Repair in Rheumatoid Arthritis, Current Status. Report of a Workshop at OMERACT 8

DÉSIRÉE van der HEIJDE, ROBERT LANDEWÉ, JOHN T. SHARP for the OMERACT Subcommittee on Repair

ABSTRACT. Repair of structural damage in rheumatoid arthritis has drawn much attention with newly available effective treatments. A workshop was held at OMERACT 8 to update current knowledge on the validity of the concept of repair and on the assessment of repair. In preparation for the workshop several studies were performed and the results were presented. This was followed by a discussion and voting on statements on various aspects of repair. A majority of participants agreed that results of the new studies strengthen the validity of the concept of repair, and that repair can be assessed on radiographs. There was less agreement on the best means of measurement and there was a plea for more extensive reporting of data, i.e., not limited to sum scores of all joints together. The conclusions of the workshop mean a big step forward in the acceptance and assessment of repair. (J Rheumatol 2007;34:884–8)

Key Indexing Terms:
RADIOGRAPHS REPAIR DAMAGE EROSIONS
JOINT SPACE NARROWING RHEUMATOID ARTHRITIS

During OMERACT 6 in Brisbane, Australia, a workshop on repair of structural damage in rheumatoid arthritis (RA) was held. Since then, major progress has been made in this field. Much work has involved inferring the presence of repair from scoring radiographs using various strategies. Many of the research questions formulated at that time have been answered, and the time has come to try to reach consensus based on this new information. During the first workshop on repair it was concluded that repair does take place, and that it is worth further investigation[1]. There is also continued interest in the use of other imaging modalities alongside traditional radiographs to shed light on the concept of repair. It was also stated that for the time being we would only consider repair of erosions, not of joint space narrowing representing cartilage, as it was unknown whether cartilage has the capacity to repair. This statement should now also be revisited. The issues that remained unanswered are discussed below.

From the Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, and CAPHRI Research Institute, University Maastricht, Maastricht, The Netherlands; and the University of Washington, Seattle, Washington, USA.

Studies 1 to 3 were supported by unrestricted grants from Abbott, Amgen, and Centocor to the OMERACT Working Group on Repair and by contributed services from BioImaging Technologies.

D. van der Heijde, MD, PhD, Professor of Rheumatology; R. Landewé, MD, PhD, Associate Professor of Rheumatology, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, and CAPHRI Research Institute, University Maastricht; J.T. Sharp, MD, Affiliate Professor of Medicine, University of Washington.

Members of the OMERACT Subcommittee on Repair: M. Boers, VU University Medical Center, Amsterdam, The Netherlands; A. Boonen, University Hospital Maastricht, Maastricht, The Netherlands; S. Einstein, BioImaging Technologies, Newtown, PA, USA; P. Emery, University of Leeds School of Medicine, Leeds, United Kingdom; G. Herborn; R. Rau; S. Wassenberg, Evangelische Fachkrankenhaus Ratingen, Ratingen, Germany; B. Weissman; C. Winalski, Brigham and Women's Hospital, Boston, MA, USA.

Address reprint requests to Prof. D. van der Heijde, Department of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands. E-mail: d.vanderheijde@kpnplanet.nl

Training and specific features of repair

Data available from a few studies among experts in radiographic scoring showed that experts agreed on the presence of repair, and that a repair judgment was based mainly on a reduction of erosion size[2]. Our studies have not shown which morphologic features considered to be specific for repair play a definitive role in readers' judgment regarding presence of repair. As there were a few potential drawbacks in the studies performed, it was decided to execute 2 additional studies[3]. In preparation, all 8 experts participating in the new studies were very experienced, reducing the possibility that a lower level of experience would negatively influence the outcome. Moreover, there was a training session to ensure that everyone agreed on the definition of the features of repair. Several of the definitions were revised, providing greater detail, and were illustrated by examples. A new set of images was selected by one investigator, who supervised blinding and did not participate in reading the images.

