Criteria for Improvement in Rheumatoid Arthritis: Alternatives to the American College of Rheumatology 20

DANIEL A. ALBERT, GRACE HUANG, GEORGE DUBROW, COLLEEN M. BRENSINGER, JESSE A. BERLIN, and H. JAMES WILLIAMS

ABSTRACT. ***Objective.*** Appropriate outcome measures are critical to estimating treatment effects in clinical trials and observational studies. The American College of Rheumatology 20 (ACR 20) is the standard measure to assess treatment effects in rheumatoid arthritis (RA) randomized controlled clinical trials (RCT). Although the ACR 20, which measures at least a 20% improvement in a number of different measures, is a very useful dichotomous measure for identifying novel treatments with potential activity, it may fail to discriminate among active treatments. In an effort to increase stringency, trials have used the ACR at 50% and 70%. We investigated the behavior of these and other scoring systems to quantify treatment outcome in several RCT in RA.

Methods. The Cooperating Clinics for the Systematic Study of Rheumatic Disease (CSSRD) database contained 2 trials with a total of 6 arms. Using raw data for each patient entered, we calculated ACR 20, 50, and 70 for each arm. We also calculated the average number of criteria met for the ACR 20, 50, and 70 and their respective areas under the curve over time.

Results. The ACR 20 failed to discriminate among active therapies; however, the ACR 50 was too stringent, and only one patient in these trials satisfied the ACR 70. The average number of criteria satisfied at the 50% level at the final trial visit discriminated well, as did the area under the curve.

Conclusion. The average number of ACR criteria met at a 20% or 50% level discriminated better than the traditional ACR criteria in these trials. More of the information is preserved by the area under the curve of the number of ACR criteria satisfied at each level because area preserves both the number of criteria and the time dependence. The area under the curve of the number of ACR criteria met is a discriminatory, specific, time-dependent, responsive, and domain-preserving metric to use as the primary outcome measure in trials of agents for the treatment of RA. These conclusions should be tested in additional data sets. (J Rheumatol 2003;31:856–66)

Key Indexing Terms:
RHEUMATOID ARTHRITIS AMERICAN COLLEGE OF RHEUMATOLOGY 20
CRITERIA IMPROVEMENT COOPERATING CLINICS

Outcome measures for rheumatic diseases have been a major focus of clinical research for decades. The original measures for disease activity and severity for rheumatoid arthritis (RA) such as the Lansbury[1] and Richie[2] activity indices and the American Rheumatism Association functional classification have been modified to generate scales that satisfy the demands of clinical epidemiology. These newer activity indices, such as the American College of Rheumatology 20[3] (ACR 20) and the Paulus criteria[4], incorporate different dimensions, each of which has been subject to scrutiny for both validity and reliability. The current iteration most commonly used is the ACR 20, which determines whether a patient has had a clinically significant response (i.e., a dichotomous decision) based on fulfilling a 20% improvement in tender and swollen joints as well as 3 of 5 remaining criteria: physician global assessment, patient global assessment, patient pain severity (by visual analog scale, VAS), functional assessment (by Health Assessment Questionnaire, HAQ, or the Arthritis Impact Measurement Scale, AIMS), and an acute phase reactant (erythrocyte sedimentation rate, ESR, or C-reactive protein, CRP).

While the ACR 20 has undergone a major effort to validate it as a discriminatory and specific outcome measure for RA, questions remain about its utility and appropriateness especially for the newer agents such as etanercept[5,6], leflunomide[7,8], and infliximab[9-11]. A primary concern is that in the search for a highly sensitive test of efficacy to promote further study of potentially useful agents, the ACR 20 may have limited utility in discriminating among efficacious agents. Further, within a dichotomous framework, some of the information in the continuous scales of the

From the Division of Rheumatology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania; and the University of Utah, Salt Lake City, Utah, USA.

D.A. Albert, MD; G. Huang, BA; G. Dubrow, MA; C.M. Brensinger, MS; J.A. Berlin, ScD, University of Pennsylvania; H.J. Williams, MD, University of Utah.

Address reprint requests to Dr. D.A. Albert, Division of Rheumatology, University of Pennsylvania School of Medicine, 5 Maloney, Suite 504, 3400 Spruce Street, Philadelphia, PA 19104-4283.
E-mail: albertd@mail.med.upenn.edu

Submitted August 6, 2002; revision accepted November 25, 2003.

components of the ACR 20 is lost. Finally, there is no assurance that the scale behaves in a linear fashion, so that the current trend of using it as a more stringent test of efficacy by demanding a 50% or a 70% improvement (ACR 50 or ACR 70) may not result in appropriate statistical test characteristics.

With these questions in mind we undertook a reevaluation of the ACR criteria examining various aggregated forms. Our intent was to determine which of the various forms of the ACR criteria provided the best ability to discriminate between efficacious therapies. To examine these quantitative properties, we utilized a data set that had a number of different multicenter clinical trials that used the same scales and measurement characteristics for each trial. While these studies are somewhat dated, our intent was to analyze the design of the outcome measures, not the agents themselves, so any appropriate data set that included trials of efficacious agents would have sufficed for our purposes. Indeed, less efficacious agents are in some sense preferable since they may better help distinguish between outcome measures that discriminate between treatments and those that do not. Specifically, we used the Cooperating Clinics for the Systematic Study of Rheumatic Disease (CSSRD) data from 2 randomized controlled trials[12,13] of 3 agents, with a total of 6 drug-to-placebo or drug-to-drug comparisons. We utilized these data to calculate the ACR 20, 50, and 70 scores; the mean number of ACR criteria fulfilled, which might be called ACR-C; and the area under the curve (AUC) over time of the number of ACR criteria fulfilled, which might be called ACR AUC.

MATERIALS AND METHODS

Data were obtained from 2 trials conducted by the CSSRD. The trials are randomized, double-blinded studies performed on adult patients with RA as the disease is defined by the ACR. The trials include oral gold (auranof) versus injectable gold (myochrys) versus placebo, and high D-penicillamine (high dpen) versus low D-penicillamine (low dpen) versus placebo. Data were used to compare the ability to differentiate active treatment from control among proposed outcome measures for defining improvement in RA. The trials are identified by publication in Table 1. The CSSRD used similar but not identical variables to those the ACR criteria call for. Specifically, the joint counts used (66 for swelling, 68 for tenderness) and the functional score were different than those prescribed for use in the ACR 20. These validated scales were used to calculate a score equivalent to the ACR 20, 50, and 70.

