Radiological progression in rheumatoid arthritis: how many patients are required in a treatment trial to test disease modification?

John T Sharp, Frederick Wolfe, Mary Corbett, Heikki Isomaki, Donald M Mitchell, Daniel E Furst, John Sibley, Michael Shipley

Abstract

Objective—To determine whether the number of patients required in a therapeutic trial that uses progression of radiological abnormalities as the outcome measure would be similar for multiple centres.

Methods—The progression of radiological damage to the fingers and wrists of patients with rheumatoid arthritis in five centres, three in North America and two in Europe, was examined. The reproducibility of repeated readings by the same and multiple observers was examined. The number of patients required in a two group trial was calculated for several combinations of power and significance.

Results—Scoring progression of radiological abnormalities in sequential films taken between 0-5 and 2-1 years was found to be highly reproducible. When the scores of a single reader were used the rate of change of radiological scores was similar in all centres. Based on the mean progression rate for all centres it was estimated that 153 patients in each group would be required to assure 90% power for detecting a 50% slowing of radiological progression at a significance of 0.05. Review of the experience in three trials showed a large variability in the radiological progression rates.

Conclusion—The progression of scores for radiological damage in rheumatoid arthritis is relatively uniform in North America and Europe and thus the number of patients required in a trial would be similar. Experience in three trials showed that patient selection is of paramount importance in setting up a successful study.

(Br J Rheumatol 1993; 32: 332–337)

Bony erosions and atrophy of cartilage are the usual results of sustained synovial inflammation, the histological hallmark of rheumatoid arthritis (RA). The detection of erosions and cartilage loss by radiographic examination is readily accomplished and methods have been developed to score these abnormalities. The reproducibility of x-ray abnormality scores has been partly defined and such scores have been found to correlate with important clinical and laboratory manifestations of the disease. In spite of the desirable features favouring the use of radiographic scores, treatment trials do not routinely incorporate their use and, when used, conclusive results are not always forthcoming. Several reasons may contribute to the under use and potential misuse of this outcome measure. One key factor is the paucity of data on the expected progression rate of radiographic abnormalities, essential data for calculating the number of patients required in a treatment trial to test the ability of new drugs to alter the underlying disease.

This has prompted us to review our experience in measuring the progression of radiological abnormalities in the fingers and wrists of patients with RA in two European centres, one Canadian centre, and two centres in the USA. The annual increase in radiographic scores was remarkably similar in four of five centres, which led us to use these data to calculate the numbers of patients required in treatment trials designed to achieve conclusive results under four different combinations of significance and power. Experience with radiological assessment in three actual trials is included to illustrate some of the problems encountered.

Methods

Patients were drawn from five arthritis study centres and three treatment trials. The study centres were the Arthritis Clinic at the Middlesex Hospital, London; the Rheumatism Foundation Hospital, Heinola, Finland; the University of Saskatchewan, Saskatoon, Canada; the Arthritis Center of Wichita, Kansas; and the Joe and Betty Alpert Arthritis Center, Rose Medical Center, Denver. In Heinola, 20 patients were randomly selected from a study population of 103 patients by a person who did not know any of the patients. All the patients from Heinola were seen in the first six months after the onset of inflammatory arthritis and met the American Rheumatism Association (ARA) and New York criteria for RA after three years of follow up. In London 14 consecutive patients were originally selected from a special study clinic at the Middlesex Hospital for patients with early disease, who were having their annual x-ray films made in August 1984. Patients were called for follow up films from a list ordered according to the time of admission to the study. Those with films made in August
1984 were in the middle of the list. Patients in the early disease study were first seen between 1966 and 1977; had an average disease duration of 7-9 months when first seen, and were classified as having possible, probable, definite, or classical RA. During 15 years of follow up 86 patients were eventually classified as having definite or classical RA by the 1958 criteria or RA by the revised criteria.\(^3\) Four of the 14 patients originally selected for this study did not fulfil the old ARA criteria for definite or classical RA or the new ARA criteria for RA, leaving 10 patients who were included in this study.\(^{34-37}\) Fifty eight patients who had two or more x ray films were selected from 100 consecutive admissions to the Wichita Arthritis Center, and 50 patients with two or more films among 103 consecutive patients from Saskatoon were selected by methods described previously.\(^{38}\) One hundred and seventy eight patients with two or more films from the Alpert Arthritis Center, Rocky Mountain Center, were chosen from 422 patients admitted consecutively in three time periods: 100 from 1975 and the rest from 1981 to 1983. Patients from Denver, Saskatoon, and Wichita were classified as having definite or classical RA by the 1958 ARA classification criteria\(^31\) or as having RA by the revised (1987) criteria.\(^33\)

