Radiography as primary outcome in rheumatoid arthritis: acceptable sample sizes for trials with 3 months follow-up
Karin Bruynesteyn, Robert Landewé, Sjef van der Linden, and Désirée van der Heijde

DISCLAIMER
The initial version of ARD Online First articles are papers in manuscript form that have been accepted and published in ARD Online but they have not been copy edited and not yet appeared in a printed issue of the journal. Copy editing may lead to differences between the Online First version and the final version including in the title; there may also be differences in the quality of the graphics. Edited, typeset versions of the articles may be published as they become available before final print publication.

Should you wish to comment on this article please do so via our eLetter facility on ARD Online (http://ard.bmjournals.com/cgi/eletter-submit/ard.2003.014043v1)

DATE OF PUBLICATION
ARD Online First articles are citable and establish publication priority. The publication date of an Online First article appears at the top of this page followed by the article's unique Digital Object Identifier (DOI). These articles are considered published and metadata has been deposited with PubMed/Medline.

HOW TO CITE THIS ARTICLE

*Replace with date shown at the top of this page - remove brackets and asterisk

Online First articles are posted weekly at http://ard.bmjournals.com/onlinefirst.shtml
Radiography as primary outcome in rheumatoid arthritis: acceptable sample sizes for trials with 3 months follow-up.

**Category:** Extended report  
**Authors:** Karin Bruynesteyn, Robert Landewé, Sjef van der Linden, Désirée van der Heijde  
**Affiliation:** Department of Internal Medicine, Division of Rheumatology, University of Maastricht, Maastricht, the Netherlands

**Corresponding author:**  
Désirée van der Heijde  
Department of Internal Medicine, Division of Rheumatology  
University Hospital Maastricht  
PO Box 5800  
6202 AZ Maastricht, the Netherlands  
Tel + 31 43 387 75026 / Fax + 31 43 387 5006  
E-mail: dhe@sint.azm.nl
Abstract.

Objective. To investigate whether plain radiographs are able to reveal changes in joint damage due to rheumatoid arthritis (RA) within a 3-month interval.

Method. One-hundred-and-eighty-eight 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 on 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).

Result. 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 1) on how the investigated drug would reduce the individual risk on progression of joint damage (by an absolute or a relative risk reduction model), 2) on how damage was scored (chronological or paired), 3) on the baseline risk and 4) whether 2-sided or 1-sided tests would be used.

Conclusion. It is possible to reliably detect changes in joint damage due to RA already within 3 months. This finding can be used to plan short-term randomised controlled trials with radiographic progression as primary outcome.

Keywords: rheumatoid arthritis, plain film, 3-month interval, progression of joint damage
**Introduction**

Prevention of structural damage is an important goal of rheumatoid arthritis (RA) therapy and recognized by the Food and Drug Administration (FDA) as a separate claim [1]. Trials have to be at least one year in duration in order to label the drug as preventing structural damage. Recent trials have shown that differences in radiological progression between an experimental drug and a control group can be detected already after 6 months [2-4]. Because phase II (dose-finding) trials are of shorter duration, often 3 months or less, structural damage has never been included as outcome measure in these trials. If it would be 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 in a very early stage. And even more, the optimal dose- and exposure range for slowing progression of radiological joint damage could then be defined. MRI 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 an interval of 3 months; however this has never been investigated. 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.

**Methods**

Radiographs of a phase II multi-centre double blind randomised placebo-controlled trial were evaluated for this study. The trial investigated the efficacy and tolerability of a new compound in RA patients during 3 months. In this trial, superiority of the drug over placebo could not be demonstrated, both regarding disease activity parameters and radiographic joint damage. Therefore we considered all patients enrolled in this study as untreated controls.

**Patients**

Patients fulfilled the ACR criteria for RA. The patients enrolled were recruited from the RA population of both general and academic rheumatology centres. All patients had to have an active poly-arthritis with a Modified Disease Activity Score (DAS28) [5] of 4.5 or more at screening. The subjects should not have failed treatment on more than 3 disease modifying anti-rheumatic drugs (DMARDs). Patients with a history of RA longer than 15 years or treated with a biological therapy during the last year were excluded from trial enrolment. Concomitant medication with stable doses of NSAIDs or oral corticosteroids (maximal 7.5 mg prednisone or equivalent) was allowed during the trial period. Intra-articular injections with corticosteroids were not allowed.

