezium and trapezoid (as one unit; multangular), right and left navicular bones, right and left lunate bones, the 10 metatarsophalangeal (MTP) joints, and the two interphalangeal joints of the big toes. Erosions are assessed in the 10 MCP joints, the eight PIP joints, right and left third, fourth and fifth carpometacarpal joints, right and left multangular-navicular joints, right and left capitate-navicular-lunate joints, right and left radiocarpal joints, the 10 MTP joints and the two interphalangeal joints of the big toes. Erosions are scored 1 if there is a discrete interruption of the cortical surface; if there is a larger defect, a score is given according to the surface of the joint involved [2–5]. Consequently, for confluent erosions, the score cannot decrease. In the hands, the maximum erosion score in a joint is 5: in the feet, it is 10.

For JSN, five grades are recognized: 0 = normal, 1 = focal or doubtful; 2 = general, <50% of the original joint space; 3 = general, >50% of the original joint space or subluxation; 4 = ankylosis. If a joint cannot be scored correctly, e.g. because of previous surgery, the last score of the joint is carried forward. The maximum number of erosions is 160 in the hands and 120 in the feet; and the maximum scores for JSN are 120 and 48, respectively. The total score is the sum of scores for erosions and JSN. The maximum total score is 448.

SENS assesses the same joints. The sites that are included in both SHS and SENS are shown in Fig. 1. In SENS, a joint is scored as affected (‘1’) if there is any erosion in the joint. A joint is scored as affected (‘1’) for JSN if the joint is scored 1 or more in the original method, this means at least focal JSN. So per joint, the score can range from 0 to 2. The number of joints in which erosions can be scored is 32 in the hands and 12 in the feet; the number of joints in which JSN can be scored are 30 and 12, respectively. Therefore, the maximum total score of SENS per patient is 86. In this study, the scores for SENS were deducted from the SHS scores. However, we also scored 12 films directly with SENS and compared these direct scores with those derived from the SHS scores.

For both methods (SHS and SENS), data from joints in hands were grouped together as one source of measurement information; data from joints in feet were separately grouped as another source of information. At the same time, data from erosions can be seen as one source of information and data from JSN as another. Relevant patient information consists of a cross-section of these four sources, i.e. erosions in hands, erosions in feet, etc. These four types of partial patient information can be found as models 1–4 in Table 1. Sources of partial information can be combined separately (like presented in models 5–9) or scores can be summed and so generate new data concerning, for example, both abnormalities in hands (model 10) or erosions in both hands and feet (model 12). Next, summations can be combined separately (model 14 and 15) or can, again, be summed for the whole patient (model 16). Models containing sources of relevant patient information can next be tested for reliability (reproducibility). We compared reliability results of both methods for each of the 16 models separately.

To be able to compare the scores of the methods for the various models, the percentage of the scores with respect to the maximum score had to be determined. This percentage of the maximum score was determined by dividing the actual score by the maximum possible score for each model.

Statistical analysis

Reliability. Reliability was tested by using generalizability theory, a random model ANOVA approach which estimates the components of variance within each
model [16]. We have used the computer program GENOVA for PCs by Crick and Brennan, which is especially suited for calculating random model variance components within analysis of variance [17, 18]. Elementary sources of variance in data are called facets in generalizability theory. Relevant facets in this study are: method (SHS vs SENS), patient identification number (1–20), type of abnormality (erosion, JSN), extremity (hands, feet), time (1–7) and number of observations (occasion 1, occasion 2). In generalizability theory, a distinction is made between fixed and random facets. The facets ‘patient’, ‘time’ and ‘number of observations’ were defined as random facets, the others as fixed facets. The facet ‘time’ was nested within the facet ‘patient’ because of unequal spacing of the radiographs (14 patients were followed for 5 yr and six patients were followed for 10 yr). The overall reliability coefficient over all facets is called the $\phi$-coefficient, showing the reliability of the methods with all the sources of variance included. Theoretically, the $\phi$-coefficient ranges from 0 (not reliable at all) to 1.00 (maximum).

So-called decision studies were made within some of the 16 models (Table 1) of both methods to estimate the number of observations needed for a specific level of intra-observer reliability. Next to this, the intraclass correlation coefficients (ICC) of some of the models (1–4, 9, 14–16) were calculated to determine intra-observer reliability. These ICC are not similar to the classical definition of ICC. The ICC calculated in this study are called G-coefficients as defined by Streiner and Norman [19]. We retained the term ICC to indicate that the results are comparable to the classical ICC.

