Reprinted from ARTHRITIS & RHEUMATISM
Published by J. B. Lippincott Company
Vol. 38, No. 6, June 1995

AMERICAN COLLEGE OF RHEUMATOLOGY PRELIMINARY DEFINITION OF IMPROVEMENT IN RHEUMATOID ARTHRITIS

DAVID T. FELSON, JENNIFER J. ANDERSON, MAARTEN BOERS, CLAIRE BOMBARDIER, DANIEL FURST, CHARLES GOLDSMITH, LINDA M. KATZ, ROBERT LIGHTFOOT, JR., HAROLD PAULUS, VIBEKE STRAND, PETER TUGWELL, MICHAEL WEINBLATT, H. JAMES WILLIAMS, FREDERICK WOLFE, and STEPHANIE KIESZAK

Objective. **Trials of rheumatoid arthritis (RA) treatments report the average response in multiple outcome measures for treated patients. It is more clinically relevant to test whether individual patients improve with treatment, and this identifies a single primary efficacy measure. Multiple definitions of improvement are currently in use in different trials. The goal of this study was to promulgate a single definition for use in RA trials.**

Methods. **Using the American College of Rheumatology (ACR) core set of outcome measures for RA trials, we tested 40 different definitions of improvement, using a 3-step process. First, we performed a survey of rheumatologists, using actual patient cases from trials, to evaluate which definitions corresponded best to rheumatologists' impressions of improvement, eliminating most candidate definitions of improvement. Second, we tested 20 remaining definitions to determine which maximally discriminated effective treatment from placebo treatment and also minimized placebo response rates. With 8 candidate definitions of improvement remaining, we tested to see which were easiest to use and were best in accord with rheumatologists' impressions of improvement.**

Results. **The following definition of improvement was selected: 20% improvement in tender and swollen joint counts and 20% improvement in 3 of the 5 remaining ACR core set measures: patient and physician global assessments, pain, disability, and an acute-phase reactant. Additional validation of this definition was carried out in a comparative trial, and the results suggest that the definition is statistically powerful and does not identify a large percentage of placebo-treated patients as being improved.**

Conclusion. **We present a definition of improvement which we hope will be used widely in RA trials.**

Recent work by our committee in concert with the international rheumatology community has led to

From the Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials, a subcommittee of the Committee on Health Care Research, American College of Rheumatology.

David T. Felson, MD, MPH, Jennifer J. Anderson, PhD: Boston University Arthritis Center, Boston, Massachusetts; Maarten Boers, MD, PhD, MSc: University Hospital, Maastricht, The Netherlands; Claire Bombardier, MD: Wellesley Hospital, University of Toronto, Toronto, Ontario, Canada; Daniel Furst, MD: Virginia Mason Medical Center, Seattle, Washington; Charles Goldsmith, PhD: McMaster University, Hamilton, Ontario, Canada; Linda M. Katz, MD, MPH: Food and Drug Administration, Rockville, Maryland; Robert Lightfoot, Jr., MD: University of Kentucky, Lexington; Harold Paulus, MD: University of California at Los Angeles; Vibeke Strand, MD: Stanford University, Stanford, California; Michael Weinblatt, MD: Brigham and Women's Hospital, Boston, Massachusetts; H. James Williams, MD: University of Utah, Salt Lake City; Frederick Wolfe, MD: Arthritis Center, Wichita, Kansas; Stephanie Kieszak, MA: American College of Rheumatology, Atlanta, Georgia.

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as a reflection of the views of the Food and Drug Administration.

Address reprint requests to American College of Rheumatology, 60 Executive Park South, Suite 150, Atlanta, GA 30329.

Submitted for publication September 2, 1994; accepted in revised form December 6, 1994.

the development of a uniform core set of outcome measures for rheumatoid arthritis (RA) trials (1). While this core set represents an advance in defining and standardizing the outcomes to be measured in RA trials, it has not changed the focus of trial reporting and analysis, i.e., average improvement for each of the outcomes measured. Usually clinical trials in RA report the average (mean or median) improvement experienced by treated patients, with the average improvement with one treatment compared with the average improvement with another.

Unfortunately, this current practice is problematic: moderate average improvement of patients undergoing a treatment may occur because all patients improved modestly or because half of the patients experienced dramatic improvement and the other half no improvement at all. Further, testing for significant results in each of 7 core set measures increases the likelihood of detecting a difference between therapies when no real difference exists (a Type I error) and makes it difficult to interpret the difference between therapies when just 1 or 2 outcome measures are significantly different (Are 2 therapies different if 1 of such outcomes shows significant differences between treatment groups? Two of 7? etc.).

The availability of a single definition of response in RA trials would resolve this problem. It would be a single primary end point for analysis. Problems associated with multiple testing would diminish. If a uniform definition of improvement were used, the percentage of patients improving could be compared across trials, with the caveat that patients in different trials are different and may not be equally likely to improve given the same therapy.

