
Arthritis & Rheumatism

Official Journal of the American College of Rheumatology

THE AMERICAN COLLEGE OF RHEUMATOLOGY PRELIMINARY CORE SET OF DISEASE ACTIVITY MEASURES FOR RHEUMATOID ARTHRITIS CLINICAL TRIALS

DAVID T. FELSON, JENNIFER J. ANDERSON, MAARTEN BOERS, CLAIRE BOMBARDIER,
MIRIAM CHERNOFF, BRUCE FRIED, DANIEL FURST, CHARLES GOLDSMITH,
STEPHANIE KIESZAK, ROBERT LIGHTFOOT, HAROLD PAULUS, PETER TUGWELL,
MICHAEL WEINBLATT, RUDOLPH WIDMARK, H. JAMES WILLIAMS, and FREDERICK WOLFE

Objective. To develop a set of disease activity measures for use in rheumatoid arthritis (RA) clinical trials, as well as to recommend specific methods for assessing each outcome measure. This is not intended to be a restrictive list, but rather, a core set of measures that should be included in all trials.

Methods. We evaluated disease activity measures commonly used in RA trials, to determine which measures best met each of 5 types of validity: construct,

The active participation of Dr. Widmark in the development of this core set of outcome measures does not signify that the FDA has altered its primary efficacy outcomes for rheumatoid arthritis trials.

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

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

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

Submitted for publication July 27, 1992; accepted in revised form January 4, 1993.

face, content, criterion, and discriminant. The evaluation consisted of an initial structured review of the literature on the validity of measures, with an analysis of data obtained from clinical trials to fill in gaps in this literature. A committee of experts in clinical trials, health services research, and biostatistics reviewed the validity data. A nominal group process method was used to reach consensus on a core set of disease activity measures. This set was then reviewed and finalized at an international conference on outcome measures for RA clinical trials. The committee also selected specific ways to assess each outcome.

Results. The core set of disease activity measures consists of a tender joint count, swollen joint count, patient's assessment of pain, patient's and physician's global assessments of disease activity, patient's assessment of physical function, and laboratory evaluation of 1 acute-phase reactant. Together, these measures sample the broad range of improvement in RA (have content validity), and all are at least moderately sensitive to change (have discriminant validity). Many of them predict other important long-term outcomes in RA, including physical disability, radiographic damage, and death. Other disease activity measures frequently used in clinical trials were not chosen for any one of several reasons, including insensitivity to change or duplication of information provided by one of the core measures (e.g., tender joint score and tender joint count). The committee also proposes specific ways of measuring each outcome.

Conclusion. We propose a core set of outcome measures for RA clinical trials. We hope this will

decrease the number of outcomes assessed and standardize outcomes assessments. Further, we hope that these measures will be found useful in long-term studies.

Clinical trials are our main source of knowledge about the comparative efficacy of drugs used to treat rheumatoid arthritis (RA). In RA trials, at least 10 measures of disease activity are usually assessed, often, many more. Since changes in disease activity (e.g., tender joint count, erythrocyte sedimentation rate [ESR]) reflect patient outcomes, disease activity measures are often called "outcome measures," or "end points." While the Cooperative Systematic Studies of Rheumatic Disease (CSSRD) and other groups that conduct multicenter trials have provided sets of disease activity measures that are used widely by other researchers, there is still a disconcerting heterogeneity in the measures assessed in RA trials. The US Food and Drug Administration's Center for Drug Evaluation and Research has also provided guidelines, suggesting 4 primary disease activity measures in RA clinical trials: swollen joint count, tender joint count, and physician's and patient's global assessments. It is surprising, however, that most clinical trials do not even incorporate all of these 4 variables (1), yet still assess multiple outcome measures.

Why is the reliance on multiple outcome measures and, moreover, the use of various sets of measures in different RA trials a problem? First, given the plethora of therapies available and under development, it is highly unlikely that all therapies will be directly compared. The use of dissimilar outcome measures to assess drugs in different trials makes it impossible to judge therapies against a common standard. Second, since statistical significance often arises by chance when multiple statistical tests are performed, the use of many outcome measures makes it likely that any therapy will be statistically significantly efficacious on some measure. Third, the absence of a core set of outcome measures allows the selection of which measures to report, and this can result in reporting only those outcomes which showed impressive results. Finally, clinical trials continue to include outcomes that are insensitive to change, even in the presence of a demonstrated treatment effect.

Our goal was to develop a specific set of disease activity measures that should be used to assess outcome in all RA clinical trials, regardless of the therapy being evaluated, whether a nonsteroidal antiinflammatory drug (NSAID), a second-line drug, a new bio-

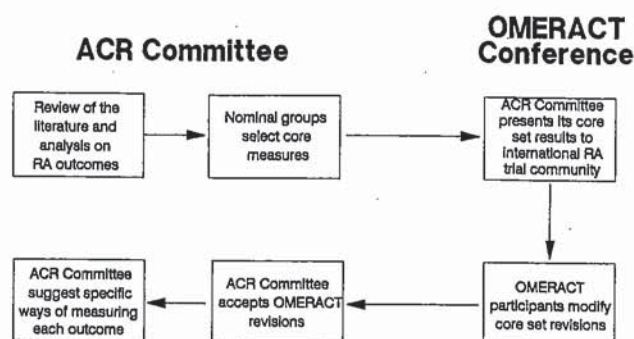


Figure 1. Process for selection of core measures of rheumatoid arthritis (RA) disease activity. ACR = American College of Rheumatology; OMERACT = Outcome Measures in Rheumatoid Arthritis Clinical Trials.

logic agent, or even a nonpharmacologic treatment. Our intent was not that the set of outcome measures would serve as a restrictive list, but rather, as a list of core measures that would be included in all trials.

