Subtle changes in individual joints result in both positive and negative change scores in a patient: results from a clinical trial in patients with rheumatoid arthritis

C Lukas,[1] R Landewé,[2] S Fatenejad,[3] D van der Heijde[4]

[1] Service d'Immuno-Rhumatologie, Hôpital Lapeyronie, Montpellier, France; [2] University Hospital and CAPHRI Research Institute, Maastricht, The Netherlands; [3] Wyeth Research, Collegeville, Pennsylvania, USA; [4] Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands

Correspondence to:
Dr C Lukas, Service d'Immuno-Rhumatologie, Hôpital Lapeyronie, 371 avenue du Doyen Gaston Giraud, 34295 Montpellier cedex 5, France;
cedriclukas@voila.fr

Accepted 15 November 2008
Published Online First 23 November 2008

ABSTRACT

Background: Radiographic progression in clinical trials is assessed by interpreting changes in total radiographic joint score, and the reliability of those scores depends on an evaluation of sum scores. It is not known how consistently changes in individual joints are identified by independent readers and in independent readings.

Patients and Methods: 7255 single joints from 178 patients who participated in the Trial of Etanercept and Methothrexate with Radiographic Patient Outcomes (TEMPO) trial were evaluated. Every image was independently scored twice according to the Sharp–van der Heijde method by two independent readers, so that four scores per joint were available. Absolute agreement and consistency of negative and positive erosion change scores across readers and readings were compared on a per-joint level, as well as on a per-patient level.

Results: The number of joints showing a change for erosion was very low in this trial: 691/7255 analysed joints had at least one non-zero change score out of four readings. Absolute agreement between readings was remarkably poor: only 12 joints showed a consistently positive or negative change in all four readings. Change scores in opposite directions in the same joint across independent readings were rare (25 joints). Frequency of opposite joint scores in the same patient (mixed change patterns) was reader dependent.

Conclusion: Substantial intra and interreader disagreement in scoring change in individual joints is common. Opposite joint scores in the same patient, however, are rare and reader dependent. Notwithstanding these subtle inconsistencies on the individual joint level, the total Sharp score is a useful and discriminatory outcome measure.

Joint damage progression as measured on consecutive plain radiographs of hands and feet is an important outcome when evaluating the course of rheumatoid arthritis (RA). Several scoring systems and modifications have been developed to quantify progression radiographically. The Larsen and Sharp methods, and their modifications, which were developed to quantify radiographic progression, are best known and applied most frequently in clinical trials.[1–3] Trials evaluating tumour necrosis factor alpha-blocking drugs in the treatment of RA have recently introduced the phenomenon of negative change, which could, among other things, indicate the repair of previously existing (erosive) damage in joints. A few trials have shown statistically significant negative average progression scores on a group level, leading to the possibility of joint repair. However, it is not known what such mean negative scores truly imply.[4 5] There is no doubt that part of the negative (but also the positive) scores is due to measurement error, but it is impossible to separate measurement error from true change. Results from recent studies have reported change scores in either direction that are so low that they could theoretically stem from changes within one or only a few joints. Currently, there are few insights into how the change scores at the patient level (patient score) reflect individual joint elements. Do negative and positive scores occur in the same patient, or does the direction of change dominate the patient score? Another unanswered question is "Are negative or positive changes in a single joint recognised independently of each other by independent readers, or in independent readings by one reader?"

There are a number of reports claiming the existence of joint repair in RA.[6–9] A subcommittee of the Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) Imaging Committee on the Healing of Erosions has conducted several exercises using selected case reports that underscore the validity of the concept of the repair of erosions.[10–12] These exercises have provided corroborating evidence regarding the validity of currently existing scoring methods in the detection of repair.

The relation between negative and positive change scores, and the reliability of both phenomena, have never been investigated at the level of single joints in a large unselected sample of patients with RA.

We have therefore evaluated the consistency of positive and negative individual joint change scores, as well as their occurrence within the same patient, in the Trial of Etanercept and Methotrexate with Radiographic Patient Outcomes (TEMPO), which is a large randomised clinical trial that showed a statistically significant negative mean change score in one of the trial arms and a statistically significant positive change score in another trial arm.[4] This trial was chosen because a large set of radiographs has been scored twice independently, by the same two readers, thus providing a unique opportunity to learn about agreement in scoring negative and positive changes in individual joints.

