EXTENDED REPORT

Radiography as primary outcome in rheumatoid arthritis: acceptable sample sizes for trials with 3 months' follow up

K Bruynesteyn, R Landewé, S van der Linden, D van der Heijde

Ann Rheum Dis 2004;**63**:1413–1418. doi: 10.1136/ard.2003.014043

See end of article for authors' affiliations

Correspondence to: Professor D van der Heijde, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands; dhe@sint.azm.nl

Accepted 18 January 2004
Published online first
22 March 2004

Objectives: To investigate whether plain radiographs can show changes in joint damage due to rheumatoid arthritis (RA) within 3 months.

Methods: 188 film pairs taken with a 3 month interval were evaluated. They were scored with (chronological) and without (paired) knowledge of the sequence of the films according to the Sharp/van der Heijde method. Changes in joint damage were analysed on a group and an individual level for different subsets of patients. Sample sizes required to detect statistically and clinically significant differences were estimated based on the percentages of patients with progression larger than the smallest detectable change (SDC).

Results: Changes in joint damage were seen by both the chronological and the paired scoring method. The percentage of patients with progression of joint damage larger than the corresponding SDCs (1.7 and 2.4) varied in the subsets from 18% to 64% if based on the chronological change-scores and from 9% to 36% using paired change-scores. Acceptable sample size estimates were seen in several subsets, depending on (a) how the investigated drug would reduce the individual risk of progression of joint damage (by an absolute or a relative risk reduction model); (b) how damage was scored (chronological or paired); (c) the baseline risk; and (d) whether a two sided or one sided test would be used.

Conclusions: Changes in joint damage due to RA can be detected reliably already within 3 months. This finding can be used to plan short term, randomised controlled trials with radiographic progression as primary outcome.

Prevention of structural damage is an important goal of treatment of rheumatoid arthritis (RA) and recognised by the Food and Drug Administration as a separate claim.¹ Trials have to be of at least 1 year duration in order to label the drug as preventing structural damage. Recent trials have shown that differences in radiological progression between a group receiving an experimental drug and a control group can be already detected after 6 months.^{2–4} Because phase II (dose finding) trials are shorter, often 3 months or less, structural damage has never been included as an outcome measure in these trials. If it were possible to detect progression of joint damage within 3 months of a phase II trial this would have clear advantages. The possible preventive effect of a drug on structural damage could then already be indicated at a very early stage. Furthermore, the optimal dose and exposure range for slowing progression of radiological joint damage could then be defined. Magnetic resonance imaging and ultrasonography are sensitive assessments and are assumed to detect changes after 3 months. Plain radiographs are assumed to be too insensitive to detect changes in structural damage within 3 months; however, this has never been investigated.

The primary aim of this study was to investigate whether the progression in 3 months can reliably be detected by measuring joint damage on plain radiographs.

PATIENTS AND METHODS

Radiographs of a phase II, multicentre, double blind, randomised, placebo controlled trial were evaluated for this study. The trial investigated the efficacy and tolerability of a new compound in patients with RA during 3 months. In this trial, the superiority of the drug over placebo could not be demonstrated, either by measuring disease activity

parameters or from radiographic joint damage. Therefore we considered all patients enrolled in this study as untreated controls.

Patients

Patients fulfilled the American College of Rheumatology criteria for RA. The patients enrolled were recruited from the RA population of both general and academic rheumatology centres. All patients had to have active polyarthritis with a modified disease activity score (DAS28)⁵ of 4.5 or more at screening. Treatment of these patients with three or more disease modifying antirheumatic drugs (DMARDs) should not have failed. Patients with a history of RA longer than 15 years or treated with a biological agent during the past year were excluded from the trial. Concomitant treatment with stable doses of non-steroidal anti-inflammatory drugs or oral corticosteroids (maximal 7.5 mg prednisone or equivalent) was allowed during the trial. Intra-articular injections with corticosteroids were not allowed.

Radiographic scoring method

Posteroanterior films of the hands and anteroposterior films of the feet were made at baseline and at the end visit. The films were scored according to the Sharp/van der Heijde method⁶ by an experienced observer (KB) who was unaware of the patient identity. The principal score used in analyses is the total score (max 448), which is the sum of the erosion score and the joint space narrowing score.