**Radiographic scoring method**

Posterior-anterior films of the hands and anterior-posterior films of the feet were made at baseline and at the end visit. The films were scored according to the Sharp/van der Heijde method [6] by an experienced observer (KB) who was blinded for 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.

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 repair of the erosion(s).

**Statistical analyses**

Patients with hand and foot films at baseline as well as at the end visit, and with an interval of 3 months ± 2 weeks were included in the analyses. Analyses were performed separately for early RA patients (RA duration less than 2 years since diagnosis) and late RA patients (RA duration of 2 years or more since diagnosis). Baseline characteristics with a Gaussian distribution were expressed as mean and SD; differences between the early RA and late RA patients were tested with independent t-tests. Non-Gaussian
distributed baseline characteristics were expressed by medians and interquartile ranges (IQR) and tested with Mann-Whitney tests. Baseline characteristics with discrete distributions were expressed as counts (%) and analyzed by continuity-corrected $\chi^2$ tests or Fisher’s exact test when appropriate. The 3-months Sharp/van der Heijde change-scores were presented by the mean, standard deviation (SD), median, full range and interquartile ranges of change-scores (IQR). 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 2 succeeding scores of the same patient that can be interpreted as a ‘real’ change beyond measurement error and was recommend as cut-of value in a consensus meeting on how to report radiographic data[7]. The intra-observer measurement error was used here to calculate the SDC, based on the 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 to calculate this SDC is given in appendix I. Differences in scores or percentages of patients with progression larger than the SDC between the chronological and paired method were analyzed with paired tests (paired t-test and the Wilcoxon test for the scores; the McNemar tests for the percentages).

To investigate whether the changes in joint damage observed in this study would be large enough to be useful as 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: early and late RA, as well as high(er) baseline risk for radiological progression of joint damage. The latter was done because randomised clinical trials often include patients with a high risk on the outcome of interest by selecting on baseline predictors for that outcome, this 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 known from literature [8-11] also operated as prognostic factor in this study we applied a logistic regression model, with correction for age, gender, RF-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-up in 3 baseline risk groups by tertiles. The sample size estimates were based on the outcome variable: ‘percentage of patients with progression larger than the SDC’ (i.e., on the progression of joint damage at the individual level). Sample size calculations based on mean group values were not performed due to skewness of radiological progression scores in RA. The sample sizes required were calculated for 3 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 shortly; for a more in-depth discussion on the concepts of these models we refer to the accompanying short paper in this journal [12]. According to the RRR model, the RRR (i.e., 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, i.e., 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 result in smaller sample sizes 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 the ARR model it is 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 2 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 patients in the experimental treatment group do not respond, irrespective of baseline risk. This results in 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 effect.
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 2-sample arcsine approximation of the binomial distribution with $\beta$ set at 0.20 and $\alpha$ at 0.05 (http://calculators.stat.ucla.edu/powercalc/). Sample size calculations were performed for two-sided statistical testing as well as for one-sided statistical testing.

Results
Two hundred and thirty-five patients had films at baseline and end of study. Of these, 188 patients, i.e., 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 early RA group and 92 days (SD 4.5) in the late RA group. The baseline characteristics of the 188 patients included are shown in table 1. The mean baseline damage scores of all patients, including those with a follow-up of less than 2.5 or more than 3.5 months, were comparable with those of the included patients. (data not shown).