Our choice of outcome measures was influenced by such complex factors as whether each measure is sensitive to change, whether each measure predicts an important end point in RA (e.g., disability), and whether, as a set, the measures are comprehensive and nonredundant. Each of these issues is, to some degree, addressable by analysis of patient data; however, critical judgment is required for weighing together the issues that affect the selection of outcomes to be measured. We therefore used a method whereby the committee critically reviewed evidence about the issues before attempting to arrive at a consensus. This technique has been shown to improve agreement among participants in consensus meetings (2). We reviewed pertinent literature and analyses prior to the meetings of the nominal groups, so as to inform them about each of the specific issues affecting the overall decision. To achieve consensus, we used a formal nominal group process, which provides an orderly procedure for obtaining information from target groups who are most closely associated with a problem area.

The initial meeting of the American College of Rheumatology (ACR) Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials was followed by an international meeting (Outcome Measures in Rheumatoid Arthritis Clinical Trials [OMERACT]) at which the set of core measures was slightly revised, a revision the ACR committee accepted by consensus. The selection process is illustrated in Figure 1.

Our work represents an advance over previous

attempts by individual investigators (3–7), including some of our group, to devise a set of outcome measures for RA clinical trials. First, rather than focusing on selected aspects of the outcome measure's performance, such as sensitivity to change (3,6,7) or concurrent validity (4), we tried to evaluate all issues that may be relevant. For example, the ability of a measure to predict important end points in RA (e.g., disability, death) has not, to our knowledge, been previously considered when selecting such measures. Furthermore, rather than a purely statistical (3,4,6,7) or a purely consensual (5) approach, we combined the two, recognizing that the selection of measures requires both evidence and judgment.

We present here the data-driven consensus process we followed. Until now, ACR diagnostic and classification studies have entailed testing of clinical diagnostic impressions in a large group of patients, with subsequent analysis of data to determine which diagnostic tests best differentiate patients who have the disease in question from those who do not. The goals and methods of our committee were different. Because the selection of measures for the core set depended on multiple qualities of these measures (e.g., their individual sensitivity to change, their predictiveness of mortality, and their lack of redundancy), we could not perform an analysis of all of these qualities simultaneously in a large data set. Furthermore, analysis of a single data set would not have been as generalizable as marshalling the results of all previous studies, including those containing large data sets.

Some issues, such as predictors of mortality, would not have been easy to address in any clinical trial data. Therefore, after reviewing data on the validity of outcome measures, we used consensus methods similar to those used to develop practice guidelines for hemodynamic monitoring (8), appropriateness criteria for postponing surgery (9), and evaluation outcomes for total hip arthroplasty (10) to arrive at consensus on a core set of measures. While space does not permit a complete review of all the data, we attempt to present sufficient information for the reader to understand what data the committee evaluated and how. In the Discussion section, we delineate the specific reasons for each selection.

METHODS

We followed a framework for the selection of clinical trial indices proposed by Tugwell and Bombardier (11), who suggested several criteria by which to choose end points for

Table 1. Methodologic framework for selection of disease activity measures in rheumatoid arthritis (RA) clinical trials*

Type of validity	Meaning
1. Construct validity	Do the results using these outcome measures change in proportion to clinical change?
2. Face validity	Credibility: Are the outcomes sensible?
3. Content validity	Comprehensiveness: Do the outcomes sample multiple domains of improvement in RA?
4. Criterion validity	Do the outcomes predict or correlate with "gold standard" measures of RA outcomes (e.g., death, disability, radiographic evidence of damage)?
5. Discriminant validity	Sensitivity to change: Do the outcomes detect the smallest clinically important improvement?

* Adapted, with permission, from *Journal of Rheumatology* (11).

clinical trials (see Table 1). For criterion validity, we chose 3 "gold standards": death, physical disability, and radiologic evidence of joint damage. To have criterion validity, an outcome measure should correlate with or predict 1 or more of these 3. Separate from the 5 validity criteria, we focused on whether selected outcome measures had high reliability and whether certain measures duplicated others, and would therefore suggest that both would not need to be assessed.

To capture all studies addressing the validity of RA outcome measures, we conducted MEDLINE searches, bibliographic reviews, and sought additional sources from experts in the field. For some issues, especially content validity and redundancy of outcome measures, there was scant literature. For these, we reviewed factor analyses of CSSRD second-line drug trials (3) and analyzed 2 large NSAID trials in RA (12).

We then held a 2-day meeting (attended by all coauthors), at which we evaluated the structured literature review of RA outcome measures and examined additional data analyses that filled in gaps in the literature. Here, because of space constraints, we present a distilled review of the evidence on validity. For criterion validity, the reader is referred to several recent reviews (refs. 13–15) and for discriminant validity, other reviews (refs. 1,3,16).

After reviewing data on the validity, reliability, and redundancy of outcome measures, the group broke into 2 equal-sized subgroups, each including experts in RA clinical trials, health services and epidemiology, and biostatistics. We followed guidelines for nominal group processes developed by Delbecq et al (17), making decisions based on literature review, data analyses, and discussion. The reliability and validity of this technique have been established through repeat evaluations using different groups (18–20) and through applying the recommendations arrived at by this

process to subsequent research (8). Each group was facilitated by an individual with content expertise and assisted by an individual with nominal group technique expertise.