The first outcome measure tested was the standard ACR definition of improvement. This definition uses 7 measures, which are quantified with a baseline score that can be compared with a final score taken at the end of a treatment or placebo trial. The first indicator measures improvement in tender joint count by scoring various degrees of tenderness to pressure and joint manipulation on physical examination; the types of tenderness are collapsed into a single tender versus nontender dichotomy for each joint. The scores for each patient were summed over 60 joints at each of their visits. The second indicator measures improvement in swollen joint count. Analogous to tender joint count, the scores of 60 joints for each patient (30 on left 30 on right) were summed at each of their visits. The third indicator measures the patient's assessment of pain. The fourth indicator measures the patient's global assessment of disease activity. The fifth indicator measures the physician's global assessment of disease activity. Each of the last 3 indicators is measured on a continuous 0 to 5 cm VAS. The sixth indicator measures the patient's assessment of physical function. The CSSRD trials validated their own functional assessment scoring[14]. Scores were collapsed into a summary scale of 0–3. The final indicator is the ESR.

We took the scores of these 7 indicators and examined various ways of using the scores that would maximize the ability to detect differences between placebo and treatment and between active treatments. The first 3 outcome measures tested were defined by the ACR. The ACR definition requires that the patient has improved by at least 20% in both tender and swollen joint counts and also has improved by at least 20% in 3 of 5 other core set indicators (patient global assessment, physician global assessment, self-reported physical disability, acute phase reactant, and patient pain assessment). This is known as the ACR 20. The ACR 50 and the ACR 70 have the same criteria as the ACR 20, but they are set to improvement percentages of 50% and 70%, respectively.

The proportion of patients who satisfied the ACR 20, 50, and 70 was compared between treated and placebo patients using the ordinary chi-square or Fisher's exact test when expected values of at least one cell count were less than 5.

Using the 7 ACR indicators, we also compared the mean number of criteria that exceeded thresholds of 20%, 50%, and 70% for each patient for the treatment and placebo groups. We first made this comparison of means at the final timepoint in each trial. For each patient, the number of indicators met at the 3 thresholds was also plotted across time and their respective areas under the curve were calculated. The average areas were found for the treatment and placebo groups. The average number of indicators met at thresholds of 20, 50, and 70 and their respective mean areas under the curve were compared between the treatment and placebo groups using both the t test and the Wilcoxon rank-sum test.

Additionally, random coefficient models were fitted to the longitudinal data for the number of criteria met over time. We included random intercepts and random slopes and examined the main effect for treatment and the treatment-by-time interaction. The random intercept in these models allows for variability in the baseline values across patients. The random slopes essentially allow each patient to have his or her own slope over time, and these slopes are then averaged across patients. An initial examination of the data suggested that the assumption of a linear trend over time was reasonable. The "baseline" for these models was actually the first followup visit, since the definition of the outcome variable is based on improvement over the true initial value for each patient.

The missing data plan for these analyses was based on the assumption

Table 1. Trials included in the analysis.

Author	Treatments in Trial		Patients Entered, n	Intervention	Duration
Ward, 1983[13]	Auranofin (oral gold)	Auranofin	77	6 mg qd	20 wk
	Gold sodium thiomalate (IM gold)	GST	81	10 mg, 25, 50 q wk	20 wk
	Placebo	Placebo	50		20 wk
Williams, 1983[12]	Low dose penicillamine	Low dose dpen	87	125 mg qd	30 wk
	High dose penicillamine	High dose dpen	86	500 mg qd	30 wk
	Placebo	Placebo	52		30 wk

IM: intramuscular.

that the usual approach to the ACR 20 is an intent-to-treat analysis. Specifically, we considered anyone who was missing data at the close of the trial as a failure on the ACR 20. For the other outcome measures, we chose a plan that is generally consistent with the philosophy that missing data, other than at baseline, should be treated as failures, whether on particular items or when a patient is missing the final visit. When a patient was missing an interim visit, but returned for the final visit, we have a different approach, as described below.

1. If a patient was missing one or more individual items from the 7 domains of the ACR 20, but not the entire visit:

(a) And the missing visit was the baseline visit, then we imputed the missing item with the overall mean at baseline taken across all patients with data for that item.

(b) And the missing visit was a followup visit, then we treated this as a failure for that item at that visit.

2. If a patient was missing data from an entire visit:

(a) But the patient *did* have data for the final visit, we ignored the intermediate visits:

(i) For longitudinal analyses, we ignored the missing timepoints and used the available data, consistent with common practice when running random coefficient models.

(ii) For the AUC, we simply interpolated between the missing visits.

(b) If a patient had no data for the final visit:

(i) For longitudinal analyses, we again ignored the missing timepoints and used the available data.

(ii) For the ACR 20, the number of criteria met, and AUC analyses, we imputed the final visit values to the baseline values for each of the 7 domains. This implies that the patient will fail to meet each of the 7 criteria.

All statistical analyses were performed using SAS version 8.1 (SAS Institute, Cary, NC, USA) or Splus version 6.0 (MathSoft Inc., Seattle, WA, USA).

RESULTS

The number of patients satisfying ACR 20, 50, and 70 criteria was obtained by subtracting the final visit score from the baseline visit score for each of the 2 indicators that constitute the ACR scoring system. Patients with more than a 20% improvement in swollen and tender joints in combination with a greater than 20% improvement in 3 of the 5 remaining criteria were designated ACR 20 responders and similarly for ACR 50 and 70. The results of this analysis of responders under the various thresholds (ACR 20, 50, 70) are shown in Table 2 for all 6 arms of the 2 trials (auranof, myochrys, low dpen, high dpen) with their respective placebo groups[12,13]. There are 6 comparisons, but the auranofin and myochrysine arms use the same placebo group and the low and high dose penicillamine use the same placebo group. Because the number of patients within each arm was not large, ranging from 50 to 87, the criteria must be efficient to discriminate between active therapy and placebo. In these 2 trials of efficacious disease modifying antirheumatic drugs (DMARD), only the difference between myochrysine and placebo was shown to be statistically significant using a 2-tailed Fisher exact test ($p < 0.01$).

The next set of analyses examined the mean number of ACR criteria (per patient) fulfilled at the 20%, 50%, and 70% level over time in each of the 6 arms of the 2 trials. This was done for 3 timepoints (Figure 1). These data permit 2 different summary analyses. First, the mean number of criteria satisfied at each level (20, 50, and 70) is shown for the *last visit* data in tabular form in Table 3. This represents the ACR-C. We present 2 different tests of statistical significance, the t test and the nonparametric Wilcoxon test. Both statistical tests give similar results. Using the average number of criteria met at the 20% improvement threshold proves to be significant in 2 of the 4 treatment versus placebo comparisons (myochrys vs placebo and high dpen vs placebo). The 50% threshold proves to be significant in all 4 of the treatment versus placebo comparisons (myochrys vs placebo, auranof vs placebo, high dpen vs placebo, low dpen vs placebo), but neither drug-to-drug comparison. The 70% threshold proves to be significant in 2 of the 4 comparisons against placebo (myochrys vs placebo, high dpen vs placebo) and one of the 2 drug-to-drug comparisons (myochrys vs auranof).