PATIENTS AND METHODS

The TEMPO trial was a 3-year study that evaluated clinical and radiographic outcomes of

patients with RA treated with methotrexate alone, etanercept alone or the combination of both drugs.[4] This analysis has used data collected during the first 2 years of the study.[13] During the reading of the radiographic images at the end of the first year, three readers scored all baseline, 6-month and 12-month radiographs in such a manner that every patient was scored by two readers. During the readings after the second year, all available baseline and 12-month radiographs were scored again by two of the three readers of the first panel. By doing so, a set of four readings per joint was available for each of the patients included in this analysis. Radiographs were scored using the van der Heijde modification of the Sharp score method.[3] This method quantifies the number and size of erosions in 32 joints of the hands and wrists and 12 joints of the forefeet, and the degree of joint space narrowing in 30 joints of the hands and wrist and 12 joints of the forefeet. Readers see all radiographs of a patient appearing on a screen grouped for the proximal interphalangeal joints, metacarpophalangeal joints, wrist and feet, score joint per joint, and decide on their joint scores by simultaneously comparing radiographs from the same patient at different time points, although they do not know the order in time (concealed time order). They do not score change directly, but they can bring change in their scores by assigning different scores to different time points. They cannot assign whether they think an observed change in a joint is due to repair or progression, because such an assignment requires knowledge about the true time order. We have demonstrated previously that readers are unable to assign the true time sequence (or to distinguish repair from progression) to pairs of single joints of hands and feet or pairs of entire radiographs, so that we assume for the remainder of this analysis that the occurrence of change in individual joints under conditions outlined above is a process not driven by the readers' presumption about the sequence of images.

The analyses provided in this report are based only on those images that were scored four times. As one of the goals of this single joint study was to gain insight into the validity of negative joint scores, and the discussion about repair involves the repair of previously existing erosions rather than the restoration of articular cartilage (joint space width), the analyses provided here are limited to erosion scores only.

Analyses

In total, change scores of 7255 single joints belonging to 178 patients were investigated. For all 7255 joints, four scores per joint were available. Of these 178 patients, 53 belonged to the methotrexate arm, 60 to the etanercept only arm and 65 to the methotrexate plus etanercept combination arm.

In a first analysis, frequencies of joints scored with negative change (improvement), positive change (worsening) or no change over time were described for each of the four readings regardless of the magnitude of the change.

In a second analysis, we investigated per reading (N = 4) whether negative and positive change scores occurred in the same patient, how frequently this phenomenon occurred, and what was the impact on total change scores.

Finally, the agreement of change scores per joint was investigated by establishing the concordance of positive and negative change scores across the four independent readings.

RESULTS

The frequency of positive and negative single joint change scores as a percentage of all 7255 single joints that were available for analysis, tabulated by reader and by reading, is shown in table 1. It is apparent that change, either positive or negative, was a very rare feature in this trial; the great majority of joints was scored as unchanged; between 1.3% and 5.8% of the joints were identified as changed readings. There was, however, intra and interreader variation: reader 1 scored a higher number of joints with change than reader 2 in both readings, and both readers assigned a change to a higher number of joints in the first reading compared with the second reading. In three of the four readings, there was a slight dominance of negative change scores over positive change scores and only reader 2 saw slightly more positive than negative change scores in one of the two readings.

We further analysed the extent to which both positive and negative single joint change scores co-occur in the same patient by aggregating single joint scores from each patient. In order to do so, three-dimensional frequency plots (histograms) were created, plotting the frequency of patients on the Y-axis, the number of joints with a positive change score on the X1-axis, and the number of joints with a negative change score on the X2-axis (fig 1). The analysis was carried out for each reading. In panel A of fig 1, some patients had no (neither positive nor negative) change in any joint, which is consistent with a sum score of zero (no change at a patient level). These patients (N = 43 for reader 1, first reading) are reflected by the highest bar at the crossing of the three inner axes of the graph (fig 1A). Patients who have one or more joints with only positive or only negative changes are depicted along one of the inner X-axes of the graph. They represent the second most frequent proportion of patients in this analysis. The three-dimensional space of the graph represents the patients who have some joints with positive changes and some joints with negative changes. The most extreme was a patient who had five joints with a negative change score and four joints with a positive change score (circle in fig 1A).