Furthermore, patients are interested in the likelihood that they themselves will improve, not in the average response of similar patients being treated. Also, a focus on which patients improve in trials could lead to investigations that characterize what types of patients improve with different therapies. Current practice does not allow this, since individual patients are not well characterized by reports of trials. Last, as will be shown below, relying on a single definition of improvement that incorporates information from several outcome measures can substantially enhance the statistical power of a trial.

EXISTING DEFINITIONS OF IMPROVEMENT

Definitions of improvement have been developed previously. First, the American Rheumatism Association (now the American College of Rheumatology [ACR]) defined remission in RA (2), but remission occurs so rarely in trials that it has not been a useful outcome measure for trials.

Using data from multicenter RA trials, Paulus et al (3) developed a definition of improvement based on a set of measures that discriminated well between active second-line drug treatment and placebo and that limited placebo response to ~5%. This definition requires response in at least 4 of 6 selected measures. These include a 20% improvement in morning stiffness, erythrocyte sedimentation rate (ESR), joint tenderness score, and joint swelling score and improvement by at least 2 grades on a 5-grade scale (or from grade 2 to grade 1) for patient and physician global assessments of current disease severity.

This definition of improvement is clinically reasonable and workable in the context of trials, but it has been used inconsistently. Although it was developed with statistical discrimination in mind, it may not correspond to the patient's or clinician's perception of clinical improvement. In addition, it relies on global severity scales that are unique to trials from the Cooperative Systematic Studies of the Rheumatic Diseases (a 5-point adjectival scale), and are not widely used elsewhere. The 5-point adjectival scale may not be as sensitive to change as a 7-point scale or a 10-cm visual analog scale (4). Furthermore, elements included in the Paulus improvement criteria do not correspond to the current core set: morning stiffness, a measure often insensitive to change, is included, and measurement of physical function is excluded. Joint counts, morning stiffness, and ESR are equally weighted in the Paulus criteria, whereas studies of clinician perception of improvement suggest that joint counts are emphasized more heavily (5).

Dutch investigators (6) have suggested an index (the Disease Activity Score [DAS]) to be used in evaluating improvement. This score, while not easy to compute, has the advantage of drawing from several different outcome measures to assess disease activity, with measures weighted toward joint counts.

The investigators in many trials have created their own definitions of improvement. For example, among 15 trials of RA treatments (other than nonsteroidal antiinflammatory drugs) published in 1992 (references available from the authors), only 6 used improvement or response criteria and each used a different definition of improvement, with only 1 using the Paulus criteria. This heterogeneity prevents comparisons of rates of improvement across trials and provides a powerful argument in favor of a standardized, widely used definition of improvement.

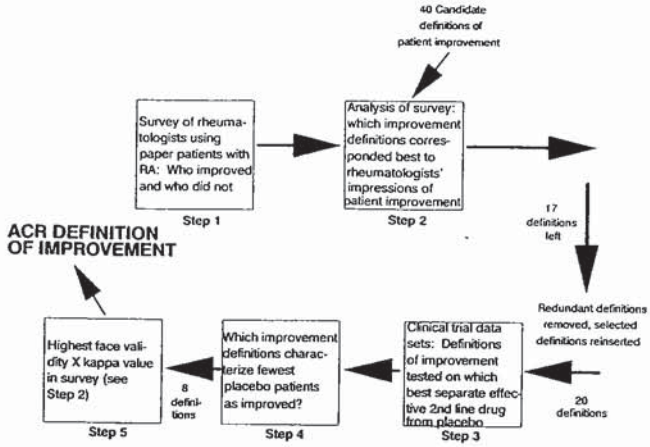

Figure 1. Process of choosing American College of Rheumatology (ACR) criteria for rheumatoid arthritis (RA) improvement.

As part of an ACR committee whose objective was to develop uniform standards for RA trial measurements, we created a definition of improvement using elements of the ACR core set. To achieve that goal, we drew on clinical impressions of which RA patients improve, to identify what measures clinicians emphasize in evaluating patient improvement. We combined this with a statistical approach similar to that used by Paulus et al (3), with additional trial data to allow comparison of a variety of improvement definitions. Our statistical approach focused on the definition of improvement that best discriminates between active drug–treated and placebo-treated patients. The overall process is depicted in Figure 1.

METHODS

Physician survey (Figure 1, step 1). The first step was to assess how rheumatologists decide whether a patient has improved. Survey studies (5) had suggested that rheumatologists regard a patient as improved if the tender or swollen joint count improves by ~20% or if other outcomes improve by a larger percent. However, earlier studies combined data on clinicians and nonclinicians, did not include all elements of the ACR core set, and did not necessarily use data from real patients.