Table 1: Baseline characteristics of the patients

<table>
<thead>
<tr>
<th>Variable</th>
<th>All patients (n=188)</th>
<th>Early RA (n=96)</th>
<th>Late RA (n=92)</th>
<th>Difference late RA - early RA</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RA duration since onset symptoms (yrs)*</td>
<td>4.5 (1.1-9.1)</td>
<td>1.2 (0.5-2.7)</td>
<td>8.5 (5.6-11.9)</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>RA duration since diagnosis (yrs)*</td>
<td>1.8 (0.2-6.7)</td>
<td>0.2 (0.05-0.7)</td>
<td>6.9 (4.2-9.6)</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>Age (yrs)†</td>
<td>56.3 (12.5)</td>
<td>57.3 (12.8)</td>
<td>55.2 (12.0)</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>DAS28 †</td>
<td>5.7 (1.1)</td>
<td>5.8 (1.0)</td>
<td>5.7 (1.2)</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>Serum C-reactive protein (mg/dl)*</td>
<td>1.5 (0.8-3.5)</td>
<td>1.1 (0.7-3.3)</td>
<td>1.8 (1.0-3.5)</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Health Assessment Questionnaire score (0 – 3) †</td>
<td>1.1 (0.7)</td>
<td>1.0 (0.7)</td>
<td>1.2 (1.1)</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Positive serum test for rheumatoid arthritis (%)‡</td>
<td>67%</td>
<td>59%</td>
<td>75%</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>DMARD naïve (%)‡</td>
<td>38%</td>
<td>69%</td>
<td>7%</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>Corticosteroids use (%)‡, ¶</td>
<td>27%</td>
<td>24%</td>
<td>30%</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Female (%)‡</td>
<td>68%</td>
<td>64%</td>
<td>72%</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Chronological damage score (0-448)*</td>
<td>7.5 (1-33)</td>
<td>3.0 (0-13)</td>
<td>21.5 (5-56)</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>Paired damage score (0-448)*</td>
<td>9.0 (2-33)</td>
<td>3.0 (0-12)</td>
<td>23.5 (6-58)</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
</tbody>
</table>

Early RA: disease duration < 2 years; Late RA: disease duration ≥ 2 years
*mean (Inter Quartile Range); difference between early and late RA patients tested with Mann-Whitney tests.
†mean (SD); difference between early and late RA patients tested with independent t-tests.
‡differnce between early and late RA patients tested with Chi-square tests; ¶ max. 7.5 mg prednisone or equivalent

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

<table>
<thead>
<tr>
<th>Patient group</th>
<th>Chronological Sharp/van der Heijde change-scores</th>
<th>Paired Sharp/van der Heijde change-scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean (sd) median (range)[IQR]</td>
<td>mean (sd) median (range)[IQR]</td>
</tr>
<tr>
<td>Early RA (n = 96)</td>
<td>0.8 (1.7) 0.0 (0 ; 8)[0 ; 1)</td>
<td>0.5 (1.5) 0.0 (- 4; 6) [0; 1]</td>
</tr>
<tr>
<td>Late RA (n = 92)</td>
<td>2.0 (3.0) 1.0 (0 ; 13)[1 ; 4)</td>
<td>1.0 (2.7) 0.0 (-10; 12) [0; 2]</td>
</tr>
</tbody>
</table>

Early RA: disease duration < 2 years; Late RA: disease duration ≥ 2 years
IQR : Inter Quartile Range

Change-scores were ≥ 0, so no clear repair 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 resulted in p-values of 0.03 and <0.0001 in the early and late RA group respectively. Progression of joint damage was the largest in the late RA group. The difference in change-scores between the early and late RA groups was statistically significant for the change-scores 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 chronological 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.

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

<table>
<thead>
<tr>
<th>Patient group</th>
<th>Percentage of patients with progression ≥ SDC*, based on the chronological change-scores</th>
<th>Percentage of patients with progression ≥ SDC†, based on the paired change-scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% (n/N)</td>
<td>% (n/N)</td>
</tr>
<tr>
<td>Early RA (N=96)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regardless of baseline damage</td>
<td>18% (17 / 96)</td>
<td>9% (9 / 96)</td>
</tr>
<tr>
<td>Low baseline damage, score ≤ 3</td>
<td>2% (1 / 50)</td>
<td>0% (0 / 52)</td>
</tr>
<tr>
<td>Medium baseline damage, score 4 – 20</td>
<td>22% (7 / 32)</td>
<td>13% (4 / 30)</td>
</tr>
<tr>
<td>High baseline damage, score ≥21</td>
<td>64% (9 / 14)</td>
<td>36% (5 / 14)</td>
</tr>
<tr>
<td>Late RA (N=92)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regardless of baseline damage</td>
<td>36% (33 / 92)</td>
<td>17% (16 / 92)</td>
</tr>
<tr>
<td>Low baseline damage, score ≤ 3</td>
<td>6% (1 / 16)</td>
<td>6% (1 / 16)</td>
</tr>
<tr>
<td>Medium baseline damage, score 4 – 20</td>
<td>25% (7 /28)</td>
<td>11% (3 / 28)</td>
</tr>
<tr>
<td>High baseline damage, score ≥21</td>
<td>52% (25 / 48)</td>
<td>25% (12 / 48)</td>
</tr>
</tbody>
</table>

Early RA: disease duration < 2 years; Late RA: disease duration ≥ 2 years; SDC: smallest detectable change (see also method section)* SDC chronological 1.7; † SDC paired 2.4

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 statistically significant in both the early RA
TABLE 4
Estimated sample sizes for 2-sided and 1-sided testing, based on the chronological placebo event rates, calculated for 5 hypothetical situations: drugs working according the RRR model (2), the ARR model (2) and a mixed model (1).

