Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 57, No. 2, March 15, 2007, pp 193–202
DOI 10.1002/art.22552

SPECIAL ARTICLE

A Proposed Revision to the ACR20: The Hybrid Measure of American College of Rheumatology Response

AMERICAN COLLEGE OF RHEUMATOLOGY COMMITTEE TO REEVALUATE IMPROVEMENT CRITERIA

Objective. **Although use of the American College of Rheumatology 20% improvement criteria (ACR20) has standardized response measurement in rheumatoid arthritis (RA) trials, the ACR20 has been criticized as less sensitive to change than are continuous measures of response, and its threshold for response (≥20%) is thought to be low. Our goal was to redefine response in RA in a manner that 1) corresponds to a clinical impression of response (clinical validity), 2) maximizes sensitivity to change, and 3) allows for calculation of the ACR20 to continue standardization of reporting.**

Methods. **We examined multiple different ways of defining response, including dichotomous definitions (patient improved versus not improved), ordinal definitions (degree of response scored on an ordinal scale), disease activity indexes, continuous definitions, and definitions that were hybrids of continuous and ordinal measures. Candidate definitions included the ACR20, ACR50, ACR70, the Disease Activity Score, the Simplified Disease Activity Index, the ACR-N, the nACR, and the European League Against Rheumatism (EULAR) response. We also tested variations on these approaches. To test clinical validity, we administered a survey involving patients from a previous trial who had various levels of improvement and asked rheumatologists whether and by how much these patients improved. To determine sensitivity to change, we collected data from 11 large multicenter trials of disease-modifying antirheumatic drugs (DMARDs) in RA comprising 3,665 patients (7 anti–tumor necrosis factor α arms, 4 conventional DMARD arms, 2 biologic arms) and ranked candidate definitions of response according to their average *P* value across trials in distinguishing active treatment from placebo or combination therapy versus single-drug therapy.**

Results. **All 135 tested measures had clinical validity based on survey responses, although dichotomous measures did not capture the range of responses (e.g., the ACR20 did not capture the extra clinical improvement between the ACR20 and the ACR50). In trial analyses, continuous measures had the best sensitivity to change. Among the best scoring measures was a hybrid measure that retained information on the ACR20, ACR50, and ACR70 and combined that with the mean percent improvement in core set measures. When comparing 2 treatments, this hybrid measure had an average *P* value much lower than that for the ACR20. If a trial needed 200 patients to have 80% power (2-sided $\alpha = 0.05$) to detect a difference between treatments if it used the ACR20, the same trial would need 108 patients if the hybrid measure were used.**

Conclusion. **We suggest use of a new hybrid measure of RA response that maximizes sensitivity to change, correlates well with rheumatologists' impressions of improvement, and preserves the ACR20.**

KEY WORDS. Rheumatoid arthritis; Trials; Outcome measures.

INTRODUCTION

Before the 1990s, reports of rheumatoid arthritis (RA) trials cited multiple primary outcomes, often with little overlap of measures from trial to trial. This made it nearly impossible to evaluate whether therapies that produced 1 or 2 positive outcomes actually demonstrated treatment effi-

Supported by grants from the American College of Rheumatology and by the NIH (grant AR-47785).

Members of the American College of Rheumatology Committee to Reevaluate Improvement Criteria in Rheumatoid Arthritis are as follows: David Felson, MD, Chair, Daniel Aletaha, MD, Jennifer Anderson, PhD, Joan Bathon, MD, Maarten Boers, MD, Claire Bombardier, MD, Hyon Choi, MD, Maxime Dougados, MD, Dan Furst, MD, Gary Koch, PhD, Robert Landewé, MD, Mike LaValley, PhD, Kaleb Michaud, PhD, Hal Paulus, MD, Theodore Pincus, MD, Jeffrey Siegel, MD, Lee Simon, MD, Josef Smolen, MD, Peter Tugwell, MD, Desiree van der Heijde, MD, Barbara White, MD, Fred Wolfe, MD, and Hui Xie, PhD.

Dr. Bombardier has received consulting fees (less than $10,00 each) from Abbott, Amgen, AstraZeneca, Bayer, Inc., Bristol-Myers Squibb Canada Company, Hoffmann-la Roche, Pfizer, Roche Products, Schering Canada, Solvay Pharma, and Wyeth-Ayerst Research, is a member of the Advisory Board and has received consulting fees (more than $10,000) for Merck & Company, and holds research grants

cacy; it also created a formidable barrier to the comparison of treatments. This situation changed with the development of the American College of Rheumatology (ACR) core set of disease activity measures (1,2). The core set required the inclusion of 7 clinical end points for all RA trials: swollen joint count, tender joint count, physician's assessment of disease activity, patient's assessment of disease activity, patient's assessment of pain, and patient's assessment of physical function, and levels of an acute-phase reactant (either the C-reactive protein [CRP] level or the erythrocyte sedimentation rate [ESR]). The core set measures still constituted 7 outcome measures, without a single primary measure.

In the early 1990s, an ACR committee used the core set to develop a single measure of improvement, the ACR preliminary criteria for improvement in RA (ACR20) (3). An ACR20 response was defined as at least 20% improvement in both the tender joint count and the swollen joint count and at least 20% improvement in 3 of the 5 other core set measures listed above.

Adoption of the ACR20 (and promulgation of a similar definition of improvement by the European League Against Rheumatism [EULAR]) (4) represented a substantial advance in measuring clinical responses in rheumatology. Trials almost immediately became standardized, with most using the ACR20 as their primary outcome measure. Investigators, appropriately or not, began to compare ACR20 response rates between therapies, and response measures were defined for other rheumatic diseases. The ACR20 also became the primary outcome used by the US Food and Drug Administration (FDA) to evaluate new treatments for RA. The ACR20 focused on improvement in individual patients rather than the mean improvement in groups of treated patients, permitting investigation of why the response to treatment of individual patients differed (5).