Second, we utilized the number of criteria fulfilled at each level for each visit to calculate the area under the curve (AUC) for each patient — the ACR AUC. The mean AUC for the number of criteria meeting ACR 20, 50, and 70 improvements are shown across time for 2 trials in Figure 2 (auranof vs myochrys vs placebo, and low dpen vs high dpen vs placebo). To generate the AUC at any given timepoint, we calculated the number of criteria satisfied at 3 points in time for each patient. We then took the AUC using the 3 points in time as boundaries for the areas, yielding a single value for each patient (at each threshold). We used a 2-sided t test and a rank-sum test on these areas, which yields the cumulative AUC for the entire trial duration (Table 4). The tests compared active treatments with their respective placebo groups. The 2 statistical tests agree on which trials were statistically significant and which were not, in all cases except 2 (myochrys vs placebo at the 50% and 70% level untransformed).

The results were analogous to the average number of criteria at the final visit, but not as sensitive, since myochrysine versus placebo was significant at the 50% and the 70% level but not at the 20% level. High dpen and low dpen were significant at all levels with both the AUC and the ACR-C approach. The plot of the AUC (Figure 2) appears to be somewhat better "behaved" than the number of criteria fulfilled because it increases linearly. We also provide a longitudinal analysis of the repeated measurements over time, including random intercepts and random slopes, which examine the treatment-by-time interaction. A statistically significant treatment-by-time interaction indicates that the slopes for the treatment effect over time differ between the active treatments and their respective placebo groups. Results from these models were similar to those examining the mean number of criteria met at the last visit.

The third analysis we performed, for display purposes only, was a contingency table comparing the percentages of patients fulfilling a given number of ACR criteria between the treatment and the control groups (Figure 3).

Table 2. Trial results using traditional ACR 20, 50, and 70. Proportion of subjects fulfilling ACR 20, 50, or 70 criteria in placebo (placebo in) or treatment groups (treatment in) for all 6 comparisons in the 2 trials versus those that did not fulfill criteria (placebo or treatment out).

	Placebo in (%)	Placebo out (%)	Myochrys in (%)	Myochrys out (%)	p
ACR 20	7 (14.00)	43 (86.00)	27 (33.33)	54 (66.67)	0.0142
ACR 50	0 (0)	50 (100)	4 (4.94)	77 (95.06)	0.2972
ACR 70	0 (0)	50 (100)	0 (0)	81 (100)	NA
	Placebo in (%)	Placebo out (%)	Auranof in (%)	Auranof out (%)	p
ACR 20	7 (14.00)	43 (86.00)	23 (29.87)	54 (70.13)	0.0397
ACR 50	0 (0)	50 (100)	3 (3.90)	74 (96.10)	0.2782
ACR 70	0 (0)	50 (100)	0 (0)	77 (100)	NA
	Auranof in (%)	Auranof out (%)	Myochrys in (%)	Myochrys out (%)	p
ACR 20	23 (29.87)	54 (70.13)	27 (33.33)	54 (66.67)	0.6399
ACR 50	3 (3.90)	74 (96.10)	4 (4.94)	77 (95.06)	1.0000
ACR 70	0 (0)	77 (100)	0 (0)	81 (100)	NA
	Placebo in (%)	Placebo out (%)	High dpen in (%)	High dpen out (%)	p
ACR 20	10 (19.23)	42 (80.77)	28 (32.56)	58 (67.44)	0.0894
ACR 50	1 (1.92)	51 (98.08)	11 (12.79)	75 (87.21)	0.0306
ACR 70	0 (0)	52 (100)	1 (1.16)	85 (98.84)	1.0000
	Placebo in (%)	Placebo out (%)	Low dpen in (%)	Low dpen out (%)	p
ACR 20	10 (19.23)	42 (80.77)	26 (29.89)	61 (70.11)	0.1653
ACR 50	1 (1.92)	51 (98.08)	6 (6.90)	81 (93.10)	0.2563
ACR 70	0 (0)	52 (100)	0 (0)	87 (100)	NA
	Low dpen in (%)	Low dpen out (%)	High dpen in (%)	High dpen out (%)	p
ACR 20	26 (29.89)	61 (70.11)	28 (32.56)	58 (67.44)	0.7044
ACR 50	6 (6.90)	81 (93.10)	11 (12.79)	75 (87.21)	0.1929
ACR 70	0 (0)	87 (100)	1 (1.16)	85 (98.84)	0.4971

"In" refers to satisfying the criterion of the threshold indicated. "Out" refers to not satisfying the criterion of the threshold indicated. No. of patients and percentage of placebo or intervention group are shown. NA: not applicable.

DISCUSSION

The management of patients with RA is critically dependent on the use of appropriate outcome measures. These measures need to fulfill various functions. They should be discriminatory and specific for the detection of novel agents with potential for therapeutic utility. In addition, they should be able to discriminate between effective agents to generate at least a rank order of effectiveness. They should be able to discern if an agent affects different dimensions of the disease differently. Finally, the measures should be responsive to modest but clinically important differences[15] and they should be interpretable in terms of clinical benefit. It is quite possible and perhaps even probable that different criteria are useful for these different purposes. With the newer, more effective agents it is essential that we have instruments that can distinguish between them in terms of degree of efficacy.

We emphasize that the measures we propose here are not the only composite scales that could be derived from these dimensions and that future developments, such as a reduced number of items, could be forthcoming. More sophisticated analytical approaches could be applied, as well. For example, one could retain the 7 individual items and apply multivariate methods that account for the correlations induced by having both multiple outcomes and multiple visits for each patient. This type of approach would have the advantage over all the others that it would provide a "penalty" for outcomes that worsen over time.