Sensitivity to change. This was measured by estimating relevant ratios of variance components from results of

---

**Table 1. The 16 models of SHS and of SENS**

<table>
<thead>
<tr>
<th>Model</th>
<th>Source of information</th>
<th>Number of parameters</th>
<th>Maximum total score SHS</th>
<th>Maximal total score SENS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Erosions in hands</td>
<td>1</td>
<td>160</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>JSN in hands</td>
<td>1</td>
<td>120</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Erosions in feet</td>
<td>1</td>
<td>120</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>JSN in feet</td>
<td>1</td>
<td>48</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>Erosions in hands, JSN in hands</td>
<td>2</td>
<td>160/120</td>
<td>32/30</td>
</tr>
<tr>
<td>6</td>
<td>Erosions in feet, JSN in feet</td>
<td>2</td>
<td>120/48</td>
<td>12/12</td>
</tr>
<tr>
<td>7</td>
<td>Erosions in hands, erosions in feet</td>
<td>2</td>
<td>160/120</td>
<td>32/12</td>
</tr>
<tr>
<td>8</td>
<td>JSN in hands, JSN in feet</td>
<td>2</td>
<td>120/48/30/12</td>
<td>30/12</td>
</tr>
<tr>
<td>9</td>
<td>Erosions in hands, JSN in hands, JSN in feet</td>
<td>4</td>
<td>160/120/120/48/32/30/12</td>
<td>32/30/12/12/12</td>
</tr>
<tr>
<td>10</td>
<td>Total hands</td>
<td>1</td>
<td>280</td>
<td>62</td>
</tr>
<tr>
<td>11</td>
<td>Total feet</td>
<td>1</td>
<td>168</td>
<td>24</td>
</tr>
<tr>
<td>12</td>
<td>Total erosions</td>
<td>1</td>
<td>280</td>
<td>44</td>
</tr>
<tr>
<td>13</td>
<td>Total JSN</td>
<td>1</td>
<td>168</td>
<td>42</td>
</tr>
<tr>
<td>14</td>
<td>Total hands, total feet</td>
<td>2</td>
<td>280/168</td>
<td>62/24</td>
</tr>
<tr>
<td>15</td>
<td>Total erosions, total JSN</td>
<td>2</td>
<td>280/168</td>
<td>44/42</td>
</tr>
<tr>
<td>16</td>
<td>Total patient</td>
<td>1</td>
<td>448</td>
<td>86</td>
</tr>
</tbody>
</table>

JSN, joint space narrowing.
mixed model repeated measurements ANOVA. For each method, Norman’s quasi-classical ICC formula for sensitivity to change was calculated with the components of ‘time’ (T) and ‘patient by time’ (P × T) [ICC = VC(T)/(VC(T) + VC(P × T)) where VC stands for variance component]. To obtain equal time periods between radiographs, so ‘time’ could be considered as a fixed facet, the six patients who had been followed for 10 yr were excluded from this analysis.

Smallest detectable difference (SDD). With the results of GENOVA, the SDD can also be calculated for the sample of 14 patients with 5 yr follow-up. This can be done for SHS and SENS separately. The variance component EMS, i.e. expected mean sums of squares of the facet time crossed with patient (T × P), is needed. The square root of this EMS gives the standard error of measurement (SEM). To decide whether there is real progression or no progression at all, one-sided testing is sufficient; because of paired observations, the results should be multiplied by \(\sqrt{2}\). For a 90% confidence interval in 14 patients, the normal range Z score is 1.282 for two radiographs. This results in the following formula: SDD = EMS(T × P) \times 1.282 \times \sqrt{2} [16]. The above formula is valid if the SDD is based on the information of two (successive) radiographs. For the analyses to determine the SDD, separate analyses were performed for each pair of successive radiographs, to obtain results that will be valid if you have two radiographs only, without the information of the complete series of seven radiographs.

Sensitivity and specificity of progression. The SDDs for SENS and SHS were used as limits to determine whether a patient showed real progression as assessed by that method. Two by two (2 × 2) contingency tables were made to assess the sensitivity, specificity and accuracy of SENS compared to the gold standard SHS. These tables were created for every period between two radiographs. The kappa statistic was calculated as a measure of agreement between SHS and SENS.

Results

The mean age of the patients at baseline was 53.5 yr (s.d. 13.0) with a mean disease duration of 2.7 yr for all 944 patients and 11 months for the 14 patients with 5 yr follow-up. Sixty-seven per cent of the patients were female and also 67% were rheumatoid factor positive. The mean sedimentation rate was 44.5 mm/h (s.d. 26.4), the mean number of swollen joints 18 (s.d. 9.7; 38 joint count).