Widespread use of the ACR20 highlighted flaws that were not noted during its development. First, as therapies improved, 20% improvement seemed like a low bar, and trials began to incorporate other thresholds for improvement such as the ACR50 and the ACR70. Second, the ACR20 was a dichotomous measure of individual response (i.e., response was either present or absent). However, incorporating information about the relative improvement among patients in a trial would not only be more informative, it would also increase the sensitivity to change of the response definition, making it easier to compare treatments. Furthermore, although the ACR20 measure was used widely, it was not used consistently. Some studies required an ACR20 response only at the end of the trial, other studies used maximal improvement at any time during the trial, and still others prespecified an area under the curve approach to the ACR20. The diversity of ways in which the ACR20 was used led to problems comparing results between trials.

Of these concerns, the main one focused on the realization that the ACR20 measure was not as sensitive to change as other, more continuously defined measures of improvement (6). For example, the ACR-N (a continuous outcome defined by the percent improvement in either the tender joint count, the swollen joint count, or 5 other outcomes) was proposed as being an improvement of the ACR20 (7), as was a definition that simply counted the number of core set measures that improved by at least 20%, the nACR. Although each of these alternatives offered advantages and disadvantages to the ACR20, it became obvious that a revision to the ACR20, with attention to its sensitivity to change, was needed. Two emerging changes in RA trials further motivated a modification of the ACR20: 1) increasing difficulty recruiting patients for trials, requiring an outcome measure that would limit sample size requirements, and 2) the availability of many new effective treatments whose comparative efficacy would not be evaluable without a new, more precise measure.

Two other challenges accompanied revision of the ACR20. Because the ACR20 is widely known and has become a standard way of communicating response rates, any new measure should be built on the ACR20 as much as possible. Also, any new measure must correlate with clinicians' impressions of improvement. With all of these sometimes conflicting goals, we set about to reevaluate the definition of response in RA trials.

METHODS

The approach we used to reevaluate improvement criteria was as follows:

1. Identify potential candidate measures of response, trying to be as comprehensive as possible. We considered all commonly used or proposed measures plus variations of these.

from Abbott Laboratories, Schering Canada, Pfizer, and BMS. Dr. Bathon has received consulting fees and/or honoraria (less than $10,000 each) from Wyeth and Centocor, and has contracts with Amgen, Biogen-Idec, Bristol Myers, Roche, and Rinat. Dr. Choi has received grant support from TAP Pharmaceuticals and serves on advisory boards for TAP and Savient Pharmaceuticals. Dr. Simon is a consultant for AAI Pharma, Abbott, Affinergy, Astrazeneca, Abraxxis, Alpha Rx, Avanir, Nuvo/Dimethaid, Neopharm, Novartis, Pfizer, Plx Pharma, Hisamatsu, LAB Pharma, Dr Reddys, Biosense, Cerimon, Leerink Swann, Alimera, Nomura, Luxor, Paraexel, Nitec, Bayer, Combinatoryx, Rigel, Chelsea, Regeneron, Genelabs, Cypress, SNBL, Skyepharma, Solace, Puretchventures, Puretch Development, Whit Mountain Pharma, TAP, Cell Therapeutics, Omeros, Jazz, Schwarz, Proethic, Takeda, Teva, Zydus, Proprius, Savient, Alder, Cure, Cellegy, Chemocentryx, McKesson, Diobex, Sepracor, Purdue, Serono, Coley, Mediimune, Altea, Neuromed, Polymerix, Talagen, and Tigenix. Dr. Tugwell has received consulting fees from Abbott, Almirall, Astra Zeneca, Aventis, Berlex, Biomatrix, Bristol-Myers Squibb, Cadeuceus Group, Centocor, Dimedix, Dimethaid, Eli Lilly, Glaxo-Welcome, Glaxo Smith Kline, Hoechst Marion Roussel, Innovus, Johnson & Johnson, Lilly Research, Medicine Group, Medicus, Merck, Merck Frosst, Novartis, Novopharm, Ortho-McNeil, Pennside, Roche, Sandoz, Scios, Searle, and Wyeth Ayerst, and organizes a biannual outcome measures consensus meeting (Outcome Measures in Rheumatology [OMERACT]), which has received financial support from numerous pharmaceutical companies. Dr. White is a past employee of Amgen, owns stock in Amgen, and has stock options in Medimmune.

The American College of Rheumatology is an independent, professional, medical, and scientific society which does not guarantee, warrant, or endorse any commercial product or service.

Address correspondence to David Felson, MD, A203, 80 E. Concord Street, Boston University School of Medicine, Boston, MA 02118. E-mail: dfelson@bu.edu.

Submitted for publication August 30, 2006; accepted in revised form September 26, 2006.

2. Determine the clinical construct validity of candidate measures. We created a survey of "paper" patients from a large anti–tumor necrosis factor (anti-TNF) trial and asked rheumatologist respondents which patients in this trial had improved and by how much. We correlated survey responses on the degree of patient improvement with the amount of response determined by candidate measures.
3. Assess the discriminant validity of candidate measures. We collected raw data from a group of large multicenter randomized trials testing second-line drug therapy in RA. We used these data to evaluate the discriminant validity (sensitivity to change) of the candidate measures.
4. Select a set of response criteria using a combination of clinical construct validity, discriminant validity, and consensus by the committee.