A concerted effort began in the 1980s to determine what outcome measures should be used for RA clinical trials. This led to a focus on joint count, ESR, and global assessment in preference to grip strength, 50-foot walking time, ring size, and other measures. This was part of a larger movement to generate useful outcome measures for treatment of many

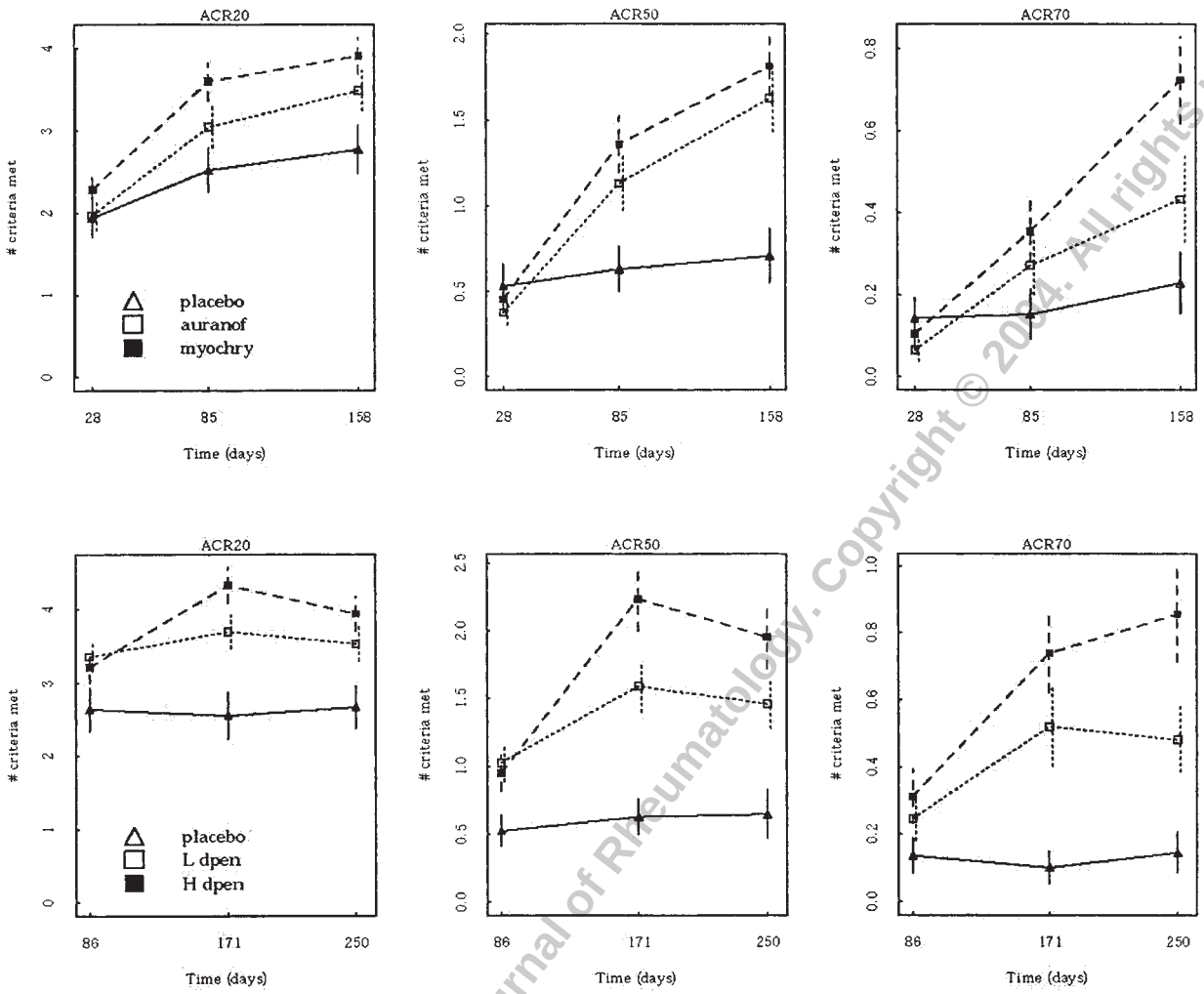

Figure 1. For each of the 2 trials arrayed on the vertical axis, each arm of the trial is plotted with the number of criteria met at the 20%, 50%, and 70% thresholds. Each value is plotted with 95% confidence intervals.

common disorders[16]. The introduction of patient-based outcome measures including functional assessment[17], patient global assessment, and pain scores[18] eventually led to a core set of disease activity measures for RA clinical trials[3] agreed upon by the ACR and OMERACT (Outcome Measures in Rheumatoid Arthritis Clinical Trials conference)[19]. Later refinements such as a reduced set of joints for a joint count[20] and a direct comparison of ACR and OMERACT proposals[21] finally led to the ACR Preliminary Definition of Improvement — the ACR 20[22].

Since that time there have been several challenges to the ACR 20, including competing criteria and different levels of improvement within the ACR dimensions. In head to head comparisons using the minocycline trial (MIRA) data set, the ACR 20 was found to have better discriminatory power than the Paulus criteria[23] but the same as the EULAR (European League Against Rheumatism) criteria[24]. Whether a 50% or a 70% level of improvement was preferable to the 20% was explored by Felson and colleagues, who determined that the 20% improvement was superior because there was a disproportionate loss of responders when the criteria were made more stringent, thus leading to a loss of statistical power using the higher stringency criteria[25]. However, Pincus and Stein questioned how clinically useful a 20% improvement is[26]. They and others[27] suggested using the continuous measure "area under the curve," which generates smaller effect sizes but better precision[27,28]. There appears to be consensus, at least in the area of pain measurement, that percentage improvement is preferable to absolute improvement since it effectively normalizes for different baseline levels of activity[29].

At present, efforts are being made to improve the measurement tools. Efforts are being exerted to develop an entirely patient based scoring system[30] and to eliminate some defects in the functional assessment scales that exhibit "floor" and "ceiling" effects[31]. There are efforts to change

Table 3. Trial results using mean number of criteria fulfilling the 20%, 50%, or 70% thresholds for each of the 6 comparisons with random coefficient models for treatment-by-time (duration of treatment in trial) interaction.

