Precision of Larsen grading of radiographs in assessing progression of rheumatoid arthritis in individual patients

M M O’Sullivan, P A Lewis, R G Newcombe, N J Broderick, D A Robinson, E C Coles, J D Jessop

Abstract
A study was designed to evaluate observer variation in the assessment of radiographic deterioration of individual patients using the Larsen grading system.

Radiographs of hands and feet of 52 patients were assessed by three observers. Each patient had paired films taken one year apart which were assessed together for change in score. To assess within-observer variation each set of films was read twice by all observers.

The average progression was 11-6 (SD 9-0). Analysis of the source of variation showed the single observer replication SD to be 3-7 but that for different observers to be 5-5.

This may be interpreted as indicating that to achieve 95% confidence of detecting a true change an increase in Larsen score of 8 is required if the same observer assesses or up to 11 if a different observer assesses.

The value of regular radiological assessments in following the course of rheumatoid arthritis and in studying the effects of second line antirheumatic treatment on disease progression has been emphasised in several reports. Various different grading systems are now being used to measure the degree of joint damage, but the Larsen grading system is being used in this department for the assessment of radiographs of patients with rheumatoid arthritis who are participating in a five year prospective clinical trial of second line antirheumatic drugs.

Radiological assessment continues to be considered the ultimate standard for assessing drug efficacy in rheumatoid arthritis, and although there has been little clear cut evidence for drug effect on radiological progression, some slowing of radiological progression has been reported.

When radiological assessment of deterioration is used to determine the need for alteration of treatment the selection of an appropriate time interval for the assessment is important. If the time interval is too short the costs escalate and the risk of a 'false positive' deterioration increases. If the interval is too long significant radiological deterioration may be missed.

The rate of radiological progression is, however, unpredictable and varies between patients. Moreover, both the assessment of significant radiological deterioration and the selection of the appropriate time interval require clinical judgment. For this study the clinicians decided that a progression of about 5 to 8 Larsen points in one year as described below would be considered a clinically significant progression. Changes less than this were considered not significant, and for larger changes there was likely to be little difficulty in establishing deterioration. This study aimed at providing an estimate of how precisely differences in this important intermediate range might be detected. In addition, we wish to present data on the rate of progression of patients with rheumatoid arthritis to indicate the percentage of patients who might be detected as having this magnitude of change in a given period.

Patients and methods
The Larsen score grades the degree of joint damage on a scale from 0 (radiologically normal joint) to 5 (maximum degree of joint destruction) with reference to a standard atlas of radiographs. The method described by Larsen, in which the wrist is scored from 0 to 5, was used and radiographs of the hands, wrists, and feet were graded giving a maximum possible total score of 210.

Fifty two patients with rheumatoid arthritis who had been participating in the Cardiff five year prospective clinical trial of second line treatment for at least one year by 1 April 1985 were studied. These patients had an entry film and a 12 month anteroposterior radiograph of both hands and both feet with a difference in Larsen score of between 3 and 16. Selection of radiographs for this study was based on independent Larsen scoring by a senior consultant radiologist experienced in bone and joint radiology and in Larsen grading. The assessors taking party in the study were three senior trainees, one in rheumatology and two in radiology, all highly competent in reading radiographs of rheumatoid arthritis but not previously experienced in using the Larsen grading system.

Within each set of films the sequence of initial and follow up was indicated as it would be in practice, and the observers were instructed to assess the initial and follow up films of the same patient together. The observers were blind to the identity of the patients and to all aspects of the study design. They were blind to the fact that the radiographs were taken one year apart, were preselected on the basis of score increases between 3 and 16, and were read twice during the study period.

Each observer conducted 26 sessions. The sets of films used for the first 13 sessions were presented again for assessment in the second 13 sessions after an interval of six weeks and without the knowledge of the observers. At each
session the sets of films were randomly allocated into blocks of four for each observer, the order within the block being altered between the first and the repeat readings of the same films. All assessments were performed in the same room with the same light box. An independent experimenter distributed and retrieved the films and collected the forms on which the results had been recorded.