An average of 3·56 (2–12) films was available for each patient spanning an average of 5·56 years (1·5 months to 22 years).

x Ray and clinical data were available from three treatment trials: two double blind trials of gold (one comparing gold with placebo and another comparing two doses of gold),\(^{39-41}\) and a pilot study of prinomide (unpublished data). Patients included in the Detroit gold study had continually active, definite, or classical RA of a duration less than five years and had not previously received gold. They were randomly assigned to treatment or placebo groups.\(^39\) The Baylor gold study patients all had definite or classical RA of six months or longer duration. When the study began all patients in the affiliated clinics fulfilling the diagnostic criteria were invited to join the study and most elected to do so. Patient pairs were chosen by matching for at least four of six variables and treatment was assigned by a coin flip for each pair.\(^40\) The prinomide trial was a 12 month, double blind, comparative study of prinomide versus prinomide plus hydroxychloroquine versus hydroxychloroquine in patients with definite or classical RA\(^41\) who had six or more swollen and tender joints and met two of the following three criteria: nine or more tender joints; one or more hours of morning stiffness; or Westergren sedimentation rate > 28 mm/hour. Patients had to have a global disease severity between 3 and 8 on a 10 point scale at two baseline visits, must have been receiving stable doses of non-steroidal anti-inflammatory drugs (NSAIDS), receiving ≤7·5 mg prednisone each day, and were free of significant haematological, renal, cardiac, pulmonary, metabolic, hepatic, and gastrointestinal abnormalities. Hand radiographs were obtained at baseline and 12 months after the start of the study.

All hand and wrist films from the arthritis study centres were scored by JTS and one or more investigators from each of the centres. In two instances readers used the modified Larsen method for scoring films.\(^4\) The senior author (JTS) and the other readers used a reported method which scores 20 finger and 14 wrist joints for erosions and 20 finger and 16 wrist joints for joint space narrowing; the total score is the sum of 70 individual scores.\(^15\) Films were identified in patient sets. Within sets, the sequence was randomly assigned and blinded from the readers.

Subsets of films were scored a second time by five readers at times that varied from one to 15 months after the original interpretation. Variation in intra-reader and inter-reader raw scores was evaluated by correlation coefficients and paired t tests. In addition, the reproducibility of determining x ray progression was tested using all available data collected by method B. In general, x ray progression rates derived from treatment trials. All scores on sequential films at intervals greater than 0·5 and less than 2·1 years that were duplicate readings by the same or different readers were chosen. The dates and therefore sequence of films were not always blinded in the first reading by JTS, but in all other readings the dates of films were hidden and the sequence was randomly assigned. Previous scores were not available to any reader. Films from a given patient were read together to ensure an accurate comparison. In many instances there were several years between the two readings by JTS. Duplicate scores by the other readers were recorded between one and 12 months apart. Sequential scores were converted to progression rates as a change in scores for each year. Agreement of progression rates derived from duplicate readings by the same and different readers was tested by paired t tests.

Scoring of films in the two gold trials was carried out using a method which incorporates erosion and joint space narrowing scores on 27 joints in each hand to give a total of 108 joint scores.\(^4\) The ‘total x ray score’ determined using 70 individual scores was previously shown to correlate closely with that using 108 scores.\(^15\)

Comparison of progression rates between centres and calculation of the number of patients required in a treatment trial were based on the x ray scores of JTS. To calculate the number of patients required in a trial, the progression rate of x ray abnormalities for each subject with two or more films was determined from the slope of scores as the change in score for each year using Medlog.\(^4\)

Statistical tests were carried out with SPSS/PC v3·1.\(^4\) Specific tests are listed in the footnotes to the tables. Data from the European centres were entered directly into SPSS files except for the x ray scores, which were entered into a MEDLOG file to use that program to determine the slope of scores over time by linear regression. Data from Saskatoon, Wichita, and Denver which have been maintained in MEDLOG databank files\(^4\) for several years were transferred to SPSS files for analysis.
The number of patients in each group in a two group, equal numbers trial was calculated from: 

\[ n = \frac{2 \times (SD)^2}{(\bar{x} - \bar{y})^2} \]

where SD is the standard deviation, \(\bar{x}\) is the difference between the two groups, \(\bar{y}\) is Student's \(t\) for the chosen \(\alpha\), \(P\) is the power desired, \(1 - P\) is the \(\beta\) error, and \(t(1-P)\) is the \(t\) for the chosen \(P\).