Step 1. Identify candidate measure. Our goal in identifying candidate measures was to be as comprehensive as possible. If we did not propose a candidate measure, it could not be ultimately selected. We used several criteria to come up with measures. We identified all currently used or recommended measures of response or disease activity in RA, including widely known measures, such as the ACR20, the ACR50, the Disease Activity Score (DAS), the Simplified Disease Activity Index (SDAI), the ACR-N, and others (Table 1). Also, we allowed for variations on these measures. For example, because some of these measures, such as the ACR20, required improvement in both the tender joint count and the swollen joint count, one variation we tested was to define response without any requirement for joint count improvement. This was called an unweighted ACR20.

We also created candidates of response that counted the number of core set measures improving. One example was the nACR (not the ACR-N), in which, for each patient, the number of core set measures that improved by ≥20% was counted. For a patient in a trial, scores for the nACR range from 0 to 7. We also tested the n2ACR, in which, for each core set measure, a patient received a score of 1 if the measure improved by ≥20% and a score of 2 if the same measure improved by 50% or more. Using the n2ACR, a patient in a trial could have a score ranging from 0 to 14.

We tested continuous measures, including the average percent improvement in all core set measures, with average defined both as the mean and the median. Other variations on this approach were evaluated.

One of the factors that has compromised the sensitivity to change of the ACR20 (8) is the requirement for improvement in the tender/swollen joint count, yet joint count improvement appears to be necessary for rheumatologists to characterize a patient as having improved (9,10). Therefore, we tested other ways of weighting joint count improvement that might not compromise sensitivity. For example, for the nACR (see above), one could doubly weight joint count improvement (i.e., ≥20% improvement in the tender or swollen joint count was counted twice, so that the nACR double-weighted score ranged from 0 to 9). Other ways of weighting joint count improvement were done in the context of new response variables (Table 1).

We created continuous measures, ordinal measures (such as the nACR), measures based on indexes of ACR activity (such as the DAS), and dichotomous measures (such as the ACR20). We also created a hybrid definition of response. A hybrid definition combines elements of 2 approaches, in this case the ACR20/50/70 approach and the continuously measured mean change in core set measures. The value of the hybrid was constrained to be within the limits of the dichotomous measures (e.g., a patient achieving ACR20 improvement but not an ACR50 response would have a hybrid score between 20 and 50. The value between 20 and 50 was determined based on the average improvement in core set measures. If a patient met the ACR50 criteria but not the ACR70 criteria, his or her score was between 50 and 70. A score of 70 or higher was defined similarly.) We tested variations on this hybrid approach, including the unweighted ACR20, ACR50, and ACR70, approaches in which the mean change was allowed to include worsening, and other approaches in which the lower limit of the score (worsening) was limited to −100 to eliminate outliers (improvement is naturally limited to +100).

Ultimately, we tested 135 candidate measures of improvement, which are listed with each of their variations in Table 1. However, presenting numeric results for all candidates involves excessive tabular data presentation. In addition, the variations on a measure performed similarly to the measure itself. Therefore, although Table 1 shows all of the candidate measures tested, we present results only for selected measures of particular interest.

Step 2. Determine clinical construct validity: the survey. To evaluate whether a candidate definition of response in RA correlated with rheumatologists' impressions of whether the patient had improved, we created a survey consisting of profiles drawn from RA patients who were receiving active treatment in a randomized trial. For each patient, we provided the values for core set measures at baseline and after 6 months of therapy (see Figure 1 for an example of a "paper" patient from the survey and the questions posed to respondents). Also, in a column after the baseline and 6-month data, we provided the percent change in each of the core set measures. For each patient, survey respondents were asked whether the patient had improved, and if respondents determined that a patient had improved, they were asked whether it was a minimal improvement, major improvement, or whether the patient's disease had gone into remission.

The paper patients chosen for the survey were selected to test the range of improvements in core set measures that might characterize different levels of clinical improvement. We selected trial patients who met criteria for ACR 0%, 10%, 20% . . . 100% improvement and varied which core set measures improved.

To examine clinical construct validity, we used rheumatologist responses to create a 4-level responder for each paper patient: no improvement, definite but not major improvement, major improvement, and remission. Using Spearman's rank correlation coefficients, we evaluated the relationship of this ordinal measure of response with the degree of response derived from each candidate measure.

Step 3. Assess the discriminant validity of candidate measures: analysis of RA trials. We collected data from 11 large, multicenter, randomized RA trials carried out