Looking at the four panels together, as well as the summarising table 2, it is obvious that in all readings (except reader 2, second reading, which was extremely "conservative") the patients with some change outnumber the patients without any change, and that in patients with an observed change those with a unidirectional change outnumber those with a mixed change pattern, but that patients with a mixed pattern of change do exist.

It is also obvious from the figures, when comparing panels A and C with panels B and D, that reader 1 in comparison with reader 2 not only assigned more joints with change (table 1) but also provided more patients with a mixed pattern, both in reading 1 and in reading 2 (table 2).

Consistency of scoring across independent readings

In the subsequent analysis we investigated the degree of agreement among readers in assigning a positive or negative change score to the same joint (table 3). This table lists the frequency at which joints are assigned a positive or a negative change score in independent readings by independent readers (the possible categories are no change (not shown), positive change and negative change). Only 12 of the 7255 analysed joints had a similar change assignment in all four readings: six with a positive change score and six with a negative change score. More joints had similar change assignments in three of the four, or in two of the four readings.

Given the very poor reproducibility at the individual joint level, we investigated whether opposite scores were being assigned to joints in independent readings (table 4). The table

Table 1 Summary of joint evaluations by reader and reading categorised by change in erosion scores

	Reader 1, first reading n (%)	Reader 1, second reading n (%)	Reader 2, first reading n (%)	Reader 2, second reading n (%)
Joint with a change	419 (5.8%)	269 (3.7%)	213 (2.9%)	98 (1.3%)
Positive change	169 (2.3%)	86 (1.2%)	90 (1.2%)	52 (0.7%)
Negative change	250 (3.4%)	183 (2.5%)	123 (1.7%)	46 (0.6%)
Joints with no change	6836 (94.2%)	6986 (96.3%)	7042 (97.1%)	7157 (98.6%)
All joints	7255	7255	7255	7255
Total erosion score, mean (SD), range	−0.76 (3.14), −14–21	−0.31 (2.49), −15–8	−0.71 (2.44), −12–15	0.08 (1.44), −8–7

Positive and negative change scores in the same patient.

lists the number of joints in which one reading, two readings or three readings assigned the same change (either positive or negative) to a single joint. It also lists per category of agreement the number of opposite scores assigned in one or more of the other readings. Note that the category of ''positive in one reading only'', which is applicable to 215 single joints, implies that there are 645 (three times 215) scores available that stem from the other readings in which this joint was not assigned a positive score. The picture is clear in that the majority of ''remaining readings'' yielded no-change scores. However, opposite results did occur at a low frequency. We identified one joint that was assigned positive change scores in three readings and a negative change score in the remaining reading. Another observation is that opposite results occurred more frequently in the case of positive change scores (4.2%, 3.8% and 7.1%, respectively) than in case of negative change scores in the majority of the readings (2.8%, 2.2% and 0%, respectively).

DISCUSSION

This single-joint analysis on radiographic progression showed that the level of agreement in assigning a positive or a negative change score to a pair of joints among readers (or at subsequent occasions) is extremely low. As a matter of fact, in the case of a change score, full agreement (similar results in four out of four readings) was obtained in only 12 of the 706 joints (1.7%), and almost complete agreement (similar results in three out of four readings) in 40 of 706 joints (5.7%).

At a first glance, these figures seem disappointing and in contrast to the reproducibility of the Sharp score and its