**Results**

Table 1 gives the general characteristics of the patients from each centre. Patients from the centres varied in terms of age, duration of disease, and the extent of damage seen by X-ray film at the time of entry, reflecting the populations from which they were chosen.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Centre</th>
<th>Middlesex (n=10)</th>
<th>Denver (n=178)</th>
<th>Harvard (n=20)</th>
<th>Kansas (n=58)</th>
<th>Saskatoon (n=50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at first visit (years)</td>
<td></td>
<td>52.0</td>
<td>54.0</td>
<td>45.5</td>
<td>50.0</td>
<td>42.0</td>
</tr>
<tr>
<td>Sex (% female)</td>
<td></td>
<td>70.0</td>
<td>76.2</td>
<td>67.0</td>
<td>70.7</td>
<td>70.7</td>
</tr>
<tr>
<td>Duration of disease at first visit (years)</td>
<td></td>
<td>11.5</td>
<td>12.7</td>
<td>8.5</td>
<td>13.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Duration of disease at last visit (years)</td>
<td></td>
<td>11.5</td>
<td>12.7</td>
<td>8.5</td>
<td>13.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Seropositive (% ever)†</td>
<td></td>
<td>80.0</td>
<td>78.4</td>
<td>89.5</td>
<td>83.1</td>
<td>85.7</td>
</tr>
<tr>
<td>Nodules (% ever)</td>
<td></td>
<td>60.0</td>
<td>52.3</td>
<td>11.3</td>
<td>45.0</td>
<td>62.5</td>
</tr>
<tr>
<td>Haemoglobin at first visit (g/l)</td>
<td></td>
<td>127.0</td>
<td>133.0</td>
<td>128.0</td>
<td>128.0</td>
<td>129.0</td>
</tr>
<tr>
<td>Antinuclear antibodies (% positive)‡</td>
<td></td>
<td>30.0</td>
<td>20.0</td>
<td>26.3</td>
<td>18.0</td>
<td>8.9</td>
</tr>
<tr>
<td>Early radiographic damage rate at first visit (mm/hour)</td>
<td></td>
<td>35.2</td>
<td>47.5</td>
<td>40.9</td>
<td>41.6</td>
<td>49.0</td>
</tr>
<tr>
<td>X Ray score at first visit</td>
<td></td>
<td>21.6</td>
<td>32.7</td>
<td>0.95</td>
<td>31.5</td>
<td>25.6</td>
</tr>
<tr>
<td>X Ray score at last visit</td>
<td></td>
<td>69.2</td>
<td>44.7</td>
<td>41.3</td>
<td>38.1</td>
<td>66.9</td>
</tr>
</tbody>
</table>

*The 10 patients from Middlesex were drawn from a larger study of 86 patients, the Heimia patients from a study group of 103, the Denver group from a database of 422, Wichita patients from a group of 100 consecutive admissions, and the Saskatoon patients from 103 consecutive admissions (see under Methods for further details).

†Tests were not standardised across centres, therefore mean titre of rheumatoid factor and antinuclear antibodies are not shown.

Rheumatoid factor was considered positive at a titre \(>180\) in Denver and Saskatoon, \(>40\) at Kansas and London, and \(>64\) at Harv. The antinuclear antibodies were considered positive at a titre \(>1/100\) in Denver, \(>1/100\) in Harv, and \(>1/80\) at Kansas, London, and Saskatoon.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Sample*</th>
<th>Parent Heimia population*</th>
<th>p Value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean joint score</td>
<td>8.85</td>
<td>9.78</td>
<td>0.51</td>
</tr>
<tr>
<td>Mean functional class</td>
<td>8.70</td>
<td>10.51</td>
<td>0.31</td>
</tr>
<tr>
<td>Mean erythrocyte sedimentation rate (mm/hour)</td>
<td>20.35</td>
<td>39.82</td>
<td>0.70</td>
</tr>
<tr>
<td>Mean C reactive protein (mg/l)</td>
<td>16.75</td>
<td>20.76</td>
<td>0.274</td>
</tr>
<tr>
<td>Mean Larsen index (total)</td>
<td>53.80</td>
<td>53.58</td>
<td>0.897</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Sample*</th>
<th>Parent Heimia population*</th>
<th>p Value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean disease activity</td>
<td>10.81</td>
<td>12.56</td>
<td>0.026</td>
</tr>
<tr>
<td>Mean duration of morning stiffness (minutes)</td>
<td>85.5</td>
<td>245</td>
<td>0.041</td>
</tr>
<tr>
<td>Mean joint score</td>
<td>10.4</td>
<td>11.2</td>
<td>0.779</td>
</tr>
<tr>
<td>Mean functional class</td>
<td>1.80</td>
<td>1.92</td>
<td>0.672</td>
</tr>
<tr>
<td>Mean erythrocyte sedimentation rate (mm/hour)</td>
<td>35.2</td>
<td>34.0</td>
<td>0.877</td>
</tr>
<tr>
<td>Mean rheumatoid factor</td>
<td>74.8</td>
<td>178</td>
<td>0.541</td>
</tr>
<tr>
<td>Mean haemoglobin (g/l)</td>
<td>127.9</td>
<td>129</td>
<td>0.689</td>
</tr>
<tr>
<td>Mean antinuclear antibody titre</td>
<td>1.0</td>
<td>8.0</td>
<td>0.386</td>
</tr>
</tbody>
</table>