Table 1. Candidate measures of response that were evaluated

Variable name	Variable definition
O'Brien test	Nonparametric test that chooses the optimal weight for each core set measure in each trial based on the variability of the measure's score during the trial (a gold standard, not suitable for a candidate measure, but a good benchmark for comparison)
O'Brien.abs	Sum of ranks of absolute changes (ac) in all 7 core measures, where ac = final − baseline for each measure
O'Brien%	Sum of ranks of percentage changes (pc) in all 7 core measures, where for each measure pc = final − baseline/baseline
O'Brien.symm%	Sum of ranks of symmetric percentage changes (spc) in all 7 core measures, where for each measure spc = final − baseline/final + baseline
Continuous definitions of improvement*	
Percent change measures (pc)	
MeanACR	Mean % improvement in the American College of Rheumatology (ACR) 7 core set measures
ACR-N	Minimum (% tender joint count [TJC], % swollen joint count [SJC], median of other 5)
Mean3ACR	Mean (% TJC, % SJC, mean of % for other 5)
MedianACR	Median of % changes in 7 core set measures
FifthACR	Fifth highest % change
ThirdACR	Third highest % change
MeanACR2	Mean of change is all core set measures/mean of all core set measures at baseline
MeanACR3	Mean of the change in each core set measure/square root of measure score at baseline
MedianACR2	Median of change is all core set measures/median of all core set measures at baseline
MeanACR_bd	Bounded version of MeanACR, restricting worsening scores for each of 7% changes to be −100%
Symmetric percentage change measures where spc = final value − baseline value/final value + baseline value	
Meanspc	Mean symmetric % improvement
Wmeanspc	Mean of 7 with symmetric % TJC and symmetric % SJC double weighted
Medianspc	Median of 7 symmetric % changes
Disease activity indexes	
%DAS	Percentage change on DAS (Disease Activity Score)
DDAS	Absolute change on DAS
SpcDAS	Symmetric percentage change of DAS
%SDAI	Percentage change on SDAI (Simplified Disease Activity Index, Smolen)
DSDAI	Absolute change on SDAI
ODAI	For each core set measure, create ordinal scale 1,2,3,4 and average them
%ODAI	Percentage change in ODAI
DODAI	Absolute change in ODAI
EULAR	European League Against Rheumatism (EULAR) response criteria, where 0 = no response, 1 = moderate response, 2 = good response
OSDAI	SDAI cut at 10, 20 (ordinal measure 0, 1, 2 values)
DCDAI	SDAI without C-reactive protein (CRP) absolute change
%CDAI	Percentage change in CDAI (SDAI without CRP)
Pcmean	(Mean of core set measures score at end − mean of measures at start)/(mean of measures at start)
Acmean	(Mean of core set measures score at end − mean of measures at start)
Spcmean	(Mean of core set measures score at end − mean of measures at start)/(mean of measures at start + mean of measures at end)
Pcmedian	Similar to pc mean but with median replacing mean everywhere
Acmedian	Similar to ac mean but with median replacing mean everywhere
Spcmedian	Similar to spc mean but with median replacing mean everywhere
Count definitions of improvement	
nACR	nACR (number of core set measures improved by ≥20%)
nACRw	Weighted by TJC and SJC, if TJC and SJC ≥20%, nACRw = nACR, else nACRw = 0
nACRdw	Double weight of TJC and SJC
nACR50, w50, dw50	Similar as nACR-type measurements, but threshold for improvement increased to 50%
nACR70, w70, dw70	Similar as above, but threshold for improvement increased to 70%

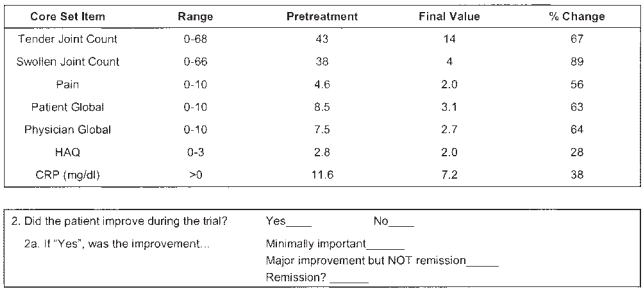
(continued)

Table 1. Candidate measures of response that were evaluated *(Continued)*

Variable name	Variable definition
n2ACR	nACR + nACR50
n2ACRw	nACRw + nACRw50
n2ACRdw	nACRdw + nACRdw50
n3ACR, w, dw	Similar as n2ACR-type measurements, but is a sum of nACR + nACR50 + nACR70 (or its variants)
Ordinal definitions of improvement	
ACRstep	0 if ACR20 = 0, 1 if ACR20 but not ACR50, 2 if ACR50 but not ACR70, 3 if ACR70
ACRuwstep	0 if not ACRuw20, 1 if ACRuw20 but not ACRuw50, 2 if ACRuw50 but not ACRuw70, 3 if ACRuw70
ACRd6wstep	0 if ACRd6w20 = 0, 1 if ACRd6w20 but not ACRd6w50, 2 if ACRd6w50 but not ACRd6w70, 3 if ACRd6w70
ACRd7wstep	0 if ACRd7w20 = 0, 1 if ACRd7w20 but not ACRd7w50, 2 if ACRd7w50 but not ACRd7w70, 3 if ACRd7w70
ACRstep3	Similar to ACRstep, but only with 20, 50 (as few achieve 70%)
ACRuwstep3	No 70%
ACRd6wstep3	No 70%
ACRd7wstep3	No 70%
Dichotomous measures of improvement	
ACR20, 30, 40, 50, 60, 70	The usual ACR20. Also tested ACR 30, 40, etc. each as dichotomous
ACRUW20, 30, 40, 50, 60, 70	ACR unweighted
ACRd6w20, 50, 70	ACR with joint counts double weighted (6 out of 9 = response)
ACRd7w20, 50, 70	ACR with joint counts double weighted (7 out of 9 = response)
MeanACR20, 30, 50	MeanACR ≥20%, 30%, 50%
SumACR-N4, 5, 6	SumACR-N ≥4, 5, 6
MedianACR20, 30, 40, 50, 60, 70	MedianACR ≥20%, 30%, 40%, 50%, 60%, 70%
MeanwACR20, 30, 40, 50, 60, 70	MeanwACR ≥20%, 30%, 40%, 50%, 60%, 70%
FifthACR20, 30, 50	FifthACR ≥20%, 30%, 50%
ThirdACR20, 30, 50	ThirdACR ≥20%, 30%, 50%
DASimp	DAS improvement (dichotomous)
LDAS1	Low disease activity based on DAS28
LDAS2	Low disease activity based on core set measurements
Hybrid definitions	
ACRhybrid	Hybrid definition: the cutoff is 20, 50, 70 based on ACR20, 50, 70; minimum (0, meanACR) if ACRstep = 0, maximum (0.2, minimum [0.5, meanACR]) if ACRstep = 1, maximum (0.5, minimum [0.7, meanACR]) if ACRstep = 2, maximum (0.7, meanACR) if ACRstep = 3
bdACRhybrid	Same as ACRhybrid but a bounded version. Replace meanACR with mean ACR_bd everywhere