We therefore surveyed rheumatologists, using "paper" patients selected from real clinical trials by stratified random sampling to include a large number of survey patients near expected thresholds for improvement (20–45% improvement in at least 3 outcomes). The 89 rheumatologists to whom the survey was sent consisted of Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) committee members, participants, and others chosen because of their considerable RA clinical and/or clinical trial experience. Sixty-eight (76.4%) returned the surveys, and all surveys returned were usable. The ages of the respondents ranged from 31 to 69 years (median 47 years), 15% were female, and the median number of hours of patient care per week was 17, with 62% of the respondents medical school based.

For each element of the core set (e.g., tender joint count), data at baseline and at 6 months were provided and the percent change was noted. We asked survey respondents whether each paper patient had improved or not. Since the survey was also designed to evaluate patient worsening, only 43 of 69 patients in the survey provided useful information on improvement. The other 26 patients were substantially below expected thresholds for improvement. As a validation of our assumptions about which patients from the survey would provide useful data regarding improvement, none of these latter 26 patients were designated as improved by more than 14% of survey respondents.

Analysis of physician survey (Figure 1, step 2). In the survey results, we focused on patients characterized as improved by at least 80% of the surveyed rheumatologists. We chose the cutoff of 80% because we were interested in patients whom almost all rheumatologists would characterize as improved. We then examined the extent to which these same patients were characterized as improved according to various possible definitions of improvement, as shown below. We also looked at the percent of false-positives, i.e., patients not identified as improved by ≥80% of rheumatologists but classified as improved by the improvement definition. We decided that all candidate definitions of improvement with chi-square values <6 (which corresponds to $P = 0.01$) or false-positive rates >25% would be excluded from further consideration. Changing these thresholds did not change the relative performance of improvement definitions in the survey.

Analysis of trial data (Figure 1, steps 3 and 4). Once the survey results had eliminated some of the possible definitions of improvement, we turned to statistical analysis of trial data. The goal was to select the improvement definition(s) that best discriminated active second-line drugs from placebo. We assembled a data set of 5 placebo-controlled trials of second-line drugs, including 1 trial of gold (7) and 4 of methotrexate (refs. 8–10 and Schmid FR et al: unpublished observations). One of these (Schmid FR et al: unpublished observations) was a small unpublished trial, and its exclusion does not affect analytic results. Since we wished to choose regimens that offered as large as possible an efficacy difference between drug and placebo, we excluded 1 auranofin arm in 1 trial, since evidence (11,12) suggests it is relatively weak. Six of the 7 ACR core set measures were included in these trials, but like many completed RA trials, 4 of the 5 trials did not include an assessment of functional status. We substituted grip strength, a measure whose change correlated moderately ($r = 0.45$ with change in Arthritis Impact Measurement Scales physical function in 1 trial [7] and $r = 0.64$ in another study [13]) and which loads with functional status in factor analyses of trial data, suggesting that it measures a similar construct (4).

The data set contained 508 patients, but 320 patients (177 active drug–treated/143 placebo-treated) remained after exclusion of patients with missing data for at least 1 element of the core set (or for grip strength). Additional analysis of the 1 trial with data on function suggests that the results would likely not have changed if such data were available in all trials. After selecting the improvement definition based

Table 1. Description of candidate definitions for improvement applied to the 43 patient profiles, and results of rheumatologist survey*