Abbreviations: ARR, absolute risk reduction; DAS28, modified disease activity score; DMARD, disease modifying antirheumatic drug; IQR, interquartile range; RA, rheumatoid arthritis; RRR, relative risk reduction; SDC, smallest detectable change

Radiographs have been scored in trials with and without knowledge of the chronological sequence of the films. So, it was decided to score the films according to both methods, with a reading interval of 2 months. For the chronological method, negative change-scores were allowed if the observer was convinced of the disappearance of the erosion(s).

Statistical analyses

Patients with films of hands and feet at baseline as well as at the end visit, and with an interval of 3 months $\pm$ 2 weeks were included in the analyses. Analyses were performed separately for patients with early RA (RA duration <2 years since diagnosis) and patients with late RA (RA duration $\geq$ 2 years since diagnosis). Baseline characteristics with a Gaussian distribution were expressed as mean (SD); differences between the patients with early and late RA were tested with independent *t* test. Non-Gaussian distributed baseline characteristics were expressed by medians and interquartile range (IQR) and tested with Mann-Whitney test. Baseline characteristics with discrete distributions were expressed as counts (%) and analysed by continuity corrected χ^2 test or Fisher's exact test when appropriate.

The 3 months Sharp/van der Heijde change-scores were presented as the mean (SD), median, full range, and IQR of change-scores. The percentage of patients with change larger than the smallest detectable change (SDC) was also determined. The SDC is a statistical concept representing the smallest difference between two successive scores of the same patient that can be interpreted as a "real" change beyond measurement error and was recommended as the cut off value in a consensus meeting on how to report radiographic data.⁷ The intraobserver measurement error was used here to calculate the SDC, based on a 95% level of agreement. For this purpose, the observer read the films of 20 randomly selected patients again, 1 month after each reading session. The formula used to calculate this SDC is given in appendix 1.

Differences in scores, or percentages of patients with progression larger than the SDC, between the chronological and paired method were analysed with paired tests (paired *t* test and the Wilcoxon test for the scores; the McNemar test for the percentages).

To investigate whether the changes in joint damage seen in this study would be large enough to be useful as an outcome measure in future clinical trials of 3 months' duration, sample sizes were estimated for imaginary trials. Sample sizes were estimated for several subsets of patients: those with early and late RA, as well as those with high(er) baseline risk for radiological progression of joint damage. The latter study was done because randomised clinical trials often

include patients with a high risk on the outcome of interest because they are selected on baseline predictors for that outcome in order to achieve a high contrast between treatment groups. Baseline damage was chosen as baseline predictor for progression of damage. To check if this prognostic factor, which has been reported,⁸⁻¹¹ also operated as prognostic factor in this study, we applied a logistic regression analysis, with correction for age, sex, rheumatoid factor status, and C reactive protein level at baseline. This confirmed that baseline damage operated as an independent prognostic factor in our study too.

The patients were split into three baseline risk groups by tertiles. The sample size estimates were based on the outcome variable: "percentage of patients with progression larger than the SDC" (that is, on the progression of joint damage at the individual level). Sample size calculations based on mean group values were not performed owing to skewness of radiological progression scores in RA. The sample sizes required were calculated for three types of drug mechanism: drugs (mainly) working according to a relative risk reduction (RRR) model, drugs (mainly) working according to an absolute risk reduction (ARR) model, or drugs working according to a mix of both. We shall describe these models briefly; for a more in-depth discussion on the concepts of these models we refer to our short paper in this journal.¹²