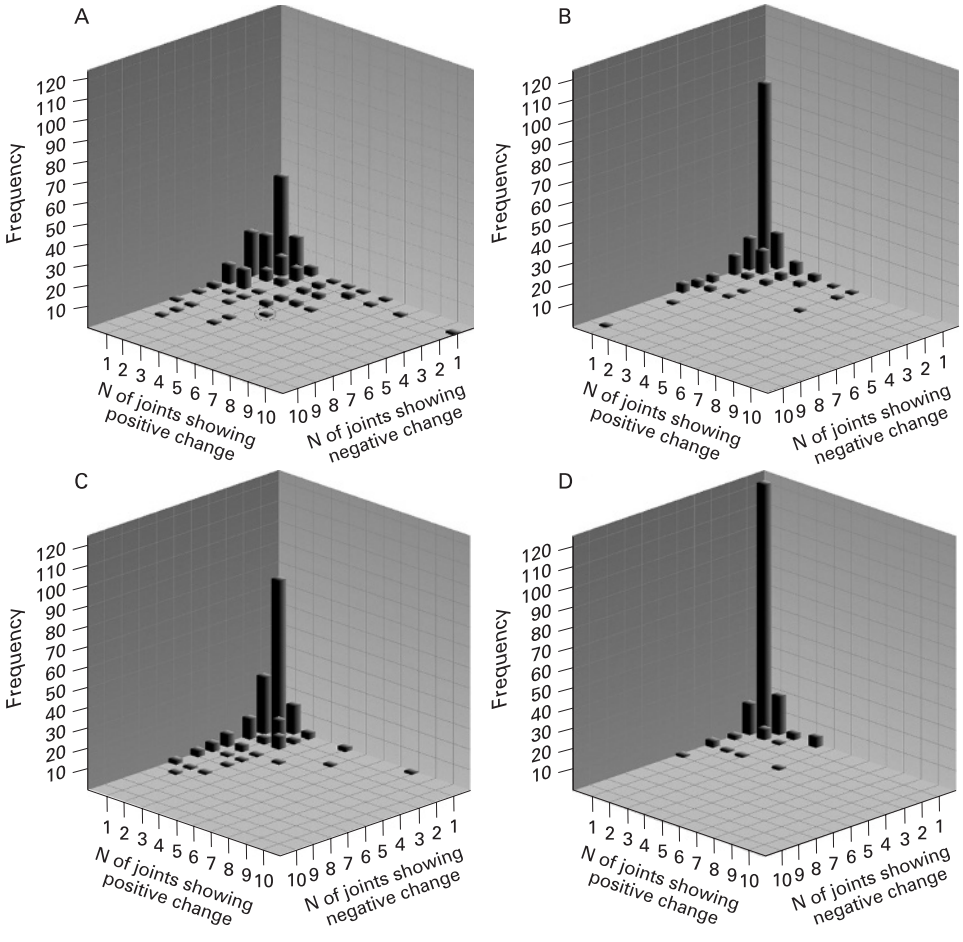

Figure 1 Number of patients with every combination of positive and negative changes in erosion score per joint, among 44 scored joints per patient. (A) Reader 1, first reading; (B) Reader 2, first reading; (C) Reader 1, second reading; (D) Reader 2, second reading.

Table 2 Proportion of patients with a particular scoring pattern, per reader and per reading separately

	Reader 1, first reading	Reader 1, second reading	Reader 2, first reading	Reader 2, second reading
% Patients with no change at all (zero positive, zero negative)	24.2	41	49.4	68
% Patients with unidirectional positive joints	13.5	11.2	15.7	15.2
% Patients with unidirectional negative joints	28.7	29.8	19.1	11.8
% Patients with a mixed pattern (see definition above)	33.7	18	15.7	5.1

modifications seen in validation studies and clinical trials.[4 14 15] Especially since modern clinical trials show only minimal progression, disagreement between readers at the joint level may have a potentially important impact on the study results. So the question is why aggregated joint scores actually do work appropriately in clinical trials. This study provides mitigating insight into how these seemingly discrepant observations can be explained, and how the van der Heijde–modified Sharp score (and probably other scoring systems) actually work in the context of a clinical trial.

First, a very small number of joints was assigned a (positive or negative) change score in at least one of the readings (table 1). The most likely explanation is that the large clinical trial database we investigated assessed therapies with confirmed efficacy for preventing the progression of structural damage. In the entire TEMPO trial, the mean change in erosion scores at one-year change was +1.68 for the methotrexate group and −0.30 for the methotrexate plus etanercept group. So one could expect a very low proportion of joints with a change assigned. More importantly, the poor absolute agreement in detecting change should be judged against the background of an extremely low previous probability of change, which may influence the performance of the readers. Suppose that only 3% of the joints are truly changed. This means that during a reading, the reader will assign ''no change'' 33 times more frequently than ''change'', which may make him reluctant to assign change. Readers will tend to assign ''no change'' in case of doubt. This hypothesis is supported by our observation that opposite scores are actually very rare (table 4); opposite scores occur at a frequency of approximately 4% or less, which is close to the average percentage of joints with change across readings, and as such are most probably due to chance occurrences (differences in judgement).