		Last Visit Data				Random Coefficient Models with Random Intercept and Slope	
Comparison	Mean No. of Criteria	SD	t Test (p)	Wilcoxon (p)	Treat (p)	Treat × Time (p)	
20 Placebo	2.66	2.07	0.0018	0.0021	0.8437	0.0016	
Myochrys	3.81	1.99					
50 Placebo	0.68	1.08	< 0.0001	< 0.0001	0.0689	< 0.0001	
Myochrys	1.77	1.52					
70 Placebo	0.22	0.51	0.0009	0.0007	0.0042	< 0.0001	
Myochrys	0.70	0.93					
20 Placebo	2.66	2.07	0.0771	0.0804	0.7769	0.0444	
Auranof	3.35	2.18					
50 Placebo	0.68	1.08	0.0011	0.0012	0.0298	< 0.0001	
Auranof	1.56	1.65					
70 Placebo	0.22	0.51	0.1564	0.3799	0.0815	0.0648	
Auranof	0.42	0.88					
20 Myochrys	3.81	1.99	0.1631	0.2011	0.5527	0.2560	
Auranof	3.35	2.18					
50 Myochrys	1.77	1.52	0.4130	0.2182	0.8790	0.3611	
Auranof	1.56	1.65					
70 Myochrys	0.70	0.93	0.0471	0.0051	0.2952	0.0502	
Auranof	0.42	0.88					
20 Placebo	2.46	2.00	0.0005	0.0006	0.7420	0.0054	
High dpen	3.80	2.21					
50 Placebo	0.60	1.19	0.0001	< 0.0001	0.5864	0.0001	
High dpen	1.88	2.14					
70 Placebo	0.13	0.40	0.0002	0.0002	0.2149	0.0003	
High dpen	0.83	1.28					
20 Placebo	2.46	2.00	0.0165	0.0187	0.5917	0.0711	
Low dpen	3.37	2.20					
50 Placebo	0.60	1.19	0.0028	0.0007	0.9696	0.0123	
Low dpen	1.39	1.64					
70 Placebo	0.13	0.40	0.0125	0.0146	0.6885	0.0144	
Low dpen	0.46	0.87					
20 Low dpen	3.37	2.20	0.1969	0.2156	0.7455	0.1721	
High dpen	3.80	2.21					
50 Low dpen	1.39	1.64	0.0904	0.2445	0.4884	0.0489	
High dpen	1.88	2.14					
70 Low dpen	0.46	0.87	0.0290	0.0758	0.5023	0.0509	
High dpen	0.83	1.28					

the way scores are calculated[32,33] and to correlate changes in activity with "objective" longterm outcomes such as radiologic damage[34] and mortality[35,36]. Attempts to extend this type of analysis to observational studies have begun[37,38]. Most recently, more sophisticated statistical modeling techniques have been applied to these issues[39-41]. Lastly, this field has been critically reviewed within the past year[42].

The data we utilized to evaluate the relative discriminatory power of the ACR criteria are appropriate for illustrative purposes because they represent 2 separate trials of 6 total arms acquired with the same methodology from multiple centers. The DMARD in these studies are somewhat less effective than currently used agents, thus the differences between treated and placebo are somewhat more subtle than might be the case for etanercept, infliximab, and leflunomide. This results in a more stringent test of the outcome measures than might be the case with more effective agents. Our study was neither an attempt to evaluate the efficacy of the drugs being administered in the trials nor an attempt to compare versions of the ACR 20 to every scoring system (such as EULAR or Disease Activity Scales). We asked, using the same measures as the ACR 20, whether there is a more effective way of using the same information to maximize the ability to discriminate between therapies

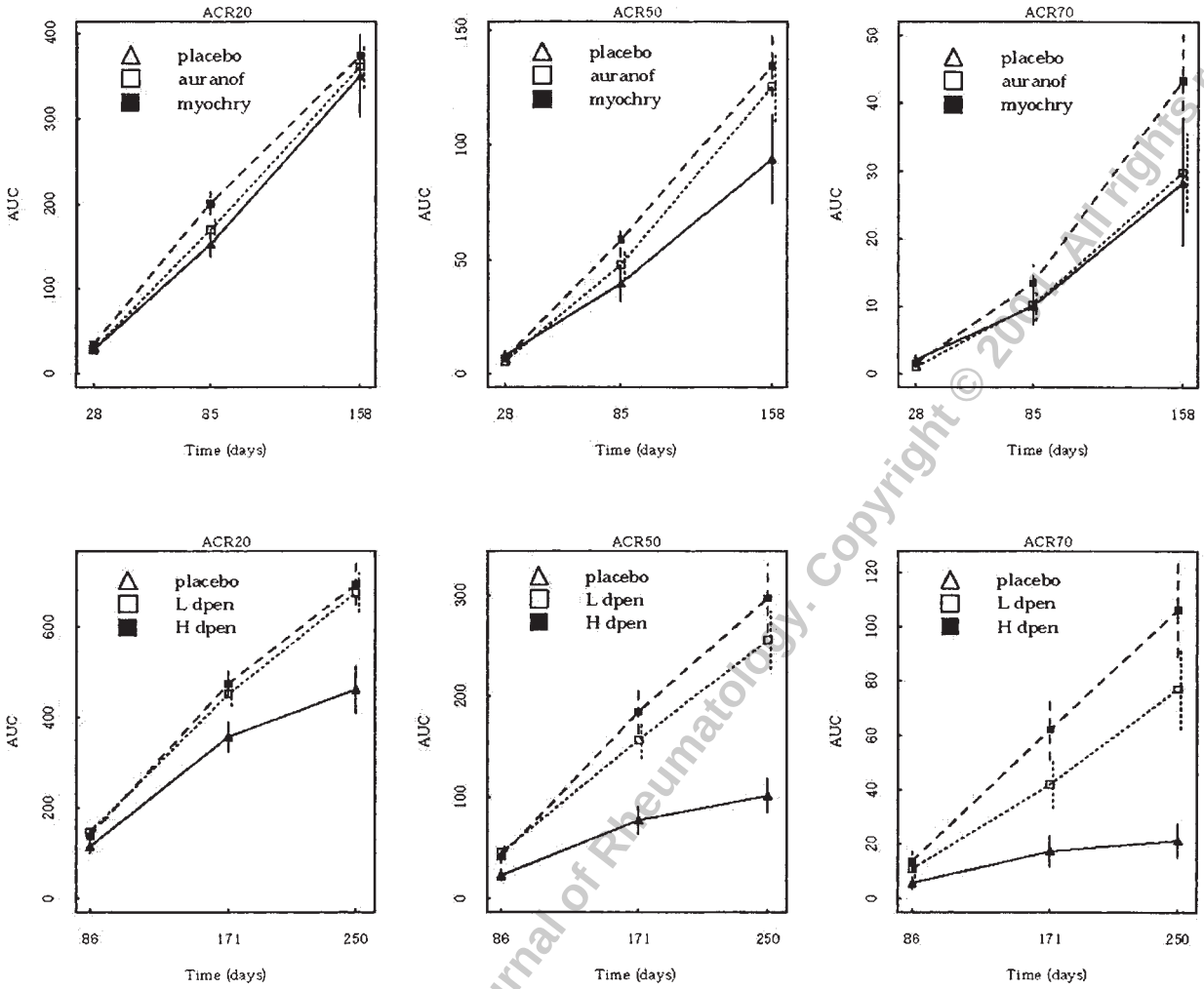

Figure 2. For each of the 2 trials arrayed on the vertical axis, each arm of the trial is plotted with the area under the curve of the criteria met at the 20%, 50%, and 70% thresholds. Each value is plotted with 95% confidence intervals.

and to determine a measure of degree of improvement. Thus, the exact extent of the dataset is secondary to our ability to use that dataset to evaluate the various scoring systems. Nevertheless, although these data come from multiple trials, each of which involved multiple centers, there may be a lack of generalizability to other sites or to newer therapies.