According to the RRR model, the RRR (that is, the reduction of the event rate in the treatment group in proportion to that in the placebo group) remains constant over the different baseline risk groups. As a consequence, the ARR (absolute reduction of negative event rate caused by experimental drug—that is, the absolute difference in the event rates between placebo and the treatment group) varies with baseline risk of the patients. If a drug acts mainly according to the RRR model, selecting patients with a high baseline risk for the progression of joint damage will mean that smaller sample sizes are needed. In the ARR model, on the other hand, the ARR stays constant irrespective of baseline risk and the RRR varies over the different baseline risk groups. Selecting patients by baseline risk may have diverse effects on the sample size required, and if a drug works according to the ARR model it has been shown that it is wise to avoid patients with a baseline risk around 50%. Note that risk reduction in the context of this article means the reduction of the number of patients with joint damage progression above the SDC. For the models two hypothetical treatment effects were evaluated. For the RRR model: a constant RRR of 50% and 75%. For the ARR model: a constant ARR of 15% and 25%. Finally, we also determined the sample sizes required for a situation in which 5% of the

Table 1 Baseline characteristics of the patients

Variable	All patients (n = 188)	Early RA (n = 96)	Late RA (n = 92)	Difference late RA – early RA (p value)
RA duration since onset of symptoms (years)*	4.5 (1.1–9.1)	1.2 (0.5–2.7)	8.5 (5.6–11.9)	<0.0001
RA duration since diagnosis (years)*	1.8 (0.2–6.7)	0.2 (0.05–0.7)	6.9 (4.2–9.6)	<0.0001
Age (years)†	56.3 (12.5)	57.3 (12.8)	55.2 (12.0)	0.24
DAS28‡	5.7 (1.1)	5.8 (1.0)	5.7 (1.2)	0.51
Serum C reactive protein (mg/l)*	15 (8–35)	11 (7–33)	18 (10–35)	0.04
Health Assessment Questionnaire score (0–3)†	1.1 (0.7)	1.0 (0.7)	1.2 (1.1)	0.04
Positive serum test for rheumatoid arthritis (%)‡	67	59	75	0.03
DMARD naïve (%)‡	38	69	7	<0.0001
Corticosteroids use (%)‡, ¶	27	24	30	0.40
Female (%)‡	68	64	72	0.30
Chronological damage score (0–448)*	7.5 (1–33)	3.0 (0–13)	21.5 (5–56)	<0.0001
Paired damage score (0–448)*	9.0 (2–33)	3.0 (0–12)	23.5 (6–58)	<0.0001

Early RA, disease duration <2 years; late RA, disease duration $\geq$ 2 years.
*Median (interquartile range); difference between patients with early and late RA tested with Mann-Whitney test; †mean (SD); difference between patients with early and late RA tested with independent *t* test; ‡difference between patients with early and late RA tested with χ^2 test; ¶max 7.5 mg prednisone or equivalent.

Table 2 Group changes in joint damage within 3 months; measured with (chronological) and without (paired) knowledge of the chronological order of the films

Patient group	Chronological Sharp/van der Heijde change-scores		Paired Sharp/van der Heijde change-scores	
	Mean (SD)	Median (range; IQR)	Mean (SD)	Median (range; IQR)
Early RA (n = 96)	0.8 (1.7)	0.0 (0 to 8; 0–1)	0.5 (1.5)	0.0 (–4 to 6; 0–1)
Late RA (n = 92)	2.0 (3.0)	1.0 (0 to 13; 1–4)	1.0 (2.7)	0.0 (–10 to 12; 0–2)

Early RA, disease duration <2 years; late RA; disease duration ≥2 years.

patients in the experimental treatment group do not respond, irrespective of baseline risk. This results in a mix of both models: increasing ARR and increasing RRR with higher baseline risk. These hypothetical risk reductions are in our opinion all clinically relevant treatment effects.

Descriptive analyses, statistical testing, and the logistic regression model were performed with SPSS, version 10.0 for Windows (Chicago, IL). All sample size calculations were performed with the power calculator from the UCLA department of statistics with the two sample arcsine approximation of the binomial distribution with β set at 0.20 and α at 0.05 (<http://calculators.stat.ucla.edu/powercalc/>, accessed July 2004). Sample size calculations were performed for two sided and for one sided statistical testing.

RESULTS

Two hundred and thirty five patients had films at baseline and at the end of the study. Of these, 188 patients—that is, 80%, had a correct interval of 3 months (range 2.5–3.5) and were included in this study. The mean follow up of the included patients was 91 days (SD 4.0) in the group with early RA and 92 days (SD 4.5) in the group with late RA. Table 1 shows the baseline characteristics of the 188 patients included. The mean baseline damage scores of all patients, including those with a follow up of <2.5 or >3.5 months, were comparable with those of the included patients (data not shown).