Hierarchical analysis of variance of the 0 to 12 month change in Larsen score by patient and by observer (regarded as a random effects factor within patients) was used to predict the degree of variation (representing imperfect reproducibility) which should be expected in the following two situations. Firstly, an observer viewing a set of films on one occasion and noting the change in Larsen score, and then subsequently repeating the assessment to give a fresh value of the change in score. Secondly, one observer viewing a set of films and noting the change in Larsen score and a different observer doing the same thing to give an independent assessment of the change in score.

The within-observer variance for the former situation was given directly by the residual mean square of the analysis of variance. The between-observer variance for the latter situation was estimated by adding to the within-observer mean square, one half of the difference between the observers-within-patients mean square and the within-observers mean square.

To assess the rate of annual progression all available data from the long term study on the difference between the entry Larsen grade and that found one year later were collected. These data included results for the population on which the experiment described earlier was based. The frequency distribution of the 0 to 1 year progressions was calculated.

Results

Table 1 presents the mean and standard deviation between 52 patients of the 42 joint total Larsen scores for the films taken during the initial work-up of the patients. It shows that despite attempts at standardisation the different observers gave somewhat different scores overall, with observer 2 giving the highest scores and observer 3 the lowest. It is also noticeable that observer 3 seemed to score closer to the other two on the repeat readings. Observer 2 showed less variation than the other two, which either implies greater consistency or possibly less ability to discriminate.

Table 2 gives the mean and standard deviation of the Larsen scores after 12 months in the study. As before, the scores rank in the sequence of observer 2, observer 1, then observer 3, who once again increased the mean score on the repetition. Observer 2 again showed the least variation. Table 3 shows the 0 to 12 month changes in Larsen scores. In these data observer 3 seems to be most consistent and observer 1 shows the smallest standard deviation.

The within-observers variance for the analysis of variance was 14-0, giving a standard deviation of 3:74. The observers-within-patients mean square was 46-1. For different, arbitrarily chosen observers, each assessing the degree of progression from 0 to 12 months using sets of films, the predicted variance was estimated as:

\[
14-0 + (46-1 - 14-0)/2 = 30-0
\]

The distribution of changes in the Larsen grades after one year in the Cardiff study for the first 258 subjects was as follows: not detectable (Larsen score 0-12) 89%, marginally detectable (13-16) 4%, clearly detectable (17+). In this instance 'detectable' means being able to detect a clinically important difference 95% of the time.

Discussion

Radiological assessments are an essential part of the management of patients with rheumatoid arthritis and may influence clinical decisions, particularly on second line treatment. The evidence that second line drugs may reduce the rate of radiological progression has been critically reviewed, and it has been suggested that more rigorous studies are required before this concept can be generally accepted.

Our study was undertaken to investigate the precision of the Larsen grading system in detecting small to moderate radiological progression in individual patients. It should be noted that the change in mean Larsen score for groups of patients as is required in a clinical trial is a different matter, though these data may,
where appropriate, be helpful in calculating statistical power and sample size.

To ensure that genuine radiological deterioration has occurred it is necessary to know what changes in score may be detected reliably and how much change is likely to be due to observer variation. The design of our study (which involved observers blindly repeating the scoring on radiographs of 52 patients) enabled us to measure the within-observer and between-observer variation on radiographs selected within a relatively narrow but critical range of between 3 and 16 increase in Larsen score. Although these arbitrary choices, this degree of change was considered clinically important as it is within the range which might influence therapeutic decisions, particularly in a trial protocol, and it is difficult to assess.

The 'true' rate of radiological progression among patients with rheumatoid arthritis is unknown, and reported rates vary from little or no progression to substantial progression. Scott et al reported a mean increase in Larsen score of 8·9 in 57 patients over one year, though the maximum increase was 34.1 Probably, published reports of extremes of progression are overrepresented owing to publication bias. Reports of extreme progressions, especially in the short term, need to be interpreted within their longitudinal context and the discrete nature of the radiological scoring system.