The written question presented to each group member was, "Which outcome measures should be used as core outcomes in RA trials?" Participants were instructed to write down as many as they wished to recommend. In the manner of a "round robin," the leader of the nominal group recorded the outcome measures specified by each participant. Equivalent disease activity measures were combined, and the definition of each was clarified. Next, each group ranked the various outcome measures, giving the highest scores to the items they most believed should be in RA trials. The leaders then summed the scores and tallied the number of times each outcome measure was ranked. The lowest-ranking measures were dropped from the list. Each measure on the residual list was discussed to clarify reasons for disagreements in ranking, and the residual-list measures were then ranked.

Ultimately, each group arrived at a list of core measures recommended for use in all trials. The 2 groups then met together to present their results. Further discussion ensued, and consensus was reached.

The ACR committee presented the results to an international conference on RA clinical trial outcome measures, OMERACT, held in Maastricht, The Netherlands, in April and May, 1992. The ACR committee later reconvened to discuss the recommendations from the OMERACT conference and to reach consensus on the final list of measures. The ACR committee then decided how each of the final disease activity measures should be assessed. Further literature review and data analyses which focused on the sensitivity to change of each different assessment technique were considered.

RESULTS

Review of the validity of RA disease activity measures. *Construct, face, and content validity.* Outcome measures in rheumatoid arthritis should have construct validity; that is, disease activity measures should change in proportion to clinical change. In addition, outcome measures in clinical trials ought to have face validity; that is, they should seem sensible. Disease activity measures currently in use generally have face validity and construct validity. The set of outcome measures should also have content validity, meaning that it should sample the broad range of clinical change in RA. Some subsets of patients may experience clinical improvement in some measures of disease activity (e.g., pain), yet not in others (e.g., swollen joints). A core set of outcome measures ought to be broad enough to capture all patients who improve.

Criterion validity. Criterion validity is the ability of a disease activity measure to predict an impor-

tant aspect of the patient's clinical status, especially a long-term outcome of importance. In the short course of a clinical trial, neither death nor the development of radiographic evidence of joint damage occurs frequently enough to be useful as an outcome measure. We reviewed the association of outcome measures with 3 important long-term end points in RA: disability, death, and radiographic damage. The following is a brief summary of these issues.

A recent review of risk factors for disability in RA (13) suggests that prior disability is a strong predictor, and that most clinical trial outcome measures are correlated with either concurrent or subsequent disability. Spiegel et al (21) found that walk time and grip strength correlated, respectively, with concurrent lower and upper extremity disability. Van der Heijde and coworkers (14) found that most conventionally assessed outcome measures in RA trials, including pain, tender and swollen joints, morning stiffness, ESR, and grip strength, correlated with physical disability. Wolfe and Cathey (22) reported that initial levels of physical function, physician's global assessment, duration of morning stiffness, and grip strength, but not tender joint count, were significant predictors of the development of physical disability in RA patients.

Some disease activity measures predict the likelihood of overall premature death better than others. Many factors associated with increased mortality among RA patients, such as extraarticular manifestations and seropositivity (15), are not clinical trial outcome measures. Of potential trial end points, Pincus et al (23) found that certain measures of functional capacity were strong predictors of prolonged survival or premature mortality, a finding that has been confirmed by other investigators (15,24,25). Steinbrocker functional class (25-27) and patient's self-report of functional status, using instruments such as the Health Assessment Questionnaire (HAQ) (28,29) and the Arthritis Impact Measurement Scales (AIMS) (24), have been used in several studies. In the largest prospective study of mortality in RA, evaluating more than 3,500 patients, Wolfe et al (30) confirmed the findings of smaller studies, reporting that, of disease activity measures, physical functional capacity was the strongest predictor of mortality. As for other clinical trial outcome measures, all of those assessed, including tender joint count, ESR, radiographic abnormalities, grip strength, and morning stiffness, were significantly correlated with risk of premature mortality. In general, these measures were significant univariate predictors

Table 2. Which clinical end points are associated with radiographic evidence of change?*

Author, year (ref.)	Correlated with radiographic change	Not correlated with radiographic change	Longitudinal study
Scott et al, 1985 (32)	ESR, CRP	Ritchie articular index (a tender joint score)	Yes
Mottonen, 1988 (33)	Number of clinically active joints, swollen joints, ESR, hemoglobin, morning stiffness, pain, grip strength	–	Yes
Fuchs et al, 1988 (34); Pincus et al, 1989 (35)	Joint deformity; swollen joints, functional status	Tender joints	No
Van der Heijde et al, 1992 (36)	Swollen joints, CRP, ESR, grip strength†	Tender joints, pain, morning stiffness‡	Yes

* ESR = erythrocyte sedimentation rate; CRP = C-reactive protein.

† Correlation with subsequent radiograph >0.30.

‡ Correlation with subsequent radiograph <0.20.

of mortality; after adjusting for functional capacity, however, most were no longer predictors. Pincus et al (31) also reported that, although the tender joint count was significantly associated with mortality risk as a univariate measure, it was not a statistically significant predictor when functional capacity was already in the model.

In reviewing factors associated with radiographic changes in the hand (or foot) or predictive of subsequent radiographic changes (see Table 2), 3 studies reported that swollen joint counts were correlated with radiographic abnormalities (33–36). Two of these studies (34–36) (one longitudinal and one cross-sectional) found that the swollen joint count was a much stronger predictor of radiographic damage than was the tender joint count. Two studies, both longitudinal (32,36), suggested that the ESR predicted the development of radiographic damage better than did a count or a score of tender joints. Mottonen (33) also commented on the predictive value of the ESR, and Dawes et al (37) found that treatment that controls levels of acute-phase reactants reduces radiographic progression. When evaluated, low grip strength values also foreshadowed radiographic abnormalities.