*Sample is the 10 patients selected for this study from the Middlesex population whose X ray films were used in determining the progression rate of radiological abnormalities. The parent population is the entire group from which the sample was drawn, including the sample.

†Values are calculated from the test comparing the sample with the patients remaining in the parent population after removing the sample. p Values are not corrected for the number of comparisons tested.

**Table 3**

Agreement in progression rates calculated from repeated reading of sequential films. Progression rates in this table were determined from the scores on sequential films from the same patient. Sequential films were selected if they had been read by two observers or twice by one observer and if the films were more than 0-5 and less than 2-1 years apart

<table>
<thead>
<tr>
<th>Readers (No of subjects)</th>
<th>Mean (SD)</th>
<th>Mean (SD)</th>
<th>Mean (SD)</th>
<th>Mean (SD)</th>
<th>Student’s paired t test</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>score differences</td>
<td>first reader</td>
<td>second reader</td>
<td>first reader</td>
<td>second reader</td>
<td></td>
</tr>
<tr>
<td>1, 3 (122)</td>
<td>51-533 (54-227)</td>
<td>74-00 (65-471)</td>
<td>22.46 (21.987)</td>
<td>11.6 (0.001)</td>
<td>0.945</td>
<td></td>
</tr>
<tr>
<td>1, 2 (30)</td>
<td>50-330 (71-51)</td>
<td>33.93 (42-46)</td>
<td>6.39 (17.93)</td>
<td>5.48 (0.001)</td>
<td>0.948</td>
<td></td>
</tr>
<tr>
<td>1, 4 (72)</td>
<td>29.29 (44-92)</td>
<td>39.06 (49-78)</td>
<td>9.76 (13-96)</td>
<td>5.93 (0.001)</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td>1, 5 (71)</td>
<td>29-352 (44-972)</td>
<td>32-851 (50-685)</td>
<td>3.47 (8-115)</td>
<td>3.61 (0.001)</td>
<td>0.903</td>
<td></td>
</tr>
<tr>
<td>1, 1 (40)</td>
<td>25-80 (24-74)</td>
<td>26-37 (27-88)</td>
<td>0.57 (3-656)</td>
<td>0.57 (0.071)</td>
<td>0.975</td>
<td></td>
</tr>
<tr>
<td>1, 0 (40)</td>
<td>25-80 (24-74)</td>
<td>31-45 (30-53)</td>
<td>3.50 (11-80)</td>
<td>3.20 (0.005)</td>
<td>0.975</td>
<td></td>
</tr>
<tr>
<td>2, 2 (122)</td>
<td>30-803 (44-514)</td>
<td>26-116 (36-998)</td>
<td>4-60 (13-564)</td>
<td>3.66 (0.001)</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>6, 3 (39)</td>
<td>39-75 (35-604)</td>
<td>42-25 (30-549)</td>
<td>2.46 (11-019)</td>
<td>1.40 (0.171)</td>
<td>0.951</td>
<td></td>
</tr>
<tr>
<td>7, 7 (41)</td>
<td>15-927 (18-301)</td>
<td>15-382 (5-911)</td>
<td>2.09 (8-959)</td>
<td>2.71 (0.010)</td>
<td>0.968</td>
<td></td>
</tr>
<tr>
<td>8, 4 (81)</td>
<td>28-976 (34-091)</td>
<td>31-488 (33-982)</td>
<td>2.51 (11-343)</td>
<td>1.40 (0.164)</td>
<td>0.945</td>
<td></td>
</tr>
<tr>
<td>9, 9* (20)</td>
<td>10-60 (10-267)</td>
<td>4-20 (4-652)</td>
<td>0-638</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1, 10* (20)</td>
<td>34-60 (33-117)</td>
<td>54-80 (40-099)</td>
<td>0-815</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Readers 7, 9, and 10 scored films by different methods from the others.

