Radiographic progression in rheumatoid arthritis

Robert Landewé, Désirée van der Heijde

University Hospital Maastricht, Department of Internal Medicine/Rheumatology, Maastricht, The Netherlands.

Please address correspondence to: Robert Landewé, University Hospital Maastricht, Department Internal Medicine/Rheumatology, P.O. Box 5800, 6202AZ Maastricht, The Netherlands. E-mail: RLAN@SINT.AZM.NL

Clin Exp Rheumatol 2005; 23 (Suppl. 39): S63-S68.

Key words: Key words: Radiographic progression, repair, rheumatoid arthritis, probability plot, structural damage.

ABSTRACT

Radiographic progression is an important outcome measure in clinical trials and observational studies with patients with rheuamtoid arthritis. In this article we describe several aspects of measuring radiographic progression. We introduce the scoring method, discuss scoring methodology and issues regarding reliability of scoring, describe the relation between disease activity, radiographic progression and physical function, and introduce the concept of repair, a novelty in the field of measuring structural changes in RA.

Introduction

Radiographic progression is nowadays an important outcome variable in clinical trials in patients with rheumatoid arthritis (RA) and in observational studies. Reasons are that radiographs of hands and feet can be easily performed and are relatively cheap (feasibility), that valid scoring methods are available and the methodology of measuring progression is standardised, that inflammatory activity in the joints leads to radiographic progression, and that radiographic damage correlates with physical function. Inflammation of the joints may fluctuate over time in individual patients, and radiographic damage may be considered a reflection of joint inflammation over time.

Plain radiography seems somewhat old-fashioned in comparison with newer imaging modalities, such as magnetic resonance imaging and power-doppler ultrasound, but the methodology of measuring radiographic progression in order to use it as an endpoint in clinical trials is still developing. New concepts emerging from this research are issues regarding sensitivity-to-change, (*or*: how long should a trial take in order to demonstrate sufficient radiographic progression for using it as a primary endpoint), the issue of repair of existing joint damage, and the issue of data presentation. In this paper we will introduce the scoring method that we use most frequently and we will briefly outline how we perform formal readings in clinical trials, we will describe how radiographic progression relates to important outcomes such as disease activity and physical function, we will elaborate on methods of presentation of radiographic data, and we will end with a brief discussion about repair.

Scoring in clinical trials and measurement error

Two major scoring systems and a number of modifications are available for scoring radiographic progression in RA: The Larsen system (1) and the Sharp system (2). A number of modifications have been described for both systems. Most landmark trials in RA now apply the van der Heijde modification of the Sharp scoring system (SvdH) (3), because this method includes both hands and feet, collects information on erosions ánd joint space narrowing, and covers a sufficiently broad spectrum of joints to provide sensitivity to change. The SvdH method was tested and approved in two previous OMERACT meetings. In brief, the SvdH method scores the presence of erosions in 16 joints of hands and wrists (graded from 0 to 5), and in 6 joints of the feet (graded from 0 to 10), and the presence of joint space narrowing in 15 joints of the hands and wrists (graded from 0 to 4) and in 6 joints of the feet (graded from 0 to 4) (Fig. 1). The maximal range is 280 units for erosion and 168 units for joint space narrowing, summing up to 448 units for the total Sharp score (TSS).

Radiographic progression in a clinical trial or an observational study is calculated by subtracting the TSSs of one patient at two subsequent time points. An important appreciation with regard to scoring radiographs is within-patient correlation. Usually, unlike damage on radiographs of different RA patients, damage on subsequent radiographs of the same patient is highly correlated (correlation coefficients > 0.9 are not

exceptional). In order to be able to distinguish the progression signal from background noise, radiographs of the same patient are scored together. In general, there are two ways of scoring: One with known time order (chronological reading), and one with concealed time order (paired reading). As a rule, radiographs from controlled trials are scored by paired reading. The most important justification is the level of blinding. Paired reading provides blinding for treatment ánd reading order, thus creating an experimental setting within the experiment. An important advantage of paired reading is that it visualises measurement error to some extent. The latter is due to technical tribulations, such as subtle differences in positioning and exposure, as well as to "true reading error" (which reflects the human limitation with regard to reproducibility). As a consequence, negative progression scores are found in clinical trials, which may reflect measurement error, though only under the premise that true negative scores do not exist (see below).