Discriminant validity. An outcome measure is sensitive to change (discriminant validity) if it can detect important clinical changes in disease activity over time. Sensitivity to change is a critical quality for RA trial outcome measures, for several reasons. Without sensitive measures of disease activity, important clinical change may go undetected. Choosing such measures in preference to less sensitive ones markedly improves statistical power, thus reducing the number of subjects required in a study. In fact, a doubling of

sensitivity to change (measured as efficiency, which equals the mean change divided by its standard deviation) reduces the number of subjects needed by 2.5–4 times. Last, measures with poor test–retest reliability are less likely to be sensitive to change because of the “noise” in their measurement.

Several recent studies (1,3,38) (see Table 3) have evaluated which outcomes are sensitive to change in RA clinical trials, comparing the change in treatment groups with the change in the control group. While each study focused on a different group of trials and even different outcomes, some common findings emerge. First, the end points most sensitive to change include global assessments by physician and patient. Second, the patient’s assessment of pain, the tender joint count, and, less so, the swollen joint count, are moderately sensitive to change. Third, in trials of second-line drugs, the ESR is sensitive to change, but that is not the case in trials of NSAIDs. Fourth, walk time, grip strength, morning stiffness, and proximal interphalangeal joint circumference are all relatively insensitive to change. Self-reported functional status was moderately sensitive to change in at least 3 trials (38–40). Studies from clinical practice have corroborated these results. In a 6-month followup study of patients attending an RA clinic, Hawley and Wolfe (41) showed that the end points most sensitive to change were pain, global assessment, tender joint count, and functional status as assessed on the HAQ instrument. Grip strength and morning stiffness were relatively insensitive to change. Also in a study of patients in a practice setting who were starting second-line drug therapy, Dixon et al (7) suggested that the Ritchie articular index (a tender joint score), patient’s

Table 3. Sensitivity to change (most sensitive to least sensitive) of disease activity measures in RA clinical trials*

Author, year (ref.); type of study		
Anderson et al, 1989 (3); 3 clinical trials of DMARDs (CSSRD)	Bombardier et al, 1991 (38); 1 large trial of auranofin†	Gotzsche, 1990 (1); 130 RA NSAID trials (metaanalysis)
Tender joint count	Patient's global assessment	Patient's global assessment
ESR	Overall health (6 days)	Pain
Swollen joint count	Keitel index	Tender joint count
Physician's global assessment	ESR	Ritchie index
Tender joint score	Pain (10-cm line) and	Morning stiffness
Platelet count	Tender joint count	Swollen joint count
Swollen joint score	Swollen joint count and	Grip strength
Grip strength	Quality of Well Being	Walk time
Patient's global assessment	Grip strength and	ESR
Pain (1-5 scale)	Rand current health	Digital joint size
Morning stiffness	Walk time	
Hemoglobin	Morning stiffness	
Functional class		
PIP circumference		
Walk time		
Measures not evaluated		
Functional status measures	Joint scores	Functional status measures
Radiograph	Laboratory other than ESR	Radiograph
	Physician's global assessment	
	Radiograph	

* Measures are listed in descending order of sensitivity to change. RA = rheumatoid arthritis; DMARDs = disease-modifying antirheumatic drugs; CSSRD = Cooperative Systematic Studies of the Rheumatic Diseases group; NSAID = nonsteroidal antiinflammatory drug; ESR = erythrocyte sedimentation rate; PIP = proximal interphalangeal joint.

† Multiple measures of pain and health status were used. Only those most sensitive to change are listed.

global assessment, and the ESR were sensitive to change over 24 weeks, whereas grip strength and digital joint circumference were not.

Using CSSRD trial data (3) and new data analysis from 2 RA NSAID trials (12), we confirmed these general findings and explored whether outcome measures were equally sensitive to change when assessing different subgroups of patients. The same end points, as noted above, were sensitive to change in those with early disease (≤ 3 years) compared with those without early disease (> 3 years). Among patients with low and moderate initial tender joint counts (≤ 20 tender joints), the tender joint count measures were less sensitive to change than in patients with higher initial joint counts. However, in both those with low and those with high initial tender joint counts, patient's report of pain, the swollen joint count and score, and patient's and physician's global assessments demonstrated sensitivity.

Other methodologic issues: reliability and redundancy. High test-retest reliability has been demonstrated for most RA outcome measures, including patient-assessed pain, global assessment by patient and by physician, physical function, grip strength, walk time, and ESR. Although one study found poor interobserver reliability for tender and swollen joint counts (21), standardization can substantially reduce the interobserver variability of this measure (42,43). Assessment of other disease activity measures is often highly reproducible, especially if standardized beforehand (43). Joint evaluations in which each joint is graded for tenderness often have lower test-retest and interobserver reliability than do simple counts in which each joint is evaluated as either tender or not tender (44). However, a greater change in effect size or relative efficiency does not always occur with joint counts.

Analyses of different data sets produced differ-

ent factor loadings, suggesting no consistent redundancy of outcome measures, except for collinearity of joint counts with their respective scores. For example, changes in the tender joint count and the tender joint score correlate highly, and in factor analyses (3), these 2 end points loaded on the same factor, suggesting they are duplicative. In factor analyses, the tender joint count and swollen joint count also usually appeared in the same factor, suggesting a high correlation between the two. However, an analysis of patients whose tender or swollen joint counts improved by 50% (see Table 4) showed a substantial lack of overlap.