Our results show that although there may appear to be good overall agreement both within and between observers, the standard deviation for the within-observer variation was 3·74 and the standard deviation for the combined within- and between-observer variation was 5·48.

These figures may be interpreted as indicating that for an individual observer to have reasonable (95%) confidence that there has been any genuine deterioration at all in a patient's radiological state a rise of at least 8 points is needed, but for that observer to convince colleagues a rise of at least 11 points is required. It is argued that if there is no expectation of an improvement in radiological score a 'one sided' rather than a 'two sided' confidence interval may be used.

As reported above, the clinicians responsible for this study felt that a change in the Larsen score of 5–8 points in one year would be on the boundary of clinical significance when taking a decision to stop a drug for lack of effect. To achieve 95% confidence that this level of change has been achieved the 8 points calculated above, which are necessary to show any change, need to be added to this 5–8 range to give 13–16 points. Thus for the clinicians in this study to be 95% confident that a treatment is having insufficient effect a change in Larsen score of 13–16 points would need to be seen, in the absence of any other corroborative evidence.

Preliminary data from the Cardiff study suggest that only 7% of patients with rheumatoid arthritis progress by more than 16 Larsen points in one year, while 4% progress by 13–16 points. It is not yet known what the association is between arbitrary change group and other clinical factors, but probably a substantial proportion of them will show other signs of lack of effect of the treatment. The small percentage of patients receiving second line treatment who show a detectable, clinically significant change in their Larsen grade scores over one year does not justify radiological monitoring of all patients more often than once a year.

The figures for observer variation which we detected are of the magnitude of change which has been regarded as clinically significant by other groups.1 3 Our results, however, relate to changes in scores for individual patients and cannot therefore be compared with changes in the mean scores in groups of patients where the inappropriate observer variation was reduced.

The correlation coefficient has been used by many workers attempting to assess the reliability of radiographic scoring systems. As Bland and Altman have so clearly shown the correlation coefficient is not appropriate for this purpose. 18 This is because it is a measure of association not of agreement and so is unaffected by one measure being offset from the other; because it is increased by having a wide range of values and decreased by a narrow range; and, most importantly, because data which show poor agreement may have high correlations with most impressive p values if the inappropriate null hypothesis of 'no agreement' is tested.

Sharp et al highlighted the problem of using the correlation coefficient as a measure of agreement when high correlation coefficients were found between individual observers but with widely differing scores of the radiographs. 19 The magnitude of the difference between scores was so large that any meaningful comparison of the absolute scores was not possible. When a standardised score was used there was a 92% chance of agreement between observers in recording a change in score if the difference between the two films was greater than 15 standardised units (equivalent to about 10 on the Larsen scale). Changes of less than 15 standardised units could not be detected reliably. The radiographs covered a wide range of disease severity and, as might be expected, there was a reduction in the coefficient of variation with increased severity of disease. Assessments of serial radiographs which are either unchanged or have large differences in scores are likely to have much less variation between observers.

In conclusion, we have estimated the observer variation in assessing change in Larsen score of individual patients. We found that changes of up to 11 might be attributed solely to observer variation and suggest that investigators should measure the precision of their assessments before attributing significance to changes in score.

We wish to thank Dr L Arnold Williams for the initial idea, his assistance with the design of the study, and for many helpful subsequent discussions. We also wish to acknowledge the help of Mr Simon G Mason for help with the study and Mrs Julie A Spiteri for typing the manuscript.

3 Dawes P T, Fowler P D, Clarke S, Fisher J, Lawton A,


Precision of Larsen grading of radiographs in assessing progression of rheumatoid arthritis in individual patients.

M M O'Sullivan, P A Lewis, R G Newcombe, N J Broderick, D A Robinson, E C Coles and J D Jessop

Ann Rheum Dis 1990 49: 286-289
doi: 10.1136/ard.49.5.286

Updated information and services can be found at:
http://ard.bmj.com/content/49/5/286

These include:

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/