Study 1 was an exercise involving 64 single joints at 2 timepoints. Study 2 involved images of the whole hand or foot that contained the individual joints from Study 1. This allowed the reader to incorporate information from other joints, as study merely of single joints eliminated the information that can be derived from an entire hand or foot. The single-joint study showed that the experts frequently agreed on the presence of repair, and that this agreement was regularly associated with a perceived change of the size of the erosions. Experts were unable to distinguish if the change in erosion size was a case of progression or repair. Specific features of repair could not be reproduced reliably and were insufficiently helpful in distinguishing progression from repair. The presentation of the entire hand or foot in Study 2 did not

improve any of the results. At OMERACT 6 it was questioned whether specific features of repair without a reduction in erosion size would be sufficient to infer repair. The results of Studies 1 and 2 did not provide positive information that repair indeed could be demonstrated by specific features. A reduction in erosion score seems to be a prerequisite to demonstrate repair.

Can repair judged by experts be picked up by negative Sharp-van der Heijde scores?

Study 3 addressed the question whether repair judged to be present by experts is reflected in negative Sharp-van der Heijde scores. The 60 radiographs of both hands and feet, including the joints scored by the experts in the previous experiments, were presented to 2 readers experienced in the Sharp-van der Heijde method[3,4]. The readers had not been involved in the discussions on repair and were not aware of the purpose of the scoring of the radiographs. The 2 independent readers found a mean overall negative change score in 23 patients assessed by Sharp-van der Heijde. In these patients, repair in the joint of interest was found by the panel in 17 patients; no repair in the joint of interest was seen by the panel in the remaining 6 patients, but the independent readers also did not apply a negative score to these joints of interest. Repair in the joint of interest was seen both by the panel and by the independent readers in 7 additional patients, but these patients had a mean overall positive change score. Therefore, in patients with total positive change scores, repair in individual joints was sometimes identified by the readers. Moreover, it became obvious that progression and repair can be seen in the same patient in different joints. The readers' negative change scores were in every case based on a reduction in the erosion score, while the positive change scores were a combination of progression in both the erosion and joint space narrowing scores. This is in agreement with the judgment by the panel, who based repair mainly on a reduction in the erosion size.

The conclusion from Study 3 was that regular readers that read with unknown time order can unconsciously identify repair in individual joints by quantifying a change in erosion size or number. This conclusion is consistent with the conclusion of Studies 1 and 2, in that a reduction in erosion size is more important than specific features of repair.

Do we need a separate scoring method for repair?

Combining the information of the 3 exercises described above, we have demonstrated that the current van der Heijde-Sharp scoring method detects repair sufficiently, and that new scoring methods with specific features of repair are not sensitive or specific enough to be useful. Based on the available data, it is most likely that repair "simply" reflects a reduction in the size of existing erosions that is detected by readers applying a scoring method, who describe a difference but are not aware that it constitutes repair because the time order is concealed. This means that the current scoring methods can be used to simultaneously assess progression and repair, which is convenient, more efficient, and less costly. However, presentation of the data as an overall change score for the entire patient might underestimate the presence of genuine repair (as might also be the case for presence of true progression). We suspect this can be solved by a different means of presentation rather than using a different scoring method, a topic currently under investigation.

How to present data on repair?

The use of cumulative probability plots is a great aid in understanding complex data[5]. Instead of group-level data, individual data from all group members are displayed. The major advantage is that the score of every patient is plotted in an orderly manner from lowest through highest observed score, and the proportion and magnitude of negative scores can be derived easily. Another application is the combination of different aspects of the same subject in one plot, for example, the scores of 2 different measures for the same patients[6]. Probability plots are only a visualization of the data, they do not replace statistical testing.

Current status of repair

A fundamental question is whether summed scores of all joints per patient do sufficient justice to the issue of repair. Based on the data summarized here, which show that positive overall change scores do not preclude negative scores in individual joints, the answer is no. However, presenting data on a joint level introduces several issues. First, we are denying the coherence of joints within a patient, which are not independent entities. Second, should the data be based on the scores of one or 2 or more readers? We know that for the total score, readers agree sufficiently well; however, for individual joints this agreement is worse. Moreover, the clinical relevance of repair in a single joint with an overall level of progression would need to be established. At OMERACT 6 it was suggested that the smallest detectable difference (SDD) could be used as a cutoff point to decide if a patient shows repair. However, on this subject the field has evolved further. It has recently been shown that the smallest detectable change (SDC), which is smaller than the SDD, is the more appropriate measure[7]. However, SDC values are still probably too large to be useful to determine repair at the level of an individual patient. Introducing presentation of data at a joint level means introducing new issues. This is currently still in the exploratory stage.