Table 2 shows the 3 month changes in radiological joint damage at a group level. Both the chronological scoring method and the paired scoring method picked up progression of radiological joint damage within the 3 month interval. For the chronological method, all change-scores were ≥0, so no clear disappearance of erosions was observed in the 3 months when the observer knew the chronological sequence of the films. The changes scored with known chronology were (statistically) higher than the changes scored without

knowledge of the chronological sequence; both the paired *t* test and the Wilcoxon test resulted in *p* values of 0.03 and <0.0001 in the groups with early and late RA, respectively. Progression of joint damage was highest in the group with late RA. The difference in change-scores between the groups with early and late RA was significant if scored with the chronological method (independent *t* test, *p* = 0.001 and Mann-Whitney, *p* = 0.002) but not if scored without information on the time order (independent *t* test, *p* = 0.15 and Mann-Whitney, *p* = 0.18).

Table 3 shows the percentage of patients with progression of joint damage larger than the SDC at 3 months. The SDC was 1.7 for the chronological scoring method and 2.4 for the paired scoring method. The differences in the number of patients with progression above the SDC between the chronological and the paired method were significant in both the group with early RA and the group with late RA (*p* = 0.04 and *p* < 0.001, McNemar test). The group with late RA contained more patients with progression larger than the SDC. This difference was significant if damage was scored with the chronological method (*p* = 0.008, χ^2 test with continuity correction), but was not when using the paired scoring method (*p* = 0.16, χ^2 test with continuity correction). If patients with a high(er) risk for progression were selected, using the baseline damage scores as prognostic factor, the percentage of patients with progression of joint damage above the SDC increased considerably. The increase in percentage of patients with progression above the SDC with increasing baseline risk was significant for both the chronological and the paired method, in the group with early RA as well as the group with late RA (*p* values < 0.0001–0.001, χ^2 tests for trend).

Table 4 and 5 present the sample sizes required to detect statistically and clinically significant differences in percentages of patients with progression larger than the SDC between an imaginary experimental intervention group and a

Table 3 Changes in joint damage within 3 months, for different risk groups based on the baseline joint damage; measured with (chronological) and without (paired) knowledge of the chronological order of the films

Patient group	Patients with progression ≥SDC*, based on the chronological change-scores % (No/total No)	Patients with progression ≥SDC†, based on the paired change-scores % (No/total No)
<i>Early RA (n = 96)</i>		
Regardless of baseline damage	18 (17/96)	9 (9/96)
Low baseline damage, score ≤3	2 (1/50)	0 (0/52)
Medium baseline damage, score 4–20	22 (7/32)	13 (4/30)
High baseline damage, score ≥21	64 (9/14)	36 (5/14)
<i>Late RA (n = 92)</i>		
Regardless of baseline damage	36 (33/92)	17 (16/92)
Low baseline damage, score ≤3	6 (1/16)	6 (1/16)
Medium baseline damage, score 4–20	25 (7/28)	11 (3/28)
High baseline damage, score ≥21	52 (25/48)	25 (12/48)

Early RA, disease duration <2 years; late RA, disease duration ≥2 years; SDC, smallest detectable change (see also "Patients and methods" section).

*SDC chronological 1.7; †SDC paired 2.4.

Table 4 Estimated sample sizes for two sided and one sided testing, based on the *chronological* placebo event rates, calculated for five hypothetical situations: drugs working according to the RRR model (two), the ARR model (two), and a mixed model (one)