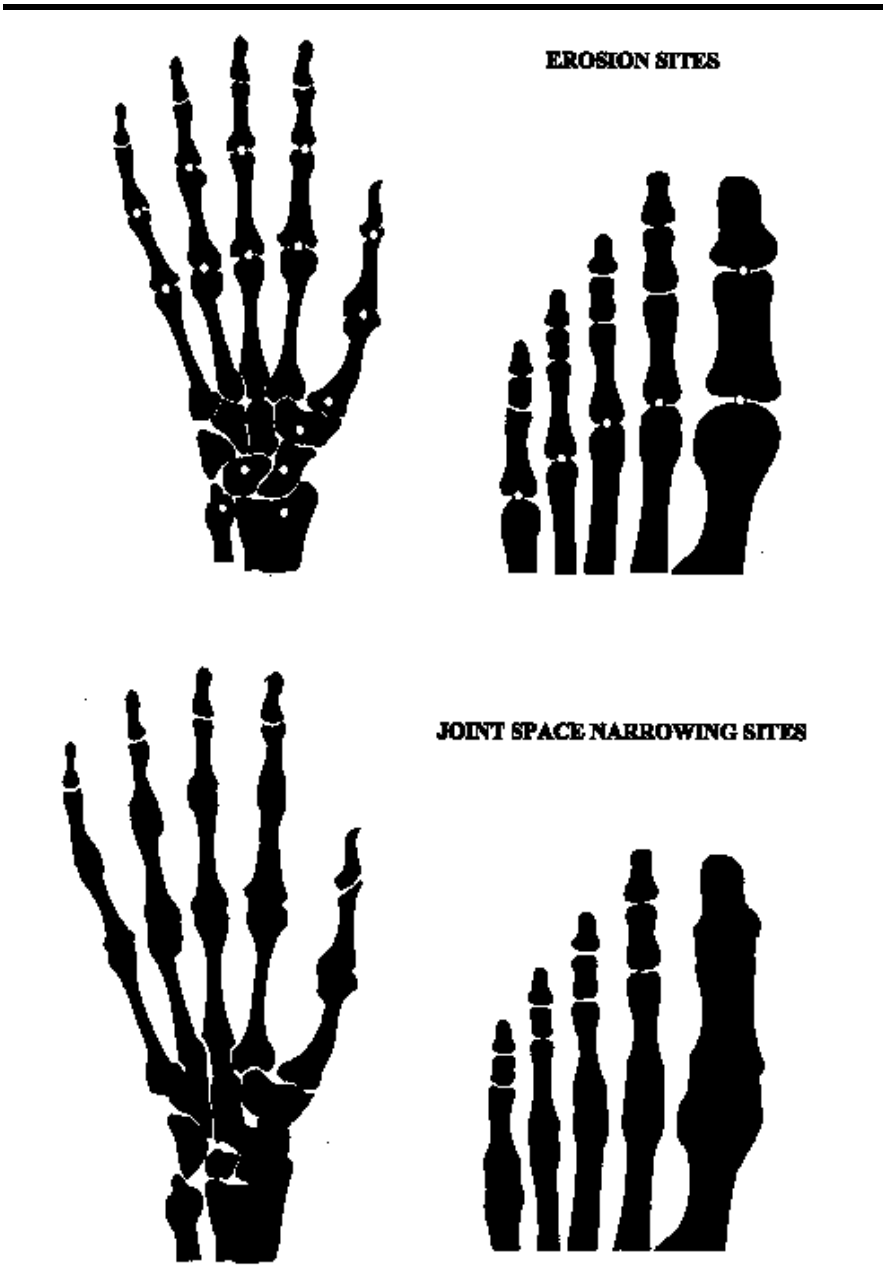

Fig. 1. Schematic representation of the scoring sites of the van der Heijde modification of the Sharp score, with regard to erosions and joint space narrowing. Erosions in the hands are from zero to 5, erosions in the feet are scored from zero to 10, joint space narrowing in the hands and feet is scored from zero to 4.