The time needed to score seven sets of radiographs of the hands and feet of one patient with SHS is ~ 25 min. SENS’ scoring time is ~ 7 min. The scores of 12 films assessed by SENS directly and deducted from the SHS scores were compared. Out of 1032 joints, 968 showed complete agreement, 30 showed abnormality in the deducted score but not in the direct SENS, and 34 in the direct SENS but not in the deducted score. These results were obtained over 2 yr after the original scoring and also take intra-observer agreement over a long period into account.

Reliability/reproducibility

The reliability coefficient (\(\rho\)-coefficient) of all 16 models ranged between 0.81 and 0.90 for SHS, and between 0.80 and 0.91 for SENS. The \(\rho\)-coefficient of model 16 by SHS and of SENS was 0.81 and 0.80, respectively, for seven radiographs and two observations. The \(\rho\)-coefficients in all 16 models of SHS varied between 0.81 and 0.90, those of SENS varied between 0.80 and 0.91. Table 2 shows the components of variance in model 16 of both methods. As could be expected, the percentages of variance components of ‘patient’ and that of ‘time nested within patient’ were by far the largest (99 and 97.7%, respectively). The high level of the latter percentage largely explains the somewhat subdued, but still very good performance of the \(\rho\)-coefficient. The fact that patients strongly differed over time, each in his/her own specific way, did much to hinder the overall \(\rho\)-coefficient reaching its maximum score of 1.00. The low variance components of ‘number of observations’, of the interaction of ‘patient’ and ‘number of observations’, and of the interaction of ‘number of observations’ and ‘time nested within patient’ indicate a high reliability of the scoring method (Table 2). The other remark that has to be made concerns the fact that the \(\rho\)-coefficient results can clearly be somewhat improved in points if one adopts other models than model 16, which summarizes over both extremities and both types of abnormalities, like the one which uses information on both types.

\[
\text{SDD} = \sqrt{\text{EMS}(T \times P) \times 1.282 \times \sqrt{2}}
\]

Table 2. Components of variance (CV) of model 16 of SHS and of SENS

<table>
<thead>
<tr>
<th>Sources of variance</th>
<th>CV in SHS</th>
<th>CV in SENS</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>51.36</td>
<td>148.32</td>
</tr>
<tr>
<td>O</td>
<td>0.04</td>
<td>0.25</td>
</tr>
<tr>
<td>T:P</td>
<td>80.87</td>
<td>234.88</td>
</tr>
<tr>
<td>PO</td>
<td>0.61</td>
<td>4.64</td>
</tr>
<tr>
<td>OT:P</td>
<td>0.70</td>
<td>3.89</td>
</tr>
<tr>
<td>Total</td>
<td>133.55</td>
<td>397.98</td>
</tr>
</tbody>
</table>

P, patient; O, observation; T:P, time nested within patient; PO, interaction of patient and observation; OT:P, interaction of observation and time nested within patient (\(n = 20\)).
of extremities and/or abnormalities apart and next to each other (e.g. model 9). The \( \phi \)-coefficients of models 1–4, which use partial patient information only, lie between 0.82 and 0.90 (SHS), and 0.82 and 0.91 (SENS). The highest \( \phi \)-coefficients are attained with ‘erosions in feet only’ (model 3) and the lowest with ‘erosions in hands only’ (model 1). Of course, the validity of the measurement system embodied by model 16 precludes such ‘improvements’ in reliability. On the whole, the reliability of model 16 seems to be quite acceptable.

In decision studies (D-studies), we could see what happens to the coefficient if we omitted any of the two observations. The \( \phi \)-coefficients in the D-studies carried out within some of the models hardly differed from the coefficients in the G-study (0.80–0.89 SHS, 0.79–0.88 SENS). The tables of the components of variance of the other models were comparable to those of tables of model 16 (results not shown). The ICC for agreement in ‘number of observations’ were very high in both methods. In models 1–4, 9, 14 and 15, the ICC for SHS were between 0.97 and 0.98, and those for SENS were between 0.89 and 0.98. The ICC of model 16 for SHS and for SENS were 0.99 and 0.98, respectively. These ICC and the \( \phi \)-coefficients in the decision studies indicate a very high intra-observer reliability in both methods.

The ICC for reliability in ‘type of abnormality’ were moderate, testing the within correlation between the scores of the erosions and the scores of the JSN. In models 14 of the SHS and of SENS, these ICC were 0.72 and 0.65. In model 9, these ICC were 0.59 and 0.69, also showing a moderate positive correlation between the erosions and the JSN in the hands, and between the erosions and the JSN in the feet separately. The ICC of the hands and feet were high in model 15 (0.81 and 0.78 for SHS and SENS, respectively), which indicates an acceptable correlation between the hand scores and the foot scores. The ICC of the hands and feet decreased in model 9 (0.68 and 0.59), showing a moderate correlation between the erosions in the hands and erosions in the feet, and between the JSN in the hands and the JSN in the feet.