<table>
<thead>
<tr>
<th>Patient group</th>
<th>Model</th>
<th>RRR constant</th>
<th>ARR constant</th>
<th>Mixed model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RRR = 50%</td>
<td>RRR = 75%</td>
<td>ARR = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
</tbody>
</table>

Patient group

Early RA
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

Late RA
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 due to a placebo event rate smaller then 25%, estimated sample size based on the maximal ARR (18%)

TABLE 5
Estimated sample sizes for 2-sided and 1-sided testing, based on the paired placebo event rates, calculated for 5 hypothetical situations: drugs working according the RRR model (2), the ARR model (2) and a mixed model (1).

<table>
<thead>
<tr>
<th>Patient group</th>
<th>Model</th>
<th>RRR constant</th>
<th>ARR constant</th>
<th>Mixed model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RRR = 50%</td>
<td>RRR = 75%</td>
<td>ARR = 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-sided</td>
<td>1-sided</td>
<td>2-sided</td>
</tr>
</tbody>
</table>

Patient group

Early RA
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

Late RA
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 chronological Sharp/van der Heijde change-score; † ARR 25% not feasible due to a placebo event rate smaller then 25%, estimated sample size based on the maximal ARR (18%)
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 due to a placebo event rate smaller then 15%, estimated sample size based on the maximal ARR (9%); ‡ARR 25% not feasible due to a placebo event rate smaller then 25%, estimated sample size based on the maximal ARR (20% and 17%).
group and the late RA group (p = 0.04 and p <0.001, McNemar test). The late RA group contained more patients with progression larger than the SDC. This difference was statistically significant if damage was scored with the chronological method (p = 0.008, χ² test with continuity correction), but was not when using the paired scoring method (p = 0.16, χ² 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 statistically significant for both the chronological and the paired method, in the early RA group as well as the late RA group (p-values <0.0001 – 0.001, χ² 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 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 un-rounded percentages, so sample sizes for the early RA group with a paired baseline damage score of 4 or higher (20.5%, 9/44) differ from those from the late RA group 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 per group would be required to demonstrate a RRR of 50% in the early RA group. A RRR of 50% in this situation means a reduction of the negative event rate in the intervention group to 9% (0.50*18%).

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

Whether progression of joint damage at the level as found in this study would provide a sufficient 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 1) on the contrast between the groups under study, 2) on the samples size of the groups, 3) on 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 4) whether 1-sided or 2-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 measurement instrument (e.g. radiological scoring method) used to detect the changes and on b) the mechanism of the risk reduction.

So, in order to statistically underscore 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 2 of the recently approved anti-rheumatic drugs, leflunomide [4] and infliximab [13], showed both a RRR around the 80% and an ARR of 14% by leflunomide and 25% by infliximab. Leflunomide reduced the percentage of patients with an erosion change-score larger than 3 units in 6 months from 17% in the placebo group to 3% in the leflunomide group. Infliximab showed a reduction of patients with progression larger the accompanying SDC in one year from 31% in the MTX group to 6% in the groups in which infliximab was added. The hypothetical RRR and ARR used in table 4 and 5 to estimate the sample sizes therefore not only clinically relevant but represent also realistic treatment effects.
Second, the number of patients included also determines the statistical power of trial. The number of patients per group that is acceptable for a phase II trial is usually around 60 per group. Table 4 and 5 show that for several patient scenarios 60 patients or less per group is sufficient. If 60 patients is sufficient depends - as mentioned - on 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, 2) which instrument is used to score the damage (with or without knowledge of the sequence of the films) and 3) whether 2-sided or 1-sided test will be used. If, for example, a drugs works according to the relative risk reduction model, the damage is scored chronologically, and 2-sided tests are used: a RRR of 75% can be determined statistically significant in early RA patients with intermediate and high baseline damage and all late RA patients irrespective of baseline damage. A 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 a 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 showed constant RRR with varying ARR for the vast majority of the treatments. How disease modifying antirheumatic drugs (DMARD) reduce the individual risk on progression of joint damage has, to our knowledge, not been investigated before. In the accompanying short paper [12], we used the data of 2 recent trials to investigate the risk reduction on progression of joint damage due to RA in these trials. Future research shall have to reveal whether other DMARDs show equal patterns of risk reduction. More knowledge on how (groups of) DMARDs reduce the risk on progression of joint damage will make it possible to further optimize the designs of studies. For phase II trials, the knowledge of the potential mechanism of action of the drug gained out of the multiple pre-clinical model systems can direct the choice for 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 is because the chronological method picked up more change in joint damage. It is known from literature that knowledge of the chronological sequence of films leads to 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 a maximum of information, thereby reducing the measurement error caused by variation of positioning of the hands and feet or variation in film quality. Results of a recent study [17] 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 would work according to a 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 approximated 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 early RA patients, irrespective of baseline damage, are lower if based on the chronological placebo event rates compared to 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 is possible in such a patient group and a reduction of 18% to 0% show statistically significant results with fewer patients than the reduction of 9% to 0%. In the situation that a drug works according a more mixed model, the chronological scoring method again would be more favorable in all settings.