Prevalence and sites of repair

The TEMPO trial had a significant proportion of patients with negative scores[8]. If we analyze these data on a joint level, however, only 4.5% of all joints scored have a negative erosion score in one of 4 readings (scored twice by 2 readers). This seems a very low percentage, but the percentage of joints

that showed progression in the erosion score (3.7%) was similarly low. Progression and repair were found with a roughly similar frequency, in proximal interphalangeal (PIP), metacarpophalangeal (MCP), and toe joints, and with a somewhat lower frequency in MCP 4 and 5 and the interphalangeal of the thumb. These findings of low frequency and wide distribution make it more difficult to apply further study, such as magnetic resonance imaging (MRI) or synovial biopsy, even though at OMERACT 6 one-quarter of the participants voted in favor of studies with synovial biopsies to relate inflammation to repair, and three-quarters proposed the use of MRI for this purpose.

When to decide that repair on a group level exists?

The testing of repair on a group level is a purely statistical concept. The null hypothesis is that the true change over time is not different from zero. To test this, we are looking for a within-group effect, and the hypothesis can be rejected if the 95% confidence interval of the mean progression score is entirely below zero (or in case of progression, above zero). It is crucial that the radiographs are read with concealed time order. This has been proposed and described by 2 of us to test if repair is present in a particular treatment arm[9]. Note that this type of information gives no insight into the real level of repair on a patient and/or joint level, which is much more difficult to discern, and different means of presentation of the results should be used to achieve this.

Do negative scores have clinical correlates?

It was clearly stated at OMERACT 6 that a correlation between negative scores (as an illustration of repair) and clinical outcomes should be established, to show that it is indeed meaningful to have repair of joint damage as compared to no progression. A further goal would be to investigate whether an early finding of a negative score translates into a later ("downstream") benefit. These are particularly difficult issues, as it took decades to be able to show that radiographic progression indeed has an unfavorable effect on patient outcomes, even though it was obvious from just looking at damaged joints on radiographs that this would have an impact on physical functioning. Recently, however, using multivariate analysis, we were able to demonstrate that positive and negative changes in a one-year period are both related to function in a dose-related relationship between one-year radiographic progression and physical function in the TEMPO trial. Patients with negative changes over a one-year period had the lowest disability scores (expressed as the one-year Health Assessment Questionnaire score), followed by those with no progression, then those with mild progression, those with severe progression showing the highest disability scores[10]. This "dose-response" relationship with function makes repair a meaningful feature, but care should be taken not to interpret this correlation as causal.

Relation between inflammation and repair in a joint

It was postulated that repair cannot occur in a joint with ongoing inflammation. It has been shown in the COBRA trial that inflammation (expressed as swelling) in a particular joint was associated with the development or the progression of damage in the same joint, strengthening the case for a causal relationship between inflammation and structural damage on a joint level[11]. To address the question whether repair can occur in joints with ongoing inflammation, we examined the data of the TEMPO trial in a per-joint analysis. For this analysis we used the information from the MTX arm and the etanercept plus MTX arm, and we pooled the data from both treatment arms. The radiographs were scored twice (with a one-year interval) by 2 readers. This resulted in 4 independently obtained scores per joint. For the present analysis we selected the PIP, MCP, and metatarsophalangeal joints. We selected all joints that showed a decrease in erosion score in at least one reading (one reader at one timepoint) with a stable score in the remaining readings ("stable/repair"). In total, 557 joints out of 11,159 (5.0%) fulfilled these criteria. We combined the data on swelling with the stable/repair condition in these same joints. Fifty-seven percent of the joints with a negative score showed improvement in swelling, and 42% of the joints showed stable swelling. Of these joints with stable swelling, 94.8% showed no swelling during the year of followup. None of the joints showed worsening in swelling.

