Reliability of scoring methods to measure radiographic change in patients with rheumatoid arthritis.

Rolf Rau and Siegfried Wassenberg

J Rheumatol 2005;32;766-768
http://www.jrheum.org/content/32/5/766.citation

1. Sign up for TOCs and other alerts
   http://www.jrheum.org/alerts

2. Information on Subscriptions
   http://jrheum.com/faq

3. Information on permissions/orders of reprints
   http://jrheum.com/reprints_permissions

The Journal of Rheumatology is a monthly international serial edited by Earl D. Silverman featuring research articles on clinical subjects from scientists working in rheumatology and related fields.
Reliability of Scoring Methods to Measure Radiographic Change in Patients with Rheumatoid Arthritis

Damage of peripheral joints following inflammation represents the most important pathology and outcome of rheumatoid arthritis. Loss of function and quality of life of the patient are clearly related to swelling and pain in early disease and to joint damage in advanced disease. To prevent or inhibit damage or its progression, which can be documented best on radiographs, is the most prominent goal of our treatment. In order to describe the status of the patient and the course of the disease, damage has to be quantified. Since there is no truly quantitative method, semiquantitative methods have been developed allowing translation of the amount of damage into a score value.

In this issue of The Journal, Guillemin and coworkers present results of a study comparing the ability of 5 different scoring methods to reliably assess joint damage and its change. Two of the methods represent composite scoring (Sharp; Sharp/van der Heijde), where erosions and joint space narrowing (JSN) are scored separately; 2 are modifications of the global method (Larsen; Larsen/Rau), with a global appraisal of joint destruction mainly relying on erosive damage; and a simplified method (SENS), where only the number of joints with erosions and/or JSN is counted.

The aim of this editorial is to discuss the merits and limitations of the study by Guillemin, et al.; to explain their results using a practice related approach, rather than focusing mainly on statistical considerations; and to mention some recent progress in evaluating and reporting studies.

As we currently do not have an established criterion for quantification of damage (an external gold standard) for comparison, the validation of a scoring method relies on the reproducibility in terms of intra- and interrater reliability. In this study, the reliability of scoring methods was evaluated by means of correlation coefficients (ICC). However, ICC are measures of association and not necessarily of agreement and strongly depend on the range of values, with extreme values having the greatest effect. In contrast, Bland and Altman proposed a method that is less influenced by extreme values, in which change of the score is related to the measurement error of the method. Standardized response mean (SRM) is the ratio of change divided by the standard deviation of change and has nothing to do with the measurement error. The smallest detectable difference (SDD) is defined as a difference that is greater than the measurement error at baseline or followup and therefore can be taken as “real.” As the main goal of a scoring method is to measure change, it may be better to use the measurement error of the change score to determine the smallest detectable change.

The results of Guillemin, et al’s comparative study are very reassuring: there was a good correlation in the baseline values, a similar amount of progression was seen with all methods (with the exception of SENS), and the differences regarding sensitivity to change and SDD were small. However, radiographs were read in a fixed order of methods (apparently at the same reading session), where the score with one method may have influenced the score with the other method; this, for example, explains the astonishing similarity of the mean scores reached with the Larsen and the Larsen/Rau methods. Radiographs were also read knowing the time sequence of the films, thereby introducing the bias of progression. Unbiased reading blinded to sequence would give a more objective judgment about the comparative performance of the methods. The study was based on a set of hand radiographs (excluding the feet) of only 20 patients with radiographically early disease, with baseline and only one followup assessment after a mean of 2.5 years. This study population can hardly “represent the total spectrum of disease damage.” The inclusion of feet would have been strongly desirable, since in many cases feet are involved earlier and show more relevant changes than hands. The spectrum of disease could be represented...
best in a study population with a high progression rate followed from early to late disease with several assessments in between.