A disadvantage of paired reading is that the progression signal is usually lower as compared to the same setting scored with chronological time order (4). The most probable explanation for this discrepancy is that knowledge of the time order may increase the level of confidence with respect to subtle changes, that are ignored in paired reading, to such an extent that it is scored as present in chronological reading. Until now, it is not clear whether the advantage of blinding counterbalances the disadvantage of a lower signal.

We have already touched the issue of measurement error. Scoring is a matter of subjective interpretation of changes that are often subtle. Subjective interpretation creates space for intra-reader variability, the phenomenon that observers score hardly if ever exactly the same twice despite an identical context. Subjective interpretation is also the most important source of inter-observer variability, the phenomenon that two readers confronted with the same set of radiographs do not provide exactly the same scores. In order to constrain measurement error, and optimise signal-to-noise ratio, several techniques are applied.

The readings are always performed by two or more readers, and their average scores are considered as the progression signal. From a theoretical point of view, the precision of a score increases by increasing the number of readers, because it eliminates all kinds of random error, operative in different directions (5). It is still a matter of debate whether a trial result improves by increasing the number of readers. Probably the most important advantage of increasing the number of readers is the better external validity (because the trial result better reflects the truth). Internal validity is not at stake with "only" one reader, as long as blinding is preserved, since measurement error is "symmetrical" in both trial arms. Undoubtedly feasibility comes into play if the number of readers exceeds 2 or 3. A better way of constraining mea-

surement error is a reading by two readers, with re-reading of the radiographs in case of inter-reader discrepancies beyond a certain threshold. A third technique, that can be applied per se in cases with too much discrepancy, or only if re-reading does not result in deflation of inter-reader variability, is adjudication by a third reader. In such a scenario, the scores of the adjudicator and the reader's score that is closest to the adjudicator's score are used to calculate the mean reader score.

Radiographic progression as an outcome measure in clinical trials

Most scoring methods and their modifications are based on assessing damage in hands and feet. Damage scored in hands and feet has been shown to sufficiently reflect damage of large, often weight bearing joints that are excluded from scoring, thus providing content validity to scoring methods involving hands and feet. There is increasing evidence that radiographic progression, as measured by assessing changes in hands and feet is associated with inflammatory activity. In older studies, time-averaged variables of acute phase reactants and time-averaged disease activity scores have shown to be related to radiographic progression over the same time period in which these variables were measured (6).

More recent studies from our group have explored the longitudinal relationship between disease activity and radiographic progression (7). Using marginal modelling by generalised estimating equations and mixed linear modelling, we were able to demonstrate a longitudinal relationship between disease activity and radiographic progression. The distinction between the older time-averaged models and the modern longitudinal models seems subtle, but the difference in interpretation is important and clinically relevant: A longitudinal relationship implies that an increase in disease activity is immediately followed with an increase in radiographic progression rate in the individual patient, and the same is *mutatis mutandis* true for a decrease in disease activity. *En passant*, it became clear that radiographic progression in the individual patient is not a linear process, which is always suggested by group analyses, but may include accelerations and decelerations, invoked by tribulations in disease activity. Embarking on this concept, we demonstrated that it is not only the absolute level of disease activity that is contributory, but also – and independently – the fluctuation in disease activity over time (7).