Results of the ACR nominal group meetings and international meeting. Each of the nominal groups chose a set of outcome measures to be used in all RA clinical trials. Reassuringly, each group chose exactly the same outcome measures (items 1–6, Table 5). When the groups rejoined, these outcome measures were selected unanimously as the official preliminary set. One nominal group suggested that imaging studies also be included for long-term trials. After further discussion, imaging was recommended in some circumstances (see Table 5). This set of measures and selected background data were then presented to the OMERACT group (an international conference on RA trial outcome measures sponsored by the International League Against Rheumatism, the ACR, and other regional and national organizations). After discussion and presentation of additional data by European and

Table 5. ACR disease activity measures for rheumatoid arthritis clinical trials: core set*

Disease activity measure
1. Tender joint count
2. Swollen joint count
3. Patient's assessment of pain
4. Patient's global assessment of disease activity
5. Physician's global assessment of disease activity
6. Patient's assessment of physical function
7. Acute-phase reactant value
For trial duration ≥ 1 year and agent being tested as a "DMARD" (see ref. 45), also perform:
8. Radiography or other imaging technique

* See Table 6 for recommended assessment techniques. ACR = American College of Rheumatology; DMARD = disease-modifying antirheumatic drug.

American investigators, OMERACT conferees selected a core set, which in addition to the 6 components noted above, contained a measure of 1 acute-phase reactant. A description of the presentations, discussions, and selection process at the OMERACT meeting will be published separately. The addition of an acute-phase reactant to the core set of measures was later unanimously accepted by the ACR committee. The core set is presented in Table 5.

DISCUSSION

In proposing this core set of outcome measures for rheumatoid arthritis clinical trials, we seek to promote uniformity in the measurement and reporting of outcomes in rheumatoid arthritis trials. Our endeavor was facilitated by the methodologic investigations into the validity, reliability, and sensitivity to change of outcomes in RA, which have been published during the last 5 years.

The set of outcome measures chosen has content validity, in that it samples from a broad range of improvement in RA, from the perspectives of both patient and clinician. It contains measures of symptoms and measures which, independent of symptoms, may reflect the biology of the disease (swollen joints, acute-phase reactant levels). Also, the set consists of individual disease activity measures which are predictive of important outcomes in RA, such as disability, radiographic abnormalities, and overall premature death. Furthermore, each outcome measure is at least moderately sensitive to change. We did not select end points that are not likely to detect clinical improvement.

One review of the sensitivity to change of RA

Table 4. Numbers of RA patients with at least 50% improvement in either the tender joint count or the swollen joint count, according to treatment with NSAIDs or second-line agents*

50% improvement in tender joint count	50% improvement in swollen joint count		
	No	Yes	Total
All patients taking NSAIDs†			
No	128	20	148
Yes	52	113	165
Total	180	133	313
All patients taking second-line drugs‡			
No	156	30	186
Yes	40	73	113
Total	196	103	299

* RA = rheumatoid arthritis; NSAIDs = nonsteroidal antiinflammatory drugs.

† Data derived from additional analyses performed on the 2 RA trials reported by Kolodny (see ref. 12).

‡ Data derived from additional analyses performed on the 3 RA trials reported by Anderson et al (see ref. 3).

outcome measures in trials of disease-modifying antirheumatic drugs which was published after our meetings, corroborates our findings, and suggests that the tender joint count, patient's report of pain and of functional status, and the ESR were the outcome measures most sensitive to change (16). Also, a recently published study of factors predicting the risk of disability in RA has confirmed that, among conventional disease activity measures, functional status is by far the most potent (46).

Some outcome measures were selected for specific reasons. For example, while the tender joint count and swollen joint count are correlated, improvement in one does not necessarily mean improvement in the other. Furthermore, while the tender joint count is generally more sensitive to change than is the swollen joint count, the number of swollen joints correlates better with later radiographic disease (either new abnormalities or progression of existing damage), and may be a better predictor of disability.

Physical disability measures were chosen because of their sensitivity to change and because they themselves are outcomes of importance. Furthermore, in RA, the patient's physical disability status strongly predicts premature mortality and eventual physical disability.

Physician's and patient's global assessments and patient's report of pain were chosen, in part, because of their face validity and, in part, because these measures are among the most sensitive to change among RA outcomes.

An acute-phase reactant should also be measured. Those most frequently measured are the ESR and the C-reactive protein (CRP) level. Although the response of the ESR to different therapies is thought to be inconsistent, it is very sensitive to change among patients treated with second-line drugs, and it is a predictor of radiographic damage. The CRP level is also very sensitive to change in disease activity.

While there were many published studies on criterion and discriminant validity, we found few published data about the redundancy of outcome measures. This was an important issue if our goal was to avoid including 2 measures of essentially the same thing. We analyzed several large data sets from different sources to assess redundancy, but our results (some of which are depicted in Table 4) may not be generalizable to all trials.

While this core set of outcomes should be included in all current trials, new RA outcome measures may become important for future trials. Mea-

surement of the patient's valuation of functional improvement (patient utilities) may be recommended. Psychosocial outcomes and economic costs of therapy, which include both actual costs of treatment and monitoring as well as toxicity costs, should be studied. In addition, reporting of the frequency and severity of medication side effects is critical, as is the description of any deaths that occur during a trial.