**Sensitivity to change**

Table 3 shows the results of repeated measurements ANOVA combined with the results of the calculation of the components of variance (model 16). Based on seven radiographs scored by one observer, the sensitivity to change is 0.88 in SENS and 0.84 in SHS.

**Smallest detectable difference**

The SDDs are calculated for both SHS and SENS for each period and are presented in Table 4. The SDD for SHS based on two successive radiographs ranges from 7 to 24, the majority around 10 (out of a maximum score of 448). The SDD for SENS ranges from 4 to 6 (out of a maximum score of 86).

**Sensitivity and specificity of progression**

Applying these values to the data set determines the sensitivity, specificity, accuracy and the percentage of falsely classified patients for every period (Table 5). Also, the kappa statistic is presented. The accuracy over all periods is satisfactory. The sensitivity falls quickly if one mismatch occurs because of the small number of patients in the study and especially the small number of patients with progression above SDD. The kappa ranges from 0.44 to 1.00 (the mean over all periods is 0.73). These values indicate acceptable agreement between the two methods. Except perhaps for the last period, there seems to be no real change in accuracy (agreement), sensitivity and specificity. This indicates that the ability of SENS to detect progression is similar to that of SHS over the 5 yr period of follow-up studied. Looking at individual patients, SENS classified the presence of progression correctly in all periods in nine patients; four patients showed a mismatch in one of the six periods and one patient showed a mismatch in three periods. This last patient showed SHS progression in four periods and SENS progression in three, but these fall in different periods.

**Discussion**

Radiographic damage in RA is the irreversible result of chronic joint inflammation. Several radiographic...
methods have been developed for evaluating the disease progression and drug efficacy in clinical trials. Only a few studies have compared these methods directly [11–13]. However, the techniques used to determine and compare the reliability and the sensitivity to change (e.g. Pearson correlation or Spearman rank correlation coefficient) are theoretically not appropriate [16]. Therefore, our results are not directly comparable.

This study showed that SENS had a reliability equal to that of SHS. The intra-observer reliability was very high in both methods and did not decrease by scoring only parts of the radiographs (only erosions, only hands, etc.). As could be expected, the greatest source of variability in scores was the diversity of patients and the diversity in the course of progression of the individual patient. The latter source of variance induced a decrease in the reliability coefficients, but given the fact that other more important sources of variability (repeated observations at each point in the progression in time) were of only minor importance, both methods fared quite well. The ICC of the facet extremities and abnormalities show that the hand scores and the foot scores, and the erosions scores and the JSN scores, do agree with each other. This agreement was not so high that one part can easily be omitted without loss of information. Fries et al. [3] described the additional information of erosions and JSN before. Also, the importance of including the foot joints in a scoring method has been described before [15]. In conclusion, both hands and feet, and erosion and JSN, should be scored to get the maximum information.

The sensitivity to change of SENS, expressed as an ICC, was very similar to that of SHS. As far as data are available, the original method of SHS seems to have a greater sensitivity to change and a higher reliability than other scoring methods, such as Larsen’s method and the CMC ratio [11, 13]. SHS has the best sensitivity to change compared to some other radiographic methods, including Sharp’s method [14]. Therefore, we used the most sensitive method available as the gold standard with which to compare SENS. SENS’ sensitivity to change can be considered as good.

Expressed as a percentage of the maximum scores, the means of SENS were higher than the means of SHS, indicating that in these patients many joints are damaged, but the amount of damage per joint is limited (Fig. 2). Van der Heijde et al. [15] made similar observations in another larger cohort of patients with a follow-up of 3 yr.

The smallest detectable difference for SHS was in most periods around 10 (out of a maximum of 448) and for SHS 5 (out of a maximum of 86). These figures are valid for this patient group and can be extrapolated to a similar group of patients as we studied. This generalizability can be compared to the results of a clinical trial that can be extrapolated to patients fulfilling the same inclusion and exclusion criteria. To have comparable figures for another patient group, a new so-called G-study (such as this) is needed. Based on the SDD, the patients were classified as having progression present in every period. The accuracy of classifying patients as having progressed with one method and also with the other was satisfactory for all periods during the 5 yr follow-up. This was also expressed in a mean kappa of 0.65 over all periods. These data indicate that SENS is approximately as good as SHS in detecting progression.

The results of this study indicate that damage to joints of patients with RA can be scored reliably with SENS during the first 5 yr. More data are needed to judge the performance of the method with longer disease duration. Also, interobserver agreement will have to be included in future studies of SENS. Because of the time profit and the results of this study, SENS seems useful in clinical practice in at least the first 5 yr of RA.

References