Our initial analysis replicated that of Felson and colleagues, who compared the ACR 20 with 2 other more stringent versions of the ACR criteria, at 50% and 70% improvement level. We agree with their conclusion that at least in our data set ACR 20 is a more efficient discriminator than ACR 50 or 70. We also agree that the reason for this is a disproportionate loss of responders at the higher stringency levels. However, the traditional ACR 20 (or 50 or 70) does not discriminate as well as any of the continuous measures we examined, which included the mean number of criteria fulfilled at a 20% or 50% level and the mean AUC at 20% or 50%. To be fair, one could develop a repeated-measures approach to analysis of the ACR 20, using data from all timepoints. This is why we also compared the mean number of criteria met at the final visit. (Recall, too, that for both the mean number of criteria at the final visit and the ACR 20, patients missing data at the final visit were treated as failures on all criteria.)

Within the mean number of criteria fulfilled, there is little to recommend 20% versus 50%, but overall, 50% appeared to function slightly better. It is, however, our recommendation that all 3 (20%, 50%, and 70%) be reported routinely, since they basically fulfill different functions. The 20% level indicates the number of criteria per person showing a minimally clinically significant improvement, 50% indicates how many criteria show a "good" response, and 70% shows how many criteria achieve a near-remission. Using all 3 thresholds would result in a clearer picture of a drug's utility.

Table 4. Trial results using mean area under the curve (AUC) of criteria fulfilled at the 20%, 50% or 70% threshold for each of the 6 comparisons with t test (parametric) and Wilcoxon (nonparametric) square root transformation and Fisher's exact test p values.

		Untransformed				Square Root Transformation			Binary (AUC = 0 vs AUC > 0)
Comparison		Mean Area	SD	t Test (p)	Wilcoxon (p)	Mean Area	SD	t Test (p)	Fisher Exact p
20	Placebo	349.99	340.27	0.6287	0.1754	17.08	7.72	0.3781	0.5576
	Myochrys	373.56	216.55			18.19	6.57		
50	Placebo	93.89	136.69	0.0755	0.0059	7.03	6.74	0.0076	0.0023
	Myochrys	134.58	119.41			10.05	5.83		
70	Placebo	28.32	65.03	0.1826	0.0244	2.72	4.62	0.0356	0.0115
	Myochrys	43.33	60.53			4.53	4.81		
20	Placebo	349.99	340.27	0.8226	0.3510	17.08	7.72	0.5822	1.0000
	Auranof	361.20	222.53			17.79	6.73		
50	Placebo	93.89	136.69	0.1993	0.1368	7.03	6.74	0.0971	0.0695
	Auranof	125.48	133.56			9.06	6.63		
70	Placebo	28.32	65.03	0.8807	0.5683	2.72	4.62	0.6708	0.7013
	Auranof	29.86	49.87			3.07	4.55		
20	Myochrys	373.56	216.55	0.7241	0.6637	18.19	6.57	0.7038	0.3578
	Auranof	361.20	222.53			17.79	6.73		
50	Myochrys	134.58	119.41	0.6519	0.2834	10.05	5.83	0.3195	0.2115
	Auranof	125.48	133.56			9.06	6.63		
70	Myochrys	43.33	60.53	0.1299	0.0519	4.53	4.81	0.0527	0.0166
	Auranof	29.86	49.87			3.07	4.55		
20	Placebo	462.33	368.98	0.0014	0.0018	19.03	10.11	0.0013	0.1983
	High dpen	693.77	422.63			24.62	9.41		
50	Placebo	101.80	120.70	< 0.0001	< 0.0001	7.41	6.92	< 0.0001	0.0023
	High dpen	297.05	312.62			14.17	9.86		
70	Placebo	21.33	44.48	0.0003	0.0004	2.15	4.13	0.0002	0.0022
	High dpen	106.01	158.47			6.72	7.85		
20	Placebo	462.33	368.98	0.0020	0.0031	19.03	10.11	0.0027	0.4721
	Low dpen	676.68	397.90			24.24	9.49		
50	Placebo	101.80	120.70	0.0001	0.0003	7.41	6.92	0.0002	0.0186
	Low dpen	255.53	263.06			13.11	9.21		
70	Placebo	21.33	44.48	0.0054	0.0031	2.15	4.13	0.0017	0.0065
	Low dpen	77.16	137.89			5.52	6.87		
20	Low dpen	676.68	397.90	0.7845	0.8853	24.24	9.49	0.7921	0.6820
	High dpen	693.77	422.63			24.62	9.41		
50	Low dpen	255.53	263.06	0.3458	0.5224	13.11	9.21	0.4624	0.5577
	High dpen	297.05	312.62			14.17	9.86		
70	Low dpen	77.16	137.89	0.2031	0.3751	5.52	6.87	0.2849	0.7618
	High dpen	106.01	158.47			6.72	7.85		

The AUC gives virtually identical results to those provided by the mean number of criteria. Again, 50% performs marginally better than 20% or 70%, but the difference is subtle and our recommendation is to report all of these criteria. There is a theoretical reason to utilize the AUC, since it captures the longitudinal effect of the drug on the outcome of interest. However, it is harder to calculate. A similar effect is also captured by the use of longitudinal analysis, which involves a different set of statistical assumptions. Any of these approaches improves the discriminatory power of the outcome measure. The result is that smaller numbers of patients need to be entered into a trial to obtain statistically significant results. Conversely, the power to detect differences between modalities is improved. In either case the efficiency of trial management is improved. An example of this is our finding that auranofin is more effective than placebo when the outcome is measured by the more discriminatory criteria.

While the contingency tables and histograms were useful to illustrate the distributions that resulted in the mean number of criteria fulfilled, they did not generate any new hypotheses or any distinctly novel insights, as might have occurred if the distributions were unusual, such as bimodal or skewed in a particular fashion.