Patient group	Placebo event rate (%)	RRR constant				ARR constant				Mixed model	
		RRR = 50%		RRR = 75%		ARR = 15%		ARR = 25%		Intervention event rate = 5%	
		2 sided	1 sided	2 sided	1 sided	2 sided	1 sided	2 sided	1 sided	2 sided	1 sided
<i>Early RA</i>											
Regardless of baseline damage	18	227	179	84	66	60	48	37*	29*	94	74
Medium + high baseline damage, score ≥ 4	35	97	76	36	28	135	106	24	19	25	20
High baseline damage, score ≥ 21	64	34	27	13	10	168	132	40	32	7	5
<i>Late RA</i>											
Regardless of baseline damage	36	92	73	34	27	138	109	25	20	23	18
Medium + high baseline damage, score ≥ 4	42	73	58	27	21	154	122	31	24	17	14
High baseline damage, score ≥ 21	52	52	41	19	15	169	133	37	29	11	9

Early RA, disease duration < 2 years; late RA: disease duration ≥ 2 years; RRR, relative risk reduction = $(1 - (\text{negative event rate intervention group} - \text{negative event rate placebo group}) \times 100)$; ARR, attributive risk reduction = $(\text{negative event rate placebo group} - \text{negative event rate intervention group})$; event rate, percentage of patients with progression of joint damage larger than smallest detectable change (SDC) based on the chronological Sharp/van der Heijde change-score.

*ARR 25% not feasible owing to a placebo event rate $< 25\%$, estimated sample size based on the maximal ARR (18%).

placebo group, based on the *chronological* (table 4) and *paired* (table 5) placebo event rates found in this study. Note that the sample sizes are calculated with the unrounded percentages, so sample sizes for the group with early RA with a paired baseline damage score of 4 or higher (20.5%, 9/44) differ from those from the group with late RA with a paired baseline damage score of 4 or higher (19.7%, 15/76). One can see that, for example, if scoring with knowledge of the chronological sequence of the films, 227 patients in each group would be required to demonstrate an RRR of 50% in the group with early RA. An RRR of 50% in this situation means a reduction of the negative event rate in the intervention group to 9% ($0.50 \times 18\%$).

DISCUSSION

This study showed that even within 3 months changes in joint damage in patients with active RA could be detected on plain radiographs. Although most patients showed no progression in these 3 months—the data were highly skewed—a substantial proportion of the patients showed unequivocal progression of joint damage. For example, 18% of the patients with early RA and 36% of the patients with late RA had change-scores above the SDC if scored with known chronology.

Whether progression of joint damage at the level found in this study would provide a sufficiently large contrast between treatment groups of a phase II placebo controlled clinical trial to detect clinically and statistically significant differences depends on the overall power of the trial. The power of a study depends on (a) the contrast between the groups under study; (b) the sample size of the groups; (c) the risk level accepted for rejecting the null hypothesis that the treatment effects are equal when this null hypothesis is in fact true (type I error, “false positive” result); and (d) whether one sided or two sided confidence intervals or p values are appropriate. The contrast between groups under study depends, besides the actual treatment effect, also on (a) the sensitivity of the measurement instrument (for example, radiological scoring method) used to detect the changes and (b) the mechanism of the risk reduction.

So, to show a clinically significant difference between treatment groups, the treatment effect of the drug under study achieved in a 3 month interval first needs to be large enough. Studies on two of the recently approved antirheumatic drugs, leflunomide⁴ and infliximab,¹³ showed for both an RRR of around 80% and an ARR of 14% for leflunomide and 25% for infliximab. Leflunomide reduced the percentage of patients with an erosion change-score larger than 3 units in 6 months

Table 5 Estimated sample sizes for two sided and one sided testing, based on the *paired* placebo event rates, calculated for five hypothetical situations: drugs working according to the RRR model (two), the ARR model (two), and a mixed model (one)

Patient group	Placebo event rate (%)	RRR constant				ARR constant				Mixed model	
		RRR = 50%		RRR = 75%		ARR = 15%		ARR = 25%		Intervention event rate = 5%	
		2 sided	1 sided	2 sided	1 sided	2 sided	1 sided	2 sided	1 sided	2 sided	1 sided
<i>Early RA</i>											
Regardless of baseline damage	9	463	365	172	135	76*	60*	76*	60*	151	119
Medium + high baseline damage, score ≥ 4	20	191	151	71	56	75	59	31†	24†	42	33
High baseline damage, score ≥ 21	36	93	73	34	27	138	109	25	20	18	14
<i>Late RA</i>											
Regardless of baseline damage	17	232	183	86	68	59	46	38†	30†	54	43
Medium + high baseline damage, score ≥ 4	20	200	157	74	58	71	56	32†	26†	45	35
High baseline damage, score ≥ 21	25	150	118	55	44	97	77	24	19	31	25