Of the 10,497 joints not fulfilling the stable/repair condition, 40.2% showed improvement in swelling, 58.4% showed stable swelling, and 1.3% showed worsening in swelling. The difference in distribution of swelling between both radiographic conditions was highly statistically significantly different (Fisher exact test $p < 0.0001$), implying that there is a relationship between the absence of inflammation and the repair/stable state[12].

A second prerequisite to be able to show repair is the presence of damage at baseline. To confirm the relationship between inflammation and the presence of damage we divided the joints into 4 groups: Group A1 showing persistent swelling but no radiographic damage at baseline. Radiographic damage is defined here as erosion if we are evaluating erosions, and joint space narrowing if we are evaluating joint space narrowing. Group A2 showing persistent swelling, but with radiographic damage at baseline. Group B1 showing either no swelling or improved swelling during followup, but no radiographic damage at baseline, and group B2 showing either no swelling at baseline or improved swelling during followup, but with radiographic damage at baseline. The 2 hypotheses were that progression would occur in both Group A1 and A2, but would be more pronounced in Group A2, and that repair could occur only in patients in Group B2.

In total, 3% of the joints fitted in Group A1, < 1% in Group A2, 81% in Group B1, and 15% in Group B2. Group A1 showed significant progression. As well, Group A2 showed progression with a higher point-estimate, but did not reach

statistical significance, probably due to the small number of joints in this group. Groups A1 and A2 combined showed statistically significant progression. Group B1 showed no change, and Group B2 showed statistically significantly negative change scores, indicating repair[12]. Analysis of the data with joint space narrowing as the dependent variable yielded the same conclusion, although the level of both progression and repair was somewhat lower. Using the total score as the dependent variable, as expected, the differences were more pronounced.

In addition, we performed analyses for correlated data, using a generalized linear model for repeated measures and a linear mixed model, with erosions, joint space narrowing, and total score as the dependent variable in 3 separate analyses. Combination treatment, swelling score, and damage at the baseline radiograph were included as independent variables. These 2 types of analyses confirmed the results described above for erosions, joint space narrowing, and for total score: repair occurred in joints with baseline damage in which swelling was improved during treatment, and preferentially if the treatment was etanercept plus MTX (Lukas C, et al, unpublished data).

Is repair of cartilage possible?

The results described above that a decrease (improvement) in joint space narrowing was preferentially seen in patients with improvement in swelling with presence of baseline joint space narrowing suggest that cartilage repair might be a real prospect. However, additional studies should focus on joint space narrowing to gain insight into the possibility of cartilage repair. It is very likely that one or more of the computerized methods of measuring joint space width discussed in the OMERACT workshop will be helpful in this area[13].

Discussion at OMERACT 8

During OMERACT 8 the data presented above were introduced and discussed, followed by voting on 10 statements with an anonymous keypad voting system. The results are presented in Table 1. Overall, there was strong agreement on the various validity aspects of repair based on the panel agreement on judging the second radiograph in time as the radiograph with the least damage, the agreement between negative scores by independent readers and the panel judgment for repair, and the almost exclusive occurrence of negative scores in joints with absent or improved inflammation. There was a less pronounced majority who judged that repair being the opposite of progression might be related to better functional outcome. Opinion was split on whether the joint space narrowing data were indicative of repair of cartilage. As well, participants were uncertain about the best way to present the data; a considerable percentage of participants were not sufficiently familiar with probability plots to judge the usefulness, but of those that were, the large majority voted for the use of probability plots, and also for the use of the 95% confidence interval to judge repair on a group level. A majority of participants wanted to see more information than just sum scores, creating a challenge about how to do this properly, taking into account measurement error on a joint level and scores from different readers. Opinion on a conventional scoring method being sufficient to assess repair if applied with concealed time order was split in 3 equal parts. There might be several reasons for this: participants felt that other imaging methods

Table 1. Results of voting on statements, expressed as percentage of participants taking part in the vote.