While 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 to its ethical implications. Table 4 and 5 showed that in many settings it matters considerably whether a 2-sided test or a 1-sided test is chosen as the base 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 [18]
recently revitalized the discussion by stating that 1-sided testing and corresponding sample size estimations can be proposed as the preferred approach if 1) the scientific hypothesis to be tested is obviously 1-sided or if 2) only a clear advantage in effect of the principal over the reference interventions should have consequences for practice. They thus argued that in placebo controlled clinical trials 1-sided testing would be adequate and even the default option. We agree with this view, especially for radiographic progression, and are therefore in favor of the 1-sided sample sizes. To give a complete overview, we presented the estimated sample sizes based both on 2-sided and on 1-sided testing.

In this study the RA patients with disease duration of 2 years or more, which were all active at inclusion (mean DAS 5.7), showed overall more progression of joint damage than the early RA group. A possible explanation for this can be that a large part of the late RA patients already have proven to be aggressive and treatment resistant (active polyartritis and substantial radiological joint damage despite DMARD use). The early RA patients that had proven to have a worse prognosis by already showing a joint damage at baseline of 21 or more, on the other hand, deteriorated more compared to the late RA patients with similar baseline damage. Further, the differences in baseline characteristics found between the early RA and the late RA group 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 to 12 months, the majority of patients showed no- or only mild progression and only subsets of patients showed substantial progression [20]. Such skewed distributions require mathematical transformation for appropriate parametric statistical testing, non-parametric statistics or the data should be analyzed in a dichotomized fashion as was done in this study. Analyses on an individual dichotomized level of a continuous outcome measure are considered less sensitive to detect differences between treatment groups. Question is whether this is also true for data that are skewed like radiological progression in a 3-month interval. The sample size estimations in this study anyhow showed acceptable numbers of patients with which clinically important dichotomous treatment effects can be detected.

In summary, it is possible to detect changes in joint damage due to RA already within 3 months with plain radiographs. This study further showed that whether this change in joint damage will be large enough to statistically underscore clinically relevant treatment differences in a placebo controlled clinical trial depends on 1) 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), 2) on how the damage will be scored (with or without knowledge of the sequence of the films), 3) the baseline risk of the patients investigated and 4) whether 2-sided or -sided tests will be used. We conclude that it is feasible to get an impression whether or not an investigational drug is capable of retarding radiographic progression in placebo controlled trials with only 3 months follow-up.

**Appendix:**

Formula of the 95% SDC based on an analyses of variance with 2 factors; the patient’s change-score ($p$, 188 levels) and the observer ($o$, 2 levels).

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

**Source of grant:** Dutch Arthritis Association
References:


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

Karin Bruynesteyn, Robert Landewé, Sjef van der Linden and Désirée van der Heijde

Ann Rheum Dis published online March 22, 2004

Updated information and services can be found at:
http://ard.bmj.com/content/early/2004/03/22/ard.2003.014043.citation

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

Topic Collections
Articles on similar topics can be found in the following collections
- Clinical diagnostic tests (1282)
- Radiology (1113)
- Radiology (diagnostics) (750)
- Connective tissue disease (4253)
- Degenerative joint disease (4641)
- Immunology (including allergy) (5144)
- Musculoskeletal syndromes (4951)
- Rheumatoid arthritis (3258)
- Epidemiology (1404)

Notes

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

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

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