Applying the same statistical techniques to the relationship between radiographic progression and physical function, we found interesting results. So far, physical function was shown to be correlated with the level of radiographic damage at the same time point in a number of studies (8, 9). Undoubtedly, such a cross-sectional interpretation provides useful information with regard to the cause and prevention of disability. But the information is indirect, and confounded by the effects of joint inflammation on physical function. In an era of highly effective biological therapies, the type of information that is needed is whether an arrest or slowing of radiographic progression (*id est*: a change in progression rate invoked by treatment) influences physical function independently of disease activity. Or in other words: does measuring radiographic progression contribute to measuring disease activity in order to explain physical function in an individual patient ? The answer is a careful yes.

We explored the longitudinal relationship between radiographic progression and physical function in a 10-year follow up cohort of patients with RA (10, 11). We modelled both the health assessment questionnaire score (HAQ), as a measure of general physical function, and grip strength as a measure of site-specific function, and found that the radiographic progression rate at a certain time point co-determined physical function, after careful adjustment for disease activity. We are now trying to confirm these data in databases from clinical trials, and the results look promising. If so, these longitudinal data will importantly add to the validity of the working hypothesis about the relation between disease activity, radiographic progression and physical function, that is proposed by many, and visualised in Figure 2. This diagram brings radiographic progression as an independent variable in the focus of treatment goals in patients with RA, especially since there is convincing evidence emerging that under certain circumstances there is a disconnect between disease activity and radiographic progression.

The credibility of such a disconnect is increased by our work in the COBRA

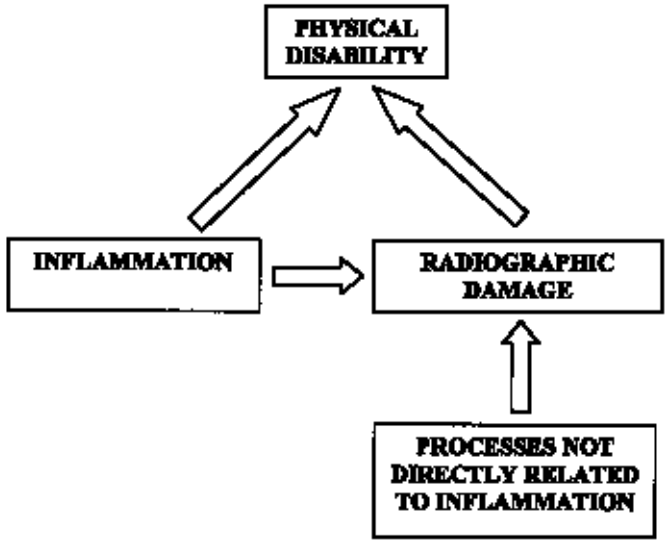

Fig. 2. Conceptual relationship between disease activity, radiographic progression and physical function in patients with rheumatoid arthritis.

database, in which we measured receptor activator of nuclear factor kappa-B-ligand (RANKL) and its naturally occurring decoy receptor osteoprotegerin (OPG) (12). RANKL is an important activator of osteoclasts, and as such implicated in bone erosions in RA, and OPG can bind RANKLso that RANKL cannot bind to its receptor RANK on osteoclasts. As such the ratio of RANKL and OPG can be considered a measure of osteoclast activating potential. We found that the RANKL/OPG-ratio, measured at baseline, independently of inflammatory activity determined long-term radiographic progression. Other sources of evidence that point to a disconnect include clinical trials with TNF-blocking drugs, that show only a marginal difference between the methotrexate (MTX)-only group and the anti-TNF only group in terms of clinical outcome, but a large difference in terms of radiographic outcome (13). We were able to formally prove the disconnect in such a trial by performing longitudinal data analysis that showed a statistically different progression rate in the group with TNF-blocking drugs as compared to the control group after careful adjustment for differences in disease activity (14). A number of those analyses is now underway in different trial cohorts.