Statement for Voting	Agree, %	Disagree, %	Don't Know, %
The fact that a panel of experienced readers can reliably assign which image in a set of two consecutive images presented to them with random time order is best, adds to the validity of repair (as many of the best films were the second in time)	86	6	8
The fact that trial readers uninvolved in the repair experiments agree almost perfectly with a panel of experienced readers with regard to the assignment of a negative score to a particular joint adds to the validity of repair	82	4	14
The fact that negative joint scores almost exclusively occur in joints that demonstrate improvement of swelling — and not in joints with persistent or worsening swelling — adds to the validity of repair	70	11	19
In terms of assessment, repair in a joint is the opposite of progression	60	19	21
The data on joint space narrowing do suggest that repair of cartilage is occurring	40	25	35
There is an indication that repair based on negative scores is independently associated with better functional outcome	54	13	33
In a RCT with reading with concealed time order, repair on a group level can be statistically demonstrated if the mean within-group progression score and the entire 95% CI is below zero	61	6	33
To best demonstrate the full information on changes in radiographic damage (progression and repair), probability plots should be used	55	2	43
To get insight in repair not only sum scores but also separate joint scores should be presented, since positive sum scores may conceal negative joint scores	75	14	12
Repair can be assessed by the conventional scoring methods, and there is no need for a separate scoring method and/or read, provided that the films are scored with concealed time order	35	33	31

RCT: randomized controlled trial.

could add extra information, that a concealed time order is not necessary, that other features should be scored on radiographs, that there should be a separate reading, and so on. The voting results were in contrast with the fact that the participants accepted negative scores in a conventional scoring method as a valid aspect of repair. The reasons for these unexpected voting results need to be explored further, but in retrospect some questions were suboptimal, because they addressed more than one topic in a single question.

In conclusion, there was broad acceptance of the existence and validity of repair among the participants of the workshop. This opens the way to incorporate assessment of repair as an outcome measure in future studies, such as observational studies, but also in the evaluation of treatment efficacy.

REFERENCES

1. van der Heijde D, Sharp JT, Rau R, Strand V. OMERACT Workshop: Repair of structural damage in rheumatoid arthritis. J Rheumatol 2003;30:1108-9.
2. Sharp JT, van der Heijde D, Boers M, et al. Repair of erosions in rheumatoid arthritis does occur. Results from 2 studies by the OMERACT Subcommittee on Healing of Erosions. J Rheumatol 2003;30:1102-7.
3. van der Heijde D, Landewé R, Boonen A, et al. Negative Sharp-scores may indeed reflect repair of existing joint damage due to rheumatoid arthritis: Results of three experiments with expert consensus. Submitted for publication.
4. van der Heijde DM, van Riel PL, Nuver Zwart IH, Gribnau FW, van de Putte LB. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. Lancet 1989;1:1036-8.
5. Landewe R, van der Heijde D. Radiographic progression depicted by probability plots: presenting data with optimal use of individual values. Arthritis Rheum 2004;50:699-706.
6. Wanders A, Landewe R, Dougados M, Mielants H, van der Linden S, van der Heijde D. Association between radiographic damage of the spine and spinal mobility for individual patients with ankylosing spondylitis: can assessment of spinal mobility be a proxy for radiographic evaluation? Ann Rheum Dis 2005;64:988-94.
7. Bruynesteyn K, Boers M, Kostense P, van der Linden S, van der Heijde D. Deciding on progression of joint damage in paired films of individual patients: smallest detectable difference or change. Ann Rheum Dis 2005;64:179-82.
8. 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.
9. van der Heijde D, Landewé R. Imaging: Do erosions heal? Ann Rheum Dis 2003;62 Suppl II:ii10-12.
10. van der Heijde D, Landewe R, van Vollenhoven R, Fatenejad S, Klareskog L. The level of radiographic damage and 2-year radiographic progression are determinants of physical function. A longitudinal analysis of the TEMPO trial [abstract]. Arthritis Rheum 2005;52 Suppl:S549.
11. Boers M, Kostense PJ, Verhoeven AC, van der Linden S. Inflammation and damage in an individual joint predict further damage in that joint in patients with early rheumatoid arthritis. Arthritis Rheum 2001;44:2242-6.
12. van der Heijde D, Lukas C, Fatenejad S, Landewé R. Repair occurs almost exclusively in damaged joints without swelling [abstract]. Arthritis Rheum 2006;53 Suppl:S512.
13. Sharp JT, Angwin J, Boers M, et al. Computer based methods for measurement of joint space width: Update of an ongoing OMERACT project. J Rheumatol 2007;34:874-83.