Criterion code	Definition type	Outcome measures							Survey results	
		Physical disability	Pain	ESR	Tender joint count	Swollen joint count	Physician global assessment	Patient global assessment	Chi-square value	No. (%) false-positives, of 25
Pa	**Paulus (4/6)**		20	20	20	20	40	40	16.4	2 (8)
Pb	**Paulus (4/6)**	20		20	20	20	40	40	14.0	2 (8)
Pc	**Paulus (5/7)**	20	20	20	20	20	40	40	18.1	0 (0)
Wa	**WHO (4/5)**		40	20	20 (≥5)	20	40†	40†	15.8	0 (0)
Wb	**WHO (4/5)**		40	20	20 (≥5) (req.)‡	20 (req.)‡	40†	40†	15.8	0 (0)
Wc	**WHO (5/7)**	20	20	20	20 (req.)‡	20 (req.)‡	20	20	18.1	0 (0)
Ea	**Equal (5/7)**	20	20	20	20	20	20	20	9.5	5 (20)
Eb	**Equal (6/7)**	20	20	20	20	20	20	20	15.8	0 (0)
Ec	Equal (4/7)	30	30	30	30	30	30	30	3.9	9 (36)
Ed	**Equal (5/7)**	30	30	30	30	30	30	30	6.0	2 (8)
Ee	Equal (4/7)	30	30	30	30 (≥4)	30 (≥4)	30	30	3.9	9 (36)
Ef	**Equal (5/7)**	30	30	30	30 (≥4)	30 (≥4)	30	30	6.0	2 (8)
Eg	Equal (3/7)	40	40	40	40	40	40	40	3.3	7 (28)
Eh	Equal (4/7)	40	40	40	40	40	40	40	2.0	1 (4)
Ei	Equal (5/7)	40	40	40	40	40	40	40	0	0 (0)
Ej	Equal (3/7)	40	40	40	40 (≥4)	40 (≥4)	40	40	4.5	6 (24)
Ek	Equal (4/7)	40	40	40	40 (≥4)	40 (≥4)	40	40	2.0	1 (4)
El	Equal (5/7)	40	40	40	40 (≥4)	40 (≥4)	40	40	0	0 (0)
Em	Equal (3/7)	50	50	50	50	50	50	50	1.8	4 (16)
En	Equal (4/7)	50	50	50	50	50	50	50	0.7	1 (4)
Eo	Equal (3/7)	50	50	50	50 (≥4)	50 (≥4)	50	50	2.9	3 (12)
Ep	Equal (4/7)	50	50	50	50 (≥4)	50 (≥4)	50	50	0.7	1 (4)
Oa	**OMERACT (4/7)**	40	40	40	20	20	40	40	7.8	2 (8)
Ob	OMERACT (5/7)	40	40	40	20	20	40	40	2.9	0 (0)
Oc	**OMERACT (4/7)**	40	40	40	20 (≥4)	20 (≥4)	40	40	7.8	2 (8)
Od	OMERACT (5/7)	40	40	40	20 (≥4)	20 (≥4)	40	40	2.9	0 (0)
Ja	**Joint count (2/2)**				20 (≥4)	20 (≥4)			12.2	6 (24)
Jb	Joint count (2/2)				50	50			0	3 (12)
Da	**DAS 3 (2 SEM)**			§	§	§			20.9	6 (24)
Db	DAS 3 (4/3 SEM)			§	§	§			16.8	8 (32)
Dc	DAS 3 (1 SEM)			§	§	§			7.7	14 (56)
Dd	DAS 4 (2 SEM)			§	§	§		§	18.7	7 (28)
De	DAS 4 (4/3 SEM)			§	§	§		§	16.8	8 (32)
Df	DAS 4 (1 SEM)			§	§	§		§	7.7	14 (56)
Dg	**Linear DAS3**			§	§	§			20.9	6 (24)
Dh	**Linear DAS4**			§	§	§		§	20.9	6 (24)
I1	Index (0.5 units)					¶			4.9	11 (44)
I2	Index (0.5 units)				¶	¶			15.0	9 (36)
I3	**Index (0.5 units)**			¶	¶	¶			20.9	6 (24)
I7	Index (0.5 units)	¶	¶	¶	¶	¶	¶	¶	15.0	11 (44)

* Definitions shown in boldface were selected for the next stage of analysis. ESR = erythrocyte sedimentation rate; WHO = World Health Organization; OMERACT = Outcome Measures in Rheumatoid Arthritis Clinical Trials; DAS = Disease Activity Score.
† Either 1 of 2 items in row may be counted.
‡ Item is required.
§ Item is in DAS.
¶ Item is in the index.

on its performance in placebo-controlled trials, we tested it in a large comparison trial data set of methotrexate and auranofin, in which methotrexate had been shown to be more efficacious (n = 274 patients with complete data) (12).

In analyzing trial data, we calculated the percentage of active drug–treated patients who were identified as improved by each candidate improvement definition and the percentage of placebo-treated patients who were characterized as improved by each definition. For each improvement definition, we also evaluated the statistical power in discriminating active drug from placebo groups.

The first stage of assessing candidate definitions entailed selecting the most statistically powerful. Of those with roughly equal power, we then chose the ones that identified the fewest placebo-treated patients as improved. Because of the imprecision of estimates, we relied further on the analysis of the comparative trial (methotrexate versus auranofin) and attempted to be generous in our estimates of equivalence, so as not to eliminate a definition of improvement because of insufficient data.

Ease of use, credibility (Figure 1, step 5). From those definitions remaining, we made our final choice. As a group

of experienced trialists, we ranked the face validity (clinical reasonableness and ease of use) of the remaining definitions on a 1–8 scale with 8 the highest, and then tabulated the ranks. Also, we returned to the rheumatologist survey and ranked each definition by its kappa statistic (another measure of agreement between the rheumatologists' impression of improvement and the definition's classification of improvement). These 2 rankings were multiplied, and the definition with the best score was selected.

RESULTS

The survey. Of the 43 "paper" survey patients that were the focus of our investigation of improvement, 18 were thought by ≥80% of the respondents to have improved and 25 were not.

We tested 40 possible criteria for improvement (Table 1). These were selected because they were used in trials, because they were recommended in publications, by members of our committee, or by the international community, or because they were variations on used or recommended definitions.

There were 7 groups of candidate improvement definitions. The first group was derived from the Paulus criteria (3) and substituted improvement in pain or physical disability for the Paulus criterias' improvement in morning stiffness. This group of definitions was referred to as Paulus.