Presentation of radiographic data

A set of data presenting progression in radiographic damage does hardly if ever have a normal, bell-shaped distribution. Often, the majority of patients shows minor or zero progression, and only a relatively small proportion of patients has significant progression. We call such a distribution skewed. Such types of distributions are difficult to describe in a comprehensible manner. Means and standard deviations as descriptive statistics may give a spurious reflection of what is really going on in the group of patients, because means and standard deviations (SD) are importantly determined by the small proportion of high scores. Medians and percentiles are often not an appropriate alternative, since they may not properly visualise treatment contrasts, especially if radiographic progression is limited to less than 50% of the patients per treatment group (median = 0).

In order to improve comprehensibility of radiographic progression data, we recently proposed probability plots as a means to show important aspects of a set of radiographic progression data (15). A probability plot is a cumulative frequency distribution that orders radiographic data from the lowest through the highest value, and plots every individual value of one treatment arm. An example of a probability plot is given in Figure 3. It compares the one-year radiographic progression scores of the two treatment groups of the COmbinatietherapie bij Reumatoide Artritis (COBRA) trial. Drop-lines reflect the median and 25/75 centiles. The mean value is by definition reflected by the area under the cumulative probability curve, and cannot be read from the plot. It is easy to see that the curve of the monotherapy group lies left to the curve of the combination therapy group, indicating that radiographic progression was worse (higher scores) in the monotherapy group. A probability plot is a means of exploratory analysis. It does not statistically test a between-group contrast, but it can serve as an adjunct to statistical testing in that it visualises directly what actually has happened in the treatment groups.

Probability plots can also show the above mentioned negative radiographic progression scores, which are often found in clinical trials but disguised in summary descriptives such as means and medians. The example of the COBRA study that we showed here is not representative since COBRA was scored with known time order, and negative scores were “forbidden”. Negative scores are the consequence of either measurement error inherent to paired reading, or so-called repair, or both, and will be described below.

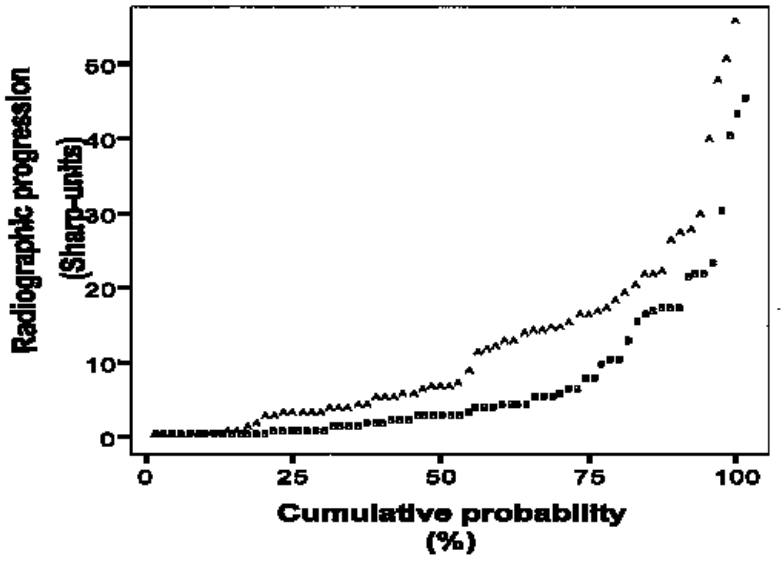

Fig. 3. Probability plots representing one-year radiographic progression in both groups of the COBRA study. Every symbol represents the score of an individual, and all scores are plotted against their cumulative probability.

Repair of existing joint damage

Repair is not a new feature. Several rheumatologists have reported this radiographic phenomenon in case reports in the literature by its more magic connotation “healing” (16,17). Healing became a conspicuous phenomenon by the appreciation of the aforementioned negative scores in clinical trials with biologicals. Although it was realised from the beginning that the occurrence of negative scores in clinical trials did not immediately indicate repair of damage, it was obvious that negative scores occurred more frequently in the treatment groups with lowest overall progression. It is however impossible to directly derive the existence of repair from negative scores from clinical trials. Figure 4 shows a theoretical repre-