Second, an aggregated score such as the van der Heijde–modified Sharp score does not give insight into the pattern of joint changes within a patient. In the pre-biologics era, with less effective treatments, the sum score was composed of positive changes in several joints, but recently a number of trials have shown an average progression of 0 units, or even negative changes. Theoretically, such mean scores around zero could be made up of joint scores with opposite change. We have shown here that this theoretical possibility indeed occurs, but at a low frequency from 2.2% to 6.7% of the 178 patients investigated, depending on the reader, with negligible impact on the total change score.

Table 3 Evaluation of consistency of change scores in independent readings

	Positive change scores N (% of total joints)	Negative change scores N (% of total joints)
In only one of four readings	215 (3.0)	295 (4.1)
In two of four readings	52 (0.7)	92 (1.3)
In three of four readings	14 (0.2)	26 (0.4)
In all four readings	6 (0.1)	6 (0.1)

In comparing two readers in the two readings, we found a difference in the tendency to assign a mixed change pattern to patients. Reader 1 was more willing to accept patients with (a low number of) opposite scores than reader 2. But regardless of the reading or the reader, the greater majority of patients with a zero sum score were assigned ''no change'' to all joints, or, in case of change, a unidirectional pattern of change. It is interesting to speculate on the nature of the mixed change pattern. Because there is a demonstrable reader effect, and because the biological plausibility of a mixed change pattern is rather low, we tend to ascribe the mixed change pattern to measurement error rather than to a true (biological) effect. Reasoning along similar lines, a unidirectional change pattern may add to the credibility of joint damage progression or repair in a patient. This was undisputed with regard to progression, but so far the concept of repair has been criticised as being a measurement artefact. Admittedly, the number of patients with a unidirectional pattern of negative change was not high, and also reader dependent, but neither was the number of patients with a unidirectional pattern of positive change, which was also reader dependent. In the absence of a gold standard, these distinguishable unidirectional patterns add circumstantially to the validity of the concept of repair (or progression).

A few limitations should be mentioned here. For reasons of plausibility, we have only focused on erosion scores in this study, and we have excluded joint space narrowing scores from the analysis, but the picture would not be different as long as we consider scoring of erosions and joint space narrowing as independent phenomena. For reasons of convenience, we have investigated change as a binomial variable (change versus no change) thus ignoring quantitative information that may have impacted the total score. In this trial, however, the change in an individual joint was 1 unit in 76% of the joints with change (data not shown), so that we considered the impact of quantification on the total score as negligible.

How does this seemingly poor reliability and these individual joint observations eventually translate into changes in the total Sharp score at the patient level (and at the trial level)?

The overall reported change score of a treatment group in a trial is the average of all individual patient scores. The individual patient score is the average of Sharp scores provided by two (or more) readers. These readers judge entire patients rather than single joints and are implicitly able to bring a pattern in the direction of change in a patient. The total Sharp score is the sum of change scores of 44 individual joints. As such, the reported change score of a group of patients is a highly aggregated composite measure, incorporating the effects of hundreds of patients, the opinions of at least two readers about thousands of joints, and factoring in the implicit direction of change. We have seen that the absolute agreement in single joint scores is (very) poor, but we have also seen that change assignments in readings are hardly if ever effaced by opposite assignments in independent

Table 4 Occurrence of opposite results in the four readings

If the change is	Total no of joints	No of times the remaining reading(s) show	
		No change (%)	Change in the opposite direction (%)
Positive in one reading only	215	618 (95.8)	27 (4.2)
Positive in two readings	52	100 (96.2)	4 (3.8)
Positive in three readings	14	13 (92.9)	1 (7.1)
Negative in one reading only	295	860 (97.2)	25 (2.8)
Negative in two readings	92	180 (97.8)	4 (2.2)
Negative in three readings	26	26 (100)	0 (0)