Our advocacy of the general use of a uniform set of outcome measures is not intended to divert attention from other essential factors in the design of RA clinical trials. These include such matters as clearly defined trial objectives, selection of appropriate patients, and sufficient duration of the trial to achieve study objectives. We have begun using this core set of disease activity measures to develop criteria for clinical improvement in RA trials and to test combinations of these measures as end points.

In conclusion, we propose a core set of disease activity measures to be used in rheumatoid arthritis clinical trials. We have also proposed specific ways of measuring each outcome. We hope that the wide acceptance of this core set of measures will lead to uniformity of measurement and reporting across rheumatoid arthritis trials.

RECOMMENDATION OF METHOD FOR ASSESSING EACH OUTCOME

The ACR committee reviewed literature and analysis and came to consensus regarding specific ways to measure each outcome (see Table 6). We hope that this will further promote uniformity in trial assessment, but acknowledge that these specific methods may not be widely used in other regions of the world.

While joint scores were not included as core outcomes, it was recommended that investigators in clinical trials assess and grade the tenderness of each joint, but ultimately report only whether each joint is tender or not tender. This encourages the observer to examine the joint more carefully than if only the presence or absence of tenderness is ascertained. The tender joint count correlates highly with the Ritchie articular index (48), which is widely used in Europe. Available information about limited joint counts suggests that they save time, with little, if any, decrease in sensitivity to change (49). If certain limited joint counts prove as capable of detecting change as do the full ACR counts, they may become appropriate substitutes.

Table 6. ACR recommendations of specific ways to assess each disease activity measure in the core set*

Disease activity measure	Method of assessment
1. Tender joint count†	ACR tender joint count (see ref. 47), an assessment of 68 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 (see ref. 47), an assessment of 66 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 (formerly, the American Rheumatism Association); AIMS = Arthritis Impact Measurement Scales; HAQ = Health Assessment Questionnaire; MHIQ = McMaster Health Index Questionnaire; MACTAR = McMaster Toronto Arthritis Patient Preference Disability Questionnaire.

† The 68 joints to be examined for tenderness are: temporomandibular (n = 2), sternoclavicular (n = 2), acromioclavicular (n = 2), shoulder (n = 2), elbow (n = 2), wrist (n = 2), metacarpophalangeal (n = 10), interphalangeal of thumb (n = 2), distal interphalangeal (n = 8), proximal interphalangeal (n = 8), hip (n = 2), knee (n = 2), ankle mortise (n = 2), ankle tarsus (n = 2), metatarsophalangeal (n = 10), interphalangeal of great toe (n = 2), and proximal/distal interphalangeal of the toes (n = 8).

‡ The 66 joints to be examined for swelling are the same as those examined for tenderness, except the hip joints are not included.

For the assessment of the patient's level of pain and the patient's and physician's global assessments of disease activity, visual analog scales and Likert descriptive scales were found to be similarly sensitive to change (50). The committee recommended that change scores or measures of change (e.g., "Have you noticed a change since starting the trial?") not be utilized in clinical trials because their validity and reliability have not yet been fully demonstrated. Furthermore, in a long trial, patients may not accurately recall their status at trial onset. Rather, the committee recommended that current pain and current global assess-

ments (e.g., "How is the patient doing today?") be ascertained both before and after treatment.

Physical function should be assessed by patient self-report. Several instruments, which either focus entirely on physical function (short HAQ, Toronto Functional Capacity Questionnaire, Lee Functional Status Questionnaire, McMaster Toronto Arthritis Patient Preference Disability Questionnaire [MACTAR]) or contain questions about physical function, have been found to be highly valid and reliable. The group recommended only those instruments which have demonstrated sensitivity to change in RA clinical

trials. At present, these instruments are the HAQ (39), the AIMS (40,51,52), the Quality of Well-Being Questionnaire (38), the McMaster Health Index Questionnaire (53,54), the MACTAR (54,55), and the Toronto Functional Capacity Questionnaire (56). For comprehensive functional status measures, the physical function portions of these instruments, and not necessarily the entire instruments, are recommended as core outcome measures. We, of course, recognize that the validity of these separate portions of instruments which assess more than just physical function may differ from the validity of the overall instrument. Of these instruments, the MACTAR (55), which focuses on selected patient-specific functional priorities, may be especially sensitive to change (54), but because it does not necessarily sample functions which may deteriorate, it may best be used to complement conventional functional status instruments. Other instruments that may yet be found to be sensitive to change in RA trials could also be used.

To assess acute-phase reactants, investigators should measure the ESR, using the method of Westergren, or the CRP level.

ACKNOWLEDGMENTS

The authors appreciate the helpful suggestions of Drs. Howard Fuchs (Vanderbilt University, Nashville, TN) and Robert Meenan (Boston University, Boston, MA).