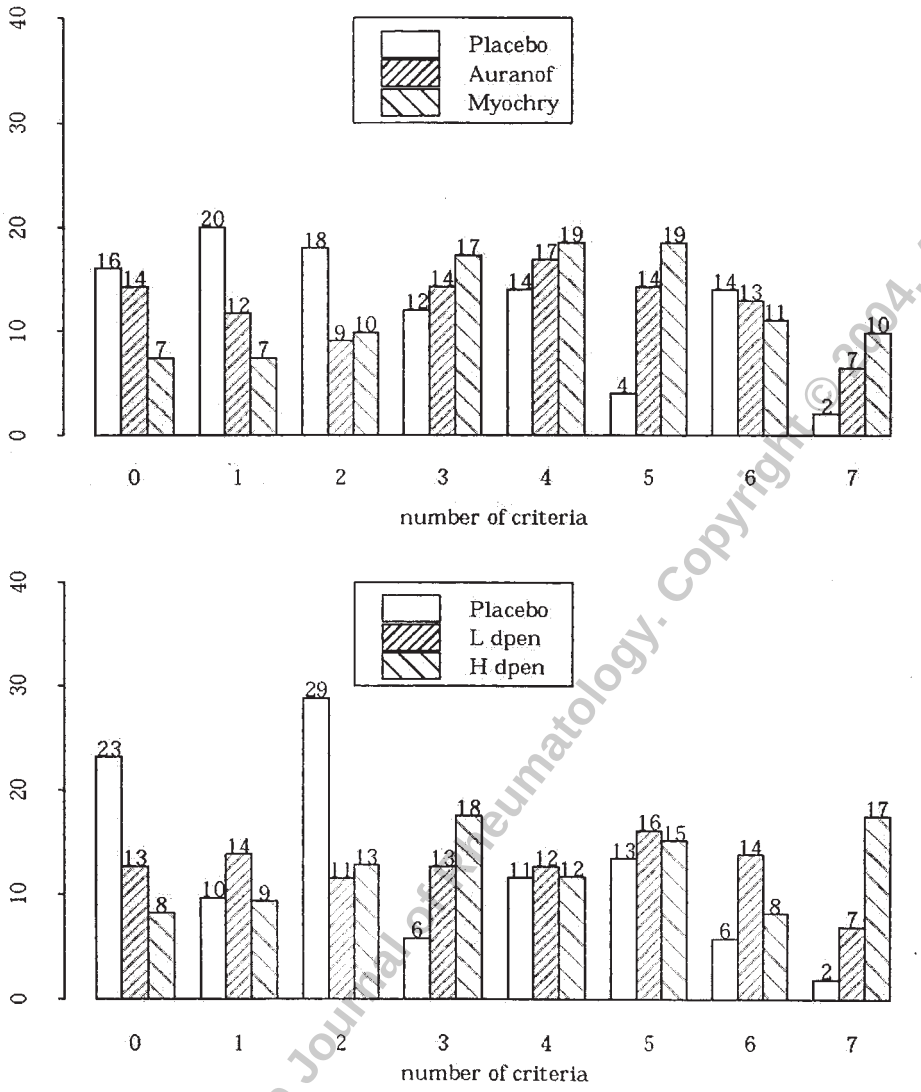

Figure 3. Histogram distribution of the number of criteria met at the 20% threshold for the placebo, auranofin, and myochrysine treated groups in the oral gold trial (top panel) and the placebo, low dpen, and high dpen treated groups in the dpen trial (bottom panel).

An interesting byproduct of this study provides a graphic illustration of the importance of discriminatory outcome measures for clinical trials. Using the mean number of ACR criteria fulfilled or AUC of the number of ACR criteria fulfilled in contrast to the ACR 20, 50, or 70 clearly indicates that auranofin and penicillamine are efficacious DMARD for the treatment of RA, although determining drug efficacy was not a primary goal of the analysis. The lack of complete data on all patients argues for some caution in interpreting these efficacy findings, although the comparisons among scoring methods remain internally valid.

In summary, these analyses suggest that the ACR 20 might profitably be amended by one of 2 continuous measures: either the mean number of criteria fulfilled or the mean area under the curve of the ACR criteria fulfilled, for the purpose of quantifying the degree of improvement of patients on a given therapeutic modality and to compare the efficacy of various modalities. The ACR 20 will remain an outstanding tool for discerning potentially useful therapies for RA. We view the different approaches as complementary — the ACR 20 helpful to discern whether the modality is efficacious and then ACR-C or area under the curve to discriminate between therapies and to give a more quantitative sense of *degree* of improvement. The consensus on the elements of the scales is critical, but our use of the various dimensions in composite criteria and indices should meet the objectives we wish to achieve. Identifying promising therapies is most appropriately done with ACR 20, whereas deciding between therapies may be more appropriately done with ACR-C or area under the curve. From a clinical stand-

point the latter issue dominates; from a pharmaceutical perspective, at least initially the former is the most important perspective. Further studies on additional data sets should be directed toward developing a reduced data set to facilitate precision and widespread utilization of this valuable outcome measurement instrument.

Finally, we believe that all trials should report their results in a similar standardized fashion. Our recommendation is that the conventional ACR 20, 50, and 70 be supplemented with one of the 2 continuous measures we propose — either the mean number of ACR criteria fulfilled at the 20%, 50%, and 70% level or the mean area under the curve at these same thresholds. This approach would generate a more comprehensive picture of the efficacy of any particular therapeutic modality.