Another group of definitions of improvement *required* improvement in the tender and swollen joint counts, as well as in a proportion of other core set elements. Because of similar recent preliminary World Health Organization recommendations developed by 1 of the authors (Dr. Furst), we designated this group of improvement definitions as WHO.

A third group (called Equal) weighted each of the core set elements equally and tested equal percent improvements in all core set elements. For example, one definition was 20% improvement in 5 of 7 of the core set elements, another 30% improvement in 5 of 7, and another 30% in 4 of 7, etc.

For the fourth group, developed from OMERACT meeting surveys (and therefore called OMERACT), we used evidence that clinicians emphasized improvement in joint counts and developed improvement definitions with ≥20% improvement in tender or swollen joint counts or at least 40% improvement in the other measures (improvements in joint count not required).

Yet another group of definitions of improvement (called Joint Count) focused only on joint count measures defining improvement as improvement in tender and/or swollen joint counts.

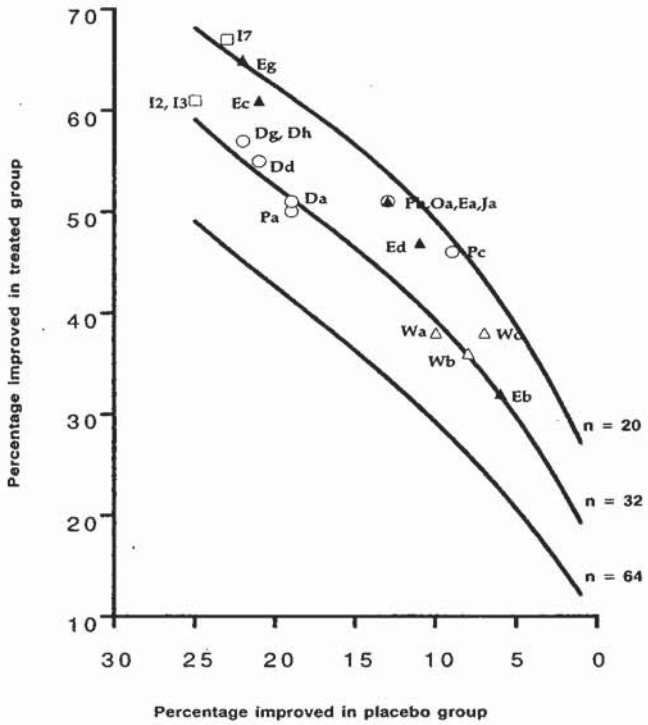

Figure 2. Performance of candidate criteria in placebo-controlled trials. See Table 1 for definitions of criterion codes.

The sixth group evaluated the recommended improvement definitions using the DAS (6), an index, and tried out different cutpoints for improvement as well as a linearized version (calculated using the linear regression estimate of log [esr] over the interval 0–50 and of the square root of tender joint count over the interval 15–45). There are 2 versions of the DAS: 1 using 2 joint count measures and the ESR and the other using the same 3 measures plus patient global assessment.

For the last group (called Index), we constructed pooled indices of improvement, dividing the change in each outcome measure by its change standard deviation (the latter derived from all trial patients) to create an effect size for each outcome measure, and then averaging effect sizes. A change of 0.5 effect size units was used as the cutoff for improvement.

Of the 40 possible definitions of improvement tested, 17 met the previously defined threshold in the survey, low false-positivity rate and high chi-square value. These 17 definitions appear in boldface in Table 1. They include all improvement definitions in groups 1 (Paulus) and 2 (WHO), and selected definitions in

Table 2. Face validity and survey performance of the 8 candidate definitions of improvement that had the most statistical power and designated the fewest placebo-treated patients as improved

Improvement definition*	Face validity rank†	Survey kappa	Face validity × survey kappa
Paulus, Pc	3.2	0.592	1.9
WHO, Wc	4.3	0.592	2.6
Equal, Eg	4.0	0.278	1.1
Equal, Ec	3.6	0.300	1.1
Equal, Ea	3.0	0.470	1.4
Equal, Eb	3.0	0.538	1.6
Omeract, Ob	3.4	0.389	1.3
Index, I7	2.7	0.516	1.4

* See Table 1 for definitions of criterion codes.
† Scored on a scale of 1–8, with 8 being the highest face validity.

each of the other groups. The WHO and Paulus groups of definitions, those using the DAS, Index 3 (with 2 joint counts), and 1 of the joint count improvement criteria all had high chi-square values, suggesting that clinical perceptions of patient improvement rely heavily on joint count improvement. Nonetheless, the tendency for the DAS and joint count improvement definitions to have high false-positive rates suggests that clinicians evaluate more than just joint count in characterizing patients as being improved.