* Percent change and symmetric percent change measures shown. Also tested were measures determined by absolute and not percent change (e.g., ameanACR = mean of actual changes in each core set measure).

since publication of the ACR20. Individual patient data were provided to investigators of the ACR subcommittee, with the goal of reevaluating the ACR20. Among the trials were placebo-controlled trials of disease-modifying anti-rheumatic drugs (DMARDs), different TNFα inhibitors, and a non–TNF-inhibiting biologic agent. We also obtained data from one comparative trial of a combination of agents versus a single agent. In most but not all trials, only 80% of the patient data were provided, with these patients selected at random; in the other trials, all patient data were provided. Based on agreement with sponsors providing the data, we have not provided identifiable trial information.



Core Set Item	Range	Pretreatment	Final Value	% Change
Tender Joint Count	0-68	43	14	67
Swollen Joint Count	0-66	38	4	89
Pain	0-10	4.6	2.0	56
Patient Global	0-10	8.5	3.1	63
Physician Global	0-10	7.5	2.7	64
HAQ	0-3	2.8	2.0	28
CRP (mg/dl)	>0	11.6	7.2	38

2. Did the patient improve during the trial? Yes____ No____

2a. If "Yes", was the improvement... Minimally important_____
Major improvement but NOT remission_____
Remission? _____

Figure 1. Example of a paper patient from the survey and example of the questions posed to respondents. HAQ = Health Assessment Questionnaire; CRP = C-reactive protein level.

We were prepared to exclude trials with ≥40% loss to followup, but no trial had this much loss. To analyze the data, we conducted intent-to-treat analyses using a last observation carried forward approach, preferentially selecting the 6-month outcome data when available. CRP was used as the acute-phase reactant when available, and the ESR was used if the CRP values were not available. Within each trial, we tested each candidate measure of response, focusing on the *P* value differentiating change

during active treatment versus change during placebo treatment, using Wilcoxon's rank sum test. For comparative trials, our *P* value was for the change in the combination-therapy arm versus that in the monotherapy arm. We defined this *P* value as our measure of discriminant validity.

In addition to the candidate measures of response tested, we added what we considered to be "gold standard" methods of differentiating treatments, variations on the O'Brien test, a nonparametric approach to comparing 2 groups assessed with multiple end points, which maximizes the discrimination between these groups and therefore the statistical difference between them (6). This test does not generate a response rate but compares treatments in terms of the percent improvement in each core set measure in each patient. This is not a measure that could be standardized across trials.

Step 4. Committee evaluation of data. First, an ACR subcommittee (the Committee to Reevaluate Improvement Criteria) and then the parent ACR Subcommittee on Classification and Response Criteria met to review preliminary results of the data analysis and survey. Based on advice from our statistical team, we focused on the leading 10–15 candidates, because it was likely that sensitivity to change dropped off substantially after that. Second, the parent committee and representatives of the FDA raised concerns that a joint count response was a necessary element of response definition. Thus, if they performed well, candidate measures that required or weighted the joint count response were considered preferable. Last, individual members of all committees emphasized the need to preserve the ACR20 in some manner.

RESULTS

Clinical construct validity: the survey. The survey was completed by 51 rheumatologists, mostly those attending the 2004 OMERACT (Outcome Measures in Rheumatology Clinical Trials) 7 meeting (Asilomar, CA), at which the survey was administered. Respondents included rheumatologists from the US, Canada, Australia, and several European countries. Of these, 2 had unusable data because too many responses were left blank. Examining the correlations of candidate measures of response with rheumatologists' impressions of the degree of improvement among paper patients (Table 2), we found that all correlated moderately well. The measures that performed poorly were primarily (but not exclusively) those that defined response in a dichotomous manner, such as the ACR70 or the DAS. This was because the range of improvement was not sampled by these dichotomous measures. For example, if a given patient had experienced major improvement or remission, the ACR20, a dichotomous measure, would label them only as improved and would not accurately characterize their more substantial improvement. Because of this limitation, dichotomous measures tended to have lower correlations (r values) with the range of patient improvement compared with ordinal measures, continuous measures, or measures that used an index approach to define response.

Table 2. Spearman's rank correlation coefficients (ranked from strongest to weakest) of mean survey response with candidate measure-defined response*