There was good agreement between the methods (except SENS), which can be demonstrated best when transforming the different scales into a normalized (0–100) scale: the baseline scores were 8%–10%, the mean progression ranged from 7% to 8.5%, and the followup reading was 15%–18% of the maximum score. These data indicate that the patients had clinically and radiographically early disease. In most scoring methods the sensitivity to change in early disease is much greater than in late disease, because the intervals on the scale are smaller, leading to an overestimation of early changes. For example, when single erosions are counted irrespective of their size (Sharp method), the sensitivity to change will be very high in early disease. However, an increase in the size of existing erosions, which is more common in later disease, cannot be counted. Moreover, 4 small erosions in a single joint give a score of 4, which is already 80% of the maximum score, and the maximum score of 5 is reached with 5 erosions or destruction of > 50% of the surface of one or both articulating bones; thus, further destruction cannot be scored. This results in a decreasing sensitivity to change and a clear ceiling effect.

With the van der Heijde modification grading the erosion size from 1 to 3, the sum of the single erosions may be 7 or 8. However, the highest score in a joint remains 5, which again reduces the sensitivity to change in later disease. In general, advanced changes are much more difficult to score than early changes. Therefore, disagreement (variability of scores) between raters and methods is much greater in late disease.

It could be expected that the reliability of the global methods was worse in this set of radiographs because of uncertainty to distinguish between grades 0 and 1, thereby increasing the measurement error and reducing the sensitiv-
ity to change. In the Larsen method, grade 1 is defined as erosion of < 1 mm, the detection of which is uncertain. The Larsen/Rau modification still includes soft tissue swelling as grade 1 (as the original Larsen method), which is difficult to identify on radiographs. Scoring soft tissue swelling may lead to a relatively high score at baseline decreasing with response to treatment. This reduces the possible score increase caused by damage progression, again resulting in a low sensitivity to change.

In spite of the methodologic differences, the progression rates were very similar, 7% with both Sharp modifications and the Larsen/Rau score, and 8% with the Larsen score. However, the SDD, based on the measurement error, was greater with both global methods.

Our improved method excludes soft tissue swelling and scores only structural damage in intervals of 20% of the joint surface, providing a more linear correlation between damage and score value from early to late disease. However, the sensitivity to change may be lower in early disease than when counting single erosions since 20% of the joint surface must be involved before the score can be increased from 1 to 2.

The SENS method scores only the number of joints that are affected by either erosion or joint space narrowing, or both, irrespective of the amount. An increase in the score is only possible if a new joint becomes involved. As the number of unaffected joints becomes rapidly smaller with time, the method becomes more and more insensitive to change; SENS performed worst in this study.

Recently, Sharp, et al analyzed 6 data sets that were scored by 11 experienced readers; they found an intrarater SDD — transformed to a 0–100 scale — of 10% for the composite method and 8% for the global method. When analyzing progression scores, global scores also performed somewhat better than composite scores. There was considerable variability in scoring among these expert readers, which contributed to Sharp’s conclusion that probably quality and consistency of the readers are more important for the result than the method used.

Assessing and reporting radiographic changes is still in development. Reading blinded to sequence has demonstrated negative scores that, if greater than the measurement error, may indicate healing. Currently the question is under investigation whether all healing can be captured with conventional scoring methods.

The introduction of probability plots has significantly improved reporting and interpretation of study results. Probability plots present all individual data demonstrating the exact amount of change for each individual patient, allowing a much better understanding of study results than with summary descriptive statistics.

In conclusion, scoring of radiographic change will remain an important and valid outcome instrument in the evaluation of treatment effects in the foreseeable future and remains open for modification and improvement.

ROLF RAU, MD, PHD.
SIEGFRIED WASSENBERG, MD.
Department of Rheumatology,
Evangelisches Fachkrankenhaus Ratingen,
Rosenstrasse 2,
D-40882 Ratingen, Germany.

Address reprint requests to Dr. Rau. E-mail: rrau@uni-duesseldorf.de

REFERENCES
3. Sharp JT, Young DY, Bluhm GB, et al. How many joints in the hands and wrists should be included in a score of radiologic...