At least 1 improvement definition from each group was included in the next stage of analysis, but 2 that met the threshold were omitted because they were duplicative (Ef is similar to Ed and Oc is similar to Oa) (see Table 1 for definitions of criterion codes). In addition, at the request of committee members and for completeness, 5 additional variations of the remaining 15 candidate definitions (2 in the Index group [I2 and I7], 1 in the DAS group [Dd], and 2 in the Equal group [Ec and Eg]) were evaluated in the next stage with the anticipation that they might do well in discriminating active drug– from placebo-treated patients, giving a total of 20. We planned that later selection of an

Table 3. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis*

Required { ≥20% improvement in tender joint count
≥20% improvement in swollen joint count
+
≥20% improvement in 3 of following 5:
Patient pain assessment
Patient global assessment
Physician global assessment
Patient self-assessed disability
Acute-phase reactant (ESR or CRP)

Disease activity measure	Method of assessment
1. Tender joint count†	ACR tender joint count, an assessment of 28 or more joints. The joint count should be done by scoring several different aspects of tenderness, as assessed by pressure and joint manipulation on physical examination. The information on various types of tenderness should then be collapsed into a single tender-versus-nontender dichotomy.
2. Swollen joint count†	ACR swollen joint count, an assessment of 28 or more joints. Joints are classified as either swollen or not swollen.
3. Patient's assessment of pain	A horizontal visual analog scale (usually 10 cm) or Likert scale assessment of the patient's current level of pain.
4. Patient's global assessment of disease activity	The patient's overall assessment of how the arthritis is doing. One acceptable method for determining this is the question from the AIMS instrument: "Considering all the ways your arthritis affects you, mark 'X' on the scale for how well you are doing." An anchored, horizontal, visual analog scale (usually 10 cm) should be provided. A Likert scale response is also acceptable.
5. Physician's global assessment of disease activity	A horizontal visual analog scale (usually 10 cm) or Likert scale measure of the physician's assessment of the patient's current disease activity.
6. Patient's assessment of physical function	Any patient self-assessment instrument which has been validated, has reliability, has been proven in RA trials to be sensitive to change, and which measures physical function in RA patients is acceptable. Instruments which have been demonstrated to be sensitive in RA trials include the AIMS, the HAQ, the Quality (or Index) of Well Being, the MHIQ, and the MACTAR.
7. Acute-phase reactant value	A Westergren erythrocyte sedimentation rate or a C-reactive protein level.

* ACR = American College of Rheumatology; ESR = erythrocyte sedimentation rate; CRP = C-reactive protein; AIMS = Arthritis Impact Measurement Scales; RA = rheumatoid arthritis; HAQ = Health Assessment Questionnaire; MHIQ = McMaster Health Index Questionnaire; MACTAR = McMaster Toronto Arthritis Patient Preference Disability Questionnaire.
† For details on which joints, see refs. 14 and 15.

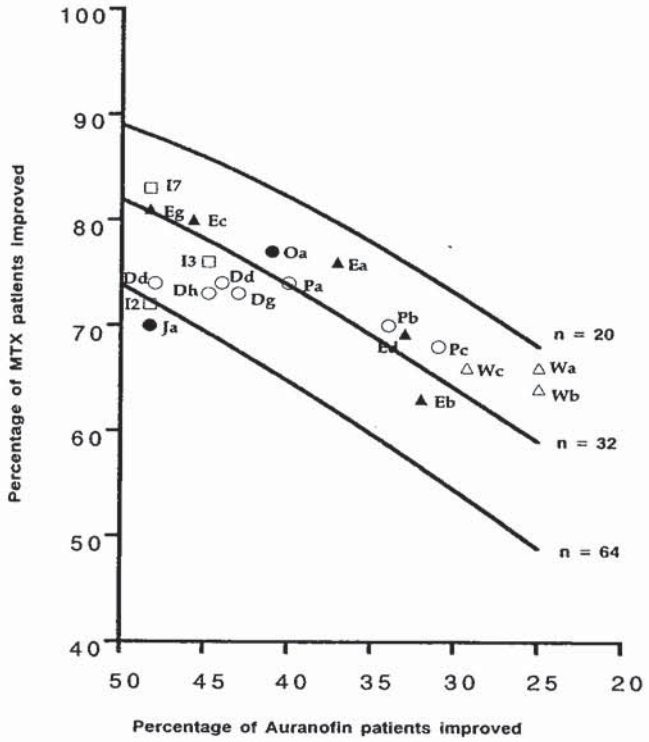

Figure 3. Performance of the newly developed criteria in comparing methotrexate and auranofin. See Table 1 for definitions of criterion codes.

improvement definition would reincorporate survey results, so that the added definitions that did not do well in the survey would be appropriately penalized.