REFERENCES

1. Gotzsche PC: Sensitivity of effect variables in rheumatoid arthritis: a meta-analysis of 130 placebo controlled NSAID trials. *J Clin Epidemiol* 43:1313-1318, 1990
2. Lomas J, Anderson G, Enkin M, Vayda E, Roberts R, MacKinnon B: The role of evidence in the consensus process. *JAMA* 259:3001-3005, 1988
3. 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
4. Bull BS, Farr M, Meyer PJ, Westengard JC, Bacon PA, Stuart J: Efficacy of tests used to monitor rheumatoid arthritis. *Lancet* I:965-967, 1989
5. Scott DL, Spector TD, Pullar T, McConkey B: What should we hope to achieve when treating rheumatoid arthritis? *Ann Rheum Dis* 48:256-261, 1989
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. *Ann Rheum Dis* 49:919-920, 1990
7. Dixon JS, Hayes S, Constable PDL, Bird HA: What are the 'best' measurements for monitoring patients during short-term second-line therapy? *Br J Rheumatol* 27:37-43, 1988
8. Technology Subcommittee of the Working Group of Critical Care: Hemodynamic monitoring: a technology assessment. *Can Med Assoc J* 145:114-121, 1991
9. McKee M, Priest P, Ginzer M, Black N: Can out-of-hours work by junior doctors in obstetrics be reduced? *Br J Obstet Gynaecol* 99:197-202, 1992
10. Liang MH, Katz JN, Phillips C, Sledge C, Cats-Baril W: The total hip arthroplasty outcome evaluation form of the American Academy of Orthopaedic Surgeons: results of a nominal group process. *J Bone Joint Surg [Am]* 73:639-646, 1991
11. Tugwell P, Bombardier C: A methodologic framework for developing and selecting endpoints in clinical trials. *J Rheumatol* 9:758-762, 1982
12. Kolodny AL: Two double blind trials of diclofenac sodium with aspirin and with naproxen in the treatment of patients with rheumatoid arthritis. *J Rheumatol* 15:1205-1211, 1988
13. Pincus T: New concepts in prognosis of rheumatic diseases for the 1990s, *Prognosis in the Rheumatic Diseases*. Edited by N Bellamy. Dordrecht, The Netherlands, Kluwer Academic, 1990
14. Van der Heijde DMFM, van Riel PLCM, van Rijswijk MH, van de Putte LBA: Influence of prognostic features on the final outcome in rheumatoid arthritis: a review of the literature. *Semin Arthritis Rheum* 17:284-292, 1988
15. Wolfe F: Rheumatoid arthritis, *Prognosis in the Rheumatic Diseases*. Edited by N Bellamy. Dordrecht, The Netherlands, Kluwer Academic, 1990
16. Van de Heide A, Jacobs WGJ, Dinant HJ, Bijlsma WJJ: The impact of endpoint measures in rheumatoid arthritis clinical trials. *Semin Arthritis Rheum* 21:287-294, 1992
17. Delbecq AL, van de Ven AH, Gustafson DH: Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Process. Glenview, IL, Scott, Foresman, 1975
18. Herbert TT, Yost EB: A comparison of decision quality under nominal and interacting consensus group format: the case of the structured problem. *Decis Sci* 10:358-370, 1979
19. Stephenson BY, Michaelson LK, Franklin SG: An empirical test of the nominal group technique in state solar energy planning. *Group Organ Stud* 7:320-334, 1982
20. Van de Ven AH, Delbecq AL: The effectiveness of nominal, Delphi, and interacting group decision making process. *Acad Manage J* 17:605-621, 1974
21. Spiegel JS, Paulus HE, Ward NB, Spiegel TM, Leake B, Kane RL: What are we measuring? An examination of walk time and grip strength. *J Rheumatol* 14:80-86, 1987
22. Wolfe F, Cathey MA: The assessment and prediction of