Early RA, disease duration < 2 years; late RA, disease duration ≥ 2 years; RRR, relative risk reduction = $(1 - (\text{negative event rate intervention group} - \text{negative event rate placebo group}) \times 100)$; ARR, attributive risk reduction = $(\text{negative event rate placebo group} - \text{negative event rate intervention group})$; event rate, percentage of patients with progression of joint damage larger than smallest detectable change (SDC) based on the paired Sharp/van der Heijde change-score.

*ARR 15% or 25% not feasible owing to a placebo event rate smaller than 15%, estimated sample size based on the maximal ARR (9%); †ARR 25% not feasible owing to a placebo event rate smaller than 25%, estimated sample size based on the maximal ARR (20% and 17%).

from 17% in the placebo group to 3% in the leflunomide group. Infliximab showed a reduction of patients with progression larger than the accompanying SDC in 1 year from 31% in the methotrexate group to 6% in the groups in which infliximab was added. The hypothetical RRR and ARR used in tables 4 and 5 to estimate the sample sizes is therefore not only clinically relevant but also represents realistic treatment effects.

Secondly, the number of patients included also determines the statistical power of a trial. The number of patients in each group that is acceptable for a phase II trial is usually around 60. Tables 4 and 5 show that for several patient scenarios 60 patients or fewer in each group is sufficient. Whether 60 patients is sufficient depends—as mentioned previously—on (a) how the drug under investigation will reduce the individual risk of the progression of joint damage (by an absolute or a relative risk reduction model; (b) which instrument is used to score the damage (with or without knowledge of the sequence of the films); and (c) whether a two sided or one sided test will be used. If, for example, a drugs works according to the RRR model, the damage is scored chronologically, and two sided tests are used: an RRR of 75% can be determined as statistically significant in patients with early RA with intermediate and high baseline damage and all patients with late RA irrespective of baseline damage. An RRR of 50% will show statistically significant differences in patients with high baseline damage but not in the other patient groups. If joint damage is scored paired instead of chronologically, only an RRR of 75% in a high baseline damage group will detect a statistically significant difference. Research on the risk reduction model of treatments outside the field of rheumatology has shown constant RRR with varying ARR for the vast majority of treatments.

How DMARDs reduce the individual risk of progression of joint damage has, to our knowledge, not been investigated before. In a separate short paper,¹² we used the data of two recent trials to investigate the risk reduction on progression of joint damage due to RA in these trials. Future research will have to show whether other DMARDs show equal patterns of risk reduction. More knowledge on how (groups of) DMARDs reduce the risk of progression of joint damage will make it possible to further optimise the designs of studies. For phase II trials, knowledge of the potential mechanism of action of the drug gained from the multiple preclinical model systems can direct the choice of type of risk reduction model used for the sample size calculations. The differences in sample size between the estimates based on the placebo event rates if scored in chronological sequence or scored in pairs without information on the sequence are found because the chronological method picks up more change in joint damage. Reported data show that knowledge of the chronological sequence of films leads to a higher proportions of patients with progression of joint damage than random reading.^{14–16} It is argued that knowledge of the chronological order provides the reader with a maximum of information, thereby reducing the measurement error caused by variation in positioning of the hands and feet or variation in film quality.

Results of a recent study¹⁷ even suggested that knowledge of the chronological sequence leads to an increase in detection of clinically relevant changes without serious overestimation of non-relevant changes. Consequently, if a drug worked according to an RRR model, the estimated sample sizes were lower if scored by the chronological method than by the paired method. However, for a drug working according to the absolute model the opposite was found: the estimated sample sizes were lower for the damage scores by the paired method. This is explained by the fact that the placebo event rates based on the chronological method approximate the 50% progression rate more closely than

those based on the paired method. One exception can be seen in table 5: in the hypothetical situations that a drug could achieve an ARR of 25%, the sample size estimates for the patients with early RA, irrespective of baseline damage, would be lower if based on the chronological placebo event rates than on the placebo events rates if scored with the paired method. The explanation for this finding is that both scoring methods show placebo event rates much smaller than 25%. Only an ARR of 9% for the paired method and 18% for the chronological method is possible in such a patient group and a reduction of 18% to 0% shows statistically significant results with fewer patients than the reduction of 9% to 0%. If a drug works according to a more mixed model, the chronological scoring method again would be more favourable in all settings.