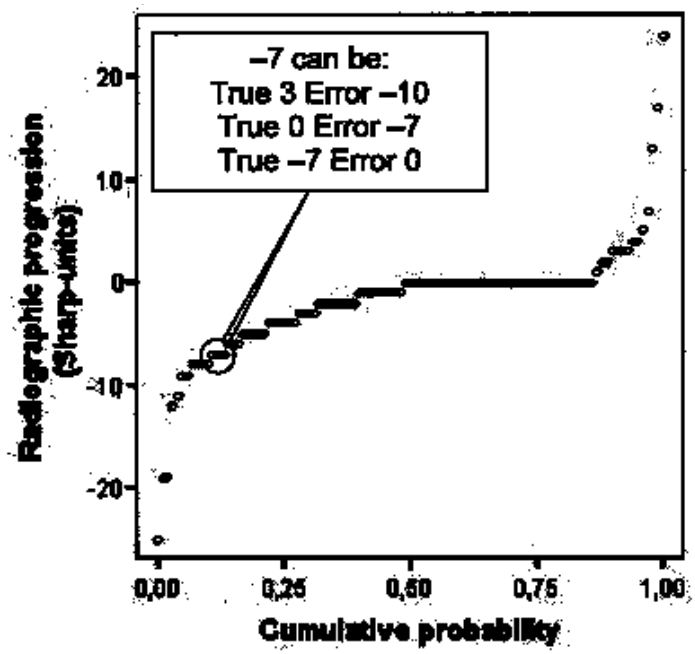

Fig. 4. Probability plot of an imaginary progression scenario. Negative progression scores do not necessarily imply repair. Every individual score represents a combination of true change and measurement error. It is impossible to distinguish both at the individual patient level.

sentation of how individual negative scores may be comprised. It can be every combination of true signal and measurement error, and it is impossible to differentiate in the individual patient.

We have proposed to interpret the likelihood of repair in the context of all within-group scores, under the null-hypothesis that true repair at the group level does not exist, and that every deviation is due to random measurement error (18). Here, the concealment of reading error, likely redundant in demonstrating between-group (treatment) contrast, becomes of utmost importance. Any knowledge about the true time order will lead to biased scores dependent on what readers expect to see, and changes due to technical imperfections will easily be attributed to either repair or progression, and scored accordingly. Under the provision of strict blinding of time sequence, the null hypothesis of no change can be rejected if within-group change is significantly higher than –, or lower than zero, as tested by a statistical method for paired observations. The scenario that progression is statistically significantly negative points to repair at a group level, and was found in the TEMPO trial (13). Repair defined in such a manner has a statistical basis, implying that its proof is dependent on statistical power in relation to the magnitude of the effect. It is possible, and maybe likely, that the same effect can be observed in clinical trials with other biologicals if tested under optimal conditions.

The statistical demonstration of repair at a group level does not mean that repair truly and irrefutably exists at the individual patient level (as is also true for the assessment of progression). Neither does the demonstration of progression at the group level preclude repair in an individual patient. We do not know how a group score of minus 0.5 Sharp-units translates to the individual patient and the individual joints, whether a negative change in TSS reflects only improvement of joint scores or is a net result of negative and positive joint scores, how many patients actually contribute to a negative change in TSS, and last but not least what the clinical meaning of negative change scores actually is.

A series of experiments has been conducted under the auspices of OMERACT, which had as a common goal to increase confidence in the existence of repair, as observed on serial radiographs. In an experiment with pairs of single joint radiographs, agreement among readers was relatively high with respect to choosing the worst image, but experts were totally unable to reproduce the true time order (19). The implication is that experts do not recognise repair as such, but actually see subtle differences in a repair-congruent time order. Translating this knowledge into the context of clinical trials (reading with unknown time order) this means that negative scores can be obtained while the readers have no idea that they are actually looking at repair. A second experiment embarked on these results, and showed that readers often thought that they were looking at specific repair signs, but that they actually did that as often in cases with progression as compared to cases with repair (20). A third exercise, in which the radiographs of hands and feet to which the single joints belonged were scored by regular trial readers, showed that the trial readers' scores of the index joints agreed very well with the majority judgement of the expert panel on those joints, also if the majority judgement was "improvement" (21). Importantly, a few index joints belonging to hands and feet that showed progression over time showed improvement according to both trial readers and expert panel. This latter finding adds to the recognition that repair may occur in occasional joints, but that change in TSS is still positive. In other words, repair may occur far more frequently than we have recognised from aggregated trial results.