readings. Similarly, mixed pattern assignments (mutually effacing effects) in individual patients are rare. So, if reader 1 assigns a positive change score to one particular joint, and reader 2 judges change in this particular joint as insufficiently clear and assigns "no change" to all joints, the total Sharp score for reader 1 will be +1 unit and for reader 2 0 units, in spite of the lack of absolute agreement, and the reported average Sharp score will be +0.5 units. If reader 1 factors in a unidirectional trend he may score two other joints positively, with consequences for his total Sharp score (+3 units) and for the grand mean score (+1.5 units), whereas reader 2 would have no reason to do that. Generally, neither reader 1 nor reader 2 would assign negative and positive change scores within the same patient (although the more sensitive reader 1 will probably do that a little bit more frequently than the conservative reader 2), so that the impact of these stochastic events is very limited. As such, subtle changes in individual joints that are not reproducibly assessed by independent readers because of differences in the level of certainty translate into subtle but quantifiable changes in a patient's total Sharp score, and eventually contribute to changes in group means. In modern trials with a very low level of true progression, scoring systems such as the modified Sharp score are instruments that challenge the level of confidence of individual readers in assigning change scores to potentially changed joints. These readers judge the joints of the entire patient, and are able to augment potential change if they are sufficiently confident. The low (biological) plausibility of opposite change scores within the patient and the lack of opposite results by other reader(s) protect the scoring system against a lack of sensitivity while there is a natural tendency to maintain specificity (conservatism in case of doubt). Importantly, it is crucial to maintain an absolute level of blinding of treatment and time order, in order to prevent any potential source of bias that may guide the reader in a spurious direction. In view of the subtle changes occurring in trials, such biased assignments could have an immediate impact on the total score.

This example clearly demonstrates that the common use of cut-off levels for progression is spurious in studies with mean progression scores close to zero. It may qualify a patient as a progressor, whereas in truth the result is the consequence of interreader disagreement.

In summary, we have shown here that, although absolute agreement among readers in individual joint scores is poor, opposing results within the same patient occur rarely. This single joint analysis explains why even very subtle changes in individual joints, assigned by one reader, translates into measurable changes at the level of change in total Sharp score and differentiation between treatment arms.

Competing interests: None.

REFERENCES

1. **Larsen A.** How to apply Larsen score in evaluating radiographs of rheumatoid arthritis in long-term studies. *J Rheumatol* 1995;**22**:1974–5.
2. **Sharp JT,** Young DY, Bluhm GB, *et al*. How many joints in the hands and wrists should be included in a score of radiologic abnormalities used to assess rheumatoid arthritis? *Arthritis Rheum* 1985;**28**:1326–35.
3. **van der Heijde D.** How to read radiographs according to the Sharp/van der Heijde method. *J Rheumatol* 2000;**27**:261–3.
4. **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.
5. **Landewe R,** Smolen JS, Keystone EC, *et al*. Radiographic inhibition of progression of structural damage: results from the RAPID 2 Trial [abstract]. *Ann Rheum Dis* 2008;**67**(Suppl II):321.
6. **Ideguchi H,** Ohno S, Hattori H, *et al*. Bone erosions in rheumatoid arthritis can be repaired through reduction in disease activity with conventional disease-modifying antirheumatic drugs. *Arthritis Res Ther* 2006;**8**:R76.
7. **Rau R,** Herborn G. Healing phenomena of erosive changes in rheumatoid arthritis patients undergoing disease-modifying antirheumatic drug therapy. *Arthritis Rheum* 1996;**39**:162–8.
8. **Rau R,** Herborn G, Wassenberg S. Healing of erosive changes in rheumatoid arthritis. *Clin Exp Rheumatol* 2004;**22**(5 Suppl 35):S44–9.
9. **Rau R,** Wassenberg S, Herborn G, *et al*. Identification of radiologic healing phenomena in patients with rheumatoid arthritis. *J Rheumatol* 2001;**28**:2608–15.
10. **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.
11. **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.
12. **van der Heijde D,** Landewe R, Boonen A, *et al*. Expert agreement confirms that negative changes in hand and foot radiographs are a surrogate for repair in patients with rheumatoid arthritis. *Arthritis Res Ther* 2007;**9**:R62.
13. **van der Heijde D,** Klareskog L, Rodriguez-Valverde V, *et al*. Comparison of etanercept and methotrexate, alone and combined, in the treatment of rheumatoid arthritis: two-year clinical and radiographic results from the TEMPO study, a double-blind, randomized trial. *Arthritis Rheum* 2006;**54**:1063–74.
14. **Boers M,** Verhoeven AC, Markusse HM, *et al*. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. *Lancet* 1997;**350**:309–18.
15. **St Clair EW,** van der Heijde DM, Smolen JS, *et al*. Combination of infliximab and methotrexate therapy for early rheumatoid arthritis: a randomized, controlled trial. *Arthritis Rheum* 2004;**50**:3432–43.