- functional disability in rheumatoid arthritis. *J Rheumatol* 18:1298-1306, 1991
23. Pincus T, Callahan LF, Vaughn WK: Questionnaire, walking time and button test measures of functional capacity as predictive markers for mortality in rheumatoid arthritis. *J Rheumatol* 14:240-251, 1987
 24. Kazis LE, Anderson JJ, Meenan RF: Health status as a predictor of mortality in rheumatoid arthritis: a five-year study. *J Rheumatol* 17:609-613, 1990
 25. Mitchell DM, Spitz PW, Young DY, Bloch DA, McShane DJ, Fries JF: Survival, prognosis, and causes of death in rheumatoid arthritis. *Arthritis Rheum* 29:706-714, 1986
 26. Allebeck P, Ahlbom A, Allander E: Increased mortality among persons with rheumatoid arthritis, but where RA does not appear on death certificate: eleven year follow-up of an epidemiological study. *Scand J Rheumatol* 10:301-306, 1981
 27. Rasker JJ, Cosh JA: Cause and age of death in a prospective study of 100 patients with rheumatoid arthritis. *Ann Rheum Dis* 40:115-120, 1981
 28. Wolfe F, Hawley DJ, Cathey MA: Clinical and health status measures over time: prognosis and outcome assessment in rheumatoid arthritis. *J Rheumatol* 18:1290-1297, 1991
 29. Leigh JP, Fries JF: Mortality predictors among 263 patients with rheumatoid arthritis. *J Rheumatol* 18:1307-1311, 1991
 30. Wolfe F, Kleinheksal SM, Cathey MA, Hawley DHJ, Spitz P: Antecedents of mortality in rheumatoid arthritis: a study of 1405 patients with RA in a community setting (abstract). *Arthritis Rheum* 34 (suppl 9):S48, 1991
 31. Pincus T, Callahan LF, Sale WG, Brooks AL, Payne LE, Vaughn WK: Severe functional declines, work disability, and increased mortality in seventy-five rheumatoid arthritis patients studied over nine years. *Arthritis Rheum* 27:864-872, 1984
 32. Scott DL, Dawes PT, Fowler PD, Grindulis KA, Shadforth M, Bacon PA: Anti-rheumatic drugs and joint damage in rheumatoid arthritis. *Q J Med* 54:49-59, 1985
 33. Mottonen TT: Prediction of erosiveness and rate of development of new erosions in early rheumatoid arthritis. *Ann Rheum Dis* 47:648-653, 1988
 34. Fuchs HA, Callahan LF, Kaye JJ, Brooks RH, Nance EP, Pincus T: Radiographic and joint count findings of the hand in rheumatoid arthritis: related and unrelated findings. *Arthritis Rheum* 31:44-51, 1988
 35. Pincus T, Callahan LF, Brooks RH, Fuchs HA, Olsen NJ, Kaye JJ: Self-report questionnaire scores in rheumatoid arthritis compared with traditional physical, radiographic, and laboratory measures. *Ann Intern Med* 110:259-266, 1989
 36. Van der Heijde DMFM, van 't Hof MA, van Riel PLCM, van Leeuwen MA, van Rijswijk MH, van de Putte LBA: A comparison of validity of single variables and composed indices for measuring disease activity in rheumatoid arthritis. *Ann Rheum Dis* 51:177-181, 1992
 37. Dawes PT, Fowler PD, Clarke S, Fisher J, Lawton A, Shadforth MF: Rheumatoid arthritis: treatment which controls the C-reactive protein and erythrocyte sedimentation rate reduces radiological progression. *Br J Rheumatol* 25:44-49, 1986
 38. Bombardier C, Raboud J, and the Auranofin Cooperating Group: A comparison of health-related quality-of-life measures for rheumatoid arthritis research. *Controlled Clin Trials* 12:243S-256S, 1991
 39. Van der Heijde DMFM, van Riel PLCM, van de Putte LBA: Sensitivity of a Dutch health status questionnaire in a trial comparing hydroxychloroquine vs sulphasalazine. *Scand J Rheumatol* 19 (suppl):407-412, 1990
 40. Meenan RF, Anderson JJ, Kazis LE, Egger MJ, Altz-Smith M, Samuelson CO Jr, Willkens RF, Solsky MA, Hayes SP, Blocka KL, Weinstein A, Guttadauria M, Kaplan SB, Klippel J: Outcome assessment in clinical trials: evidence for the sensitivity of a health status measure. *Arthritis Rheum* 27:1344-1352, 1984
 41. 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* 5:130-136, 1992
 42. Klinkhoff AV, Bellamy N, Bombardier C, Carette S, Chalmers A, Esdaile JM, Goldsmith C, Tugwell P, Smyth HA, Buchanan WW: An experiment in reducing interobserver variability of the examination for joint tenderness. *J Rheumatol* 15:492-494, 1988
 43. Bellamy N, Anastassiades TP, Buchanan WW, Davis P, Lee P, McCain GA, Wells GA, Campbell J: Rheumatoid arthritis antirheumatic drug trials. 1. Effects of standardization procedures on observer dependent outcome measures. *J Rheumatol* 18:1893-1900, 1991
 44. Thompson PW, Hart LE, Goldsmith CH, Spector TD, Bell MJ, Ramsden MF: Comparison of four articular indices for use in clinical trials in rheumatoid arthritis: patient, order and observer variation. *J Rheumatol* 18:661-665, 1991
 45. Paulus HE, Scott DL, Edmonds JP: Classification of antirheumatic drugs. *Arthritis Rheum* 35:364-365, 1992
 46. Leigh JP, Fries JF: Predictors of disability in a longitudinal sample of patients with rheumatoid arthritis. *Ann Rheum Dis* 51:581-587, 1992
 47. The Cooperating Clinics Committee of the American Rheumatism Association: A seven-day variability study of 499 patients with peripheral rheumatoid arthritis. *Arthritis Rheum* 8:302-334, 1965
 48. Ritchie DM, Boyle JA, McInnes JM, Jasani MK, Dalakos TG, Grieverson P, Buchanan WW: Clinical studies with an articular index for the assessment of joint

- tenderness in patients with rheumatoid arthritis. *Q J Med* 37:393-406, 1968
49. Egger MJ, Huth DA, Ward JR, Reading JC, Williams HJ: Reduced joint count indices in the evaluation of rheumatoid arthritis. *Arthritis Rheum* 28:613-619, 1985
 50. Langley GB, Sheppard H: Problems associated with pain measurement in arthritis: comparison of the visual analogue and verbal rating scales. *Clin Exp Rheumatol* 2:231-234, 1984
 51. Anderson JJ, Firschein HE, Meenan RF: Sensitivity of a health status measure to short-term clinical changes in arthritis. *Arthritis Rheum* 32:844-850, 1989
 52. Meenan RF, Gertman PM, Mason JH: Measuring health status in arthritis: the Arthritis Impact Measurement Scales. *Arthritis Rheum* 23:146-152, 1980
 53. Chambers LW, MacDonald LA, Tugwell P, Buchanan WW, Kraag G: The McMaster Health Index Questionnaire as a measure of quality of life for patients with rheumatoid disease. *J Rheumatol* 9:790-784, 1982
 54. Tugwell P, Bombardier C, Buchanan W, Goldsmith C, Grace E, Bennett KJ, Williams HJ, Egger M, Alarcon GS, Guttadauria M, Yarboro C, Polisson RP, Szydlo L, Luggen ME, Billingsley LM, Ward JR, Marks C: Methotrexate in rheumatoid arthritis: impact on quality of life assessed by traditional standard-item and individualized patient preference health status questionnaires. *Arch Intern Med* 150:59-62, 1990
 55. Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Hanna B: The MACTAR patient preference disability questionnaire: an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. *J Rheumatol* 14:446-451, 1987
 56. Bombardier C, Ware J, Russell IJ, Larson M, Chalmers A, Read JL: Auranofin therapy and quality of life in patients with rheumatoid arthritis. *Am J Med* 81:565-578, 1986