We are currently working on trial databases in an attempt to show the consistency of repair in single joints across different (more than 2) time points, in order to corroborate the level of evidence that repair is a true – and not only a statistical - phenomenon.

Scoring radiographic progression in individual patients

An important disadvantage of the scoring methods for clinical trials is the fact that they require significant training, and that scoring according to these methods is very time consuming, mak-

ing these techniques unfeasible for routine clinical practice. In order to overcome these limitations, we have developed Simplified Erosion and Narrowing Score (called "SENS"), that is entirely based on the van der Heijde modification of the Sharp score (22). It exploits the same joints of hands and feet, but only asks for the presence or absence of erosions (bimodal answer modality) and/or joint space narrowing per joint, thus arriving at a sum score of 86 in stead of 448. The SENS was shown to be reliable with respect to intra-and interreader reliability, and is sensitive to change. Its decisive advantage is its feasibility in clinical practice.

Conclusion

Conventional plain radiography of the hands and feet is still vivid in the determination of the course of RA and the effects of treatment. It can serve as an outcome parameter in clinical trials that investigate the potential of new drugs to preserve structural integrity of the joints, and it can be used as a reflection of disease activity in clinical care. Appropriate scoring methods are pivotal in order to quantify radiographic progression, especially in clinical trials. Scoring radiographs is "work of man", and measurement error is a serious concern in the interpretation of radiographic results of clinical trials. And one should realise that these same limitations apply to other imaging techniques, probably even to a greater extent. Developments in methodology and robust epidemiological research in this field have learned that measurement error is manageable, and that subtle new effects such as repair may emerge from trial results.