Variable	Spearman's correlation	*P*
%SDA1	0.96569	< 0.0001
%CDAI	0.93778	< 0.0001
n3ACRdw	0.92795	< 0.0001
n2ACRdw	0.92328	< 0.0001
%DAS	0.89903	< 0.0001
nACRdw50	0.89547	< 0.0001
n3ACR	0.87737	< 0.0001
n2ACR	0.86971	< 0.0001
nACR50	0.86144	< 0.0001
nACRdw	0.84058	< 0.0001
ACR-N	0.83657	< 0.0001
ACRhybrid	0.83150	< 0.0001
bdACRhybrid	0.82231	< 0.0001
%ODAI	0.80802	< 0.0001
ACRstep	0.79135	< 0.0001
nACR	0.78574	< 0.0001
MedianACR	0.77829	< 0.0001
n3ACRw	0.77121	< 0.0001
ACRuw50	0.76574	< 0.0001
FifthACR50	0.76574	< 0.0001
ACRstep3	0.76277	< 0.0001
nACRw	0.75884	< 0.0001
n2ACRw	0.75826	< 0.0001
ACR30	0.74588	< 0.0001
ACR20	0.73140	< 0.0001
ACR40	0.71090	< 0.0001
LDAS1	0.68017	< 0.0001
EULAR	0.66271	< 0.0001
nACRw50	0.65902	< 0.0001
MeanwACR50	0.64574	< 0.0001
ACRuw30	0.62228	< 0.0001
FifthACR30	0.62228	< 0.0001
ACR50	0.60604	< 0.0001
ThirdACR50	0.58918	< 0.0001
ACRuw20	0.58586	0.0001
FifthACR20	0.58586	0.0001
ACR70	0.57165	0.0002
MeanACR	0.55360	0.0003
MeanACR_bd	0.55338	0.0003
MeanACR30	0.50842	0.0011
LDAS2	0.48516	0.0020
MeanACR20	0.46735	0.0031
MeanwACR30	0.46149	0.0035
MeanwACR70	0.45426	0.0042
MedianACR20	0.43411	0.0065
DASimp	0.35198	0.0302

* For definitions, see Table 1.

Discriminant validity: trial data. We turned to analysis of RA trial data, displaying candidates in order of the most sensitive to change to the least sensitive to change (Table 3). For each candidate measure, we produced an average rank of its sensitivity to change compared with all other candidate measures (average rank) and a weighted average rank that weights each trial by its study size. The minimum/maximum rankings range from 1 to 135, because we tested a total of 135 candidate measures, although (because of space limitations) we show only measures of interest. The measure with the greatest sensitivity to change was a

data-driven measure designed to be our gold standard, based on the O'Brien test.

The top scoring measures were those that counted the number of core set measures that improved by 20%, 50%, and 70%. This group of measures includes the n3ACR and the n3ACRdw (doubly weighted for joint count). The n3ACR definitions characterize a patient as having improved by 20%, 50%, or 70% in each of the core set measures, in which they would get, respectively, a score of 1, 2, or 3 for that core set measure. The n3ACR count for a given patient would range from 0 to 21, with a score of 21 being achieved if the patient had at least 70% improvement in all core set measures.

Next in the ranking of candidate measures was the hybrid measure, the bdACRhybrid. This measure combined the ACR20, ACR50, and ACR70 scores with a patient's mean improvement in core set measures. Because the patient in Figure 1 from the survey met the ACR50 criteria but not the ACR70 criteria, that patient had a bdACRhybrid score between 50 and 70. The patient's mean percent improvement in core set measures was 57.9, which is the bdACRhybrid score. The scores for other patients are presented in Appendix A.

The bdACRhybrid variable is limited to an overall score of −100 to +100. Given that when core set measures improve there is a drop in score, these measures cannot improve more than 100%, thereby setting an upper bound. This particular hybrid measure was also limited to −100 (maximal worsening). We tested unbounded versions of the hybrid, in which the score was allowed to decrease according to the mean percent worsening, but these hybrid versions were not as sensitive to change as the bounded version, perhaps because they permitted outlier values to have a greater effect.

When we examined other candidate measures of response, we found that their discriminant validity was less than that of the above-described measures (Table 3). These included commonly recommended measures, such as the EULAR criteria for improvement and the ACR-N. Dichotomous measures of improvement, such as the ACR20, were farther down the list, and the worst scoring measure tested was one of the measures of low disease activity. Both measures of low disease activity scored poorly, suggesting that low disease activity should not be used as a primary outcome measure in trials.

The difference between the best and the worst candidate scoring measures was substantial. For example, in a trial of conventional DMARDs, the worst scoring measures showed no significant difference between the efficacy of second-line drug and placebo, whereas the top scoring measures showed a significant difference between drug and placebo ($P < 0.001$). In general, the difference between the worst and the best groups of measures in terms of P values ranged from 10^{-3} to 10^{-6}, which represents a huge difference between the ability of the most and the least sensitive candidate measures to detect differences between therapies. Thus, the efficacy of some drugs versus placebo (or modest efficacy differences between 2 treatments) would not be detected if candidate measures at the bottom of the list were used. Our results also translate to smaller sample sizes needed in trials to detect the same treatment effects with statistical significance. For example, to achieve 80% power with a significance level of 0.05, a trial using the ACR20 would need 200 patients in order to detect a specific difference in treatments. Using the hybrid ACR response measure, only 108 patients would be needed to detect the same difference with the same power.

The committee considered data on the sensitivity to change of different candidate measures. Initially, because dichotomous measures had worse construct validity than did ordinal or continuous measures and were consistently among the worst performing candidate measures of response, these were rejected in our deliberations. We focused instead on the 10–15 measures that performed best. Additional concerns that were considered were understandability of the candidate measure, its compatibility with the ACR20, and its emphasis on joint count improvement without compromising sensitivity to change.

The options considered by the ACR committee and its parent committee included variations on the n3ACR and the bdACRhybrid. Because the other measures had worse sensitivity to change and did not preserve the ACR20/50/70 approach and/or they were not understandable, they were eliminated from consideration. Ultimately, the Committee to Reevaluate Improvement Criteria, working with the parent committee, selected the bdACRhybrid (which we will call the "hybrid ACR response measure") for the following reasons:

1. It was highly sensitive to change (ranked close to the top of all measures tested).
2. It incorporated a requirement for joint count improvement without compromising sensitivity to change.
3. It preserved the ACR20/50/70; the proportion of patients improving by these amounts is easily determined from data in which the bdACRhybrid is used.
4. It could be understood readily.