Second scores of films by the same readers related well to the original scores, as judged by correlation coefficients greater than 0.45-5, but a drift in mean scores shown by paired \(t\) tests was often observed for the raw scores (table 2). Scores between different readers using the same method also showed high correlation but raw scores did not show close agreement.

Scores by the Larsen method for the three year films from Heimia (reader 9) correlated poorly with the readings by JT. There was a better relation between the two methods for the eight year films from Heimia (reader 10). When progression rates for intervals 0-5 and 0-1 years were examined for reproducibility there was satisfactory agreement (table 3).

The selected patients from Heimia closely reflected the population from which they were drawn with respect to joint scores, functional index, C reactive protein, and Larsen index (table 4).

Although the patients from Wichita, Saskatoon, and Denver were chosen from consecutive admission to those centres, not all selected patients had X ray films. When those with films available were compared with the total group from which they were chosen, no difference was found for multiple variables. The addition of more
patients from the Denver group to this study did not alter this comparison.

Progression rates of x ray scores were fairly similar for four of the five centres, varying from 4·792 to 5·207 (table 6). The progression rate for the Middlesex group was considerably higher, but, owing to the small number of patients and the variance, was not significantly different from any of the others. The numbers of patients in each group required to participate in a two group, one year study comparing a single treatment to placebo when calculated for each centre's data at 90% power for finding a treatment effect of 50% at a significance level ≤0·05 varied between 80 and 190. When data from all the centres were combined, 153 patients would be required for a trial to achieve 90% power to detect a 50% reduction in radiographic progression at a significance level of 0·05. To test for a 70% reduction in progression only 78 subjects would be required in each group for the same power and significance. If a greater degree of certainty is desired 259 subjects would be required to achieve 95% power at 0·01 significance for 50% and 135 subjects for 70% reduction of progression (figure).

In contrast with the study centres, the progression rates and standard deviations observed in the trials showed much more marked variation (table 7). Consequently, the number of patients required in a trial to obtain firm results based on these figures also showed marked variation. From 46 to 392 subjects are needed for a 90% probability of showing a 50% reduction in progression at a significance of 0·05.

Discussion

The rate of development of radiological damage expressed as the change in score of abnormalities for each year was similar in five geographically widely separated centres. Four of five progression rates differed by less than 10%. The more rapid progression rate seen in the fifth may represent sampling variation due to small numbers. As small numbers from several centres are regularly used in multicentre treatment trials we considered it appropriate to include the Middlesex sample in the calculations presented here. We conclude that the diagnosis of RA is relatively uniform, the disease runs a similar course, and treatment effects, if any, are similar in these European and North American centres.

Raw x ray scores on films from a large number of patients with a broad spread of severity were closely paralleled by second scores recorded by the same and different readers. Means, paired t tests, and plots of scored data, however, showed that the exact reproduction of scores was not regularly achieved, a finding consistent with previous observations.

To test whether the estimates of numbers of patients required in a trial can be expected to apply when films are read by other observers, progression rates were calculated on sequential films obtained in a six month to two year interval. Under these conditions of reading films and calculating progression of x ray damage that closely simulate trial conditions, acceptable agreement was observed between multiple readers using the same method to score films and on repeated scoring by the same reader. Thus it appears probable that the numbers of patients estimated for the different trial conditions listed in table 6 would be broadly applicable for films read by the Sharp method, assuming that selection of patients sampled the same population of patients represented in the five centres studied.

It is not so clear that the predictions of patients required in a trial would apply in a study using Larsen scores as data were not available to test the reproducibility of multiple readings by that method. The much lower correlation coefficients observed between the Larsen and Sharp scores for the Heinola films than observed in multiple sets of films read
multiple times by the Sharp method suggest using caution in extrapolating these findings to a trial using Larsen scores.