References

1. LARSEN A, DALE K: Standardized radiological evaluation of rheumatoid arthritis in therapeutic trials. *In* DUMONDE DC and JASANI JK (Eds): *Recognition of Anti-Rheumatic Drugs*. Lancaster, MTPPress 1977: 285-92.
2. SHARP JT, LIDSKYMD, COLLINS LC, MORELAND J: Methods of scoring the progression of radiologic changes in rheumatoid arthritis. Correlation of radiologic, clinical and laboratory abnormalities. *Arthritis Rheum* 1971; 14: 706-20
3. VAN DER HEIJDE D: How to read radiographs according to the Sharp/van der Heijde method. *J Rheumatol* 1999; 26: 743-5.
4. BRUYNESTEYN K, VAN DER HEIJDE D, BOERS M *et al.*: Detecting radiological changes in rheumatoid arthritis that are considered important by clinical experts: influence of reading with or without known sequence. *J Rheumatol* 2002; 29: 2306-12.
5. FRIES JF, BLOCH DA, SHARP JT *et al.*: Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial. *Arthritis Rheum* 1986; 29: 1-9.
6. VAN LEEUWEN MA, VAN RIJSWIJK MH, SLUITERWJ *et al.*: Individual relationship between progression of radiological damage and the acute phase response in early rheumatoid arthritis. Towards development of a decision support system. *J Rheumatol* 1997; 24: 20-7.
7. WELSING PM, LANDEWE RB, VAN RIELPL *et al.*: The relationship between disease activity and radiologic progression in patients with rheumatoid arthritis: a longitudinal analysis. *Arthritis Rheum* 2004; 50: 2082-93.
8. DROSSAERS-BAKKER KW, DE BUCK M, VAN ZEBEN D, ZWINDERMAN AH, BREEDVELD FC, HAZES JM: Long-term course and outcome of functional capacity in rheumatoid arthritis: the effect of disease activity and radiologic damage over time. *Arthritis Rheum* 1999; 42: 1854-60.
9. SCOTTDL, SMITH C, KINGSLEY G: Joint damage and disability in rheumatoid arthritis: an updated systematic review. *Clin Exp Rheumatol* 2003; 21 (Suppl. 31): S20-7.
10. ØDEGARD S, LANDEWE R, VAN DER HEIJDE D, KVIEN TK, MOWINCKEL P, UHLIG T: Radiological damage is asociated with impaired grip strength independently of disease activity: A longitudinal 10-year analysis in patients with rheumatoid arthritis. *Ann Rheum Dis* 2005; 64 (Suppl. III): 184.
11. ØDEGARD S, LANDEWÉ RBM, VAN DER HEIJDE D, KVIEN TK, MOWINCKEL P, UHLIG T: Radiological damage contributes to physical disability: a 10-year longitudinal analysis of a cohort of patients with rheumatoid arthritis of short disease duration. *Ann Rheum Dis* 2005; 64 (Suppl. III): 182.
12. GEUSENS PMB, VAN DER HEIJDE D, Sj vdL, GARNERO P, LANDEWE R: Serum levels of receptor activator of nuclear factor kappa-beta-ligand (RANKL) and osteoprotegerin (OPG) independently predict long-term radiographic progression in patients with early rheumatoid arthritis. *Arthritis Rheum* 2003; 48 (Suppl. 1): s458.
13. KLARESKOG L, VAN DER HEIJDE D, DE JAGER JP *et al.*: Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: Double-blind randomised controlled trial. *Lancet* 2004; 363: 675-81.
14. LANDEWE R, VAN DER HEIJDE D, BURMESTER G, PEREZ JL, SPENCER-GREEN G: Radiographic improvement in clinical responders in the early treatment of recent-onset rheumatoid arthritis: Subanalysis of the PREMIER study. *Ann Rheum Dis* 2005; 64 (Suppl. III): 442.
15. LANDEWE R, VANDER HEIJDED: Radiographic progression depicted by probability plots: presenting data with optimal use of individual values. *Arthritis Rheum* 2004; 50: 699-706.
16. RAU R, WASSENBERG S, HERBORN G, PERSCHELWT, FREITAG G: Identification of radiologic healing phenomena in patients with rheumatoid arthritis. *J Rheumatol* 2001; 28: 2608-15.
17. RAU R, SANDER O, WASSENBERG S: Erosion healing in rheumatoid arthritis after anakinra treatment. *Ann Rheum Dis* 2003; 62: 671-3.
18. VAN DER HEIJDE D, LANDEWE R: Imaging: Do erosions heal ? *Ann Rheum Dis* 2003; 62 (Suppl. 2): ii10-2.
19. LANDEWE R, VAN DER HEIJDE D, BOONEN A *et al.*: Providing readers with films of the entire hand or foot compared to single joints does not improve the ability to discriminate between progression or repair of radiographic damage in RA. *Ann Rheum Dis* 2005; 64 (Suppl. III): 183.
20. VAN DER HEIJDE D, LANDEWE R, WINALSKI CS *et al.*: Erosion repair in rheumatoid arthritis is recognised by expert readers but cannot be distinguished from progression by specific features. *Ann Rheum Dis* 2005; 64 (Suppl. III): 183.
21. VAN DER HEIJDE D, LANDEWE R, BOONEN A *et al.*: Negative modified Sharp-scores truly reflect repair in joint damage: a validation study. *Ann Rheum Dis* 2005; 64 (Suppl. III): 186.
22. VAN DER HEIJDE D, DANKERT T, NIEMAN F, RAU R, BOERS M: Reliability and sensitivity to change of a simplification of the Sharp/van der Heijde radiological assessment in rheumatoid arthritis. *Rheumatology (Oxford)* 1999; 38: 941-7.