Analyzing trial data. Using the previously described set of 5 placebo-controlled clinical trials, we evaluated the proportion of active drug–treated patients designated as improved and the proportion of placebo-treated patients as not improved for each of the remaining definitions of improvement (see Figure 2). Curves of equal power (isopower lines) are superimposed on the plot. Any 2 points on the same isopower curve are definitions with equal discriminating power, i.e., the trial sample sizes needed for those 2 definitions to detect differences between active drug– and placebo-treated patients as significant (2-tailed $\alpha = 0.05$, power 80%) are the same. In the lowest curve 64 patients per treatment group are needed, while for the other 2 lines, sample sizes of 32 and 20 per group, respectively, are required. For example, Equal definition Eb and DAS definition Da have similar discrimination in these trial data, but they differ in the proportion of placebo-treated and active drug–treated patients they identify as improved, with Da identifying more of both placebo-treated and active drug–treated patients as improved. The 2 candidate definitions discriminating best between active and placebo treatments were 2 that did not perform well in the physician survey, Index definition I7 and Equal definition Eg.

Definitions with the most power, that designated the fewest placebo-treated patients as improved, were chosen (see Figure 2). These were Paulus definition Pc, WHO definition Wc, Equal definitions Ea, Eb, Ec, and Eg, OMERACT definition Oa, and Index definition I7. Most candidate definition groups remained represented in this final list, although definitions of improvement based solely on joint count improvement and those based on the DAS were eliminated. These latter definitions had less power than the ones selected and were especially likely to characterize placebo-treated patients as improved.

We then scored each of the 8 remaining candidate definitions of improvement for face validity and multiplied the face validity score by the survey kappa score (Table 2). This procedure identified 1 definition that clearly scored better than the others, and this definition, WHO definition Wc, was selected as the definition for improvement (Table 3). It should be noted that not only did this definition do well in the survey (chi-square 18.1, no false-positives [Table 1]), but, in the analysis of trial data, it discriminated well between placebo and active treatment and identified few placebo-treated patients as improved.

Next, we tested this definition in another clinical trial data set, a multicenter trial of methotrexate versus auranofin. In this trial, mean improvements in individual measures were, in general, much greater for methotrexate-treated patients than for patients receiving auranofin (12). The definition selected and others like it in the WHO series performed as well as or better than any other types of definitions in discriminating between methotrexate and auranofin (Figure 3). As in placebo trials, joint count– and DAS-based definitions identified as improved a large percentage of patients who received the weaker therapy. The Equal definition and the Paulus definitions characterized more methotrexate-treated and more auranofin-treated patients as improved than did the definition selected.

DISCUSSION

Based on this analysis using several different approaches to evaluating potential definitions of improvement in RA, we suggest that improvement for clinical trial patients be defined as ≥20% improvement

in tender and swollen joint counts and ≥20% improvement in at least 3 of the following 5 ACR core set measures: pain, patient and physician global assessments, self-assessed physical disability, and acute-phase reactant. Our work suggests that this definition corresponds closely to clinicians' impression of patient improvement since it emphasizes joint counts, and furthermore, it discriminates powerfully between active and placebo treatment, identifying few placebo-treated patients as being improved.

This definition of improvement provides a single outcome measure that can be used in all RA trials. The definition of improvement can characterize the response of individual patients to therapy, and using it, investigators can profile those likely to respond to a therapy.

Our analyses suggest that this definition of improvement increases the power of clinical trials since it draws on information from multiple different outcome measures. Therefore, the sample size needed to demonstrate differences between therapies may decrease, making it possible for some trials that previously would have been considered to be underpowered to have sufficient patients to compare treatments. For example, for the comparative trial analyzed in Figure 3, between 20 and 32 patients per treatment group would be required using this improvement definition (80% power, $\alpha = 0.05$, 2-sided), versus at least 80 patients per group if the trial were analyzed in the current and traditional way, evaluating 1 of the 7 core set measures. Ultimately, if the improvement criteria are widely used in a standardized manner, it may be possible to rank the efficacy of different therapies based on the percentage of patients who improve.

Since our data analysis focused on defining improvement based on the differences between end-of-trial and start-of-trial scores, we recommend that patients be evaluated as improved or not improved based on their scores at trial's end (or at the time they drop out) compared with entry scores.

Until now, improvement criteria have often relied on changes in joint count to determine whether a patient has improved. Compared with more comprehensive measures, definitions that depend only on joint count generally do not discriminate as well between active drug–treated and placebo-treated patients, and usually identify more placebo-treated patients as being improved. We hope that our definition of improvement satisfies a middle ground in that it relies heavily on joint count improvement while incorporating data from other measures.

There are limitations both to our approach to defining improvement and to our definition. First, our analysis of how well improvement definitions distinguished active drug–treated from placebo-treated patients was limited by the absence of functional status data in our data sets. We had to rely on grip strength instead. Analyses with smaller data sets that did contain functional status suggest that the results would have been similar. Nonetheless, it is essential that these improvement criteria be validated with data sets that contain information on functional status change. In general, validation in other prospectively measured data sets would be of great value.