Two problems not addressed in this calculation are the duration of the hypothetical trial and the number of readers who score films. It is suspected that relatively prolonged treatment will be required to establish an effect due to the slow evolution of radiological damage. As the error in scoring films does not increase proportionally with the increase in total score (the coefficient of variation decreases as the total score increases), increasing the observation period reduces the variance due to error relative to the variance in progression.14 In this respect it should be noted that the progression rates used for the estimates of the number of subjects were based on slopes of scores for multiple films which were sometimes over extended times. This might reduce the error in determining the progression rate, which would mean that the numbers of subjects required in a study are underestimated and some caution should be used in interpreting these figures as exact numbers. This effect may not be large, however, as the squared mean of the ratios of standard deviations to means for the δ scores in the duplicate readings of reader 1 (table 3) was only 1·14 times the square of this ratio for the data derived from the slopes (table 6). Understanding the full impact of varying the observation period will require additional study.

It has been suggested that having multiple readers score films independently would significantly reduce the number of subjects required but, in fact, in a study that used a set of films with a broad range of progression rates, the effect of using multiple, trained readers was small.27 45 As the spread of progression scores would probably be much smaller in an actual trial, the variance due to error is likely to be greater. Consequently, the effect of multiple readers is not predictable and will have to be established in a trial.

What effect did treatment have on the radiological progression rate in these centres, and how did that affect the estimated numbers required in a trial? If the patients in the study centres, who were receiving what was believed at that time to be the optimum treatment, had had increased x ray progression rates without that treatment, there would still have been no effect on the number of patients required in a trial unless there was a greater effect on the variance of the progression rates. The data derived from the treatment trials illustrate the importance of patient selection. Patients in the Detroit gold study, which showed slowing of radiological progression, were actively selected for severe disease and the study encompassed a two year period.40 This made drawing firm conclusions from a relatively small number of patients possible.

The patients in the Baylor study receiving two doses of gold were chosen from all the patients in three affiliated clinics who had active disease and had not received a course of gold treatment previously.40 41 As gold had not been used regularly in those clinics before the study, the patients who enrolled were representative of the clinic populations. The slightly lower mean progression rate seen in this group (when adjusted for scoring method) may represent the effectiveness of gold, an effect which has been shown in multiple studies.40 46 48

The strikingly lower progression rate seen in the prinomide pilot study is in contrast with the other groups. There is reason to suspect that this pattern of patient selection prevails for all trials of drugs that are perceived to have no disease modifying potential. Selection of patients for trials is greatly influenced by (a) the patient's opinion of the risks, (b) the doctor's views of the potential effectiveness of the drug, and (c) the doctor's concepts of how best to fulfill his ethical obligations in advising patients on participation.49 51 During the window of opportunity after synovitis has developed and before major damage has occurred, doctors feel obligated to use established and previously tested treatments that are believed to slow the progression of disease and patients are usually unwilling to try unproved treatments. The Detroit gold study, which continued for two years, was conducted before the effectiveness of gold was established. A placebo controlled study continuing for two years was acceptable then, but would not be today.

This leads to the conclusion that new strategies for conducting effective trials must be developed. Are there reliable surrogates for rapid radiological progression which would allow short, pilot trials of promising drugs?52 56 Drugs that show promise then could be tried in long term studies on patients recruited because of prognostic evidence of severe, rapid progression of disease. Truly promising drugs could be used as alternative treatments to already accepted, slow acting disease modifying drugs without encountering the serious ethical problems of treating patients with a placebo for two or more years.49 51 One study design, for example, would compare three groups of patients: (a) those receiving a new drug; (b) those receiving an established drug of proved effectiveness; and (c) those receiving a combination of the two. As the introduction of a new, slow acting drug depends on answering two questions—(a) is the new drug as effective as the old 'standard' treatment and (b) is the new drug no more toxic than the 'standard' treatment—a comparison with a standard, slow acting disease modifying drug directly reaches the central issues and the answer to a third question—would the new drug have additive effects if used in combination—would provide a bonus. Although analysis of studies to test whether a new, putative, effective drug is the equivalent of an established drug poses some difficulties, these are manageable and the dilemmas faced by prospective patients and their doctors would be greatly reduced so that recruitment of patients with severe disease should become practical.

If we can clarify our objectives and agree on a successful strategy for achieving them,
measurement of radiological progression can and should be an important outcome measure when testing new disease modifying drugs. Hopefully, this will hasten the development of potent new drugs to supplant some of the pitifully weak drugs now used, drugs that are all too often ineffective and much too often seriously toxic.

Supported in part by a grant from the National Institutes of Health, NIHAR2193, clinical research funds of Rose Medical Center and the Rheumatism Research Foundation, Helsinki, Finland.