Although the ethical debate on rheumatological phase II trials often focuses on the use of placebo controls, estimations of the required sample size are usually not evaluated for their ethical implications. Tables 4 and 5 show that in many settings it matters considerably whether a two sided or one sided test is chosen as the basis of sample size estimations. One sided testing is mostly considered unacceptable because it does not account for the possibility that the reference group might be better. Knottnerus and Bouter recently revitalised the discussion by stating that one sided testing and corresponding sample size estimations can be proposed as the preferred approach if (a) the scientific hypothesis to be tested is obviously one sided or (b) only a clear advantage in effect of the principal over the reference interventions should have consequences for practice.¹⁸ They thus argued that in placebo controlled clinical trials one sided testing would be adequate and even the default option. We agree with this view, especially for radiographic progression, and are therefore in favour of the one sided sample sizes. To give a complete overview, we have presented the estimated sample sizes based both on two sided and on one sided testing.

In this study the patients with RA with disease duration of 2 years or more, who all had active disease at inclusion (mean DAS28 5.7), showed overall more progression of joint damage than the group with early RA. A possible explanation for this may be that a large proportion of the patients with late RA already have disease which is aggressive and treatment resistant (active polyarthritis and substantial radiological joint damage despite DMARD use). The patients with early RA who have already proven to have a worse prognosis with joint damage at baseline of 21 or more Sharp units, on the other hand, deteriorated more than the patients with late RA with similar baseline damage. Further, the differences in baseline characteristics found between the group with early RA and the group with late RA were understandable from the pathophysiological mechanism of joint damage by RA.

It is known that radiological progression scores show a highly skewed distribution pattern. In studies over a period of 6–12 months, most patients showed no progression or only mild progression and only subsets of patients showed substantial progression.^{19–20} Such skewed distributions require mathematical transformation for appropriate parametric statistical testing, non-parametric statistics, or the data should be analysed in a dichotomised fashion as was done in this study. Analyses on an individual dichotomised level of a continuous outcome measure are considered less sensitive to detect differences between treatment groups. The question is whether this is also true for data that are skewed, such as radiological progression over 3 months. The sample size estimations in this study showed acceptable numbers of patients in whom clinically important dichotomous treatment effects could be detected.

In summary, it is already possible to detect changes in joint damage due to RA within 3 months with plain radiographs. This study further showed that whether this change in joint damage will be large enough to statistically underline clinically relevant treatment differences in a placebo controlled clinical trial depends on (a) how the drug under investigation will reduce the individual risk on the progression of joint damage (by an absolute or a relative risk reduction model); (b) how the damage will be scored (with or without knowledge of the sequence of the films); (c) the baseline risk of the patients investigated; and (d) whether a two sided or one sided test will be used. We conclude that it is feasible to get an impression of whether or not an investigational drug can retard radiographic progression in placebo controlled trials with only 3 months' follow up.

Authors' affiliations

K Bruynesteyn, R Landewé, S van der Linden, D van der Heijde, Department of Internal Medicine, Division of Rheumatology, University of Maastricht, Maastricht, The Netherlands

APPENDIX

The formula of the 95% SDC based on an analyses of variance with two factors: the patient's change-score (p, 188 levels) and the observer (o, 2 levels).