Scoring of the hybrid ACR response measure is shown in Table 4.

DISCUSSION

Using a data-driven consensus approach, a committee constituted by the ACR has revised the ACR definition of response and recommended a new hybrid measure that incorporates a patient-specific definition of improvement, allowing for the computation of ACR 20%, 50%, and 70% improvements, yet also incorporating a continuous measure of change. This measure addresses conflicting needs in defining improvement; that is, developing a measure that optimally detects modest differences between therapies, and retaining a standard measure that does not discard the widely used ACR20.

The hybrid ACR response measure is intended for use as an outcome measure in clinical trials. Its use enables modest treatment differences to be detected as statistically significant and allows trials to be carried out with fewer subjects than are required using current outcome measures.

Because drug developers may rely on this measure to perform smaller pivotal trials than are currently undertaken to gain FDA approval, use of the hybrid ACR re-

Table 3. Ranking of the discriminant validity among 135 candidate definitions in rheumatoid arthritis trials*

Candidate definition	Rank of definition among all tested using weighted average of trial data	Lowest (best) rank for candidate among all trials	Highest (worst) rank for candidate among all trials
O'Brien%†	1	1	27
n3ACR†	3	1	52
n3ACRdw	4	2	50
bdACRhybrid†	7	1	50
MeanACR_bd†	8	1	54
n2ACR	9	4	61
n2ACRdw	10	5	58
MeanACR2	15	6	60
ACRhybrid†	18	2	49
%DAS†	20	1.5	89.5
%SDAI†	27	1	122
MeanACR	28	5	85
MedianACR	31	5	89
nACR†	32	3	80
nACRdw	33	1	78
nACR50	34	4	66
nACRdw50	35	6	73
%CDAI	45	21	86
%ODAI	46	20	70
MedianACR2	49	20	79
n3ACRw	50	8	105
ACR-N†	53	14	93
EULAR†	54	7	111
n2ACRw	55	12	102
nACRw	62	7	98
ACRstep (ACR20,50,70)	64	12	97
ThirdACR30	69	1	131
MedianACR30	70	17	130
ACRstep3	72	51	103
DASimp†	74	5	126
ThirdACR50	77	36	113
ACRuw30	78	13.5	111.5
FifthACR30	79	13.5	111.5
MeanwACR30	83	5	116
ACRuw20	87	7	131
FifthACR20	88	7	131
MeanACR20	90	22	122
MeanACR30	91	11	129
MedianACR20	92	39.5	125.5
nACRw50	96	31	129
MeanwACR50	97	41	129
ACR20†	101	40	122
LDAS1	107	18	131
ACR30	110	76	118
ThirdACR20	113	11	130
ACR40	116	84	128
ACRuw50	118	61	118
FifthACR50	119	61	118
ACR50†	123	68	133
ACR70†	130	18	133
LDAS2	135	82	133

* Most widely clinically used measures and other representative results are presented. For definitions, see Table 1.
† Candidates of special interest.

sponse measure may make safety data less complete at the time of drug approval.

Disease activity states and response measures are not the same, and the former may be useful in clinical practice and even as a measure of the attained state in trials (e.g., low disease activity, remission). Our effort does not supplant

Table 4. Scoring method for the Hybrid American College of Rheumatology (ACR) response measure*

ACR status	Mean % change in core set measures			
	<20	≥20, <50	≥50, <70	≥70
Not ACR20	Mean % change	19.99	19.99	19.99
ACR20 but not ACR50	20	Mean % change	49.99	49.99
ACR50 but not ACR70	50	50	Mean % change	69.99
ACR70	70	70	70	Mean % change

* 1) Calculate the average percentage change in core set measures. For each core set measure, subtract score after treatment from baseline score and determine percentage improvement in each measure. Next, if a core set measure worsened by >100%, limit that percentage change to 100% (a − 100% bound). Then average the percentage changes for all core set measures. 2) Determine whether the patient has achieved: ACR20, ACR50, or ACR70. 3) Using the table above, obtain the Hybrid ACR response measure. To use the table, take the ACR20, ACR50, or ACR70 status of the patient (left column) and the mean percentage improvement in core set items; the Hybrid ACR score is where they intersect in the table.

valuable efforts by other groups of investigators to define different states of activity in RA, including low disease activity and remission. We do, however, suggest that the current definitions of low disease activity (and, by extension, remission) should not be used as primary outcome measures in trials, because they are among the least sensitive to change of the measures we evaluated. Our data also suggest that the available disease activity state measures do not perform as well in trials as does the new hybrid ACR response measure.

Our approach has some important limitations. First, we made several critical assumptions that affected our choice. We accepted the validity and high performance of the core set of outcome measures. Other core set measures might yield a different definition of response. Second, although our list of candidate measures was large, we may not have considered certain approaches that might have altered our decisions. Third, even with the assumptions, we may have produced a different answer using a different set of trial data, although our trials were large and varied enough that we suspect the ranking of candidate measures would not have changed much. Fourth, our candidate measures may rank slightly different if a different analytic approach (e.g., parametric analysis) were used, although in such an approach, outliers would be likely to compromise power to a greater extent.

Although requiring improvement in both the tender joint count and especially the swollen joint count can compromise the sensitivity to change of outcome measurement in RA, we believe there are creative solutions to this problem. We have proposed one solution, the hybrid ACR response measure, which requires improvement in the swollen and tender joint counts to reach a certain threshold score. Combined with the mean percent improvement in core set measures, one can derive the ACR hybrid measure score (see Table 4). To make it clear how this works, we provide additional tables in which we compute the new recommended response measure (see Appendix A).