REFERENCES

1. Lansbury J. Clinical studies with an articular index for the assessment of joint tenderness in patients with rheumatoid arthritis: theoretic and clinical considerations. Arthritis Rheum 1958;1:505-22.
2. Ritchie DM, Boyle JA, McInnes JM, et al. Clinical studies with an articular index for the assessment of joint tenderness in patients with rheumatoid arthritis. QJM 1968;37:393-406.
3. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729-40.
4. Paulus HE, Egger MJ, Ward JR, Williams HJ. Analysis of improvement in individual rheumatoid arthritis patients treated with disease-modifying antirheumatic drugs, based on the findings in patients treated with placebo. The Cooperative Systematic Studies of Rheumatic Diseases Group. Arthritis Rheum 1990;33:477-84.
5. Weinblatt ME, Kremer JM, Bankhurst AD, et al. A trial of etanercept, a recombinant tumor necrosis factor receptor:Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate. N Engl J Med 1999;340:253-9.
6. Moreland LW, Schiff MH, Baumgartner SW, et al. Etanercept therapy in rheumatoid arthritis. A randomized, controlled trial. Ann Intern Med 1999;130:478-86.
7. Strand V, Tugwell P, Bombardier C, et al. Function and health-related quality of life: results from a randomized controlled trial of leflunomide versus methotrexate or placebo in patients with active rheumatoid arthritis. Leflunomide Rheumatoid Arthritis Investigators Group. Arthritis Rheum 1999;42:1870-8.
8. Tugwell P, Wells G, Strand V, et al. Clinical improvement as reflected in measures of function and health-related quality of life following treatment with leflunomide compared with methotrexate in patients with rheumatoid arthritis: sensitivity and relative efficiency to detect a treatment effect in a twelve-month, placebo-controlled trial. Leflunomide Rheumatoid Arthritis Investigators Group. Arthritis Rheum 2000;43:506-14.
9. Elliott MJ, Maini RN, Feldmann M, et al. Repeated therapy with monoclonal antibody to tumour necrosis factor alpha (cA2) in patients with rheumatoid arthritis. Lancet 1994;344:1125-7.
10. Elliott MJ, Maini RN, Feldmann M, et al. Treatment of rheumatoid arthritis with chimeric monoclonal antibodies to tumor necrosis factor alpha. Arthritis Rheum 1993;36:1681-90.
11. Kavanaugh A, St. Clair EW, McCune WJ, Braakman T, Lipsky P. Chimeric anti-tumor necrosis factor-alpha monoclonal antibody treatment of patients with rheumatoid arthritis receiving methotrexate therapy. J Rheumatol 2000;27:841-50.
12. Williams HJ, Ward JR, Reading JC, et al. Low-dose D-penicillamine therapy in rheumatoid arthritis. A controlled, double-blind clinical trial. Arthritis Rheum 1983;26:581-92.
13. Ward JR, Williams HJ, Egger MJ, et al. Comparison of auranofin, gold sodium thiomalate, and placebo in the treatment of rheumatoid arthritis: a controlled clinical trial. Arthritis Rheum 1983;26:1303-15.
14. Egger MJ, Ward JR, Karg MB, Williams HJ, Reading JC. Reliability and validity of the CSSRD functional assessment survey in rheumatoid arthritis. Cooperative Systematic Studies of Rheumatic Diseases. Arthritis Care Res 1995;8:21-7.
15. Bellamy N. Clinimetric concepts in outcome assessment: the OMERACT filter. J Rheumatol 1999;26:948-50.
16. Anderson JJ, Felson DT, Meenan RF, Williams HJ. Which traditional measures should be used in rheumatoid arthritis clinical trials? Arthritis Rheum 1989;32:1093-9.
17. Hawley DJ, Wolfe F. Sensitivity to change of the Health Assessment Questionnaire (HAQ) and other clinical and health status measures in rheumatoid arthritis: results of short-term clinical trials and observational studies versus long-term observational studies. Arthritis Care Res 1992;5:130-6.
18. Ward MM. Clinical measures in rheumatoid arthritis: which are most useful in assessing patients? J Rheumatol 1994;21:17-27.
19. Tugwell P, Boers M, Baker P, Wells G, Snider J. Endpoints in rheumatoid arthritis. J Rheumatol 1994;21 Suppl 42:2-8.
20. Smolen JS, Breedveld FC, Eberl G, et al. Validity and reliability of the twenty-eight-joint count for the assessment of rheumatoid arthritis activity. Arthritis Rheum 1995;38:38-43.
21. Buchbinder R, Bombardier C, Yeung M, Tugwell P. Which outcome measures should be used in rheumatoid arthritis clinical trials? Clinical and quality-of-life measures' responsiveness to treatment in a randomized controlled trial. Arthritis Rheum 1995;38:1568-80.
22. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727-35.
23. Pillemer SR, Fowler SE, Tilley BC, et al. Meaningful improvement criteria sets in a rheumatoid arthritis clinical trial. Minocycline in Rheumatoid Arthritis Trial Group. Arthritis Rheum 1997;40:419-25.
24. van Gestel AM, Anderson JJ, van Riel PL, et al. ACR and EULAR improvement criteria have comparable validity in rheumatoid arthritis trials. American College of Rheumatology European League of Associations for Rheumatology. J Rheumatol 1999;26:705-11.
25. Felson DT, Anderson JJ, Lange ML, Wells G, LaValley MP. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? Arthritis Rheum 1998;41:1564-70.
26. Pincus T, Stein CM. ACR 20: clinical or statistical significance? Arthritis Rheum 1999;42:1572-6.
27. Pham B, Cranney A, Boers M, Verhoeven AC, Wells G, Tugwell P. Validity of area-under-the-curve analysis to summarize effect in rheumatoid arthritis clinical trials. J Rheumatol 1999;26:712-6.
28. Hurst S, Kallan MJ, Wolfe F, Fries JF, Albert DA. Methotrexate, hydroxychloroquine, and intramuscular gold in rheumatoid arthritis: relative area under the curve effectiveness and sequence effects. J Rheumatol 2002;29:1639-45.
29. Farrar JT, Portenoy RK, Berlin JA, Kinman JL, Strom BL. Defining the clinically important difference in pain outcome measures. Pain 2000;88:287-94.
30. Wolfe F, Pincus T. Listening to the patient: a practical guide to self-report questionnaires in clinical care. Arthritis Rheum 1999;42:1797-808.
31. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis

Rheum 1999;42:2220-30.
32. Boers M, Verhoeven AC, Markusse HM, et al. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet 1997;350:309-18.
33. Paulus HE, Bulpitt KJ, Ramos B, Park G, Wong WK. Relative contributions of the components of the American College of Rheumatology 20% criteria for improvement to responder status in patients with early seropositive rheumatoid arthritis. Arthritis Rheum 2000;43:2743-50.
34. van der Heijde DM, van Riel PL, van Leeuwen MA, van't Hof MA, van Rijswijk MH, van de Putte LB. Prognostic factors for radiographic damage and physical disability in early rheumatoid arthritis. A prospective follow-up study of 147 patients. Br J Rheumatol 1992;31:519-25.
35. Krause D, Schleusser B, Herborn G, Rau R. Response to methotrexate treatment is associated with reduced mortality in patients with severe rheumatoid arthritis. Arthritis Rheum 2000;43:14-21.
36. Callahan LF, Pincus T, Huston JW 3rd, Brooks RH, Nance EP Jr, Kaye JJ. Measures of activity and damage in rheumatoid arthritis: depiction of changes and prediction of mortality over five years. Arthritis Care Res 1997;10:381-94.
37. Wolfe F, Lassere M, van der Heijde D, et al. Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. J Rheumatol 1999;26:484-9.
38. Wolfe F. Critical issues in longitudinal and observational studies: purpose, short versus long term, selection of study instruments, methods, outcomes, and biases. J Rheumatol 1999;26:469-72.
39. Albert DA, Aksentijevich S, Hurst S, Fries JF, Wolfe F. Modeling therapeutic strategies in rheumatoid arthritis: use of decision analysis and Markov models. J Rheumatol 2000;27:644-52.
40. Kobelt G, Eberhardt K, Jonsson L, Jonsson B. Economic consequences of the progression of rheumatoid arthritis in Sweden. Arthritis Rheum 1999;42:347-56.
41. Tilley BC, Pillemer SR, Heyse SP, Li S, Clegg DO, Alarcon GS. Global statistical tests for comparing multiple outcomes in rheumatoid arthritis trials. MIRA Trial Group. Arthritis Rheum 1999;42:1879-88.
42. Molenaar E, van der Heijde D, Boers M. Update on outcome assessment in rheumatic disorders. Curr Opin Rheumatol 2000;12:91-8.