$$\pm 1.96 \times \sqrt{\sigma^2(o) + \sigma^2(p \times o, e)}$$

REFERENCES

- 1 **Guidance for Industry.** *Clinical development for drugs, devices and biological products for the treatment of rheumatoid arthritis (RA).* Rockville: Center for Drug Evaluation and Research, Food and Drug Administration, 1997:1–56.
- 2 **Boers M, Verhoeven AC, Markusse HM, van de Laar M, Westhovens R, van Denderen JC, et al.** Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. *Lancet* 1997;**350**:309–18.
- 3 **Bathon JM, Martin RW, Fleischmann RM, Tesser JR, Schiff MH, Keystone EC, et al.** A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis. *N Engl J Med* 2000;**343**:1586–93.
- 4 **Sharp JT, Strand V, Leung H, Hurley F, Loew-Friedrich I.** Treatment with leflunomide slows radiographic progression of rheumatoid arthritis: results from three randomized controlled trials of leflunomide in patients with active rheumatoid arthritis. Leflunomide Rheumatoid Arthritis Investigators Group. *Arthritis Rheum* 2000;**43**:495–505.
- 5 **Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL.** Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis Rheum* 1995;**38**:44–8.
- 6 **van der Heijde D.** How to read radiographs according to the Sharp/van der Heijde method. *J Rheumatol* 2000;**27**:261–3.
- 7 **van der Heijde D, Simon L, Smolen J, Strand V, Sharp J, Boers M, et al.** How to report radiographic data in randomized clinical trials in rheumatoid arthritis: guidelines from a roundtable discussion. *Arthritis Rheum* 2002;**47**:215–18.
- 8 **Uhlig T, Smedstad LM, Vaglum P, Moum T, Gerard N, Kvien TK.** The course of rheumatoid arthritis and predictors of psychological, physical and radiographic outcome after 5 years of follow-up. *Rheumatology (Oxford)* 2000;**39**:732–41.
- 9 **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.
- 10 **Combe B, Dougados M, Goupille P, Cantagrel A, Eliaou JF, Sibilia J, et al.** Prognostic factors for radiographic damage in early rheumatoid arthritis: a multiparameter prospective study. *Arthritis Rheum* 2001;**44**:1736–43.
- 11 **Drossaers-Bakker KW, Zwinderman AH, Vlieland TP, Van Zeben D, Vos K, Breedveld FC, et al.** Long-term outcome in rheumatoid arthritis: a simple algorithm of baseline parameters can predict radiographic damage, disability, and disease course at 12-year followup. *Arthritis Rheum* 2002;**47**:383–90.
- 12 **Bruynesteyn K, Wanders A, Landewé R, van der Heijde D.** How the type of risk reduction influences required sample sizes in randomised clinical trials. *Ann Rheum Dis* 2004;**63**:1368–71. doi: 10.1136/ard.2003.014035 [published Online First: 1 July 2004].
- 13 **Lipsky PE, van der Heijde DM, St Clair EW, Furst DE, Breedveld FC, Kalden JR, et al.** Infliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. *N Engl J Med* 2000;**343**:1594–602.
- 14 **Ferrara R, Priolo F, Cammisia M, Bacarini L, Cerase A, Pasero G, et al.** clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from the Grisar Study. *Ann Rheum Dis* 1997;**56**:608–12.
- 15 **Salaffi F, Carotti M.** Interobserver variation in quantitative analysis of hand radiographs in rheumatoid arthritis: comparison of 3 different reading procedures. *J Rheumatol* 1997;**24**:2055–6.
- 16 **van der Heijde D, Boonen A, Boers M, Kostense P, van der Linden S.** Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. *Rheumatology (Oxford)* 1999;**38**:1213–20.
- 17 **Bruynesteyn K, van der Heijde D, Boers M, Soudan A, Peloso P, Paulus H, 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.
- 18 **Knottnerus JA, Bouter LM.** The ethics of sample size: two-sided testing and one-sided thinking. *J Clin Epidemiol* 2001;**54**:109–10.
- 19 **Strand V, Landewé R, van der Heijde D.** Using estimated yearly progression rates to compare radiographic data across recent randomised controlled trials in rheumatoid arthritis. *Ann Rheum Dis* 2002;**61**(suppl II):ii64–6.
- 20 **Landewé R, Boers M, Van Der Heijde D.** How to interpret radiological progression in randomized clinical trials? *Rheumatology (Oxford)* 2003;**42**:2–5.