We noted earlier that one major source of measurement heterogeneity in RA trials has been the varied ways in which the ACR20 has been defined. There are a variety of ways in which the hybrid ACR measure could be used based on timing of measurement. In another report, we shall propose an approach to timing of the hybrid ACR measure in trials so as to encourage more uniformity in measuring response in RA trials.

In summary, we suggest a revision to the ACR20 that creates a new hybrid outcome measure, a measure combining the ACR20, the ACR50, and the ACR70 and a continuous score of the mean improvement in core set measures. This new measure has much greater statistical power to distinguish the efficacy of treatments than do the currently recommended measures of response, including the ACR20.

AUTHOR CONTRIBUTIONS

Dr. Felson had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study design. Felson, Xie, LaValley, Anderson, Furst, Boers, Pincus, Dougados, Bombardier, Simon.

Acquisition of data. Felson, Xie, Anderson, Wolfe, Furst, White.

Analysis and interpretation of data. Felson, Xie, LaValley, Anderson, Koch, Wolfe, Michaud, Furst, Boers, Bathon, Tugwell, Pincus, Dougados, van der Heijde, Smolen, Aletaha, Siegel, Choi, Paulus, White, Landewé, Simon.

Manuscript preparation. Felson, Xie, LaValley, Koch, Michaud, Furst, Boers, Smolen, Aletaha, Siegel, Choi, Paulus, Bombardier, Landewé.

Statistical analysis. Felson, Xie, LaValley, Koch, Furst.

Background indices. Pincus.

REFERENCES

1. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al, and the Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729–40.
2. Boers M, Tugwell P, Felson DT, van Riel PL, Kirwan JR, Edmonds JP, et al. World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol Suppl 1994;41:86–9.
3. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. American College of Rheumatology: preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.
4. Van Gestel AM, Prevoo ML, van 't Hof MA, van Rijswijk MH, van de Putte LB, van Riel PL. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis: comparison with the preliminary

American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. Arthritis Rheum 1996;39:34–40.
5. Anderson JJ, Wells G, Verhoeven AC, Felson DT. Factors predicting response to treatment in rheumatoid arthritis: the importance of disease duration. Arthritis Rheum 2000;43: 22–9.
6. Anderson JJ, Bolognese JA, Felson DT. Comparison of rheumatoid arthritis clinical trial outcome measures: a simulation study. Arthritis Rheum 2003;48:3031–8.
7. Siegel JN, Zhen BG. Use of the American College of Rheumatology N (ACR-N) Index of Improvement in Rheumatoid Arthritis: argument in favor. Arthritis Rheum 2005;52:1637–41.
8. Pincus T, Strand V, Koch G, Amara I, Crawford B, Wolfe F. An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from that of placebo as effectively as the American College of Rheumatology 20% response criteria (ACR20) or the Disease Activity Score (DAS) in a rheumatoid arthritis clinical trial. Arthritis Rheum 2003;48:625–30.
9. Goldsmith CH, Boers M, Bombardier C, Tugwell P, and the OMERACT Committee. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. J Rheumatol 1993;20:561–5.
10. Van der Heijde DM, van 't Hof MA, van Riel PL, Theunisse LA, Lubberts EW, van Leeuwen MA. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. Ann Rheum Dis 1990;49:916–20.

APPENDIX A. AMERICAN COLLEGE OF RHEUMATOLOGY HYBRID SCORE OF 3 TRIAL PATIENTS

Patient 1. The patient did not meet the criteria for ACR20 (swollen joint count worsened). Therefore, the Hybrid ACR score must be <20, even though the mean change of all scores is 42.8. The Hybrid ACR score for this patient is 19.99.

Core set item	Range	Pretreatment	Final value	% change
Tender joint count	0–68	43	14	67
Swollen joint count	0–66	38	44	−16*
Pain	0–10	4.6	2.0	56
Patient global	0–10	8.5	3.1	63
Physician global	0–10	7.5	2.7	64
Health Assessment Questionnaire	0–3	2.8	2.0	28
C-reactive protein level (mg/dl)	>0	11.6	7.2	38

* Swollen joint count has worsened.

Patient 2. The patient met the criteria for ACR50 but not ACR70. Therefore, the Hybrid ACR score must be between 50 and 69.99. The mean percentage improvement is 73.29%. The ACR Hybrid score (limited by the failure of the patient to reach ACR70) is 69.99.

Core set item	Range	Pretreatment	Final value	% change
Tender joint count	0–68	43	14	67
Swollen joint count	0–66	38	14	63
Pain	0–10	4.6	1.0	78
Patient global	0–10	8.5	1.1	87
Physician global	0–10	7.5	2.7	64
Health Assessment Questionnaire	0–3	2.8	1.0	64
C-reactive protein level (mg/dl)	>0	11.6	1.2	90

Patient 3. The patient met the criteria for ACR50 but not ACR70. Therefore, the Hybrid ACR score must be between 50 and 69.99. The mean percentage change is 52.29, and the ACR Hybrid score is 52.29.

Core set item	Range	Pretreatment	Final value	% change
Tender joint count	0–68	43	14	67
Swollen joint count	0–66	38	19	50
Pain	0–10	4.6	2.0	56
Patient global	0–10	8.5	3.1	63
Physician global	0–10	7.5	2.7	64
Health Assessment Questionnaire	0–3	2.8	2.0	28
C-reactive protein level (mg/dl)	>0	11.6	7.2	38