In addition, the use of one single measure to evaluate the response to therapy in rheumatoid arthritis may be overly simplistic. Some treatments affect joint count improvement more than improvement in acute-phase reactants, and others do the opposite. To ignore the spectrum of improvement induced by a particular treatment would be a mistake, and we recommend that the change in each outcome still be reported, but that the primary outcome for trials be improvement as reported here.

In summary, we suggest a definition for improvement in rheumatoid arthritis that corresponds closely to rheumatologists' own impressions of patient improvement and also discriminates between active drug– and placebo-treated patients, which suggests that its use will enhance the statistical power of future trials.

ACKNOWLEDGMENT

The authors are indebted to members of the ACR Committee on Health Care Research for their critical comments.

REFERENCES

1. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, Furst D, Goldsmith C, Kieszak S, Lightfoot R, Paulus H, Tugwell P, Weinblatt M, Widmark R, Williams HJ, Wolfe F: The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 36:729–740, 1993
2. Pinals RS, Masi AT, Larsen RA, and the Subcommittee for Criteria of Remission in Rheumatoid Arthritis of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee: Preliminary criteria for clinical remission in rheumatoid arthritis. Arthritis Rheum 24:1308–1315, 1981
3. Paulus HE, Egger MJ, Ward JR, Williams HJ, and the Cooperative Systematic Studies of the Rheumatic Diseases Group: Analysis of improvement in individual rheumatoid arthritis patients treated with disease-modifying antirheumatic drugs, based on the findings in patients treated with placebo. Arthritis Rheum 33:477–484, 1990

4. Anderson JJ, Felson DT, Meenan RF, Williams HJ: Which traditional measures should be used in rheumatoid arthritis clinical trials? Arthritis Rheum 32:1093–1099, 1989
5. Goldsmith CH, Boers M, Bombardier C, Tugwell P: Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. J Rheumatol 20:561–565, 1993
6. Van der Heijde DMFM, Van't Hof MA, Van Riel PLCM, Theunisse Lam, Lubberts EW, van Leeuwen MA, van Rijswijk MH, Van de Putte LBA: Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. Ann Rheum Dis 49:916–920, 1990
7. Ward JR, Williams HJ, Egger MJ, Reading JC, Boyce E, Altz-Smith M, Samuelson CO Jr, Willkens RF, Solsky MA, Hayes SP, Blocka KL, Weinstein A, Meenan RF, Guttadauria M, Kaplan SB, Klippel J: Comparison of auranofin, gold sodium thiomalate, and placebo in the treatment of rheumatoid arthritis: a controlled clinical trial. Arthritis Rheum 26:1303–1315, 1983
8. Weinblatt ME, Coblyn JS, Fox DA, Fraser PA, Holdsworth DE, Glass DN, Trentham DE: Efficacy of low-dose methotrexate in rheumatoid arthritis. N Engl J Med 312:818–822, 1985
9. Furst D, Koehnke R, Burmeister LF, Kohler J, Cargill I: Increasing methotrexate effect with increasing dose in the treatment of resistant rheumatoid arthritis. J Rheumatol 16:313–320, 1989
10. Williams HJ, Wilkens RF, Samuelson CO Jr., Alarcón GS, Guttadauria M, Yarboro C, Polisson RP, Weiner SR, Luggen ME, Billingsley LM, Dahl SL, Egger MJ, Reading JC, Ward JR: Comparison of low-dose oral pulse methotrexate and placebo in the treatment of rheumatoid arthritis: a controlled clinical trial. Arthritis Rheum 28:721–730, 1985
11. Felson DT, Anderson JJ, Meenan RF: The comparative efficacy and toxicity of second-line drugs in rheumatoid arthritis: results of two metaanalyses. Arthritis Rheum 33:1449–1461, 1990
12. Weinblatt ME, Kaplan H, Germain BF, Merriman RC, Solomon SD, Wall B, Anderson L, Block S, Irby R, Wolfe F, Gall E, Torretti D, Biundo J, Small R, Coblyn J, Polisson R: Low-dose methotrexate compared with auranofin in adult rheumatoid arthritis: a thirty-six–week, double-blind trial. Arthritis Rheum 33:330–338, 1990
13. Van der Heide A, Jacobs JWG, Van Albada-Kuipers GA, Kraaimaat FW, Geenen R, Bijlsma JWJ: Physical disability and psychological well being in recent onset rheumatoid arthritis. J Rheumatol 21:28–32, 1994
14. Fuchs HA, Pincus T: Reduced joint counts in controlled clinical trials in rheumatoid arthritis. Arthritis Rheum 37:470–475, 1994
15. American College of Rheumatology Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials: Reduced joint counts in rheumatoid arthritis clinical trials. Arthritis Rheum 37:463–464, 1994